Identification and Interpretation of Figurative Language with Computational Semantic Models

Aaron Gerow

March 2014

Declaration

This thesis has not been submitted as an exercise for a degree at this or any other university. It is entirely the candidate's own work. The candidate agrees that the Library may lend or copy the thesis upon request. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Aaron Gerow

Summary

This thesis is about the automatic extraction of metaphors as they appear in English text. This task is important to research in information retrieval, corpus linguistics and computational linguistics. The work was motivated by theories of metaphor comprehension and statistical semantics and contributes to areas of natural language processing (NLP) and information extraction where figurative language continues to present a challenge. Chapter 2 reviews related psychological and computational work and provides a foundation for a method described in chapter 3. Chapter 4 describes my implementation of this method – a system called *MetID*. Chapter 5 evaluates MetID on three increasingly difficult tasks: identification, interpretation and extraction of figurative language. The final chapter describes the contribution of this research, contextualising it in light of the research goals and concludes with a discussion of future work.

Methods and techniques of the project were inspired by research on how people comprehend metaphors, by linguistic research in how metaphor is used in text, and by NLP techniques for extracting particular types of metaphor. The goal was to build and test a system for automatically finding and providing interpretations of figurative language. A central task is representing word associations that account for the semantics of figurative language. Specifically, three types of lexical models were evaluated: WordNet, distributional semantic models and co-occurrence likelihood estimation. The method also uses a number of heuristics that typically mark linguistic metaphor, such as selectional violation and predication. The system can be used to analyse individual phrases, a corpus (which can simultaneously be used to build the lexical model) or a collection using pre-built models. The output is a ranked list of candidate metaphors by which to interpret a statement. For example, analysing "my heart is on fire" produces the interpretation AFFECTION AS WARMTH. The system attempts to account for two common forms: noun- and verb-based metaphors. Evaluation results suggest that the method performs significantly above chance on noun-based statements but not for verb-based. The choice of lexical model has a significant effect when analysing noun-based statements, but not verbs. The results on an interpretation task, which were validated with participant ratings, found that 1) noun-based statements were more easily interpreted, 2) the system was better at interpreting figurative statements than literal statements and 3) in some configurations, the system's scores correlate strongly to participant ratings. Additionally, an interesting interaction was found: the literal / non-literal distinction mediated the role of a statement's grammatical form when considering the quality of interpretation. Last, a case study was used to aid a corpus-based terminological analysis of the word *contagion* in finance and economics where it has been adopted with a number of figurative features.

Acknowledgements

First, thanks to Professor Khurshid Ahmad for supervision throughout – his generosity with time, energy and material support goes unmatched among peers. Thanks to Professors John Barnden and Sam Glucksberg for mid-way examination and advice. Thanks to Professors Carl Vogel, Mark Davies, Yorick Wilks, Mark Keane, Fred Cummins and Dr. Hua Huo for ideas and pointers at various times. Thanks also to Dr. Ekaterina Shutova, Professor Andrew Goatly and Professor Thomas Landauer for providing various data, as well as the Trinity Centre for High Performance Computing for computing resources. Thanks to my peers in the school, especially my office-mates at Trinity and my former class-mates at UCD. This degree was financially supported by a Trinity Postgraduate Studentship from the University of Dublin, three Trinity Travel Grant awards and Enterprise Ireland grant #CC-2011-2601-B: GRCTC.

Thanks are also due to friends and family beyond my academic life – a life which often subsumed most of my time and energy. I am grateful to friends in Chicago, Tacoma and Dublin: to Justin, Rosie, Nicky, Maria, Fearghal, everyone at the castle, at PLU and at UCD. With hindsight, I have come to greatly appreciate my friends' tolerance, patience and optimism with me and my work, which at many points, must have seemed an obsession. Thanks to my parents, Bill and Susan, for their support and perpetual pride, despite my departure from gainful employment to cross the Atlantic in pursuit of vague academic aspirations. And lastly, thanks to Kristin (soon-to-be Dr. Hadfield) for her absolute and endless help, advice, patience and love for the last three wonderful years in Dublin. I can't imagine my life without Kristin, nor can I imagine my future with anyone else.

Contents

Chapter 1

Introduction

1.1 Motivation: Computation, Language & Metaphor

The systematic analysis of language has been around long before the availability of automatic, computational methods. Some of the earliest computational analysis of communication examined patterns, habits and trends in language use. The advent of the internet and growth in computing resources – retrieval, storage and processing power – have enabled large-scale analyses of text. The wealth of available data has encouraged increasingly advanced techniques to automatically find and extract meaning from text. Despite advances in automated text analysis, one continued difficulty is in analysing certain types of figurative language.

Figurative communication is important to the way people converse with one another and understand the world around them. Many observed metaphors are common across languages, cultures and types of communication. The genesis of many words often points to a literal sense being re-applied in a figurative way to a new domain or idea. Certain words like "rise" and "fall" are consistently used without any reference or analogues to changes in height. Other concepts, like temperature, movement and weight, are commonly used to describe feelings, quantities and complexity. The work presented in this thesis explores strategies for automatically identifying and interpreting certain forms of figurative language. This work combines three areas of research: cognitive linguistics, computational semantics and natural language processing (NLP). It is an attempt to address technical questions about linguistic phenomena and seeks to extend the state of the art to better account for certain types of metaphor. This thesis will motivate and present a method for finding metaphors in text, using corpus-based semantic models in combination with legacy resources and NLP techniques. The key question is this: can computational semantic models be combined with NLP tools and techniques, to accurately find and interpret figurative statements?

Early views on the role of metaphor in communication and thought held that they were poetic decoration of otherwise literal language. Such a view does not address questions about how metaphors are created and processed. Modern theories propose that metaphor is integrated with human experience and thought – an integration that beckons further inquiry [92, 132, 171]. Why do people apply some metaphors so consistently? What accounts for systematic relationships among metaphors? How do people identify and make sense of metaphor in discourse? How are metaphors used in explanations of new, complicated or abstract concepts? What are the lexical properties of metaphor and do they relate to discourse in a way that they might be automatically extracted and interpreted? Some of these questions have been addressed in psychological and linguistic research [75, 80, 235]. The current work is motivated by linguistic inquiry, grounded in experimental theories, and seeks to extend the fields of NLP, information extraction and statistical semantics to better address the semantics and pragmatics of figurative language.

Metaphor research attempts to help application-driven fields like text summarisation, document retrieval and information extraction. These fields have begun to focus on a number of questions in lexical semantics, often implementing analogues to mental procedures for extracting and processing meaning in text. For finding meaning in text, there are two commonly manifested problems: sparse or incomplete data, and assessing the contribution of extra-textual information like context or outside knowledge. The sparsity problem has lead to proposed solutions ranging from reformalisations of similarity and relatedness, to statistical normalisation and approximation [139, 144, 164]. The problem of contextual knowledge, on the other hand, is somewhat of an open question in lexical semantics. While there are other approaches, the current research takes a text-centric view to semantics. That is, while an ontological structure to meaning, with which to inform semantics in language, may be proposed, it would be of little use to computers until it is built and verified. Instead, building a semantic understanding from the text, in a ground-up fashion, provides methods that are tractable, empirical and language independent.

The project described in this thesis is one in NLP, but it deviates from the increasingly common hybrid of NLP + machine learning work-flow of data gathering, feature selection and classification to fit a model with observed data. Instead, testing representations will allow the method itself to be tested rather than the features or learning algorithms. This goal stems from the motivations of the research: to test corpus semantic models against figurative language. Thus, the theoretical question is this: can NLP techniques be combined with general purpose semantic models to accurately find and interpret figurative language.

1.2 Goals

There are two goals of my research reported here. First, with regards to figurative language, a number of researchers have shown that what is commonly called metaphor is more accurately defined as a range of phenomena – linguistic and conceptual [76, 80, 134, 135]. Linguistic metaphor (metaphors expressed in language) has many forms and appears not to be entirely distinct from other linguistic phenomena such as analogy, ellipsis and metonymy [10, 46]. Ostensibly, these forms of figurative communication are defined by different features. Exploring these differences, in particular with respect to lexical features, is the over-arching aim of this research. To the extent that figurative language is available to lexical analysis, doing so systematically will help provide a data-driven foundation of metaphor use. Second, this work contributes to computing and NLP. Though the project is inspired in-part by psycholinguistics and statistical semantics, a number of a NLP tasks have been explored, implemented and evaluated to produce a system designed to identify and interpret the use of metaphor in raw text.

Figurative Language

The first main goal is to explore the boundaries of various types of figurative language. The hallmark of a metaphor is the mapping of one concept onto or by another [92, 128, 132, 133]. Different research describes this in different ways, but the common element is that in a metaphor, a concept is being made sense of by another, the relationship being in some way figurative. However, to differentiate figurative and literal statements, more information is needed – both linguistic and conceptual. There is more than one way to classify linguistic metaphors; Walter Kintsch offers four types of metaphor [122]:

1.	Simple metaphors ($Noun_1$-is-$Noun_2$).	*My lawyer is a shark.*
2.	Simple analogy-based metaphors.	*She blew up at me.*
3.	Complex analogy-based metaphors.	*The universe is a computer.*
4.	Literary metaphors.	*We are the eyelids of defeated caves.*

These types of metaphor help narrow the scope of the proposed metaphor identification system. The fourth type, literary metaphors, is perhaps inviable without operationalising people's intuitions about the of symbolism, metaphysics and identity in interpretations of language. The third type, which Kintsch calls ”complex analogies”, also may require extensive background knowledge to interpret accurately. In the example above, people need to know what it is that computers do and how it could be used to relate them to the universe. Complex analogies like the example are thought to a rely on a set of features (functional, associative, semantic, etc.) that can be aligned to make sense of one concept (computers) in terms of another (the universe). The first two types of metaphor can indeed be complex in ways similar to the third type, but Kintsch proposes that their semantics are less reliant on non-lexical features. Given the surface similarity (evident between examples 1 and 3) it would be difficult to categorise an observed metaphor in the way Kintsch proposes. Instead, the current project attempts to address two surface forms of figurative language: noun- and verb-based metaphors. It is likely, however, that types 1 and 2 listed above are more easily detected and interpreted computationally, regardless of their surface similarities to more difficult kinds of figurative language.

Taking the first two types of metaphor above, their differences can be further specified. The first is typical of noun-based metaphors where one *thing* is compared by asserting it *is* another. The example above uses the stereotypically predatory nature of sharks to exaggerate an analogous aspect of being a lawyer. The second example, simple analogy-based metaphors, is a verb-based statement in which “blew up” implies something like “got very angry”. These two types of metaphor appear to be available to a lexical analysis, and as such, they are the focus of this thesis.

NLP

In addition to exploring the use of figurative language, this project is equally concerned with the technical task of automating its identification and interpretation in text. A defining feature is that figurative language tends to evade interpretation with brittle accounts of lexical semantics in which words have a meaning and their relationships are instantiated by connections dictated by the words themselves [197, 230]. This conception of semantics has been helpful and productive in areas like semantic networks, the semantic web and lexical semantic modeling [48, 53, 167]. However, the range of figurative language is unlikely to be captured in its entirety with such a model.

Given the success of computational models in NLP, and corpus-based semantic models in particular, this project attempts to push the boundaries of existing work in computation metaphor processing. To that end, much of this work is a comparative exercise between various types of semantic models. This research will look specifically at WordNet, a semantic network produced by lexicographic research, distributional semantic space models such as latent semantic analysis (LSA), and purely statistical models based on term co-occurrence likelihood estimation. Three types offer a spectrum of approaches. Precisely which models perform best and why will be addressed in the last two chapters but, preliminarily, the computational task has three potential outcomes:

1. The models fail to detect or accurately interpret figurative language in text.

2. The models capture the presence of figurative language in text and provide accurate interpretations.

3. The models detect some kinds of figurative language with variable accuracy of interpretations.

Though the third outcome is the most likely, considering the literature on semantic models, the first two are possibilities. Should the first outcome be the case, a new goal would be set for semantic models: to accurately represent figurative meaning. Should the second be the case, it would further support the models against critiques. In the third case, more particular conclusions will be in order. Which configurations succeed or fail, in what circumstance, in what way and why? What properties of figurative language contribute to metaphor evading a computational analysis? Are these models inherently flawed, or can they be revised to better account for the breadth of such language? Answers to these questions may be technical in nature and will perhaps precipitate new strategies in computational models of meaning.

1.3 Relationship to Other Work

This section reviews some of areas related to this project. A more in-depth review of the literature will be presented in the next chapter, but below is a review of the potential contributions to corpus linguistics, computational semantics, NLP and information extraction.

Corpus Linguistics

Corpus linguistics is the sub-field of linguistics that adopts a language-as-data approach to researching language use, change and theory [59, 129]. Corpus linguistics differs from the Chomskyan or "internalist" approach of analysing constraints placed on language by formal syntactic structures [20, 21, 36]. Instead, corpus linguistics utilises language use as data in which to find patterns, consistencies and changes. Corpus linguistics has been enabled by advances in computational analysis as well as data access, retrieval and storage. Though dictionary makers and lexicographers inspired the field, Kucera and Francis' *Computational Analysis of Present-Day American English* [129] is often cited as a one of the first corpus linguistics publications. Kucera and Francis present a computer-assisted analysis of what is now known as the Brown Corpus [59], and it exemplifies two central tasks: corpus construction and systematic analysis.

Corpus construction is about developing a sample of language, which is in some way representative of language use. Corpora are often built for specific genres, domains, publication types or readership levels to investigate language use in specific contexts [29, 129, 152]. An important kind of corpus is a diachronic collection, where texts are organised over time, usually with some uniformity from one period to another. Diachronic corpora are a relatively recent advancement which has inspired the idea of a "monitor corpus": a corpus that can track changes in language use over time [43].

Corpus analysis starts with frequency observations at different levels of linguistic description (words, word forms, stems, lemmas, phrases, etc.) and may employ NLP techniques such as dependency parsing and part-of-speech tagging. During the course of the current project, a number of corpora were used to build and compare semantic space models and provide various statistical information such as common predications and selectional preferences. From one perspective, this project is a computational branch of corpus linguistics, seeking to combine its data-driven approach to language with cognitive linguistic theory and NLP techniques.

Computational Semantics

At the centre of the method employed in this project is a set of lexical models which will be used to relate observed words with a set of seed terms derived from corpus linguistic research. For this work, a semantic model has a single purpose: to associate words. The crucial part, which is addressed in different ways by different models, is that word association can mean something different depending on morphological, grammatical, lexical and sentential context. The oldest, and perhaps most simple semantic model is a dictionary, often in which multiple senses of a word can be found (for example "river *bank*" vs. "investment *bank*"). Respecting this contextually dependent aspect of natural language is non-trivial for computers.

The first corpus-based semantic models, like Salton's *vector space model*, did not delineate word senses, separate phrases or compose multi-word terms [192]. Nonetheless, these prototypical models provided foundational methods which have been adopted almost uniformly in models that build a "semantic space" from frequency observations[1]. Exploring specifically what features a model must represent to accurately find and interpret linguistic metaphor is a central goal of this project. In pursuing this, improved strategies for computational modeling may be proposed.

NLP & Information Extraction

NLP and information extraction are, tasked with automatically making sense of naturally occurring text. This project is concerned with semantic modeling as it relates to non-literal statements, but many of the tools and techniques used and developed here are enabled by and contribute to work in NLP. Not only is this work aided by advances in parsing and POS-tagging, it also implements a number of NLP-style solutions to problems like selectional preference induction, analysing predications and word clustering. Three types of models will be used for building clusters and associating words: an explicit model, semantic space models and a statistical method of co-occurrence likelihood estimation. However, the system is designed to be neutral with respect to what models are used. This addresses an important experimental goal of testing different strategies to address the analysis of figurative language.

1.4 Structure of the Thesis

This thesis presents the motivation, design, implementation and evaluation of a computational system for processing figurative language. Chapter 2 reviews the foundations of figurative language in terms of use and understanding of metaphor, and its properties evident in language. This includes relationships among types of figurative language and how the linguistic properties of metaphor can aid its automatic identification and interpretation. An overview of existing computing and NLP work addressing figurative language is reviewed placing the proposed method in a unique position to 1) combine metaphor identification and interpretation tasks and 2) analyse both noun- and verb-based metaphors in a unified manner. The third chapter describes the overall method the implementation of which is in chapter 4. The method's design combines the use of legacy resources and linguistic findings with computational / NLP tools and techniques. Chapter 4 describes the system, called *MetID*, that was built to allow a comparison of different semantic models' ability to process figurative language using a word-clustering strategy [199, 200]. The system's modular design consists of a structural module, a word-clustering module (achieved using a range of corpus-based semantic models) and a series of post-hoc, conditional heuristics. Chapter 5 contains an evaluation of MetID. Because a number of valid configurations are possible, the first section evaluates word-clusters built with different text collections, semantic models and their variants. For the first evaluation of language, a subset of the valid configurations are used to address an idealised identification task in which MetID is used to pick figurative statements from a set of literal-figurative

[1] For example, the proposal to use log-entropy normalisation to improve the performance of TF-IDF is ubiquitous among distributional semantic space models.

pairs. The best performing configurations (which perform at about 73% accuracy) are used in a subsequent interpretation task where the system's output is evaluated against the results of an online user-study. The third and final evaluation is based on a case-study [71] that used MetID to aid a terminological analysis of *contagion* as it is used figuratively in finance and politics. The final chapter discusses the findings of this research as they relate to the synthesis of linguistic metaphor theory, statistical semantics and NLP. This includes technical alternatives to MetID's architecture that could yield gains in performance, how grammatical structure and conceptual representations impact metaphor processing, the role of non-lexical information in metaphor interpretation and areas of future work.

1.5 Key Contributions & Findings

The main contribution of this work is the MetID system, which implements a number of NLP techniques and represents a unique combination of work in figurative language (cf. Goatly; [92]) and statistical NLP (distributional semantic models). The novelty of the system is a cluster-based approach that can use interchangeable lexical models (WordNet, distributional models and co-occurrence likelihood estimation). MetID extends the clustering approach to metaphor identification in two important ways. First, the word association method (a lexical model) is interchangeable, allowing comparative analysis. Secondly, the system augments word association information with clusters' intrinsic quality metrics, allowing it to better know when results are sub-optimal. The system also implements measures of selectional violation [182, 200, 230] and predication [15, 158]. Additionally, instead of choosing a best answer (or interpretation) the system uses a rank-based algorithm allowing a number of possible interpretations for a single statement. Moreover, performance with certain lexical models show the system operationalises two dominant theories of metaphor comprehension: feature mapping and category matching. That is, often, the word associations given by the lexical models represent featural and categorical information about concepts.

The goal of this project is to operationalise the identification and interpretation of metaphor in a unified manner. However, these tasks are not strictly defined and there are no gold standards to measure a method's success. Instead, the system is first evaluated on idealised identification and interpretation tasks. The first is designed to test the system's ability to differentiate literal and non-literal statements with respect to the grammatical form, the lexical model and the corpus used to train the model. Results on this task suggest that noun-based statements are considerably easier to process and that the choice of lexical model is significant. The results show that WordNet and a distributional model (called COALS) performed relatively well, at about 73% accuracy. In the second task, automatically generated interpretations were rated by participants in a sensibility and paraphrasing task. The results show that noun-based, figurative statements were the most accurately interpretable. Additionally, in figurative statements, the grammatical form plays a larger role than when analysing literal statements. Lastly, the lexical model was again important to the system's performance, the best of which were about 38% above chance. In some configurations, scores from the system correlate strongly with participants' ratings, implying that when using some lexical models, the system's interpretations are tempered by their respective scores. A third evaluation is a case study, applying MetID to terminological analysis [71]. In a corpus-driven project, the system was used to provide an overview of common figurative concepts related to the term *contagion* as it is used in finance and economics. The *contagion* case-study exemplifies the role of computational methods in corpus analysis, especially in language change and figurative terminology.

Figurative language presents a challenge because it involves a range of contextual and stereotypical knowledge, making it a difficult task for statistical models of meaning. The current work adapts and extends NLP techniques to better address figurative language. While this work is not a comprehensive solution for the myriad forms of metaphor, it embodies a novel combination of existing resources and state-of-the art computational techniques.

1.6 Previous Publications

Below are abstracts and brief annotations of my publications to-date. These seven papers represent the results of my current research and a previous degree in cognitive science that were published during this PhD.

Gerow, Aaron and Keane, Mark T. (2011) Mining the Web for the "Voice of the Herd" to Track Stock Market Bubbles. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI '11), Barcelona, Spain, 16-22 July, 2011.*

> We show that power-law analyses of financial commentaries from newspaper websites can be used to identify stock market bubbles, supplementing traditional volatility analyses. Using a four-year corpus of 17,713 online, finance-related articles (10M+ words) from the Financial Times, the New York Times, and the BBC, we show that week-to-week changes in power-law distributions reflect market movements of the Dow Jones Industrial Average (DJI), the FTSE-100, and the NIKKEI-225. Notably, the statistical regularities in language track the 2007 stock market bubble, showing emerging structure in the language of commentators, as progressively greater agreement arose in their positive perceptions of the market. Furthermore, during the bubble period, a marked divergence in positive language occurs as revealed by a Kullback-Leibler analysis.

This paper was the product of empirical findings during the first phase of research for a master's degree. In developing and examining a corpus of financial texts, we found that power-laws were a helpful way of characterising the diversity of word-usage over time. Fluctuations in these power-laws initially appeared quite non-random, but when we inspected changes within POS distributions, we found a distinct correlation with the markets from 2006 to 2010. Heartened by this, we extended the work to a sentiment analysis of the same words and phrases comprising the power-law analysis [67].

Gerow, Aaron and Keane, Mark T. (2011) Identifying Metaphor Hierarchies in a Corpus Analysis of Finance Articles. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (Cogsci '11), Boston, MA, USA, 20-23 July, 2011.*

> Using a corpus of over 17,000 financial news reports (involving over 10M words), we perform an analysis of the argument-distributions of the UP- and DOWN-verbs used to describe movements of indices, stocks, and shares. Using measures of the overlap in the argument distributions of these verbs and k-means clustering of their distributions, we advance evidence for the proposal that the metaphors referred to by these verbs are organised into hierarchical structures of super-ordinate and subordinate groups.

Here, we explored how clusters of UP- and DOWN-verbs can reveal similarities in metaphorical instances of those words. The key finding was that the words "up" and "down" often stand alone, and that spatial instances like "rise", "fall", "lift" and "drop" clustered together while more dramatic instances like "soar" and "plummet" form a third cluster. This phenomenon was shown to be more or less symmetric for both UP- and DOWN-verbs. The work was a bifurcation of my master's thesis in cognitive science [68].

Gerow, Aaron and Keane, Mark T. (2011) Identifying Metaphoric Antonyms in a Corpus Analysis of Finance Articles. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci '11), Boston, MA, USA, 20-23 July, 2011.*

> Using a corpus of 17,000+ financial news reports (involving over 10M words), we perform an analysis of the argument-distributions of the UP and DOWN verbs used to describe movements of indices, stocks and shares. In Study 1 participants identified antonyms of these verbs in a free-response task and a matching task from which the most commonly identified antonyms were compiled. In Study 2, we determined whether the argument-distributions for the verbs in these antonym-pairs were sufficiently similar to predict the most frequently-identified antonym. Cosine similarity correlates moderately with the proportions of antonym-pairs identified by people (r = 0.31). More impressively, 87% of the time the most frequently-identified antonym is either the first- or second-most similar pair in the set of alternatives. The implications of these results for distributional approaches to determining metaphoric knowledge are discussed.

This paper used the results of a human experiment in which people paired antonyms for UP- and DOWN-verbs and showed that a distributional representation, measured by cosine similarity, tended to correctly pick human responses. This finding is interesting with regard to cognitive linguistic theories of metaphor, which propose systematicity between metaphorical words. This study showed that this systematicity, which is used by humans to generate antonyms, is also realised in the distributional structure of such words [69].

Gerow, Aaron and Ahmad, Khurshid. (2012) Diachronic Variation in Grammatical Relations. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India, 10-14th December, 2012.*

> We present a method of finding and analyzing shifts in grammatical relations found in diachronic corpora. Inspired by the econometric technique of measuring return and volatility instead of relative frequencies, we propose them as a way to better characterize changes in grammatical patterns like nominalization, modification and comparison. We examine a corpus of NIPS papers and report a number of trends which manifest at the token, part-of-speech and grammatical levels. Building on frequency observations, we show that shifts in lexical tokens overlook deeper trends in language, even when part-of-speech information is included. Examining token, POS and grammatical levels of variation enables a summary view of diachronic text as a whole. We conclude with a discussion about how these methods can inform intuitions about specialist domains as well as changes in language use as a whole.

This paper introduced the use of two second-order statistical methods for the analysis of diachronic corpora: return and volatility. There has been significant work in tracking various forms of linguistic data through diachronic corpora, but it often uses first-order analyses of frequency data, averages and standard deviation. But using return and volatility, this paper shows how examining the *changes* in time-series data can help make observations about trends in language use. We examined grammatical relationships of five key terms and present a summary analysis of noun keywords in the NIPS corpus [70].

Gerow, Aaron; Ahmad, Khurshid, and Glucksberg, Sam. (2013) The Concept of Contagion in Finance: A Computational Corpus-based Approach. In *Proceedings of the 19th European Symposium on Languages for Special Purposes (LSP 2013), Vienna, Austria, 7-10 July, 2013.*

> In everyday communication, figurative language is used to express emotions, value judgments and beliefs as well as to blend and create new concepts. In finance, metaphors often present messages of alarm in a soothing tone to overlook the cause of a problem and focus, instead, on a solution. The concept of contagion has recently entered discourse on international systems of regulation, finance and economics. We trace the emergence and successive use of the term in medicine and biology, to the social sciences and into the language of finance. We examine the use of *contagion* at word and grammatical levels and show how various patterns are used to elaborate particular features and diminish others. First, to look at the onset of the term, we track its use in *Annual Reviews* articles and journals of finance. We then present a corpus-based analysis of 38 US Congress documents and compare them to medical reports from the World Health Organization and the Center for Disease Control. The results show that some lexical-pragmatic properties are carried over from the biomedical context while others are not, which has implications to the special purpose language of

> finance and politics. In another analysis, we present a computational method based on word-clustering in WordNet, to analyze how the context of contagion signals various metaphors (or *root analogies*) in the congressional corpus (Goatly, 2011). The results of this newly developed method show patterns in the metaphorical domains which contagion makes use of in figurative contexts. We conclude that the nature of *contagion*'s use in finance and politics is more complex than term-borrowing as it establishes a range of lexical, pragmatic and metaphorical properties.

The paper is the result of preliminary computational explorations using MetID, the system implemented in this thesis to investigate the use of *contagion* in financial texts. The findings, which are presented in chapter 5 of this thesis, exemplify the computational corpus linguistic approach enabled by MetID. The study shows how the system can be used to address questions about concept and category creation in the lexicon. Additionally it provides a good introduction to the methodological relevance of a computational corpus-based approach to language change in the lexicon and in language use [71].

Keane, Mark T. and Gerow, Aaron (2014) It's Distributions All The Way Down!. To appear in *Behavioral and Brain Sciences*, 37 (1), 2014.

> The textual, big-data literature misses Bentley, O'Brien, & Brock's (Bentley et al.'s) message on distributions; it largely examines the first-order effects of how a single, signature distribution can predict population behaviour, neglecting second-order effects involving distributional shifts, either between signature distributions or within a given signature distribution. Indeed, Bentley et al. themselves under-emphasise the potential richness of the latter, within-distribution effects.

This paper is a peer-reviewed commentary on an article by Bentley, O'Brien and Brock [18]. The target article describes how four distributions of decision-making habits can be used to characterise various strategies groups adopt. Our commentary underscores the subtlety and novelty of distributional movements, both as first-order changes *between* distributions as well as movement *within* a distribution. Our earlier paper, [67], is an example of this type of analysis, and exemplifies methods that have only recently been enabled by data and intuitions regarding population-level behavior.

Gerow, Aaron; Ahmad, Khurshid, and Glucksberg, Sam. (2014; under review) Contagion in Finance: A Computational Corpus-based Approach. *Fachsprache / The International Journal of Specialized Communication*, under review.

> In everyday communication, figurative language is used to express emotions, value judgments and beliefs as well as to blend and create new concepts. In finance, metaphors often present messages of alarm in a soothing tone to overlook the cause of a problem and focus, instead, on a solution. The concept of *contagion* has recently entered discourse on international systems of regulation, finance and economics. We trace the emergence and successive use of the term in medicine and biology, to the social sciences and into the language of finance. We examine the use of *contagion* at word and grammatical levels and show how various patterns are used to elaborate particular features and diminish others. First, to look at the onset of the term, we track its use in *Annual Reviews* articles and journals of finance. We then present a corpus-based analysis of US congressional documents and compare them to medical reports from the World Health Organization and the Center for Disease Control. The results show that some lexical-pragmatic properties are carried over from the biomedical context while others are not, which has implications for the specialist language of finance and politics. In another analysis, we present a computational method based on word clustering in WordNet, in order to analyse how the context of *contagion* signals various metaphors in the congressional corpus [92]. The results show patterns in the metaphorical domains that *contagion* makes use of in figurative contexts. We conclude that the nature of *contagion*'s use in finance and politics is more complex than term-borrowing as it establishes a range of lexical, pragmatic and metaphorical properties.

The paper is an elaboration of the computational, corpus-analysis of *contagion* in finance and economics and is currently under review. The findings are based on those presented in the LSP 2013 paper above, [71], but with more explanation of the related literature, methods and implications. This paper is the first to explain and use the system presented in this thesis, *MetID*, and exemplifies its role in corpus-based analysis.

Chapter 2

Background & Related Work

2.1 Introduction

This research explores computational methods for automatically identifying, extracting and interpreting the use of figurative language in text. Automation requires clearly articulated methods and procedures, which can be implemented using properly curated data. The goal is to apply and extend current techniques in information extraction to better address figurative language. The inspiration for this work is in psycholinguistics and experimental psychology on one hand, and corpus linguistics and statistical semantics on the other. The combination of corpus and cognitive linguistics, computational semantics and natural language processing will be evidenced in the methods and implementation chapters. Here, I review the foundations of metaphor use and comprehension, previous computational work in the area and models of lexical semantics which may be helpful in designing a system for identifying and interpreting metaphor[1].

This section will review corpus linguistic work on metaphor, which provides a starting point for computational work. Section 2.2 provides a catalog of different kinds of figurative language and common classifications. I then turn to metaphor in language, its meaning and structure which will lead to an introduction of theories of metaphor use and comprehension. Computational work relating to metaphor will be reviewed in section 2.5 which will motivate the use of corpus-based semantic models, to be reviewed in the concluding section.

2.1.1 Corpus-Based Approaches to Metaphor

Some metaphor research has adopted corpus linguistic methodology [30, 34, 35, 45, 46]. Corpus linguistics is a field historically undertaken by lexicographers tasked with developing dictionaries and thesauri. It primarily consists of systematically sampling discourse to find patterns and habits of language users, and to track changes in lexica. In the last three decades, corpus-based techniques have become more widely available with increased access to literally endless text on the internet.

[1] Here, the term *semantics* refers to meaning in the lexicon (tokens of language use) [40], as opposed to truth indication or truth marking [36]. Though conceiving of semantics in these terms may be controversial, discussion at this level informs increasingly comprehensive theories of communication, memory and cognition – without regard to internalist considerations. For more on this idea and debate, see [20, 21, 37, 38].

The accessibility of data, combined with growing computational power, has enabled other research to adopt corpus-based methods for studying language.

Corpus-driven approaches to figurative language have also become increasingly popular [45, 46, 88]. Recent work has explored proposals that metaphors structure knowledge and thought by providing mappings between concepts – a theory which can be tested with corpus linguistic methods. This includes examining the "super-structure" of figurative language evident in linguistic instantiations of metaphor [45]. Other research has shown that properties of metaphors exhibit constraining relationships to their linguistic instances [93, 128]. For example, subject domains often show preference for certain metaphors – using organisms as companies, economies and countries [46, 168]. Another example is how metaphors of quantity-change tend to be realised as single verbs like *rise* and *fall* [69]. The latent structure of linguistic metaphors may also be available to corpus analysis: "collocating" metaphors may offer a measure of systematicity [45, p.219], a property proposed to define links between metaphors [133], but which appears less common in some domains [45, p.260]. The corpus approach is ostensibly well poised to address such discontinuities by tracking the development of linguistic metaphors and their conceptual counterparts.

Corpora are the first piece in developing a systematic analysis of language. The second is a method that exploits the availability of text from which to extract meaning and information and is capable of mitigating the effects of noise, sparsity and incompleteness of the data. A central question is: how can we most effectively represent meaning in text so that it can be analysed in large quantities over large collections? Here we find a sub-field known as *statistical semantics*, which is concerned with modeling lexical meaning using co-occurrence observations. The statistical (or "distributional") approach has made contributions, both theoretical and functional, to cognitive psychology, corpus linguistics and NLP (see [217] for a survey, [184] for a comparative case-study and [20] for an internalist critique). Broadly stated, distributional semantics is the analysis of lexical patterns as they relate to one-another in a series of documents or a stream of text, to construct a representation of lexical semantics. The typical unit of analysis is a word, and the typical data-structure is a matrix built from colligates and various contextual features (lexical, morphological, grammatical, etc.). Work in the field will be reviewed in more depth in section 2.6.

2.2 Catalog of Metaphors

Recall the four types of metaphor: noun-based, verb-based, analogical and complex. There are indeed more (and more useful) ways to delineate metaphor. This section reviews some ways to organise metaphor in relation to its linguistic instantiations and figurative language in general.

2.2.1 Noun- & Verb-based Metaphors

The distinction between noun- and verb-based metaphors is useful for automatic interpretation because it offers a surface-level distinction based on part-of-speech (POS). Metaphor is perhaps most simply explained using noun-based metaphors because the conceptual structure is evident in the linguistic instantiation. However, verb-based metaphors are perhaps the most common type and come in a range of forms [133, 122, 212]. Typically, the defining feature of a verb-based metaphor is that an object is doing something which it cannot literally do. Verb-based metaphors take a feature from a verb and apply it to an object. Take (i) for example:

(i) The boy got on his bike and *flew* home.

Here, *flew* could be lexically translated to *quickly went*, the metaphor being one in which fast movement is made sense of as flying. Because it is cumbersome to explain concept combinations apart from the instantiations, metaphors are denoted as TOPIC = VEHICLE or TOPIC AS VEHICLE, where the TOPIC is the object of interpretation and VEHICLE is the means by which it is interpreted. The example in (i), is said to "instantiate" the metaphor FAST = FLYING. Verb-based metaphors have been analysed in cognitive psychology [212] and computational linguistics [15, 69, 158, 201]. In computational research it has been proposed that verb-based metaphors constitute a violation of *selectional preference* [41, 199, 200, 230], which can intuitively be defined as an object's preference for selecting certain actions. (i) is a good example of such a violation, because boys (and bikes) tend not to fly. Selectional preference and violation will be expanded on later.

2.2.2 Metonymy & Metaphor

There is an interesting and complex distinction between instances of linguistic metaphor and metonymy. Metonymy is a semantic relationship between a thing and a referent where the referent is itself related to, but not exactly or entirely, the thing itself. Metonymy is best illustrated by example:

(ii) *The White House* is in talks with *the Kremlin*. (Meronymic metonymy)
(iii) We need a few more *hands* for this job. (Synecdoche)
(iv) *The ham sandwich* is waiting for their bill. (Ellipsis / synecdoche)
(v) Lend me your *ear*. (Partial functional metonymy)

Sentences (ii-v) are four kinds of metonymy (the metonymic referent is italicised). The common feature is that they each refer to something other than the actual referent. In (ii) *The White House* and *the Kremlin* are meronyms, presumably for representatives of America and Russia respectively. (iii) is a type of metonymy called synecdoche, which is defined by Kenneth Burke as a piece, or sub-part of something is used to refer to the whole [28]. In this example, *hands* refers to people, as in "helping hands". Like (iii), (iv) is also a form of synecdoche in which *ham sandwich* refers to a patron who, presumably, ordered a ham sandwich. Instead of a part of a person, the synecdoche is made by choosing a sub-aspect or sub-feature of a person, here as the food they ordered, to refer to them. The last example, (v), is similar to (iii) and (iv), but is perhaps closer to a typical metaphor. Here, *ear* cannot refer to a person, as *hands* do in (iii); the intended interpretation is more likely a request for attention – the *function* of ears. The metonymy, then, is between part and function: ear and listening. These forms of figurative language range from clear forms of metaphor to forms of figurative language that are not easily understood in the topic / vehicle structure of metaphor.

Metonymy can often function like metaphor, especially when the metonymic relationship crosses a conceptual boundary. Given the multitude of forms in which metaphor can occur, one could conclude that metonymy is a sub-set of metaphor. However, some counter-examples show that this is not always the case:

(vi) Please pass *the mustard.*

(vii) Can we hail *a taxi*?

Consider (vi) in which *mustard* likely refers to a container of mustard instead of the mustard itself. This form of metonymy is so common that it is nearly unavoidable in daily communication. Likewise, in (vii), people literally hail the *driver* of a taxi, not the *taxi* itself. These subtle figurative references seem to beckon redefining the meaning of common terms. For instance, defining a taxi to include a person, not just a car, would make it literal to hail one. Such an argument is beside the point: neither (vi) nor (vii) are metaphorical, because neither uses one concept to make sense of another, but they are clear instances of metonymy. The fact that some cases of metonymy appear to be straight-forward metaphors, like (iii) and (iv), while others are nearly literal, has lead researchers to propose a cline between metaphor to metonymy [10, 45, 46]. Still, many metaphors are not forms of metonymy and it is these types that will be explored in the current research.

2.2.3 Sense-based, Complex & Other Metaphors

Peter Stockwell proposes nine types of metaphor commonly found in literature (table 2.1) [206]. Stockwell offers categories of multi-word metaphors which are distinct from one another. While some are lexically instantiated, like pre-modification or paritive[2] / genitive[3] statements, others are grammatical metaphors such nominalisation of verbs and verb pre-modification. Also note compounds and blends are examples of morphological metaphors like “mind-scape” or “techno-babel”. This set of metaphors exemplifies the breadth and diversity available to figurative language.

The proposed research on identifying and interpreting figurative language in text will focus on two specific types: lexically instantiated noun- and verb-based metaphors. While this decision is motivated by its viability, it is also inspired by a generalisation of metaphor: that metaphor consists of mapping one concept onto another. Even when a linguistic metaphor is complex, subtle or metonymic, at some level it combines two concepts. Because noun and verb metaphors make this mapping more explicit (noun metaphors to a greater extent) they are a good place to start. Though this may seem like a simplistic starting point, computational literature has tended to separate noun-based metaphors from verb-based, examining one or the other (compare [15] to [200]). The current project explores a unified, lexical approach for both noun- and verb-based metaphors.

[2] $[_{DP}$ DET + of + $[_{DP}$ DET + N*]] in English.

[3] The syntactic case marking a possessive relationship between nouns.

Type	Part of Speech	Example
Simile, **analogy** and **extended metaphor**	Noun	The brain is like a city.
		It's oldest parts are surrounded by developments in its later evolution.
Copula constructions	Noun	The brain is a city.
Apposition and other **parallelisms**	Adjective	The brain, that teeming city...
		Into my mind, into my mental cityscape...
Paritive and **genitive** expressions	Noun	Paris is the city of my mind.
		In the streets and on the corners of my mind...
Pre-modification	Adjective	The urban brain
	Verb	A thinking city
Compounds and **lexical blends**	Morph	Mind-scape
	Morph	Metromind
Grammatical metaphor	Verb	The city considered the problem.
	Verb	The city sleeps.
Sentence metaphor (including negation)	Noun	This is the nerve-centre of the body.
Fiction and **allegory**		(A narrative in which psychoanlytical archetypes are figured as city land-marks and inhabitants.)

Table 2.1: Peter Stockwell's types of metaphor [206], examples of CITY AS MIND and MIND AS CITY.

2.3 Language & Metaphor

Ostensibly, metaphors are artefacts of language, akin to similes and poetic imagery, where a concept is expressed in terms of another. For example, one could say "my heart is on fire" to express intense feelings. Many metaphors are used to exaggerate, hide, highlight, broaden or otherwise change meaning in ideas. This intuition is what lead Locke to denounce the use of figurative language, stating that, "all the artificial and figurative application of words eloquence hath invented, are for nothing else but to insinuate wrong ideas [...]"[4]. However, there appear to be many concepts which cannot help but be described with metaphor – not the least are vague, abstract or fleeting concepts like emotions and dreams. Concepts like quantities, position, movement and temperature are expressed metaphorically more often than not [92, 128]. Additionally, there are systematic correspondences between different metaphors, such as feeling "up", "warming up to" or "boiling with rage". What makes these metaphors similar and how do people use this similarity? Are there mental, linguistic or communicative constraints on the process of interpreting metaphors? And are these processes informed by learning to communicate in a given language or a particular culture? Contemporary theories of metaphor seek to answer these fundamental questions.

One of the first cognitivist theories to address the question of metaphor comprehension was *substitution theory* which proposed that metaphors are understood by substituting figurative terms with literal counterparts, rendering a normative interpretation [92]. However, substitution theory fails to account for a number of properties of metaphor, such as the use of metonymy, the lack of distinct or sometimes *any* literal counterparts, peoples' speed of comprehension, cross-cultural

[4] *Essay concerning Human Understanding*, Book 3, chapter 10, page 105.

correspondences and systematic relationships among metaphors. *Interaction theory* was offered to address the weaknesses of substitution theory [17, 19]. Interaction theory introduced the idea that metaphors have a topic concept, analogous to the subject of a sentence, and a vehicle concept by which the topic is understood. It has been argued that the metaphorical nature of an utterance is marked by an interaction between the topic and vehicle, where they are likened to one-another by attribution or analogy [92]. The tensional theory holds that metaphor is marked by an emotional or logical tension between the vehicle and topic concepts. This tension is resolved by a language user's higher-level cognition. Another proposal is the *comparison theory*, which holds that topics and vehicles undergo a process of comparison [168], lexical and conceptual, after which a mapping or transference process projects aspects of the vehicle onto the topic, hilighting and hiding certain other features. None of these theories, however, addressed the systematicity between metaphors, nor did they address why (or how) so many metaphors are grounded in embodied cognition.

One influential theory of metaphor is *conceptual metaphor theory* (CMT) [132, 133]. CMT begins with the premise that metaphor is not purely linguistic. Instead, there is considerable evidence that metaphors have a conceptual super-structure (the elements of which are so-called *conceptual metaphors*) that accounts for the abundance and relatedness of metaphors in language. CMT appears to account for systematicity and polarity in metaphors across concepts, domain, language, culture and even modality [24]. It also defines metaphor as a concept-mapping, unrelated to linguistic instantiation. CMT is particularly helpful to the current research because it makes testable claims about the relationships between conceptual and linguistic metaphors. Further, because CMT proposes the existence of supra-metaphors, it provides a good foundation for an identification task. That is, CMT describes what exactly it is that *would* be identified in an identification task: a conceptual metaphor. CMT attempts to explain metaphor comprehension by proposing a set of mental objects (conceptual metaphors) that are called on to interpret linguistic metaphors. This model of metaphor processing leaves some open questions about how conceptual metaphors are constructed, related and put to use. One symptom of this weaknesses is CMT's lack of support for the results of experimental studies; it is not clear what mental processes or constraints are supplied by the structure of conceptual metaphors [81]. Cognitive psychologists, such as Sam Glucksberg, Dedre Gentner and others have carried out studies relating metaphor processing, use and development [24, 75, 76, 82, 118, 211, 235]. These enquiries have resulted in two dominant views of metaphor comprehension: category matching (cf. Glucksberg) and analogical reasoning (cf. Gentner), which will be compared in section 2.4.

2.3.1 Meaning in Metaphor

The key feature of linguistic metaphors is that they challenge a literal semantic interpretation. Take this example:

(viii) My butcher is a **surgeon**.

which is intended to be interpreted as "My butcher is **very good**" [171]. The sentence elicits an equation between *surgeon* and *very good*. The resulting interpretation is based on prior conceptual knowledge – namely about what butchers do, what surgeons do, and the expectation that they are different. Now note how transposing the nouns does not simply reverse the interpretation:

(ix) My surgeon is a **butcher**.

(ix) lends a different interpretation than (viii) – that the surgeon is bad at her job. The fact that butchers are not necessarily bad precludes a metonymic interpretation. Instead, a valid interpretation must commit to a metaphorical relationship between butchers, surgeons and features of their respective occupations. In this way, (viii) and (ix) use the same background information, but to different ends.

While (viii) and (ix) are relatively easy to interpret, some metaphors are more complicated. Consider (x):

(x) It will take **a lot** of patience to finish this [thesis].

At first glance (x) may not seem like a metaphor at all, but note the quantising of emotion. Lakoff and Johnson propose that this is an example of an *ontological metaphor*, more precisely referred to as a *quantifying metaphor* [133]. Ontological metaphors make normative claims about concepts, which in this example is EMOTION AS QUANTITY. Other examples from [133] include (xi) and (xii):

(xi) There is **so much** *hatred* in the world.

(xii) You have **too much** *compassion* for them.

As noted in the previous section, metaphors can use specific semantic relationships. For example, antonymy appears to be consistent in spatial metaphors of quantity and change [69]. These instances of spatial metaphor have also been shown to cluster in ways that correlate with the structure implied by CMT [68]. An important feature of linguistic metaphors is that their interpretation is not always available to a lexically constructed semantics. The semantics of metaphor often draw on syntagmatic, phrasal and circumstantial cues to guide an interpretation [92]. Making use of contextual and abstract information is a non-trivial computational task, but researchers have begun to address the interplay between symbolic theories of lexical semantics and embodied representations [146, 147, 149, 180]. The aim of such research is to explore ways to account for meaning in text. Linguistic metaphor offers a unique phenomenon with a complex, and sometimes under-defined semantics, that is tightly connected to non-linguistic conceptual thought. This makes metaphor a rather difficult phenomenon to address computationally, but one that is unique and interesting.

Contributing to a growing body of work on statistical models of lexical semantics, the current research seeks to combine such models with NLP techniques to test the bounds of computational metaphor processing (see [201], for an example). To lay the foundations for this computational undertaking, the next section will describe the structure of metaphor as it is instantiated in language and communication.

2.3.2 Structure of Metaphor

A metaphor consists of two pieces: a *topic* and a *vehicle*. In CMT these are concepts, instantiated as terms in linguistic metaphor. Consider (xiii):

(xiii) My lawyer is a shark. [81]

Here, *lawyer* is the topic and *shark* is the vehicle. The defining feature of metaphor is that a topic concept is being figuratively understood by using the vehicle concept. In (xiii), the interpretation is to understand a lawyer as if they were a shark – presumably having features of aggression and predation. Because topics are the objects of understanding, there appear to be few constraints on what they may be [92, 128]. Common topics include emotion ("*she* was deeply **moved**"), morality ("*he*'s a **straight** shooter") and economy ("**growth** of the *economy*"). Unlike topics, vehicles tend to be constrained by how well they work with a topic, how often they are used and how specifically they can be applied. Common vehicles include anatomy ("the **eye** of the *storm*"), plants and animals ("the **fruits** of our *labour*") and economics ("**spend** your *time* wisely") [128].

Lakoff and Johnson [133] survey metaphors used to convey things like numbers "rising", institutions "growing", emotions "flowing" and lovers "moving forward." In the same work, the authors develop a theory of *entailment* between conceptual metaphors. Entailments are constraints and implications that extend from the use of a particular metaphor, such as EMOTION AS LIQUID, that activate or require other metaphors like EMOTIONS ARE OBJECTS or perhaps FEELING AS TEMPERATURE. The organisation of metaphors, which is not unique to CMT or Lakoff's work, implies a network of interrelated metaphors. Consider the metaphors ANGER IS HEAT and EMOTION IS LIQUID. Combining them, the metaphor ANGER IS HOT LIQUID can be derived, which makes interpreting the following metaphor reliant on both:

(xiv) I was boiling with rage.

Lakoff's *Master Metaphor List* and Andrew Goatly's *Map or Root Analogies* are two resources which interrelate metaphors by entailments.

Goatly's theory of *root analogies*[5] proposes that there are irreducibly primitive metaphors on which most others rely [92]. The map is organised in two sets of target (topic) and source (vehicle) domains (table 2.2). Stressing that linguistic instances of are not chosen arbitrarily, Goatly shows that many metaphors draw on root analogies like HUMAN = PLANT, SPACE = TIME and SIMILARITY = PROXIMITY. Table 2.3 shows different kinds of linguistic metaphor and their defining elements as they affect potential methods of interpretation (which will be discussed in the following section). Figure 2.1 is one of Goatly's examples that shows the relationships between root analogies used to interpret ANGER = HOT FLUID IN CONTAINER [93]. Note that some relationships are not bi-directional. Goatly argues that the instantiation of a metaphor in language is constrained by conceptual features as well as concerns about linguistic processing and communicative efficacy. In other words, it is not enough to provide a taxonomy of linguistic-conceptual mappings, instead research should also be guided by linguistic constraints like grammar, convention and morphology.

[5] An interactive "map of root analogies" is available at http://www.ln.edu.hk/lle/cwd/project01/web/rootanalogy.html; 11 January, 2013.

	Source Domains (Vehicles)		**Target Domains (Topics)**
1.	Activity & Movement	A.	Things & Substances
2.	Human, Sense & Society	B.	Human / Animal Body & Sense
3.	(Living) Things & Substances	C.	Activity & Movement
4.	Values, Qualities & Quantities	D.	Space & Place
5.	Emotion, Experience & Relationships		
6.	Thinking, Communication		

Table 2.2: Top level source (vehicle) and target (topic) domains which organise the map of root analogies. Each sector on the map corresponds to an intersection of a source and target domain, where a number of constituent metaphors are found.

Example	**Unconventional Elements**	**Interpretive Elements**				
		Vehicle	**Topic**	**Actual referent**	**Actual referent**	**Similarity / Analogy**
He put his back against ***the suitcase***	Reference	suitcase	rock			Similarity
The building was ***a barn***	Reference	barn	=cathedral	cathedral		Similarity
the sardine tin *of life*	Reference Colligation	sardine tin	=life	life		Similarity & Analogy
John is ***a pig***	Reference Colligation	pig	greedy	John		Similarity
the ***naked*** *shingles (of the world)*	Reference Colligation	naked	uncovered	shingles	body	Analogy
the air was ***thick***	Reference Colligation	thick	?	air	solid/liquid	Analogy
her son had been ***damaged*** *in a crash*	Colligation		son	object		Similarity

Table 2.3: Various types of metaphor with topics underlined, vehicles in bold-face and their defining feature(s): unconventional colligation or reference. The topic and vehicle are listed, some of which are implied outside the context, producing differences between the actual and implied concept. The last column, based on the preceding features, is the expectation of how such a metaphor is processed (see section 2.4). Adapted from [92].

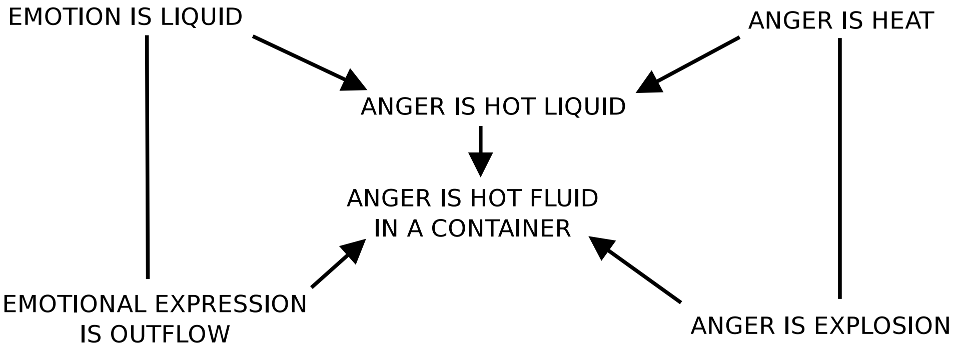


Figure 2.1: Root analogies, in the four corners, used to interpret ANGER = FLUID IN CONTAINER as the metaphors in the middle. Adapted from [93].

2.4 Theories of Metaphor Comprehension

Though both CMT and root analogies account for the ubiquity, interrelatedness and systematicity of common metaphors, neither explicitly address experimental evidence about how people create, use and comprehend metaphor [132, 133]. Some lexicalised or so-called “dead” metaphors may be understood colloquially as idioms while novel metaphors rely on more complex processes. Two theories have emerged addressing comprehension – the analogical process of structure-mapping (cf. Gentner) and category matching (cf. Glucksberg). As we will see, different metaphors appear to be processed in different ways, which has lead to hybrid theories.

2.4.1 Structure Mapping

In a model developed by Dedre Gentner and colleagues, known as structure-mapping, metaphors are interpreted by a process of featural comparison and projection. The structure-mapping model, which is based on analogical modeling, holds that analogies are made sense of by structurally aligning features [25, 76, 77, 235]. To process a metaphor, the structure of mappings and entailments are aligned between the topic and the vehicle, after which those left over are projected onto the topic, re-representing it in terms of the vehicle. This process of alignment proposes discrete systematic mappings (similar to CMT’s systematicity assumptions [132]) which may be abstracted to compensate for differences (similar to French and Hofstadter’s “conceptual slippage” [61]). The result of the alignment process is a structure with some elements apparent in the vehicle concept but not the topic. These remaining elements are projected to create potential interpretations of the topic, the more systematic and apt of which are kept.

CMT and parts of Goatly's theory of root analogies, claim that metaphor comprehension is an embodied process, reliant on abstract conceptual metaphors. Structure-mapping appears to be a good candidate for explaining how conceptual metaphors are compared. Also supporting the theory are results from analogy making and comprehension, which have found the a structure-mapping process accurately models results from controlled experiments [25, 76, 77, 211]. However, structure-mapping fails to account for cases where a topic is indirectly implied or obscured, which is common in some linguistic metaphors [81].

2.4.2 Category Matching

Sam Glucksberg and colleagues developed a theory of metaphor comprehension that involves category decisions as opposed to similarity judgements [80, 81, 82]. Instead of using the predicate-structure, arguments and entailments of metaphors, the category matching theory holds that metaphors are understood by category decisions for the topic and vehicles. The previously mentioned metaphor

(xv) My lawyer is a shark.

is understood by attributing concepts from the vehicle *shark* to the topic *lawyer*. Generally, the categorisation process is the attribution of the vehicle's super-ordinate category features to the topic. Here, a shark is a member of the *predator* category, a category "lawyer" is made a part of, thus making sense of the metaphor LAWYER = PREDATOR. In the process of matching a category, both the topic and vehicle provide constraints: the vehicle offers a set of categories, some of which are ruled out by the topic, based on relevance and systematicity. A category matching process has been shown to accurately predict how "is-a" metaphors, like (xv), are interpreted [81].

Category decisions have been found to invoke different cognitive processes than similarity decisions – especially in semantics [81, 98]. Though category matching fits well with experimental findings, systematicity and embodied coherence are not entirely accounted for [24]. Category matching has also been criticised for downplaying the role of the topic [25], despite offering a more explicit means by which *new* features are projected onto to the topic.

2.4.3 Hybrid Theories

Because there is experimental evidence for both structure mapping and category matching procedures in metaphor comprehension, it is apparent that some combination of the two processes must be involved in metaphor comprehension. Three views have been proposed to reconcile the two dominant theories: the conventionality view, the aptness view and the interpretive diversity view.

Conventionality View. The conventionality view [25], proposes that the conventionality of the vehicle mediates the comprehension mechanism: categorisation or comparison. This view claims that metaphors are initially processed by comparison (structure mapping) but metaphors with more conventional vehicles are processed by category matching. Conventionality refers to how well associated the metaphor's figurative meaning is to its vehicle [25, 110, 112]. The metaphor (xiii) is conventional because *shark* provides a salient property to *lawyer*. Alternatively,

(xvi) A fisherman is a spider.

is novel because it requires a comparative process to attribute aspects of spiders to fishermen. This view stresses the repetitious, figurative use of vehicle terms. The defining feature of a novel metaphor is a creative vehicle, one which is not often used figuratively. While the conventionality view offers a resolution to conflicting support for structure-mapping and category matching, it has been criticised for relying too heavily on features of the vehicle, and ignoring other aspects of the metaphor [83].

Aptness View. Glucksberg and Haught [83, 84] offered the aptness view, which holds that the aptness of a metaphor mediates its comprehension; apt metaphors are processed by categorisation while less apt metaphors resort to a comparison process like structure-mapping. Aptness is defined as the vehicle's ability to invoke metaphoric categories that capture salient features of the topic. For example, the metaphor

(xvii) The model is a rail.

is apt, and therefore processed by categorisation, because the salient feature of a rail (being thin) is aptly applicable to the stereotypically slender body of a model. In less apt metaphors the comparison process is invoked by a failure to quickly[6] find a salient category for the vehicle. The metaphor in (xvi) is not apt because *spider* offers relatively few applicable features to fishermen, thus, a comparison process is used to provide its interpretation, perhaps about how fishermen catch fish in nets like spiders catch insects [222].

Interpretive Diversity View. Akira Utsumi offers a third view to reconcile category matching and structure-mapping [220, 221]. The interpretive diversity view states that the *richness* of metaphoric categories invoked for the topic will mediate the comprehension process. Without appealing to lexical conventionality or the aptness of a vehicle, this view claims that the diversity of potential interpretations will mediate the comprehension process; diversely interpretable metaphors will be processed by categorisation while less diverse metaphors will be processed by comparison. Diversity of a metaphor refers to its semantic breadth: the quantity and uniformity of features invoked by the vehicle. Take the following metaphor as an example:

(xviii) My memories are money.

While (xviii) may appear neither conventional nor apt, requiring an attribute to mediate the comprehension process, in the interpretive diversity view, this phrase will be processed by categorisation. This is because *money* applies a relatively large number of properties to the concept of *memories*. Conversely, (xvi) is less diversely interpretable because *spider* invokes a relatively small number of properties.

[6] There is actually some evidence that both procedures, category matching and structure-mapping, may be activated in parallel, the winner being the process that finishes first [25, 84].

While this conception of a diversely interpretable metaphor being one in which different processing is required is intuitively plausible – it may be subsumed by the idea of aptness. While the original research into the aptness view did not use the term "interpretive diversity" it could certainly be argued that if a metaphor is diversely interpretable, then it is not apt. Utsumi makes an example-based distinction between the two ideas (compare (e) with (f) in table 2.4). In his examples, which are designed to highlight the different between interpretive diversity and aptness, it is still not entirely clear on what grounds (e) is apt. More generally, because the two views make similar claims about the default processing mechanism (category matching) if interpretive diversity and aptness are not unique, then interpretive diversity appears to be a special case of aptness.

	Conventionality View		**Aptness View**		**Interp. Diversity View**	
Example Metaphor	**Conventional**	**Process**	**Aptness**	**Process**	**Diversity**	**Process**
(a) *My job is a jail*	Yes	Categ	Yes	Categ	Yes	Categ
(b) *A gene is a blueprint*	Yes	Categ	Yes	Categ	No	Comp
(c) *My memories are money*	Yes	Categ	No	Comp	Yes	Categ
(d) *Birds are airplanes*	Yes	Categ	No	Comp	No	Comp
(e) *A goalie is a spider*	No	Comp	Yes	Categ	Yes	Categ
(f) *The supermodel is a rail*	No	Comp	Yes	Categ	No	Comp
(g) *A child is a snowflake*	No	Comp	No	Comp	Yes	Categ
(h) *A fisherman is a spider*	No	Comp	No	Comp	No	Comp

Table 2.4: Example metaphors, with defining features as proposed by the three hybrid views of metaphor comprehension. Adapted from [222].

2.5 Identification & Interpretation

There have been a number of computational explorations of metaphor ranging from analogy-making [61, 75, 235], to solving analogies [215, 225], finding idioms and stereotypes [22, 95], answering metaphorical questions [55, 156, 157], modeling categorical and analogical comprehension [121, 122, 222], finding conceptual metaphors in text [15, 158] and finding verb-based metaphors [199, 200, 201]. The breadth of research on metaphor is, in part, due to different goals, strategies and theoretical foundations. For example, Robert French and Douglas Hofstadter offer the *Tabletop* model of analogy-making [61], which they present as a cognitive model of creativity. In artificial intelligence, John Barnden's ATT-Meta project seeks to model metaphor understanding using contextualised reasoning and formal logic [7, 9]. Because NLP tasks are often concerned with making sense of naturally occurring text, many projects begin with *identifying* figurative language [15, 22, 41, 158, 182, 201, 218]. It is this strand of research the current project seeks to contribute to.

2.5.1 Strategies

There are three main strategies that address metaphor identification: sense-tagging, interpretation resolution and mapping. These three strategies loosely correlate to three tasks, respectively: word-sense disambiguation, semantic parsing and word clustering. While the work presented here draws on tools and techniques from each strategy, it is most similar to mapping / clustering strategy.

Sense Tagging

Sense-tagging, which consists of annotating a text with word-sense information, is one strategy used to identify figurative language. Here, the unit of analysis is a word, stem or lemma. Sense tagging exploits advances in word-sense disambiguation [119, 237] and co-reference resolution [141, 142, 179]. The goal of the task is to resolve which sense of a word is being used, given the context in which it occurs. Take the example, "That's a horse you can *bank* on." Here, the word *bank* is being used figuratively to mean something like "assuredly bet on". Firstly, this *bank* is a verb, which gives some indication of its sense. Additionally, *horse* is the indirect object of the verb *bank* – which signals an unusual sense of the word. With this, and perhaps other information, the sense implied by this instance of *bank* may be judged to be figurative, where *banking* is likened to investing as a bet.

Sense tagging usually relies on a pre-existing set of word-sense options, like entries in a dictionary [56, 172]. As we will see, a number of metaphors can be identified by inferring the correct sense of a word in context. This type of identification has been adopted by NLP projects [22, 218], however, the strategy will overlook metaphors spanning multiple words, phrases or sentences. The strategy may also overlook novel metaphors – those which are not common enough to have a specific word-sense. Take for example (xv), "My lawyer is a shark." Here, neither *lawyer* nor *shark* are likely to have a figurative sense in a dictionary, which makes the sense-tagging approach unhelpful.

Interpretation Resolution

A second strategy is to assume that something is a metaphor and try to resolve its interpretation. For example the lawyer/shark metaphor would not resolve correctly without making use of some mechanism for getting aspects of sharks and correctly applying them to lawyers. A resolution strategy has pedagogical advantages because the strategy itself has to implement a procedure for interpreting a metaphor as such. This strategy has been adopted by some NLP projects in question-answering systems [55, 156, 157, 215] as well as models comparing metaphor comprehension mechanisms [221, 222]. They have also helped validate theories of how people comprehend metaphors and have been used to analyse the use of metaphor in various kinds of text [15, 158].

However, without an adequate set of tests for both metaphorical and literal statements, a decision about if (and how) to correctly resolve a statement may be under-informed. Compared to sense-tagging, the strength of resolution-based systems is that they are inherently geared to finding and interpreting *novel* metaphors. Moreover, if the goal is to test a computational method of comprehending metaphors, adopting an interpretation resolution strategy will be fruitful, whereas if the goal is first to classify statements as literal or figurative, it may not be.

Mapping

A third approach to identifying metaphors in text is to find concepts which are lexically or grammatically related, and attempt to project their comparison onto a set of a mappings like conceptual metaphors or root analogies. This type of work uses linguistic research on how metaphors are instantiated in language [92, 128] as well as how metaphors comprise a set of a fundamental supra-metaphors [92, 132, 133]. The technique here is to maximise the likeness of a pairing found in text to a known metaphorical pairing. The degree, then, of similarity is a measure of how likely the text is an instance of the given metaphor. This technique can make use of a syntactic, grammatical and lexical patterns in the text, as well as semantic information, such as selectional preferences, pre-modifications and category assertions. The strength of the mapping approach is that it does not rely on a database of word-senses, nor on the assumption that a given statement is figurative. Also, the unit of analysis can be a pair of words, a relational triple, a phrase or a sentence.

The mapping approach requires some pre-existing knowledge about how metaphors commonly map topic and vehicle domains. In the literature, Lakoff's *Master Metaphor List* has been used [158], while others have either built their own [198] or used more general resources to indirectly represent the desired information [22, 218]. In addition to the set on which to map linguistic metaphors, this approach relies on solving an important sub-problem: word-word associations. To map a textual occurrence (presumably lexical) to a metaphorical concept, an association must be made between the observed lexica and the metaphorical terminology. This can be done in a number of ways, ranging from using a structured resource like WordNet [145], to using a database of word associations [124, 232], or using vector-space models [201, 209, 217].

2.5.2 Models of Metaphor Interpretation

MIDAS

One of the first computational models to address metaphor interpretation was James Martin's Metaphor Interpretation, Denotation, and Acquisition System, "MIDAS" [156, 157]. MIDAS assumes that metaphor is inherent in language use and is not anomalous. MIDAS uses a network of semantic and syntactic information about words and a hierarchy of conceptual mappings. Using these two resources, the system interprets a metaphor by building a local path of reasoning about words found in certain constructions. Because the model uses a structured set of metaphors, it is capable of determining if one is conventional or novel. In addition to interpreting observed metaphors, MIDAS can strengthen the paths it uses in the network to prioritise them for later use. In this way, given enough training, MIDAS has the potential to extend the set of metaphorical mappings and get better over time. One weakness is that it relies heavily on hand-coded knowledge in the semantic network. Arguably, MIDAS is more of semantic network that can learn to derive "correct" interpretations for metaphorical sentences than a model of interpretation.

Met*

Met*, which later became Met5, is a system to differentiate literal, metaphorical, metonymic and anomalous language [55]. Building on Daniel Fass' theory of collative semantics [54] and Yorick Wilks' preference semantics [230], met* uses a database of interrelated verbal and nominal "sense frames", that contains semantic information. These frames are used as a reference to determine if a grammatical argument is a metonymic reference, a metaphorical comparison or an anomaly. The use of semantic frames also assumes a lexical semantics can be adequately modeled as mental objects. Also, it does not appear people use a back-off or "default" mechanism to process figurative statements in the way met* does [81, 82]. From a computational perspective, the strategy is intractable with large amounts of data because it requires that word information be hand-coded in the semantic frames. This makes an approach like met* inadequate to meet our goals of the corpus-based approach.

ATT-Meta

John Barnden's ATT-Meta (ATTitudes and Metaphor-based reasoning) is a reasoning system for interpreting non-literal statements [8, 9]. ATT-Meta's strategy incorporates the idea of a "metaphorical view", which is a framework from which a statement can be accurately understood. By employing what the developers call common-sense logic in the form of graded-truth propositions, ATT-Meta assumes a statement is true, and by applying or changing the frame, it adjudicates a solution for processing a given statement. The logical changes the system makes to accommodate a statement constitutes its interpretation. ATT-Meta also implements a system of graded-reasoning which allows propositions to be held with a variable degree of certainty. The system represents a different strategy than MIDAS and met* because instead of using a semantic network, in which to explore figurative paths, ATT-Meta *simulates* metaphorical reasoning. The metaphor-ness is neither in the input or output, instead, it is represented by the logical steps and changes needed to resolve the statement. ATT-Meta relies on some hand-coded knowledge in the form of predicate and proposition logic and does not have a built-in system for learning new metaphors. However, it appears to be the only metaphor processing system which can *reason* in a metaphorical sense.

LSA-Based Simulations

A simulation by Walter Kintsch, based on *latent semantic analysis* (LSA) [121, 122], addresses some of the questions left open in MIDAS and met*. Kintsch constructs a semantic space using LSA and models the spreading activation between words to extract class-inclusion relationships for noun-based metaphors[7]. WordNet is then mined for possible semantic relations between topic-vehicle pairs that are ranked by prominence in the spreading activation network. Word-order constraints, which are known to have an effect on metaphor comprehension [82], are accounted for by using Kintsch's Constructive Integration model which adds a steaming, word-order component to LSA [122, 123]. The strength of this simulation is that LSA encodes multiple word-

[7] Kintsch notes that this simulation only addresses metaphors involving nominal topic and vehicle concepts.

senses as single entries in a semantic space, making a representation-level, unified distinction between figurative and literal language – one that is more plausible than sense interpretations in a knowledge-base of examples.

In an effort to test three theories of metaphor processing, the conventionality, aptness and interpretive diversity views, Akira Utsumi presents two models based on LSA [222]. The two models are meant to simulate category matching and structure mapping respectively. The first, *Categ*, which implements a category matching procedure similar to Kintsch's model described above, computes LSA vectors for the neighbors of the topic-term. From these, the closest words to the vehicle-term are used to build a centroid vector, which is used as the category. Note that this category may not be an actual word, but instead an abstract representation comparable to other vectors in the semantic space. The second algorithm, *Compa*, simulates structure mapping. *Compa* builds a set of intersecting topic and a vehicle neighbors: the set of overlapping neighbours for each word. Then the centroid vector is calculated between the topic and *all* vectors from the first step. After showing these two algorithms predict experimental findings for categorisation and comparison, Utsumi integrates them to implement the conventionality, aptness and interpretive diversity procedures. Using stimuli from other experiments [25, 83, 221], he finds that the aptness and interpretive diversity views are plausible.

Because Utsumi's models were built to validate cognitive theories, and not to actually find and interpret metaphors, they differ from other computational metaphor systems. The simulations only use novel, noun-based metaphors [222, p.274] despite reporting a spectrum of stimuli from novel to conventional [222, p.281]. The work does, however, exemplify the role of computational modeling in testing and verifying experimentally grounded theories. To this end, LSA shows a particular strength as a tool for similar research.

2.5.3 Models of Identification

Given the different ways to identify metaphor in text, there have been a number of computational projects addressing the task. Some current projects seek to bring automatic metaphor processing to a wider user-base, at a production level. This section surveys a few projects and publications which exemplify the diversity of techniques in metaphor identification.

MIP

The metaphor identification procedure (MIP) was developed to help language learners identify figurative language [177]. MIP proposes the following steps:

1. Read the entire document.

2. Determine the lexical units used in the document (open class words).

3. For every lexical unit, determine how it applies to an entity, relation or attribute evoked by the context of the word.

4. Determine if the unit has a more basic interpretation (more concrete or precise, historically older or related to bodily activity).

5. Given its relations, (3), if the unit has a more basic interpretation elsewhere, (4), then mark it as metaphorical.

Because MIP is geared toward informing a reader, not a computer, it is vague with respect to its operationalisation. For instance, the first step is important, because it gives a reader a sense of the discourse-level pragmatics, which can affect what kinds of linguistic metaphors are used [92, 128, 206]. It is not entirely clear if the intention of MIP is to be a step toward an automated procedure, but it provides a systematic way for people to detect figurative words. MIP is slightly more complex than a sense-tagging strategy, because it incorporates a degree of etymology and domain-specific knowledge. However, this makes it harder to operationalise for exactly the reasons it helps language learners.

MIP is important to the current project because it highlights the differences and similarities between a procedure for people to understand figurative language and one for computers. In doing so, it provides a touchstone for automation. For example, step 1 of MIP is to read the entirety of a document to get a "sense" of how words are used. This step is similar to building a semantic space for a corpus. The second step is relatively straight-forward for a computer. Steps 3 and 4 are perhaps the hardest to automate, making use of contextual knowledge, but as we will see, they have been addressed to varying degrees in computational projects.

TroFi

TroFi (Trope Finder) is a feature-based NLP model that can classify the use of verbs as literal or figurative – an example of the sense-tagging approach [22]. TroFi builds clusters around a seed set of nouns and verbs using an unsupervised algorithm. Then, TroFi uses features like proximate words, POS-tags and SuperTags[8] to train word-sense classifiers that "vote" on new senses. The authors report an average F-score of 64%, which is significantly higher than a baseline of about 25%. Though the results have room for improvement, TroFi exemplifies the NLP and sense-tagging approach to finding figurative words. In addition to the results, the authors have made their seed clusters available as the *TroFi Example Base*[9], which has been used in other sense-tagging approaches to metaphor identification [218]. TroFi takes a strictly NLP approach to figurative language, which makes the project attractive for tasks in information extraction and search engine design, but it departs from cognitivist theories of metaphor processing with its use of classifiers as opposed to semantic representations.

[8] SuperTags are a kind of high-level semantic tag in a structure similar to a localised parse-tree [5].

[9] http://www2.cs.sfu.ca/˜anoop/students/jbirke/; 22 January, 2013.

CorMet & CMI

Zachary Mason's *CorMet* is a system that calculates systematicity between noun-terms occurring in certain grammatical relations [158]. The relations provide a list of potential metaphors, the terminology of which is expanded using search engine queries. The expanded list is pruned by selectional preference, association and polarity and compared against Lakoff's master metaphor list [134]. Mason reports 72% precision in identifying correct conceptual metaphors, but does not report recall or F-score. CorMet's biggest limitation is the restriction to noun-terms in certain grammatical relations which overlook common lexicalised metaphors like "*falling* stocks" or "*rising* tensions".

A similar model is *computational metaphor identification* (CMI) [15]. CMI identifies computational metaphors by calculating asymmetric information transfer among synonyms from WordNet. The candidate metaphors are pruned by grammatical heuristics and systematicity calculations, similar to those used in CorMet. CMI uses WordNet again to calculate the closest superordinate categories and to pick a conceptual metaphor that most likely underlies the linguistic instance. CMI's developers note that the method is slightly intractable, taking a long time to generate results, as well as requiring a directed search for topic and vehicle domains.

Noun and Verb Clustering

A recently proposed technique for identifying metaphor is based on noun and verb clustering [199, 200]. Drawing on Wilks' theory of preference semantics [230], which implies that violations of semantic preference hallmark figurative language, the clustering technique operationalises the search for such violations. A seed set of topic and vehicle terms are first extracted from a subset of the British National Corpus (BNC) [198]. Then, using a spectral clustering algorithm, the seeds are expanded from their original set in subject- and object-verb relations to include similar words found in a given corpus. This set is used to determine the metaphor in phrases containing any of the terms mapping one cluster to another. The authors report significantly higher F-score in finding metaphors than a WordNet baseline.

The current project is most similar to this clustering approach, but with two important extensions. The first is the use of distributional and corpus-based semantic models for the word clustering operations. Second, seeds will be used from previous linguistic research into how metaphors are created and used [92], as opposed to the less structured set taken from the BNC.

2.5.4 Metaphor Processing in Real-World Systems

The projects described above address aspects of finding, extracting or simulating the interpretation of metaphors. Recently, a new task has emerged in the fields of computational creativity and natural language generation: metaphor as a service. By operationalising the *creation* of metaphor, researchers have built web-services that can mine concepts from text and create metaphors [226], provide figurative mapping using stereotypical knowledge [225] and generate idioms using the Google *n*-gram corpus[10] [165]. These systems build up conceptual knowledge using semi-structured text to extract typical lexical and semantic relationships among target words. Using statistical analysis of relationships found between topic and vehicle domains, the system *Metaphor Eyes*[11] can construct new, creative combinations with conventional knowledge. While computational creativity has been a field on the edge of artificial intelligence for some time, the use of NLP and data-mining techniques to generate concept combinations is a new and exciting field. These systems are available online as web-services, effectively providing on-demand creativity in the form figurative language.

Two large-scale metaphor processing projects have recently received investment from the American Intelligence Advanced Research Projects Association (IARPA): *METAL* [231][12] [13] and *MetaNet*[14] [15]. These projects both seek to accurately and flexibly automate the processing of metaphor for text-analysis systems. Because linguistic metaphor uses a plethora of contextual and cultural cues to produce new and novel interpretations, from a linguistic viewpoint, it can often obscure, complicate or highlight a language user's intention. These projects seek to build repositories and techniques to allow a system to extract metaphors from unannotated, naturally occurring text in English, Persian, Russian and Chinese. Detecting the use of metaphor and providing immediate interpretations is a big task and the METAL and MetaNet projects represent a long-term commitment to that goal. These projects exemplify what could become the first production-level metaphor processing systems, as they could be used for intelligence gathering and analysis, not just for academic research.

2.6 Lexical Semantic Models

A central task in identifying figurative language is modeling relationships between words. This task is the core of the word-clustering strategy adopted by the current project, described above. In terms of a computer system, a way to relate words to one another is required – namely between observed words in input text and terms from a set of seeds. There are a number of ways to accomplish this, ranging from purely statistical methods, to semantic-space models and information-theoretic formalisations of similarity. This section reviews three strategies for relating words: ontology-based resources, co-occurrence likelihood estimation and semantic space construction.

When defining word-word relations, people often appeal to definitional similarity, citing similar entries in a dictionary or shared synonyms in a thesaurus [187]. This assumption is not unwarranted, but it is only one way words can be related. Words can also share associations, featural similarity, function, grammaticality or specific semantic relations like synonymy, antonymy, hyponymy, etc. Dictionaries are a common source of structured information about words and offer

[10] http://ngrams.ucd.ie/idiom-savant/; 6 August, 2013.

[11] http://ngrams.ucd.ie/metaphor-eye/; 6 August, 2013.

[12] http://www.theatlantic.com/technology/archive/2011/05/why-are-spy-researchers-building-a-metaphor-program/239402/; 6 August, 2013.

[13] http://www.ihmc.us/news/20120529.php; 6 August, 2013.

[14] http://www.icsi.berkeley.edu/icsi/gazette/2012/05/metanet-project; 6 August, 2013.

[15] https://metanet.icsi.berkeley.edu/metanet/; 6 August, 2103.

one way to address word-similarity. However, using a dictionary, which usually contains definitions as well as grammatical and semantic information, relies on the research of lexicographers that can be outdated, under-informed, incomplete and in some cases, simply incorrect. Dictionaries and thesauri can be a good resource for a computational project, but usually lack consistent, formal and reliable structure.

To address inconsistencies and the lack of computer interoperability with dictionaries and thesauri, George Miller and associates developed WordNet [167]. Described as "a lexical database for English", WordNet is a semantic network of lexical entries with information about word-sense, part-of-speech, semantic type, frequency information, semantic relationships[16] and brief definitions. In essence, WordNet is a well-organised, strictly formatted dictionary which can be integrated with computer systems. WordNet version 3.1 will be used as one semantic model with which to help identify and interpret figurative language in text.

From a computational perspective, it is often desirable to reduce outside requirements and build word information from the ground up. One such strategy is to estimate the probability that words will co-occur, based on patterns observed in a set of documents. One of the first methods to operationalise this concept of "statistical semantics" is called *term-frequency inverse document-frequency* or *TF-IDF*. Intuitively, TF-IDF ranks a document's relatedness to a given word if the word occurs with disproportionate frequency in the document than it does in the whole collection [109]. Building on TF-IDF is a metric of word association called *point-wise mutual information* (MI) [39]. MI measures the proportion of two words co-occurring verses them occurring independently. As we will see, MI is a formalisation that can be used to associate words without appeal to an external lexicon. Recent, work in information extraction has developed new techniques which are faster, more accurate and more scalable than MI at estimating term co-occurrence [104, 105, 139, 164]. These co-occurrence estimation methods will be used as a statistical model of word association to help identify metaphor in natural text.

2.6.1 Distributional Semantic Models

The distributional approach, known variously as *vector space* or *semantic space* modeling, was first envisioned by Zellig Harris when he noted that words that occur near one-another often share a degree of similarity. Distributional models begin by using word frequencies to build a distribution vector. These distributions are combined to build a matrix of context-word frequencies, which can then be transformed and analysed to metricate relationships among words. One strength of the distributional approach is that it is ground-up and avoids the problems of a lexicon-based approach. For this reason, distributional semantic models are ostensibly better suited to addressing tasks of a subtle, vague or nuanced nature – problems such as figurative language.

Distributional semantic models have had success in simulating cognitive processes. LSA was initially introduced as a model of associative memory [136], but has also been used to model similarity in Shakespearean prose [148], conventionality in metaphor [147] and in information extraction [214]. Other models, such as the *hyperspace analogue to language* (HAL), have been

[16] WordNet 3.1 encodes relationships for synonymy, hyponymy, troponymy, meronymy and pertainymy.

used to cluster words based on conceptual and functional categories [149] as well as semantic categories [216]. More recently [184] showed that distributional models can cluster noun-concepts as effectively as feature-based models built with human responses. Using grammatical rules, the *Strudel* model can extract property-based concept descriptions from raw text [14] and appears to generalise to a range of tasks [12]. Specific to metaphor, Walter Kintsch and Akira Utsumi showed that LSA can simulate different mechanisms and strategies of metaphor comprehension [122, 222].

Models of distributional semantics differ in motivation, techniques, and of course, results, but they all share two traits. First, they use words as the unit of analysis. Second, they realise some degree of context (see [217]) which can be a number of things: a sliding window of adjacent words, a document, a sentence or phrase, a syntagmatic or syntactic relation, or an increment of text. Different models implement different strategies with regards to context – to be reviewed below.

Document-based Models

Document-based distributional models use a corpus of documents to contextualise a word's feature vector. Documents can be the results of web search queries, paragraphs of a larger text, or articles in a journal or news publication. Document-based models are popular because they exploit a naturally authoritative segmentation – realised in the choice of documents. That is, documents, paragraphs or sentences are all definitively related by the author and come with an inherent coherence. Because decisions about how best to segment text are not easy to make, they are sometimes left to the researcher. Prominent examples of document-based distributional models are LSA [136], *explicit semantic analysis* (ESA) [65] and Salton's *vector space model* [192][17].

Co-occurrence Models

Co-occurrence models define feature-vectors within the local context of a target word. This can be a sliding window, part-of-speech templates or dependency structure. Co-occurrence models have progressed, in part, due to advances in probabilistic parsers and machine learning algorithms [160]. The theory remains that words in similar context are similar themselves – co-occurrence models simply advance the notion of context. HAL [27] and the *correlated occurrence analogue to lexical semantics* (COALS) [188] are examples of co-occurrence models, both of which will be used in the current project.

[17] The term "vector space model" was originally used to describe Salton's particular model, but is often used more generally to describe semantic models which use feature vectors to represent words.

Approximation Models

Approximation models are an extension of co-occurrence models where, instead of explicitly tracking context, it is statistically approximated. Approximating co-occurrence can reduce computational complexity and increase the scalability of co-occurrence models [115], but the methods are numerous and diverse. *Bound encoding of the aggregate language environment* (BEAGLE) [111], a model which has seen success in recent research [184], *incremental semantic analysis* (ISA) [12] and *random indexing* [191] are representative examples of approximation models. Chapter 4 will describe implementations in greater detail.

2.7 Summary

This chapter introduced the complexities associated with a computational analysis of figurative language. Building on corpus-linguistic and psychological work concerning the meaning and comprehension of metaphors, the current project will combine and extend existing computational techniques of identifying and interpreting non-literal statements in text. The project adopts the strategy of mapping terms found in text to paired seed clusters constituting root analogies. This approach combines the CorMet and CMI methods described above [15, 158], which look for conceptual metaphors, with the clustering method of associating observed words with a set of metaphorical terms [199, 200]. This will involve addressing a number of NLP tasks, perhaps most importantly is to use a lexical model to build clusters around the terminology of root analogies. This task is one that may be best addressed using corpus-based semantic models. In addition to testing various models, the project also uses other techniques like dependency parsing, POS-tagging and selectional preference induction. This combination of corpus-based models and state of the art NLP tools will provide a testable system for addressing figurative language in naturally occurring text.

The remainder of this thesis will describe the methods, implementation, evaluation and contribution of a system called *MetID*. MetID attempts to identify and interpret both noun- and verb-based metaphor in raw text. The requirements are informed by the overlapping areas of research reviewed in this chapter, prioritising corpus-based techniques and considerations of computational feasibility.

Chapter 3

Methods

3.1 Introduction

This chapter outlines methods and resources required to identify, extract and interpret non-literal statements in raw text. These methods have been implemented in a system called *MetID*, which is described in the next chapter. The method is designed to operate on a stream of text drawn from any source: textbooks, magazines, newspapers, blogs and so on. The text will be pre-processed to normalise its format and remove non-linguistic markers (section 3.3). The resulting cleaned text can be used for two purposes: to build corpus-based semantic models and as the input to analyse for figurative statements. For the semantic models, the text is processed according to each model's specification. For the analysis, the text's morpho-syntactic frequencies and lower-level features (lemmas, parts-of-speech, grammatical relations, etc.) constitute the items to be analysed.

This approach to finding figurative statements uses word-clusters to rank potential interpretations. These clusters are built with statistical evidence provided by the various lexical models. Three types of model will be used, but the method is neutral with regard to what model is used, so long as it provides a score of relatedness between words. The first type of model is a lexicographic, database called WordNet. WordNet is more structured than general purpose dictionaries and provides a base-line to compare corpus-based models. The second type are the semantic space models, where co-occurrence distributions are used to construct high-dimensional spaces in which entries are related by a vector similarity measure (see section 2.6.1). The last type of model is based on estimating co-occurrence likelihood where the likelihood of two words co-occurring is a probability over prior observations in a set of documents (see section 4.4.5).

The method identifies potential metaphors by ranking how likely it is a given statement is an instance of a root analogy. A high-scoring statement, constitutes its identification and its corresponding root analogy is the interpretation – thus combining the identification and interpretation tasks. Potentially figurative word pairs are extracted and matched with entries in an external thesaurus of figurative concepts. Figure 3.1 shows the structure of the proposed method described in the rest of this chapter. The external resource provides two pieces of information: figurative terminology and mappings between topic and vehicle concepts. This information will allow the system to build clusters around the terminology and to pair those clusters using the mappings.

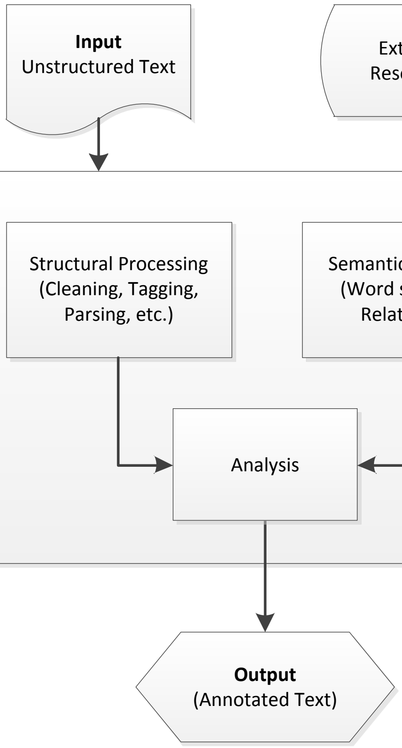

Input
Unstructured Text
External Resources
MetID System
Structural Processing (Cleaning, Tagging, Parsing, etc.)
Semantic Processing (Word similarity / Relatedness)
Analysis
Output
(Annotated Text)

Figure 3.1: The three main components of the proposed method are structural text processing, semantic processing and analysis modules. Structural processing examines the input and yields items to analyse for figurative content. The semantic processing module can be any lexical model that relates words observed in the input to entries in the external thesaurus. The analysis component combines the output of the structural and semantic modules to search for pairs of words that may comprise a metaphor.

After reviewing assumptions undergirding the proposed method, an overview of the architecture will be presented in the following section. The text processing sub-system will be introduced in section 3.3. Third, the three types of lexical model used in the semantic module will be reviewed. Last, the analysis will be explained with a focus on how the semantic and text processing modules are used in the paired search algorithm. The last section will also review some heuristics that augment the core analysis, along with their motivation and application. This chapter will conclude with a summary of how the proposed system will be implemented and evaluated.

3.1.1 Assumptions

The clustering approach makes two general assumptions about modeling language and meaning. First, that word meaning (ie. lexical semantics) can be modeled by corpus data without appealing to authoritative resources like dictionaries. The second assumption, which is the central proposal of conceptual metaphor theory, is that there are fundamental, extra-linguistic metaphors used to understand metaphor in text.

Corpus-based semantic models assume that text is evidence of meaning. This assumption has been supported by cognitive linguistic research [121, 146, 147, 180], but is also motivated by concerns of operationalising an algorithm in a computer system. The more independent the semantic model is from the intuitions of lexicographers and from a conception of "representative language", the more independent the method will be from the data. This method was developed with the assumption that metaphors are fundamental to conceptualisation and cognition – and that language, as with any concept, is a common way of instantiating them. It does not assert any theoretical claim about how these fundamental metaphors are created or processed – something conceptual metaphor theory attempts.

3.2 Architecture

The core algorithm focusses on using word-clusters to relate observed terms in the input to candidate metaphors (topic-vehicle pairs). Figure 3.2 shows this procedure where the topic and vehicle seed terms are used to build clusters of related words. The extent to which a pair of observed words is included in a pair of clusters, is proportional to the likelihood the observed terms are an instance of that root analogy. Inclusion in a cluster is measured as the inverse distance from the nucleus to a clustered word, which varies between models. This formulation affords two computational advantages: it produces a measurement of likelihood and it allows a pair of observed terms to instantiate more than one metaphor. Those it requires an exhaustive search of candidate metaphors, the results can be qualified by both their rank-order as well as individual scores for different candidate metaphors. This will allow two methods of testing the output: in terms of the score, and in terms of the top-scoring candidates. The initial seed terminology will be provided by a figurative thesaurus, *Metalude*, described in section 3.5.

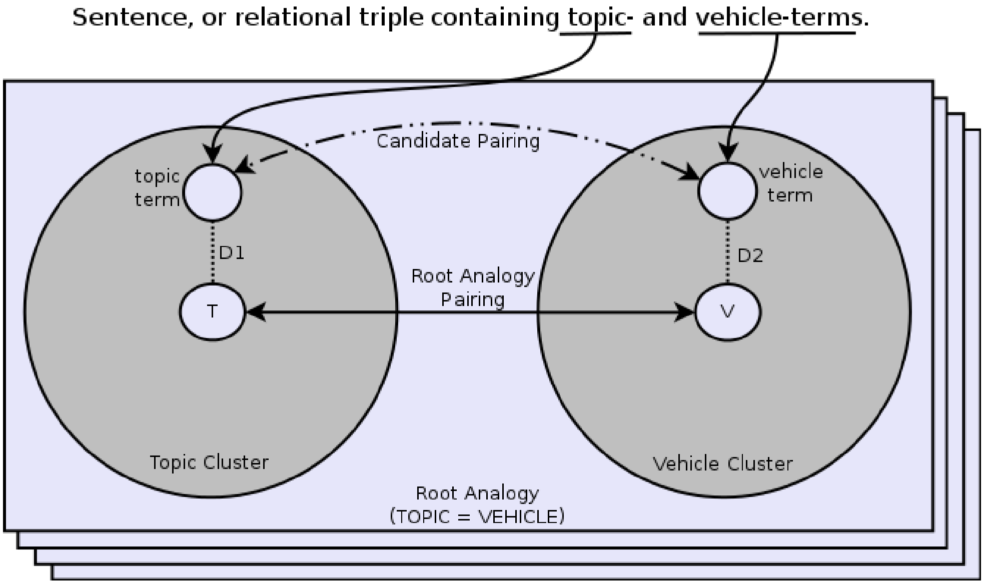

Sentence, or relational triple containing topic- and vehicle-terms.
Candidate Pairing
topic term
vehicle term
D1
D2
Root Analogy Pairing
T
V
Topic Cluster
Vehicle Cluster
Root Analogy (TOPIC = VEHICLE)

Figure 3.2: The core algorithm ranks candidate metaphors based on the distances D_1 and D_2, which correspond to the distance from a pair of identified topic and vehicle terms, to the their respective nuclei, T and V. The best candidate minimises the average of these distances. Note that the nuclei-pairing will be given by a database called *Metalude* and constitutes a root analogy [92]. The clusters for each nucleus, however, are built using an interchangeable lexical model. The score for a given input is the average of D_1 and D_2.

The word clustering task is different from some other word clustering tasks. Instead of using co-occurrence observations to find the best clusters, the system builds lists of the most related words or "nearest neighbours" to the nuclei. This task has been addressed in corpus- and non-corpus-based lexical models such as WordNet, LSA and others [3, 22, 184, 199]. To prioritise corpus-based models and to account for the variable relation between a clustered word and its seed, distributional semantic models (DSMs) appear to be a good option. In addition to being corpus-based, distributional models like LSA are known to represent associations beyond lexical semantic relations [12, 27, 111, 113, 136, 150, 188, 217]. It may be the case that DSMs are categorically the wrong strategy to account for metaphor. For this reason the method was designed to use *any* model that provides a measure of word relatedness. In addition to the DSMs, two other models will be explored: WordNet [167] and a purely statistical model based on co-occurrence likelihood estimation [104, 105]. These models are detailed in section 4.4.

The cluster generation and paired search are combined with a set of heuristics and comprises the semantic and analysis modules. The other piece of the system is concerned with extracting meaningful units of analysis from the input text. These items can be sentences, phrases, relational triples (as parsed dependencies) or just a pair of words. One goal of the system is to allow a phrase to instantiate more than one metaphor, which is in part addressed by the ranking approach instead of choice-based procedure. Additionally, by deriving multiple items to analyse from a given input, the method will not only look for every possible root analogy in *Metalude*, but it will look for them in different places. This means that a sentence can instantiate multiple metaphors in more that one way: by a term-pair providing more than one high-ranking candidate metaphor and by different items of analysis in the same input. These items are the result of a structural processing module that is outlined in the next section.

3.3 Text Processing

Before raw text can be analysed or used in a lexical model, it must undergo a series of processing steps. The first step is pre-processing and commonly involves removing control sequences (such as HTML mark-up), normalising the format, stripping punctuation, lexicalising symbols and discarding non-words. Pre-processing can be done quickly and efficiently using a series of scripts that perform each operation in sequence, resulting in a cleaned version of the input.

The second phase in text processing is to isolate separate sentences. This process, known as sentence chunking, will allow dependency parsing and POS tagging. The accuracy of sentence chunkers is about 95% and the best open-source tools can process thousands of words per second [142]. After chunking, the text-processing procedure bifurcates into two work-flows: one for using the text in the semantic models (see section 3.4) and one for analysing the text for candidate metaphors. For use in a corpus-semantic model, stop-words are removed. This consists of discarding words with little or no semantic information such as determiners, pronouns and connectives and has been shown to increase accuracy on a number of tasks [113, 137, 150, 188]. To prepare the input for analysis, instead of removing stop-words, the system annotates the text with POS information and parses each sentence into dependency trees [64]. This will allow some of the heuristics to operate on grammatical information.

3.4 Semantic Processing

The goal of the semantic analysis is to operationalise word relatedness. Computational semantic models will be used to build clusters around a set of seed terminology given by the figurative thesaurus *Metalude*. As discussed above, the system uses three types of model: a structured thesaurus-like database (WordNet), semantic space models (LSA and others) and a co-occurrence estimation model. Note the lexical model is arbitrary with respect to the overall method: it does not rely on a specific model. With the exception of WordNet, which will be used as a database, the models will produce persistent clusters for use during analysis. In addition to programmatic concerns about speed and storage, this will allow models' clusters to be independently tested and compared across different corpora, models and similarity measures. We wish to preserve each model's conception of relatedness as much as possible because relatedness means something different in, for example WordNet, than it does in LSA. This will allow each model to be qualified as it was designed to be used. This also means that clusters built with different models may be different in terms of the semantics they embody, thus affecting the performance of the whole system when using different lexical models.

3.5 Analysis

The central component of the analysis is matching the results of text processing with the paired clusters from the semantic processing. Because the clusters represent topic-vehicle mappings given by the figurative thesaurus, this search is slightly more complicated than for single words. Figure 3.2 shows the paired cluster search for an arbitrary input. This procedure will compute a score for each possible mapping in the thesaurus: namely the average of distances D_1 and D_2. Note that the input for this search need not be a sentence: it could also be parse tree, a relational triple or a pair of words. Figure 3.3 shows an example of the paired search for the input “My heart is on fire.” This analysis operationalises a semantic decision: the degree to which an observed word is included in a cluster. This decision is analogous to how close an *observed* topic is to a *known* topic concept (and likewise for vehicles). The intuition is that if the system sees two terms that might be a topic and vehicle respectively, then the degree to which they are included in a pair of clusters is how likely it is they instantiate that particular root analogy.

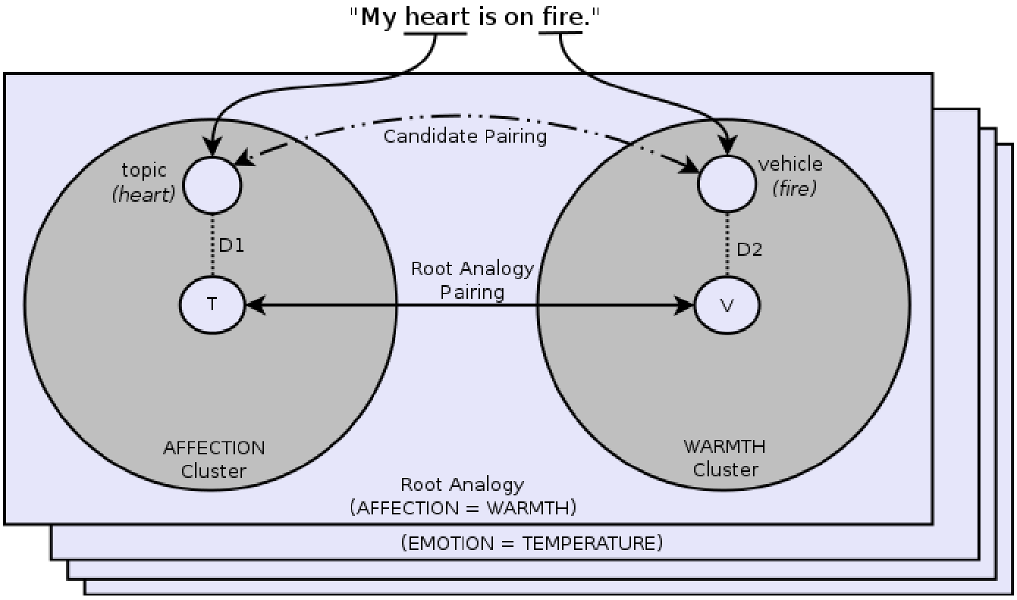
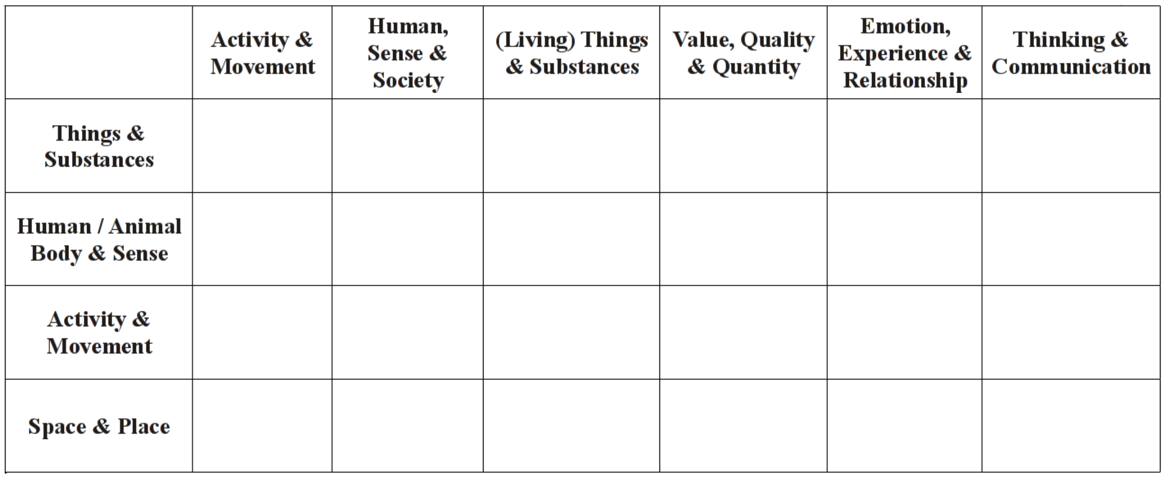


Figure 3.3: A candidate metaphor for the sentence “My heart is on fire” in which the metaphor’s topic, T, is AFFECTION and its vehicle, V, is WARMTH. The result for analysing a sentence, phrase or relational triple, is a list of candidate metaphors. Here the top-ranked candidate is AFFECTION = WARMTH because it minimises two distances: D_1 (*heart* → AFFECTION) and D_2 (*fire* → WARMTH).

Metalude

The analysis combines the outputs from the text and semantic processing components as well as a set of paired topic-vehicle concepts from a figurative thesaurus. The system will use *Metalude*[1], which consists of common mappings Andrew Goatly calls *root analogies*[92, 93]. *Metalude* organises metaphors in the *map of root analogies* and is based on so-called *conceptual transfers* (for example EMOTION = LIQUID ↦ *thinking & communication* are *things & substances*) [92, p.48]. Because Goatly's findings include a number of linguistic properties governing the creation, use and comprehension of metaphors, using *Metalude* will allow an analysis of linguistic metaphor. Figure 3.4 shows *Metalude*'s layout – the map of root analogies.



	Activity & Movement	Human, Sense & Society	(Living) Things & Substances	Value, Quality & Quantity	Emotion, Experience & Relationship	Thinking & Communication
Things & Substances						
Human / Animal Body & Sense						
Activity & Movement						
Space & Place						

Figure 3.4: *Metalude*'s map of root analogies [92], that proposes most linguistic metaphors are organised by these topic and vehicle concepts. In each cell are found a number of root analogies corresponding to the broader organisation given by the organising topic and vehicle concepts (columns and rows, respectively).

Heuristics

After the candidate metaphors have been ranked by how close an observed term-pair is to each pair of topic-vehicle clusters, a series of heuristics is applied. These heuristics will operate on the score, which up to this point, is only informed by the paired cluster search. Because every candidate metaphor will receive a score, no matter how low, the heuristics will account for various properties of figurative language without disregarding its defining feature: mapping a vehicle concept onto a topic.

The method uses three types of heuristics: lexical, grammatical and cluster-based. The lexical heuristics look for word patterns thought to mark, highlight or predict the use of a metaphor in text [92]. There is also a lexical heuristic which will penalise copula-style "is-a" metaphors if the predicate is a *literal* synonym or hyponym of the object (ie. the statement of a categorical fact as

[1] http://www.ln.edu.hk/lle/cwd/project01/web/rootanalogy.html; 1 February, 2013.

opposed to a figurative assertion). The grammatical heuristics account for syntagmatic features of linguistic metaphor. These include using a corpus-based measurement of selectional violation and predication strength to promote statements that violate statistical norms of observed text [182, 229, 230]. The last type of heuristic concerns the clusters with which the candidate metaphor was identified. If a statement fits closely to a topic-vehicle cluster pair (built with the semantic model) it will come with a high score. However, the clusters themselves can be independently qualified, the results of which will either promote a candidate metaphor (if the clusters are relatively good) or penalise it (if the clusters are relatively poor). Qualifying clusters will be described in section 4.4.6.

3.6 Summary

The proposed method consists of three main components. The first is a sequence of structural operations to extract meaningful items of analysis from unstructured text, as well as prepare it for use in a lexical model. The second piece is the lexical model, which is arbitrary as long as it offers a measure of word-word relatedness. This module will take seed terms from an external thesaurus and build clusters of nearest neighbors. The figurative thesaurus, *Metalude*, will provide two main pieces of information: the seed set of figurative terms and mappings between them. Together, the information from *Metalude* constitutes a set of root analogies [92] and will be used as the candidate metaphors. The final component is the analysis, where the paired clusters from the lexical model are ranked by how close a pair of observed terms are to a pair of nuclei. This paired cluster search will be run on each item of analysis (from the structural component) and produces a score for each candidate metaphor. This score is proportional to how well the observed terms fit each candidate metaphor. At this point, the semantic model has provided the only the input for the scores, which is the average distance between the observed topic to the candidate topic, and the observed vehicle to the candidate vehicle (figures 3.2 and 3.3). The last step is to apply a series of heuristics that use lexical, grammatical and cluster information to augment the scores.

This method addresses the goals of the current project. First, it allows testing of different lexical models (structured, semantic spaces, statistical, etc.). Second, and perhaps more importantly, the method allows a sentence, phrase or pair of words to instantiate more than one metaphor. In fact, the method forces this to happen because it will rank candidate interpretations by score. This score will be augmented with heuristics that either promote or penalise the likelihood that the given input is indeed figurative. As we will see in the next chapter, which describes an implementation of this method, the heuristics provide a significant amount of information to resulting interpretations. After describing the implemented system, MetID, its performance will be evaluated on two tasks of metaphor identification and interpretation, after which a case-study applying the method to terminological research will be presented.

Chapter 4

Implementation: The MetID System

4.1 Introduction

MetID is an implementation of the method describe in chapter 3 and consists of three main components: a structural module, a semantic module and a module for analysis. This chapter describes the design and development of MetID and will explain each module in detail. The system is implemented as a set of Ruby scripts that interact with a relational database. The main script, which performs the analysis, accepts a series of statements and will provide a spreadsheet-like text file containing the results. These results (described in the appendix B) can be perused manually, or analysed by other scripts. MetID and its MySQL database are available online at www.scss.tcd.ie/˜gerowa/metid.

Figure 4.2 elaborates on the method introduced in the previous chapter, and will frame discussion throughout this chapter. The structural processing, word clustering and metaphor identification components are implementations of the text processing, semantic processing and analysis modules (figure 3.1). The chapter will begin by introducing the work-flow from the input's perspective which includes a description of the structural module. Next, text-processing for the semantic and analysis modules will be described (section 4.3). The semantic module will be explained in general (section 4.4), and each lexical model will be discussed in section 4.4.6. Last, the implementation of the lexical, grammatical and cluster-based heuristics will be explained in section 4.5.

4.2 Work-flow

MetID consists of three main components: structural, semantic and analysis sub-systems. The structural module is used to extract items of analysis to relate to the semantic analysis. This process consists of extracting words, parts-of-speech and grammatical relations. Take, for example, the following sentence:

(xix) My heart is a fire that burns with love.

(xix) instantiates the metaphor FEELING AS TEMPERATURE and more specifically PASSION AS HEAT. Before MetID can look for associations between *heart* and FEELING or *fire* and TEMPERATURE, it must decompose the sentence into more primitive structures. To do this, the system applies a series of structural operations[1].

POS Tagging. The first thing the system does with a sentence or phrase, is tag it with part-of-speech (POS) information. MetID uses the Stanford Tagger[2]. After tagging (xix), the nouns *heart*, *fire* and *love* are identified, as well as the verb *burns*. POS information is used in the heuristics.

Dependency Parsing. In addition to POS-tagging, MetID parses a sentence into collapsed dependency relations using the Stanford Parser[3] [33]. These dependencies are organised as a parse-tree that realises recursive linguistic structures like phrasal verbs, prepositional phrases, embedded clauses and so on. MetID will use this dependency structure to inform the grammatical heuristics.

Stemming. Where appropriate, word stems are used in place of words' observed forms. MetID uses a Ruby implementation of the Porter stemming algorithm [176][4]. Stemming the input allows the system to relate observed words to any of their forms in the lexical models.

At this stage of analysing (xix), the system has a set of items to analyse for potentially figurative content. These include the sentence as a whole, a predication between *heart* and *fire* (extracted using the POS-tags) and a set of dependencies from the parser. This list is then pruned to include only open-class words and dependencies that relate two open-class words[5]. Lastly, all root-nodes, determiners, anaphora, pronouns, quantifiers, relative modifiers and coordinations are removed. Next, MetID analyses each item using the paired cluster search and the heuristics. In our example, the set of open-class words in the full statement is the first item to be analysed. Table 4.1 lists each item to be analysed and from which operation the information was retrieved.

For each potential topic-vehicle pair, MetID searches the candidate topic-vehicle clusters from *Metalude*, minimising the average distance between the identified term and its candidate nucleus (D_1 and D_2 in figures 3.2 and 3.3). In this example, the system first looks for cluster-pairs which

[1] "Structural" means without respect to semantics.

[2] http://nlp.stanford.edu/software/tagger.shtml; 10 February, 2013.

[3] http://nlp.stanford.edu/software/lex-parser.shtml; 20 February, 2013

[4] https://github.com/aurelian/ruby-stemmer; 10 February, 2013.

[5] The list of closed-class words is available at https://www.scss.tcd.ie/˜gerowa/metid/stopwords.txt.

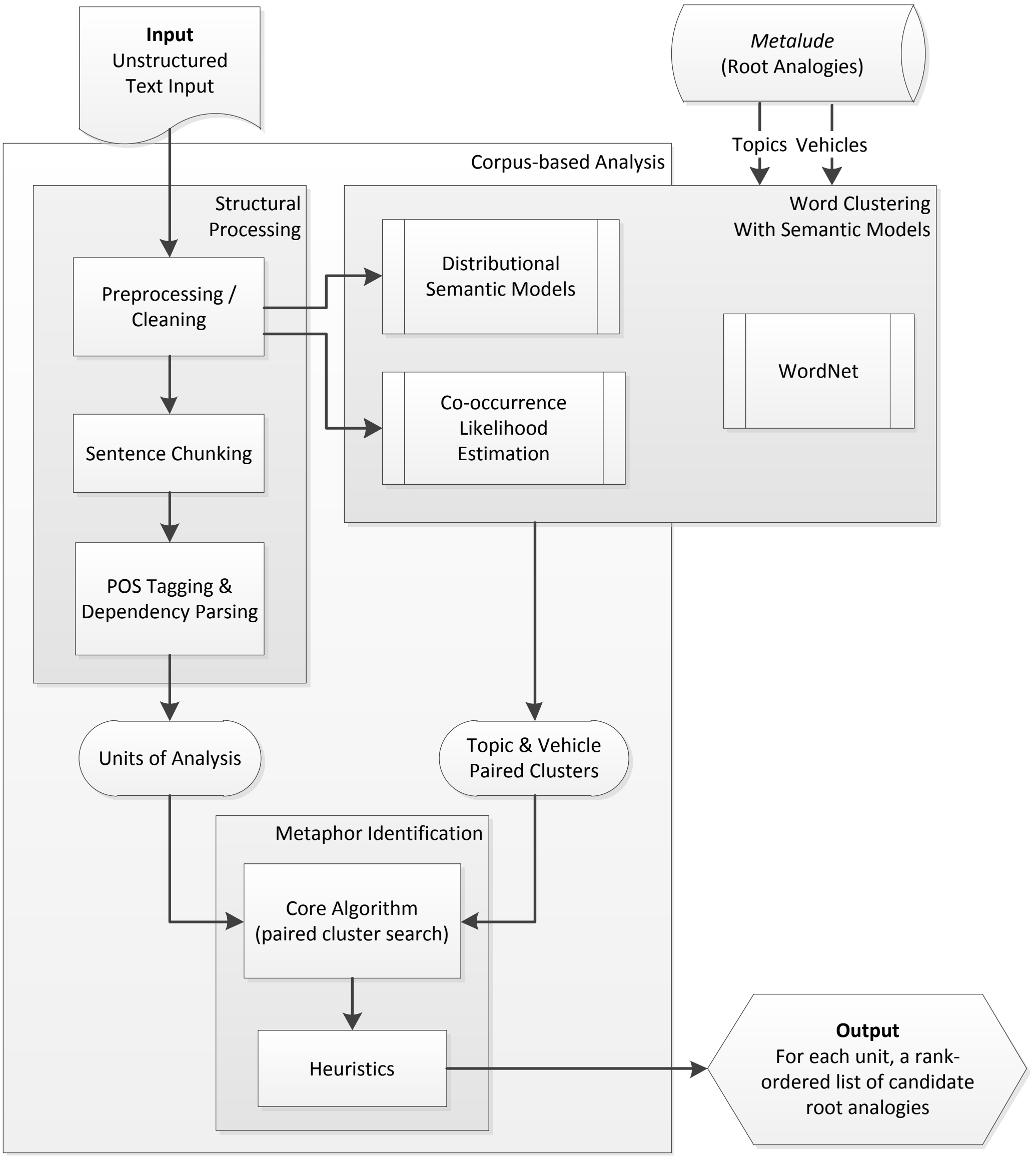

Input
Unstructured Text Input
Metalude
(Root Analogies)
Topics
Vehicles
Corpus-based Analysis
Structural Processing
Preprocessing / Cleaning
Sentence Chunking
POS Tagging & Dependency Parsing
Word Clustering With Semantic Models
Distributional Semantic Models
Co-occurrence Likelihood Estimation
WordNet
Units of Analysis
Topic & Vehicle Paired Clusters
Metaphor Identification
Core Algorithm (paired cluster search)
Heuristics
Output
For each unit, a rank-ordered list of candidate root analogies

Figure 4.1: Architecture of the MetID system. The two sub-systems for structural and word clustering are done within the confines of the text itself, unless WordNet is used as the lexical model. The structural analysis (left side) extracts syntactic and lexical information. The word clustering sub-system uses an interchangeable lexical model to build clusters around seed terminology from the figurative thesaurus *Metalude*. The last piece is the analysis module (bottom), which performs the paired cluster search and applies a set of heuristics. The output is a list of candidate metaphors ranked by scores from the metaphor identification sub-system.

Item	Potential Topic	Potential Vehicle	Relationship	Source
1*a*	heart	fire	NA	Stemmed sentence
1*b*	heart	burn	NA	Stemmed sentence
1*c*	heart	love	NA	Stemmed sentence
1*d*	fire	heart	NA	Stemmed sentence
1*e*	fire	burn	NA	Stemmed sentence
1*f*	fire	love	NA	Stemmed sentence
1*g*	burn	heart	NA	Stemmed sentence
1*h*	burn	fire	NA	Stemmed sentence
1*i*	burn	love	NA	Stemmed sentence
1*j*	love	heart	NA	Stemmed sentence
1*k*	love	fire	NA	Stemmed sentence
1*l*	love	burn	NA	Stemmed sentence
2	heart	fire	Predication	POS-tagged sentence
3	fire	heart	Nominal subject	Dependency parse
4	burns	fire	Nominal subject	Dependency parse
5	burns	love	Preposition (with)	Dependency parse

Table 4.1: Items to analyse in sentence (xix). Each item contains a potential topic, a potential vehicle and in some cases, the relationship in which they were observed.

contain *heart* and *fire* respectively in item 1*a*. The algorithm is an exhaustive search which means the system calculates the distances for every item for every candidate pair, and sorts the results. This is done for every ordered-pair of words, however, some of them (and potentially all) may not yield any candidate metaphors. This occurs, for example with item 1a, when there are no paired clusters where *heart* is in the topic cluster and *fire* is in the vehicle cluster. If this happens, MetID will allow unpaired clusters for topic-vehicle mappings but will apply a penalty to the resulting score (see section 4.5).

Now MetID has a list of candidate metaphors for item 1 – the sentence – in descending order by the average of the topic-term / topic-nucleus and vehicle-term / vehicle-nucleus distances[6]. A candidate metaphor consists of a candidate topic and vehicle, expressed as TOPIC = VEHICLE. The items of analysis, in this case the cleaned sentence, are made up of identified topics and vehicles. The clusters, then, constitute the relationship between an identified topic and the candidate topic (the nucleus) and the identified vehicle and candidate vehicle.

After the current item is used to build the list of candidate metaphors, heuristics are applied. Some of the heuristics are not relevant because the item is the set of all word-pairs in a sentence, not a relational triple like items 2 - 5. The ones that may apply in this case include a bonus if the stem of the identified topic matches that of the candidate topic and/or if the identified vehicle stem matches the candidate vehicle. After the heuristics are applied, the results are sorted by their new scores and written to the output file[7], and MetID moves to the next item.

[6] Distancing in the semantic models is defined from 0 (completely dissimilar) to 1 (completely similar). Methods for measuring distance will be covered in section 4.4.4.

[7] This second sort is a technical vestige of an earlier version of the system that did not implement any heuristics. It does, however, allow a threshold to be set by which to prune low-scoring candidates coming out of the core module to skip the heuristics module for candidates that are unlikely to end in a high score. This feature was implemented in

The second item is `pred(heart,fire)` refers the predication of *fire* and *heart*. The process for each item is similar, except that with relational triples, the first step is simplified because there is only one ordered pair of words. MetID ranks the candidate metaphors by the best topic-vehicle pair that minimises the within-cluster distance for *heart* and *fire*. Again, if no paired clusters are found, it will allow unpaired matches with a penalty to resulting score. After the initial list is created, the heuristics are applied. In addition to the those mentioned for item 1, two more heuristics may apply to predications: the predication strength bonus and the hypernym penalty (described in section 4.5.2). The predication strength heuristic uses a set of predications, extracted from two large corpora[8], to assess how unlikely the observed predication is. The second heuristic, specific to predications, is the direct hypernym penalty. In this case, if the predicate *fire* is a hypernym of *heart*, the system would apply a penalty to the score. MetID uses WordNet for its hypernym database, regardless of which lexical model is configured. In this instance, *fire* is not a hypernym of *heart*, so no penalty is applied. If, for example, this item had been `pred(lawyer, professional)`, this heuristic would have applied a penalty to account for the fact that there is nothing figurative about saying a lawyer is type of professional.

The remaining three items are the relational triples from the dependency parse: `nsubj(fire, heart)`, `nsubj(burns,fire)` and `prep_with(burns,love)`. Each of these are processed like the first two items, except that a heuristic for selectional preference violation may be applied. Selectional preference is a measure of how likely an argument is for a root in a given relation. Although this association has been traditionally used in noun-verb relations, MetID implements a generalised version that includes subject-verb, object-verb, adjective modification and noun modification. This heuristic is described in section 4.5.2, but in this example, `nsubj(fire,heart)`, would constitute a selectional violation because hearts do not tend to be fires. This heuristic is a weighted bonus, similar to predication, and is proportionally applied relative to the degree of violation.

MetID was designed to allow grammatically unrelated word-pairs to constitute a metaphor, a decision motivated by theoretical and technical considerations. To take the proposal of contemporary theories, that metaphors are not just superfluous linguistic flourishes, then were MetID to rely exclusively on linguistic structure to identify potential metaphors, it would risk precluding a variety of metaphors. Also, there are more relations that signal figurative language than those extracted from a grammatical analysis, such as semantic and pragmatic relations. Thus, expanding the scope of analysis to include such signals, will result in broader coverage of potential metaphors. Unfortunately, state-of-the-art semantic parsers (not to mention an almost complete lack of pragmatic or discoursive parsers) are not yet reliably accurate. Syntactic parsers, like those used in MetID, offer a first step toward exploiting relational structure in language to identify metaphor. In future work, were semantic parsers more robust, one can imagine discarding the bag of words analysis in favour of a set of higher-level relations.

MetID but not used in reporting results.

[8] The BNC and enTenTen collections.

After each item is analysed and subjected to the relevant heuristics, the results are written to the output file. By default, each item of is recorded with its 20 top-scoring candidate metaphors. This verbosity addresses the goal of allowing a statement to instantiate multiple metaphors. However, a user may choose to disregard the item-wise distinction and reorder the output to get the highest scoring candidates without regard for which item instantiated the metaphor. At the end of execution, a user is left with a file containing the data mentioned thus far, as well as a transcript of the execution. Appendix B provides a detailed example run of MetID.

4.3 Text Pre-Processing

Texts are pre-processed to prepare them for use in the semantic models and in the structural module. The pre-processing routines ensure a degree of normalisation and help reduce various kinds of noise, such as non-semantic variation (punctuation, regional or genre conventions, etc.). Keeping with recent literature in distributional semantic models, MetID adopts the steps used in COALS [188, p.9]:

1. Remove non-text characters (HTML tags, images, table lines, etc.)
2. Remove non-standard punctuation and separate other punctuation from adjacent words.
3. Remove words over 20 characters[9].
4. Split words joined by certain punctuation and remove other punctuation from within words.
5. Convert to upper case.
6. Lexicalise symbols.[10]
7. Lexicalise special patterns.[11]
8. Discard documents with fewer than 80% valid words to assure the text is in English.
9. Discard duplicate articles with a hashing algorithm.
10. Split hyphenated words that are not in a dictionary but whose components are.
11. Remove a common set of stop-words.

[9] [188] finds that this helps mitigate made-up or mis-spelled words.

[10] For example, @ becomes <AT> and % becomes <PERCENT>.

[11] For example, http://www.example.com becomes <URL> and nobody@nowhere.com becomes <EMAIL_ADDRESS>.

For this excerpt of the TASA corpus, the input (top) results in the normalised output (bottom):

> Who were the first Americans? Many, many years ago, perhaps 35,000 years ago, life was very different than it is today. At that time, the earth was in the grip of the last ice age. There were few people anywhere in the world, and none lived in the Americas. People did live in Asia, however. And some of them wandered into north America. The firstcomers did not know they had found a new continent.

⇓

```
AMERICANS ?
, YEARS AGO , 35 , 000 YEARS AGO , LIFE VERY DIFFERENT TODAY .
TIME , EARTH GRIP ICE AGE .
PEOPLE WORLD , LIVED AMERICAS .
PEOPLE LIVE ASIA , .
WANDERED NORTH AMERICA .
KNOW NEW CONTINENT .
```

4.4 Word Clustering

Building clusters around *Metalude*'s terminology is the core of MetID's approach to identifying figurative language. These clusters embody the models with which they were created. That is, the lexical models are built, used to create the clusters and discarded[12]. The models used for building clusters fall into three categories: WordNet, distributional semantic models (DSMs) and a co-occurrence likelihood estimation (COLE) model. The DSM and COLE models are corpus-based, in that they build word associations by analysing collections of text. This means that corpus-based models will produce different associations depending on what corpus is used and how it is structured.

4.4.1 Seed Terminology

After obtaining access from Andrew Goatly, the data from *Metalude* was scraped from its website[13]. A total of 594 root analogies were extracted, and saved in a relational database. Topic and vehicle terms were separated from each mapping and those consisting of multi-word terms were condensed to single words. The resulting set contained 582 topic-vehicle pairs, consisting of 487 unique terms and 479 unique stems. These terms, which are listed in appendix C, table C.1, make up the words around which clusters were built.

[12] Saved semantic space files for each model are available at https://scss.tcd.ie/~gerowa/metid/.

[13] http://www.ln.edu.hk/lle/cwd/project01/web/internal/database.html; 2 February, 2013.

4.4.2 Hold-out Procedure

MetID can be used to look for figurative language in the same corpus it uses to build a semantic model. Because the clusters are built using a corpus, they can potentially become representative of a metaphor as if it were a literal relationship. For example, the term *contagion* is strongly related to the concept of DISEASE, which is part of the motivation for using it to describe problematic relations in finance and economics. However, this figurative use of *contagion* would be obscured in a corpus in which it was *only* used in this way. For this reason – which is analogous to model over-fitting in classification tasks – requires a hold-out procedure to separate a corpus for building a model (training) and analysing it (testing).

Each of the text collections are made up of documents, which allows for a simple hold-procedure following the customary 7-to-10 ratio commonly used in machine learning tasks [102, 233]. To train a "held-out" model, a random 70% of a collection's documents were selected and the remaining 30% were used for analysis. Henceforth, a "held-out" model is one built on 70% of the documents in the collection. To examine whether this hold-out procedure was helpful or necessary, both held-out and full-corpus clusters were built. The procedure is the same for both types, and they were stored in the same way. Note that this hold-out procedure is not applicable to the WordNet model, because clusters are not built from a corpus.

4.4.3 Building Clusters: WordNet

WordNet is a structured lexicon of English words made available as a computer database [56, 167]. The current release, version 3.1, contains entries for approximately 155 thousand open-class nouns, verbs, adjectives and adverbs. It encodes different word-senses (including multi-word terms) with small definitions called glosses. For nouns and verbs, WordNet contains the following semantic relations: hypernymy / hyponymy, synonymy, antonymy, polysemy, meronymy, holonymy and troponymy. WordNet has additional information for entailment, pertainment, verb frames, attributes, morphological forms, coordinate terms and familiarity. Nouns are organised in a hyponym tree and verbs in a troponym tree (see figure 4.2). Table 4.2 reviews WordNet's coverage.

***risk*-[noun#1]**	***risk*-[verb#1]**
hazard, jeopardy, peril, **risk**, endangerment	**risk**, put on the line, lay on the line
→ danger	→ try, seek, attempt, essay, assay
→ causal agent, cause, causal agency	→ act, move
→ physical entity	
→ entity	

Figure 4.2: Examples of the hypernym tree for the first sense of *risk*-[noun] (left) and the troponym tree for the first verb-sense of the same word (right). Observe the multi-word terms, as well as the increasing abstractness at the top-levels of the ontology.

WordNet was initially chosen to provide a kind of baseline to the other models because it explicitly represents word-senses and semantic relationships. This is important in relation to distributional semantic models where word-sense information is encoded abstractly in a semantic space. WordNet is maintained by lexicographers who decide what word senses to include and how they interrelate. WordNet was included to provide an explicit semantic alternative to the corpus-based, distributional methods in other models, some of which have been shown to outperform WordNet [136, 184, 199, 200]. As we will see, however, in the tasks used to evaluate MetID, WordNet performs comparably well to the best DSMs and provides consistently good coverage.

POS	**Unique Entries**	**Synsets**	**Total Word-Sense Pairs**	**Monosemous Words & Senses**	**Polysymous Words**	**Polysemous Senses**
Noun	117,798	82,115	146,312	101,863	15,935	44,449
Verb	11,529	13,767	25,047	6,277	5,252	18,770
Adjective	21,479	18,156	30,002	16,503	4,976	14,399
Adverb	4,481	3,621	5,580	3,748	733	1,832
TOTAL	155,287	117,659	206,941	128,391	26,896	79,450

Table 4.2: Coverage in WordNet 3.1.

Similarity & Relatedness in WordNet

Similarity and relatedness are not inherent properties of WordNet, which relates entries by sense, part-of-speech, short definitions and specific semantic links. To address word relatedness in WordNet, researchers have developed measures that use its structure and information (see [173] for a review). There are two types of association measures for WordNet: similarity and relatedness. Similarity measures metricate the *information content* of the least common subsumer (LCS) between two words[14]. That is, similarity between two words is informed by their shared semantic information evident in the hierarchical structure. Alternatively, relatedness measures associate entries in a less strict sense – often using sense, gloss and frequency information. Table 4.3 summarises common measures of word association in WordNet.

Perhaps the most commonly used similarity measure in WordNet is Lin similarity [143, 144, 145, 173, 200, 201]. Lin similarity is similar to Resnik and Jiang & Conrath measures in that it uses the hypernym tree to find the information content of the LCS between two words, A and B. The measure scales the sum contribution of the LCS by the independent information from each word. Lin defines this similarity in a generalised form as:

$$Similarity_{IT-Lin}(w_1, w_2) = \frac{2 \times I(F(w_i) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \tag{4.1}$$

where $I(x)$ is the information content of x, and $F(x)$ is the "feature vector" for x. To metricate the information-theoretic Eq. 4.1 in WordNet, it is defined as:

[14] An LCS is the nearest, more general term shared by two entries in the hyponym or troponym tree.

Measure	Definition	Description
Gloss Vectors [172]	$\cos(T_{A*}, T_{B*})$	$\{T\}$ is a co-occurrence table of glosses $A \cup B$. Effectively measure gloss overlap between two entries.
Path Traversal	Distance $A \rightarrow B$	Common baseline measure using only hyponym / hypernym links.
Wu & Palmer [236]	$\frac{2 \cdot depth(LCS)}{depth(A)+depth(B)}$	Combines path-length and LCS contributions.
Resnik [182]	$-\log p(LCS)$	$p(x)$ is the relative frequency of observing x.
Jiang & Conrath [108]	$\frac{p(LCS)^2}{p(A)p(B)}$	Extends Resnik's measure to accommodate mutual co-occurrence.
Hirst & St-Onge [101]	Directed distance $A \rightarrow B$	Follows any semantic link, but with respect to its direction.
Leacock & Chodorow [140]	$-\log \frac{dist(A,B)}{2D}$	D is the depth of the entire hypernym tree, and $dist$ is the path distance in the same tree.
Lin [143]	$\frac{\log p(LCS)^2}{\log(p(A)p(B))}$	Information-theoretic version adapted to WordNet.
Lesk [4]	$w_1(A \cap B) + w_2(\langle A' \rangle \cap \langle B' \rangle)$	Weighted sum of the gloss overlap between A and B as well as their neighbors A' and B'

Table 4.3: Association measures used in WordNet, ordered chronologically by their publication. In the definitions, A and B refer to entries and LCS refers to the least common subsumer.

$$Similarity_{WN-Lin}(A,B) = \frac{\log p(LCS)^2}{\log(p(A)p(B))} \tag{4.2}$$

Lin's formulation attempts to apply generally to any feature-based representation. With WordNet, the feature vector is an entry's synset. To illustrate this metric, Lin uses the example of associating *hill* and *coast*, shown in figure 4.3. Each word shares some information with the others, but does so depending on the words' location in the hypernym tree.

Given the success of Lin similarity, both in WordNet [143, 144] and other areas like the semantic web and formal ontologies [48, 173, 186], MetID uses this measure for relatedness in WordNet[15]. Because it requires the compared words to share POS-class, the system uses a type of reverse-stemming where a word is stemmed after which the best[16] of all its possible forms is chosen as the closest neighbour. The WordNet::Similarity Perl module[17] was used with WordNet version 3.1 for Unix and the abstract hypernym root-nodes in WordNet::Similarity were enabled to allow comparison of entries that do not share a top-level hypernym [173].

When configured to use WordNet, MetID's work-flow is different than for the corpus-based semantic models: instead of searching a set of saved clusters generated by the models, MetID will calculate the pair-wise similarity between every word in the item of analysis and the seed terms from *Metalude*. For each of the 479 unique seed terms, an exhaustive pair-wise comparison in WordNet typically takes less than 30 seconds, which is comparable to the database queries required for the corpus-based models.

[15] MetID can be configured to use any similarity metric available in the WordNet::Similarity module.

[16] The "best" choice maximises the resulting similarity.

[17] http://wn-similarity.sourceforge.net/; 5 February, 2013.

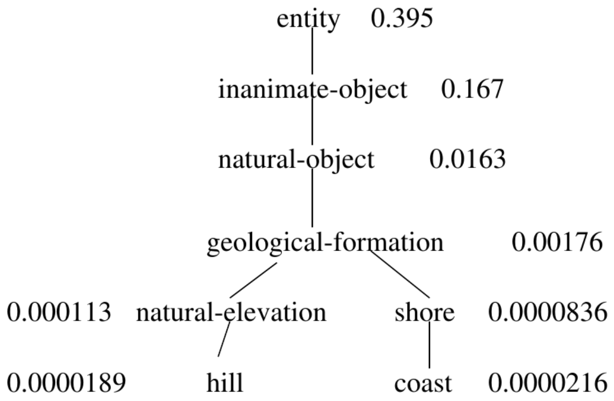


Figure 4.3: Lin's example of the similarity between *hill*-[noun] and *coast*-[noun] in WordNet. The numbers correspond to the probability (as of WordNet 2.0) that a randomly selected word is subsumed by that class. These probabilities are used to calculate shared information across sub-classes. Taken from [144].

4.4.4 Building Clusters: Distributional Semantic Models

The distributional semantic models (DSMs) implemented in MetID were picked to exemplify a range of strategies. Four models were chosen: LSA, HAL, BEAGLE and COALS. Each of the models will be reviewed in this section and an example from [137] is presented in appendix D. MetID uses each DSM in the same way: to create clusters of nearest-neighbours around the seed terminology given by *Metalude*. The clusters are stored in a database, which allows a user to select which model to use at execution-time.

Model	**Context**	**Default Association Method**	**Notes**
Context Region			
LSA	Document	Cosine	Uses SVD
Context Word			
HAL	Ramped window	Euclidean	
COALS	Ramped window	Correlation	SVD optional
Incremental			
BEAGLE	Sentence	Cosine	

Table 4.4: DSMs used in MetID.

Context Region Model: LSA

Latent semantic analysis (LSA) was developed by psychologists working on semantic memory [136]. In LSA, words are encoded as high-dimensional vectors representing their usage in documents or paragraphs of a corpus. The word-document co-occurrence matrix is first normalised by the log-entropy of each row. Next, it is decomposed to a lower-dimensional representation using singular value decomposition (SVD). The resulting *n*-dimensional semantic space allows word-pair comparison as well as word-document and document-document comparison.

LSA has been successful in modelling a variety of psycholinguistic findings and has been used in search engine design, document comparison and summarisation (see [138, 207, 214] for reviews). However, it has been criticised in cognitive modeling for employing implausible algorithms, like cosine similarity and SVD [180]. Another critique of LSA is that the arbitrary choice of dimensionality (often set empirically to 300) can have a large effect on accuracy for specific tasks [184]. Moreover, one number of dimensions often works well for one task but not for another (compare [214] with [121]).

Context Word Models: HAL & COALS

The *hyperspace analogue to language* (HAL) is a model that represents word-meaning as a word-word co-occurrence matrix [27]. HAL builds this matrix using a ramped window to construct a 140,000 dimension semantic space (see figure 4.4). HAL does not reduce the dimensionality of this space, instead, it is limited to the 70,000 most relevant entries determined by the column's log-entropy. Tabulating this matrix yields the 70,000 × 70,000 element space in which each word is a point. In this hyper-dimensional space, a word's distance from another is analogous to their semantic relatedness and is usually measured using a Minkowski distance such as Euclidean. HAL is partly motivated to address concerns of cognitive plausibility, avoiding algorithms like SVD. In fact, the only technical decisions HAL requires are the co-occurrence window-size, its weighting and the choice of how many words to represent (70,000 by default). HAL does not uses word-order information or sentence boundaries, only the windowed co-occurrence observations and the developers of HAL do not say how grammatical boundaries may impact performance. However, HAL and LSA are perhaps the two most cited distributional models in terms of correlation to experimental data [147, 162, 184].

The *correlated occurrence analogue to lexical semantics* (COALS) is inspired by HAL and LSA and implements a number of refinements [188]. COALS is a word co-occurrence model in which meaning is built using a windowed co-occurrence matrix. Unlike LSA, COALS does not rely on segmented text to build these vectors. After building this matrix, all but the top n most frequent words (14,000 by default) are discarded. The matrix is then normalised by computing pair-wise correlations, after which negative correlations are discarded and the remaining are square-rooted. Optionally, the space's dimensionality can be reduced with SVD. Using the resulting matrix, similarity between words is computed as the correlation between their vectors. It is interesting that the COALS developers found that correlation, not cosine or a Minkowski distance, was the best performing measure for vector similarity. Some recent work comparing distributional models has shown that COALS is more accurate on a number of tasks than other models [184, 188].

Word:	The	girl	likes	the	boy	with	the	long	hair
Weight:	1	2	3	4	*t*	4	3	2	1

Figure 4.4: An example 4-word, ramped window. In HAL and COALS, context is defined around a target word, t, becoming decreasingly relevant the farther away a co-occurring word is found.

Incremental Model: BEAGLE

Bound encoding of the aggregate language environment (BEAGLE) seeks to account for word-order by learning both surface and latent structure in n-gram and transition data [111, 113]. The BEAGLE developers propose four signals available in an input stream: co-occurrence (contextual preference as n-gram frequencies), transition (incremental observations of word-order), indirect or "latent" co-occurrence (reduced dimensionality vectors of n-gram frequencies) and latent transition information (approximated word-order information). BEAGLE learns both co-occurrence and sequential information and builds on the approach of HAL. BEAGLE uses a process called as *circular convolution* to encode each form of information in a single matrix. BEAGLE represents the best-of-breed model in terms of cognitive plausibility because it encodes the two surface signals, co-occurrence and order information, as well as their latent counterparts. Interestingly, BEAGLE does not perform as well as COALS and HAL on some clustering tasks [184].

Matrix Preprocessing

Before an initial co-occurrence matrix is transformed for use as a representation, it is common to normalise it by some criterion [109, 136, 192]. Two common criteria are row-length and log-entropy, which are used in LSA, HAL and COALS. The intuition behind normalisation is that models should compensate for documents' propensity for co-occurrence patterns which do not reflect topical or semantic information. For example, the past tense form of *say* is abundant in news reports, but does not reflect a given report being about "saying things". The developers of LSA note, that though stop-words are usually removed, entropy normalisation can help compensate for closed-class words' contribution to a semantic space. Log-entropy normalisation has also become common in document retrieval algorithms like TF-IDF and Okapi BM25 [109, 185][18]. Table 4.5 lists normalisations commonly applied to DSMs' initial co-occurrence matrix.

Measuring Similarity in Semantic Spaces

The goals of DSMs is to construct word representations which embody meaning in a unified data structure. Interacting with these representations usually involves comparing words to one another – ie. comparing vectors. Five measures of vector similarity are given in table 4.6: Euclidean distance, cosine similarity, the extended Jaccard coefficient, Pearson's correlation coefficient and Spearman's rank coefficient. Some models specify the intended function to be used, while others

[18] There is some mystique surrounding matrix preprocessing – especially log-entropy normalisation. Karen Spärck Jones was quoted in her obituary to admit she "didn't really know" why the process helped TF-IDF rankings.

Criterion	Normalisation Equation	Notes
Row / Column sum:	$\frac{X_{a,b}}{\sum X_{a,j}}$ or $\frac{X_{a,b}}{\sum X_{i,b}}$	Based on row representations.
Row / Column length:	$\frac{X_{a,b}}{\sqrt{\sum X_{a,j}^2}}$ or $\frac{X_{a,b}}{\sqrt{\sum X_{i,b}^2}}$	Based on row representations.
Correlation:	$\frac{(\sum\sum X_{i,j})X_{a,b}-\sum X_{a,j}\cdot\sum X_{i,b}}{\sqrt{\sum X_{a,j}\cdot((\sum\sum X_{i,j})-\sum_{a,j})\cdot\sum X_{i,b}\cdot((\sum\sum X_{i,j})-\sum X_{i,b})}}$	Matricised Pearson's correlation.
Row-Entropy:	$\frac{\log(X_{a,b}+1)}{-\sum\frac{X_{a,b}}{\sum X_{a,j}}\log(\frac{X_{a,b}}{\sum X_{a,j}})}$	Used in TF-IDF.

Table 4.5: Various normalisations commonly applied to a co-occurrence matrix. Each equation applies to a matrix $\{X\}$ of rows $a \in X_{*,b}$ and columns $b \in X_{a,*}$. All sums operate on row or column indices, i and j, respectively.

specify a class of functions. For example, LSA is designed to be used with cosine similarity whereas HAL can be used with any Minkowski distance, defined as:

$$(\sum_{i=1}^{n} |q_i - p_i|^{\lambda})^{1/\lambda} \tag{4.3}$$

where $\lambda \geq 1$ to satisfy triangle inequality, assuring the distance is a metric.

Function	Notes
Euclidean distance: $\sqrt{\sum_{i=1}^{n} \lvert q_i - p_i\rvert^2}$	Distance measure in which higher is less similar (farther away). A common Minkowski distance (Eq. 4.3 with $\lambda = 2$).
Cosine similarity: $\frac{p\cdot q}{\lVert p\rVert\lVert q\rVert}$	Compares direction of two vectors regardless of magnitude: 1 is identical, -1 is completely different.
Extended Jaccard coefficient: $\frac{p\cdot q}{\lVert p\rVert^2+\lVert q\rVert^2-p\cdot q}$	Ratio between bit-wise union and intersection where higher is more similar [107].
Pearson's correlation: $\frac{n\sum p_i q_i-\sum p_i\sum q_i}{\sqrt{n\sum p_i^2-(\sum p_i)^2}\sqrt{n\sum q_i^2-(\sum q_i)^2}}$	1 is completely linear, -1 is completely nonlinear.
Spearman's rank: $\frac{\sum_i (p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum_i (p_i-\bar{p})^2\sum_i (q_i-\bar{q})^2}}$	Monotonic version of Pearson's correlation: 1 is complete related by a monotonic function and -1 is completely unrelated.

Table 4.6: Functions commonly used to compare vectors in a semantic space. In the equations, p and q are discrete frequency distributions.

DSM Variants

Table 4.7 shows the variants of distributional models used in MetID. Note that not every model is compatible with every similarity function. Also, keep in mind that more models and variants can be included – MetID is neutral with respect to how the clusters were created. This will allow a comparative analysis of lexical models, independent of the overall method. It also means that different lexical models can be used in different situations. For example, WordNet will later be used in a case-study of financial language because it provides relatively good coverage when using a smaller corpus.

Model	**Compatible Similarity Functions**	**Notes**
LSA-100	cos pear euc spear	Reduce to 100 dimensions
LSA-300	**cos** pear euc spear	**Reduce to 300 dimensions**
LSA-400	cos pear euc spear	Reduce to 400 dimensions
LSA-500	cos pear euc spear	Reduce to 500 dimensions
HAL	cos pear **euc** jac	**Compare all dimensions**
HAL-400	cos pear euc jac	Compare the 400 best dimensions
HAL-1400	cos pear euc jac	Compare the 1,400 best dimensions
COALS-SVD 100	cos pear euc spear	Reduce to 100 dimensions
COALS-SVD 200	cos pear euc spear	Reduce to 200 dimensions
COALS-SVD 800	cos pear euc spear	Reduce to 800 dimensions
COALS-800	cos pear euc jac	Retain the 800 most frequent words
COALS-14000	cos **pear** euc jac	**Retain the 14,000 most frequent words**
BEAGLE-128	cos pear euc spear	128 permutations
BEAGLE-256	cos pear euc spear	256 permutations
BEAGLE-512	**cos** pear euc spear	**512 permutations**
BEAGLE-1024	cos pear euc spear	1024 permutations

Table 4.7: The distributional semantic models used to create clusters in MetID. All compatible similarity functions were used for each model. Bold-face denotes the default (published) configuration for each model.

4.4.5 Building Clusters: Co-occurrence Likelihood Estimation

Co-occurrence likelihood estimation (COLE) refers to the task where, given a set of documents, we wish to predict how likely it is terms will co-occur in an unseen document. COLE-based models offer a statistical approach to building word clusters without a representation scheme (like a semantic space). Three COLE models were explored for use in MetID [105]. The first is based on mutual information (MI) and the second two are based on language models (LMs) [153] The language model methods are the result of ongoing work with Dr. Hua Huo, who was a visiting scholar at Trinity under the supervision of Prof. Khurshid Ahmad. Using the models for word-clustering is the result of my adaptation of the document-indexing system developed by Dr. Huo. The LM-based models, however, did not generate viable clusters and were not tested in MetID (see appendix A section A.1.1 for further explanation). The next section introduces the MI-based method, the likelihood estimations which MetID uses as the relatedness between two words.

Mutual Information (COLE-AMI)

Mutual information (MI) is a statistic for computing the interdependence between two terms, t_1 and t_2, that has been used to extract collocations in text [202]. Intuitively, MI measures how much more often than chance two terms co-occur, or how much information they "mutually contribute" [39]. MI assumes that each term occurs with independently random probability and was first defined as

$$\mathrm{MI}(t_1,t_2) = \log_2 \frac{p(t_1,t_2)}{p(t_1)p(t_2)} \tag{4.4}$$

where 4.4 $p(t_1)$ and $p(t_2)$ are each term's independent prior probability of occurrence and $p(t_1,t_2)$ is their co-occurrence prior [39].

Using MI to estimate term co-occurrence has some drawbacks. The first is the unilateral co-occurrence problem: it ignores the case where only one term is present in a sequence. The second relates to rare occurrences: when $p(t_1,t_2)$ and $p(t_1)$ or $p(t_2)$ are very small, $\mathrm{MI}(t_1,t_2)$ can still be relatively large, despite the posterior likelihood of the sequence. This will result in infrequent words, that occur in only a few isolated places, receiving an over-estimated relevance. Further, the original formulation only works for two terms. To generalise MI for multi-term co-occurrence, Zhang and Yoshida proposed *augmented mutual information* (AMI) [241]. AMI is defined as:

$$\mathrm{AMI}(t_1,t_2,...,t_n) = \log_2 \frac{p}{(p_1-p)(p_2-p)...(p_n-p)} \tag{4.5}$$

where $(t_1,t_2,...,t_n)$ is an n-term sequence and p_n is short-hand for the probability $p(t_n)$.

One strength of AMI is that it is formulated as a product of probabilities, which can be estimated without exhaustive observation. This makes the method computationally tractable over large collections. Using maximum likelihood estimation, AMI can be defined for n terms' frequency observations in a document, D:

$$\mathrm{AMI}(t_1,t_2,...,t_n) = (n-1)\log_2|D| + \log_2 f - \sum_{t=1}^{n} \log_2(f_i - f) \tag{4.6}$$

where $|D|$ is the size of the document, f is the frequency of the sequence $(t_1,t_2,...,t_n) \in D$, and f_i is the frequency of the ith term. $\log_2|D|$ measures how much AMI will increase when a term is added to the sequence. Because $\log_2|D|$ can dominate the equation, it is often scaled by a constant $\alpha \propto n$. Because f_i may equal f, a correction constant, β, is added. This results in multi-term co-occurrence statistic using AMI defined as Eq. 4.7:

$$\mathrm{AMI}(t_1,t_2,...,t_n) = (n-1)\alpha\log_2|D| + \log_2 f - \sum_{t=1}^{n} \log_2(f_i - f) + (n-m)\beta \tag{4.7}$$

where m is the number of terms which are less frequent than the whole sequence and β is a constant by which to scale the frequency of terms that occur as many times as the sequence in which they occur [105]. β is used to diminish the impact of low frequency terms found in equally low frequency sequences such as multi-term proper nouns.

4.4.6 Testing Clusters

The intention of using clusters instead of raw similarity scores to associate words, is to account for some of the inherent underlying semantic and conceptual properties of individual words. Abstract words, such as “thing”, “way” or “item”, are vague and tend to elicit diverse associations [124]. By building sets of associated words around the seed terms from *Metalude*, the clusters can be intrinsically qualified to account for words’ abstractness using their similarity distributions. That is, words that are proximate to predominately high-frequency words, can be thought of as less well-defined compared to words that are proximate to words that range in frequency. To measure this quality, the associative neighbourhood of a word needs to be bounded. To achieve this, MetID adopts a nearest-neighbours approach where the 200 closest words are computed for a seed term, using their relative frequencies to calculate two cluster-based quality metrics: purity and entropy.

Word clustering is an NLP task that has been addressed using both linguistic and non-linguistic strategies [3, 116, 182, 199, 201, 208]. A common task in this regard is to find the *best solution* – finding a partition that maximise an objective function of the resulting clusters. In agglomerative and spectral clustering, potential solutions are measured by extrinsic scores like F-measure or Rand measure [240], which use an external gold-standard to compare class-cluster ratios[19]. Clusters can also be qualified in terms of their internal composition, without an external reference. Two commonly used intrinsic measures of cluster quality are purity and entropy [116, 240]. Purity and entropy measurements use the ratio of relative frequency to class frequency. Purity is a measure of how much of a cluster consists of the most common class. Entropy, on the other hand, measures how evenly the items are dispersed among classes.

Purity for a cluster, C with a vocabulary V_C, is defined:

$$\text{Purity} = \frac{1}{|C|} max(n_w \in V_C) \tag{4.8}$$

where n_w is the frequency of word w. As currently implemented, MetID sets $|C|$ to be 200.

Entropy, or *normalised Shannon entropy*, measures the diversity of classes in a cluster and is inversely proportional to a cluster’s quality [116, 184]. If a model predicts word w has a set of neighbours, $\hat{T}$, then the entropy of $\hat{T}$ will be high when the relative frequencies of $t \in \hat{T}$ are uniform. Conversely, if the frequencies of clustered words are consolidated in a few classes, the cluster’s entropy will be low. Entropy for cluster, C, is defined over each word in its vocabulary, $w \in V_C$:

$$\text{Entropy} = -\frac{1}{\log |V_C|} \sum_{w \in V_C} \frac{n_w}{|C|} \log \frac{n_w}{|C|} \tag{4.9}$$

where n_w is the frequency of w and $\frac{n_w}{|C|}$ is the relative frequency for w in the cluster.

[19] For our purposes, classes are words, the number of which is set to 200, and the frequencies are relative word frequencies. The 200-word cluster size was chosen because it is large enough to be inclusive of a spectrum of associations but is still below the minimum frequency threshold for all model configurations.

Figure 4.5 shows three example clusters with their respective purity and entropy calculations. Note that high purity is a positive attribute, indicating a good cluster, while high entropy is negative. Purity is naturally normalised from 0 to 1 and the normalised version of Shannon entropy is used, which is also defined from 0 to 1.

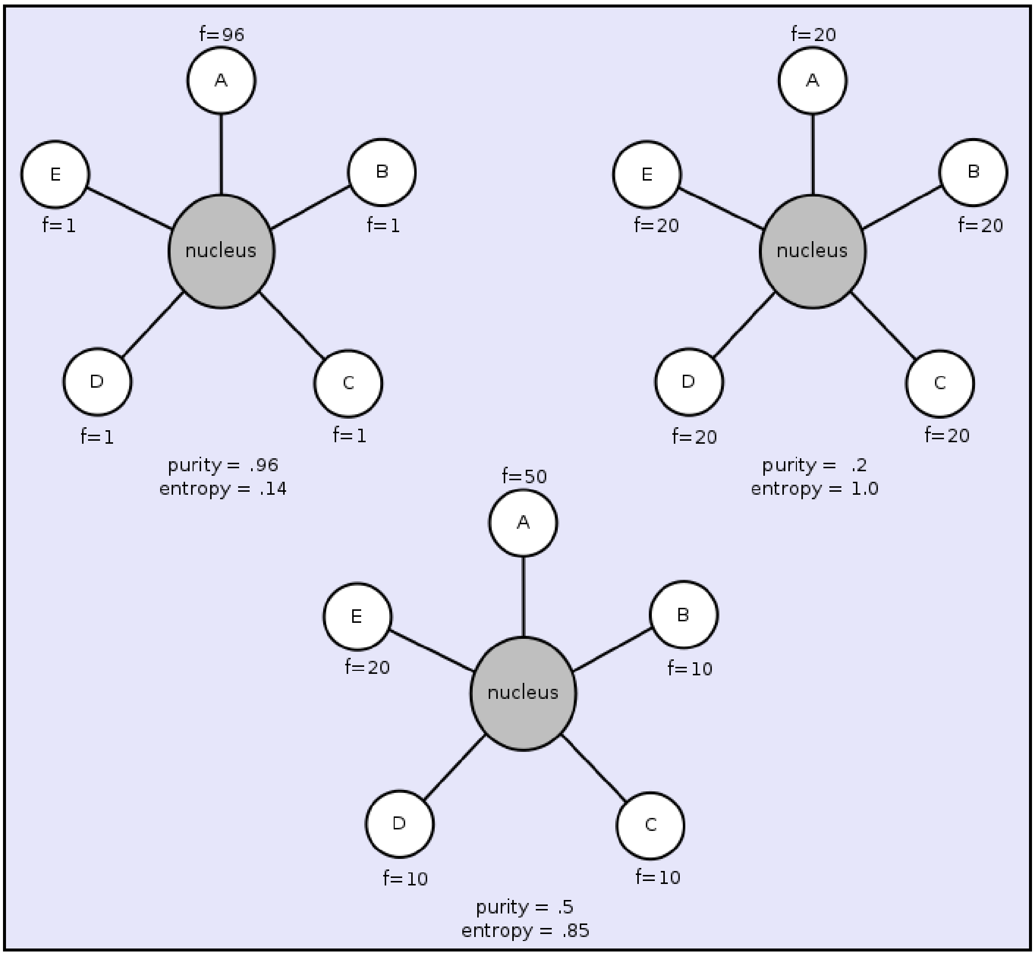


Figure 4.5: Three example clusters, each with five classes A, B, C, D and E. Clockwise from the top-left: the first has high purity and low entropy because it is made up of 96% one class (A) while the rest are uniformly distributed. The second (top right) has low purity because all classes make up one fifth of the total, while the entropy is high because the contributing classes evenly distributed. The last cluster (bottom) has moderate purity because class A makes up 50%, but has relatively high entropy given the dispersion of frequencies over the remaining classes. Note that MetID uses relative frequencies for all classes.

4.5 Implementation of Heuristics

In addition to the paired cluster search algorithm, which uses a lexical model to match figurative terms from *Metalude* with items found in the input, MetID also implements a set of heuristics (see table 4.8). The heuristics are applied as a series of conditional bonuses and penalties corresponding to various lexical, grammatical and cluster-based cues. The lexical heuristics pertain to the words constituting or found near a potential metaphor. The grammatical heuristics make use of known properties of figurative language, such as predication and violations of selectional preference. The cluster metrics are designed to account for the intrinsic quality of clusters built by the semantic module. Lastly, there are two heuristics that are procedural in nature, and account for a model's coverage: a penalty for words not found and a penalty if MetID fails to find a candidate with a mapping given by *Metalude*. Throughout this section, the terms in table 4.9 are used to refer to the various pieces of analysis.

Heuristic	**Type**	**Description**
Non-word	Penalty	The identified topic or vehicle, are not valid words.
WN Synonyms	Bonus	The identified topic or vehicle is a synonym of the candidate counter-part.
Marker	Bonus	The sentence contains a co-text cue or marker.
Unpaired Metaphor	Large Penalty	Could not find a metaphor with a pairing given by *Metalude*.
Predication	Large Bonus*	If the unit is a predication, and the identified vehicle predicates the topic.
Selectional Violation	Large Bonus*	If the identified topic and vehicle are in a relationship which violates the selectional association of the root word.
Hypernym	Penalty	If the identified vehicle and topic are nouns, and the vehicle is a hypernym of the the topic.

Table 4.8: Heuristics implemented in MetID.

*The predication and selectional violation bonuses are scaled proportional to scores calculated over a reference corpus (see section 4.5.2.

Each heuristic is either a bonus or a penalty applied to the initial score from the paired cluster search (see section 3.5). Because the scores are normalised from 0 to 1, bonuses and penalties shift the score either a quarter or half-way from its original value to 1 (bonus) or to 0 (penalty). A bonus is defined as

$$score := score + \frac{1 - score}{2} + c\frac{1 - score}{4} \tag{4.10}$$

where c is either 0 or 1 for a regular bonus (bringing the score half way to 1.0) or large bonus (bringing the score three-fourths of the way to 1.0, respectively. A penalty is defined as:

$$score := \frac{score}{c} \tag{4.11}$$

where c is either 2 or 4 for a regular penalty or a large penalty respectively.

Term	**Description**	**Example(s)**
Unit (of analysis)	The extracted piece of input which is being, or going to be analysed.	"Our hearts are on fire." `pred(heart,fire)`
Identified topic	The stem of the unit that may be the topic term of a metaphor.	*heart*
Identified vehicle	The stem of the unit that may be the vehicle term of a metaphor.	*fire*
Candidate topic	A topic from *Metalude*.	AFFECTION
Candidate vehicle	A vehicle from *Metalude*.	WARMTH
Candidate metaphor	The topic-vehicle pair from *Metalude*.	AFFECTION AS WARMTH
Topic cluster	The cluster of words (built by the lexical model) around the candidate topic.	Neighbours of AFFECTION
Vehicle cluster	The cluster of words (built by the lexical model) around the candidate vehicle.	Neighbours of WARMTH
Relational triple	Two words in a grammatical relation.	`nsubj(are,hearts)`
Stem	The canonical root of a given word.	*hearts* ⇒ *heart*
Lemma	The canonical root of a given word with respect to its POS.	*hearts* ⇒ *heart*-[noun]

Table 4.9: Terms used in the explanation of MetID's analysis sub-system. The examples refer to the input sentence "Our hearts are on fire." and its metaphor AFFECTION AS WARMTH.

4.5.1 Lexical Heuristics

The lexical heuristics are concerned with identifying markers that signal a potential metaphor. They account for properties of the identified topic- and vehicle-terms as well. Though a key feature of metaphor is that it is not purely lexical, these heuristics prioritise "marked" metaphors [92]. Lexical heuristics are defined as such because they deal directly with observed words as opposed to relationships, constructions or the properties of the word clusters.

Direct Matches, Non-Words & Synonyms

If an identified topic matches the candidate topic or the identified vehicle matches the candidate vehicle, a normal bonus is applied. If both identified terms match the candidates, a large bonus is applied. This is to assure that more obvious metaphors are promoted above less obvious ones. For example, if MetID is analysing the statement "time is money", the interpretation TIME AS MONEY should be promoted over the less precise, albeit correct TIME AS COMMODITY. Though direct matches are not common, especially for novel metaphors, this heuristic assures more accurate ranking of obvious mappings. Table 4.10 contains some examples of direct matches.

There are two edge-cases in comparing identified terms with their candidate's counterparts: the identified word may not be represented in the semantic model or the two words may be synonyms. The first case is covered by a heuristic that applies a normal penalty if an identified topic or vehicle term is not found in the lexical model (ie. it is not found in *any* cluster). This penalty is

Input Text	Topic	Vehicle	Candidate Metaphor	Bonus
Time is money	time	money	TIME AS MONEY	Large
Love is a journey	love	journey	LOVE AS MOVEMENT	Regular
Love is a journey	love	journey	AFFECTION AS JOURNEY	Regular
My car drinks gasoline	car	drink	DRINKING AS CONSUMING	Regular
Angry words are weapons	word	weapon	COMMUNICATING AS VIOLENCE	None
She exploded at me	she	explode	ANGER AS EXPLOSION	None

Table 4.10: Examples of how the direct match heuristic is applied to various input.

only applied once for each identified topic and vehicle. However, if neither term is in the semantic model, no candidate metaphors will be produced, regardless of what heuristics are applied.

The second edge-case occurs when an identified term is synonymous with the candidate's. MetID respects a notion of synonymy dictated by the lexical model. Because the similarity scores for every model are normalised from 0 to 1, the system simply applies a regular bonus if the similarity between the identified topic and candidate topic is 1, or the same is true for the vehicles. In the distributional models, synonymy is nearly impossible to achieve given the sensitivity of vector similarity measures. As implemented, WordNet is the only model that can make effective use of this heuristic. However, the direct match heuristic operates on stems, which was partly motivated to help compensate for the rarity of synonymy in the DSMs[20]. To avoid redundancy, this heuristic is not applied if the direct match heuristic was applied.

Hypernymy

A defining feature of metaphorical category assertions, which are commonly found in noun-noun copula constructions, is that they propose a figurative hypernymic relationship. The example, "my lawyer is a shark" assigns *shark* as a hypernym of *lawyer*, a figurative relationship. To discourage MetID from scoring literal class-inclusion statements as figurative, it applies a large penalty if the identified topic and vehicle terms are both nouns and the vehicle-term is a hypernym of the topic. To do this, regardless of which lexical model is configured, WordNet is used as the hypernym database. Because the top levels of WordNet's hypernym tree are relatively abstract, containing entries like "entity", "agent" and "living being", the top two levels are not considered when MetID looks for hypernyms.

There are some cases where literal hyponymy can be found in a genuinely figurative statement. For example "Boys$_1$ will be boys$_2$", though idiomatic, is a metaphor eliciting specific features of boys$_1$ and applying to the set of boys$_2$. Examples like this, however, are likely very infrequent, yielding fewer false positives for this heuristic than legitimate applications. In practice, the hypernym penalty assists ruling out expressions of literal category relations.

[20] Alternatively, one could imagine setting a threshold for similarity above which synonymy was assumed. Setting this threshold, however, would be an empirical matter and specific to each model and similarity measure.

Lexical Cues

Two kinds of contextual cues are included as heuristics, both from Andrew Goatly's *The Language of Metaphors* [92]. Goatly finds that a number of cues commonly mark the use of metaphor. The first are a set of strings that often signal metaphor. These strings, listed in table 4.11, tend to exaggerate, diminish, locate or hide the use of an upcoming metaphor. If one of these strings is found in the sentence or phrase being analysed a large bonus is applied. The second kind of lexical cue, which Goatly calls co-text markers, often signal the use of figurative language, but sometimes to diminish its novelty, highlight an aspect or to motivate the reader to process it in a particular way. If one these strings, shown in table 4.12, is found, a normal bonus is applied. In both cases, the heuristic is applied only once, even if more than one marker is found.

String	**Common Function**
metaphorically speaking	Mark a metaphor
figuratively	Diminish a previous metaphor
utterly	Exaggerate an upcoming metaphor
completely	Exaggerate an upcoming metaphor
so to speak	Mark a previous metaphor
as it were	Mark a previous metaphor

Table 4.11: Strings Goatly identifies as marking the use of a metaphor.

String		
metaphor*	figurativ*	trope
literal*	really	actually
in fact	simpl*	fairly
just	absolut*	fully
complete*	quite	thorough*
utterly	veritabl*	regular*
in a way	in one way	a bit of
half-*	practically	almost
not exactly	not so much * as	* if not *
in both senses	meaning	in more than one sense
import*	symbol*	sign
type	token	instance
example	a (sort | kind) of	(curious | strange | odd | peculiar | special) * (sort | kind) of
like (a | the)	as a *	the * of (a | the)
the * equivalent of	as if	(seemed | sounded | looked | felt | tasted) (as (though | if))
as though	!	(could | might) say
delusion*	illusion	if * (could | would | might | imagine | suppose)
hallucinat*	mirage	phantom
fantas*	unreal	

Table 4.12: Lexical markers that often signal the use of metaphor [92]. For presentation purposes, these are not regular expressions, but the * and | symbols are analogous to POSIX globs.

4.5.2 Grammatical Heuristics

Grammar plays a significant role in how metaphors are used in language and research suggests some constructions signal the presence of figurative relationships [92, 128, 133, 206]. Some NLP projects have used grammatical analyses to help detect the use of metaphor, particularly in verb-based metaphors [22, 182, 200, 201]. For example, the selectional violation heuristic described here is based on an approach that seeks to find figuratively applied verbs. The predication heuristic is inspired by the selectional violation strategy, but with respect to noun-noun predications.

Selectional Violation

Selectional preference violation is a measurable effect based on Yorick Wilks' theory of preference semantics [229, 230]. Wilks proposed that a property of lexical semantics is the emergence of certain preferences that constrain the use of words in certain relationships. Intuitively, this can be described as subjects "preferring" to verb, or that objects prefer to be verbed. For example, cars tend to *drive*, people tend to *say* and doors tend to *open* and *close*. Measuring the strength of these preferences is known as selectional preference induction [41, 52, 182, 199, 200]. The selectional strength of a word can be measured as the uniformity of its arguments in a given relationship. With a pair of words, their *selectional association* can be measured in a particular relationship using a mutual information approach: by the ratio of observing two terms outside a relationship to that of the them occurring in the relationship [39, 153]. MetID measures selectional association as

$$s\_assoc(w_1, w_2, r) = \log \frac{f_{rel}(w_1, w_2, r)}{f_{rel}(w_1, *, r) f_{rel}(*, w_2, r)} \tag{4.12}$$

where f_{rel} is the relative frequency of w_1 and w_2, and r is a grammatical relationship. This allows selectional association between any pair of words that occur in any relationship to be measured. MetID only uses scores for subject-verb, object-verb, noun-modification and adjective-modification as these relationships are known to be semantically productive [230]. To compensate for variation in relative frequencies, the scores are expert normalised[21] from 0 to 1.

Because selectional association is based on observable data, the choice of corpus has an impact on the scores. As such, scores were computed for every corpus (those listed in table 5.1) as well for all corpora combined[22]. Generally speaking, the bigger the corpus, the more smoothly distributed the scores will be, hence also computing scores for all corpora together[23]. Table 4.13 shows the selectional association scores for the word *person* as a subject, extracted from the TASA corpus.

The selectional violation heuristic is implemented as a scaled bonus. For an observed word-word-relation triple, the heuristic accumulates points for the selectional association score being below the median, mean, 1 SD + mean or unobserved for the given root[24]. Intuitively, this is similar asking, on a scale from 0 to 4, how "interesting" is this particular association, given previously

[21] Expert normalisation divides all values in a set by the maximum value.

[22] Note that inter-corpus scores are not comparable, due to the within-corpus normalisation.

[23] MetID allows the user to configure a "selectional corpus" at execution time but where not otherwise noted, the TASA corpus was used for this heuristic.

[24] The root word refers to w_1, however, its position changes depending on the relationship.

Word 1 (w_1)	**Word 2** (w_2)	$f_{rel}(w_1,w_2,\texttt{nsubj}) \times 10^{-5}$	$f_{rel}(w_2) \times 10^{-5}$	$s\_assoc(w_1,w_2,\texttt{nsubj})$
person	overweight	1.2	5.1	4.75
person	citizen	2.9	2.5	4.70
person	unconscious	0.8	4.0	4.58
person	misuse	0.6	3.2	4.53
person	immune	0.6	3.8	4.35
person	sue	1.4	0.0	4.24
person	interview	0.6	4.9	4.10
person	faint	0.6	4.9	4.10
person	injure	0.8	7.0	4.03
person	deaf	0.6	5.3	4.02
person	drink	4.4	0.5	3.94
person	drown	0.8	8.5	3.84
person	alcohol	0.6	6.6	3.80
person	ill	1.7	9.4	3.71
person	alert	0.6	8.5	3.55
person	swallow	1.0	4.9	3.50
person	mature	1.0	5.1	3.49
person	smoke	1.7	6.2	3.41
person	react	3.8	9.7	3.40
person	consume	1.0	7.5	3.34

Table 4.13: The 20 most strongly associated arguments for the nominal subject *person* ($f_{raw} = 2,732$; $f_{rel} = 0.005831\%$) in the TASA corpus, measured by Eq. 4.12 prior to normalisation. Intuitively, the score is a measure of how likely it is that the argument (w_2) is "person". Similar scores were computed for `dobj`, `nmod` and `amod` relationships.

observed cases? This heuristic is applied to a given triple $\langle w_1,w_2,r\rangle$, t, and set of scores S for $s\_assoc(w_1,*,r)$ as follows:

$$\begin{aligned}
&0 \text{ if } s\_assoc(t) \geq SD(S) + mean(S)\\
&1 \text{ if } mean(S) \leq s\_assoc(t) < SD(S) + mean(S)\\
&2 \text{ if } median(S) \leq s\_assoc(t) < mean(S)\\
&3 \text{ if } s\_assoc(t) < median(S)\\
&4 \text{ if } t \text{ has never been observed}
\end{aligned}$$

times one-forth of a large bonus. That is, if 4 is the case, a full large bonus is applied, whereas if 0 is the case, the score remains unchanged. Note that the selectional association scores are distributed as an unbounded ascending power-law, which means that the median is consistently below the mean[25].

[25] This distribution was observed in all instances of a random sample of 25 high-frequency words.

Predication

Predication is the affirmation, assertion or assignment of one thing about another. A prominent linguistic example is the ontic *is-a* declaration, as in the example "my lawyer **is a** shark." Theoretically, such a declaration can take any form, but in English, the copular construction is a defining feature of linguistic predications (see table 4.14 for examples). Predications are a common way of instantiating noun-based metaphors such as the lawyer / shark example [92, 128, 133]. Making use of predications as a heuristic for identifying figurative assertions consists of finding common instances with which to compare input. To address this, a score of predicative strength was developed, similar to selectional strength described in the previous section. This score, *predication strength*, will be used to rank a word's predicates by how common they are, which in turn will be used apply bonuses to novel predications.

`predicate`	`predicate_of`
... the *rest* are local **people**.	... great *people* who are a **pleasure** to ...
... million *people* are poor **people** like Gole.	... *people* were **hysterical**.
The *players* are **people** who are not out to ...	... and the *people* who are **carriers** ...
My older *kids* are fantastic **people**.	... descriptions of *people* are **shorthand** ...
The *emotions* are so high, **people** have to ...	... scare off some *people* who are **activists** ...
... planet *Earth* is about 1 billion **people**.	... overpowered by 6 *people* is a little **rash**.
... *cottages* are most **peoples'** idea of ...	*People* are **Afghans** first ...
Intimacy is the way **people** find happiness.	... *people* who are **victims** of tyranny.
Those *academics* were **people** who could ...	... add *people* who are not **members** ...

Table 4.14: *People*-[noun] in a sample of predicate relations in the BNC. The root word is shown in bold-face for each relation (this is the predicate in the left column and the predicate of another word in the right).

Predication strength was calculated similarly to selectional association. Two sets of predications were extracted, using two reference corpora: the enTenTen web corpus (described in section 5.2) and the British National Corpus (BNC) [29, 120]. These collections were chosen because they both contain a number of predicate relations, are freely available (enTenTen) or have been used for similar tasks (BNC). The method of identifying and scoring predicate relations was the same for each corpus. First, Sketch Engine [120] was used to extract the top 5,000 most common nouns in the corpus. Then, for each noun, all `predicate` and `predicate_of` relations were extracted with their constituent arguments and frequencies. These relations are defined by regular expression templates over POS-tagged versions of each collection[26]. This resulted in a list of word-predicate pairs, a sample of which are shown in table 4.15. Note that any word can predicate or be the predicate of a given noun, not just the 5,000 most common nouns. After this list was constructed, the predication strength of the root was calculated based on a measure similar to MI [39, 153, 202]. For a given word-predicate pair, (w, p), $p\_strength$ is defined as

[26] In BNC notation, `predicate` and `predicate_of` are defined as this POS template: `any_noun rel_start? adv_aux_string_not_be copular adv_string long_np` and in TreeTagger format, `predicate` and `predicate_of` are defined over the POS template: `"NN.?.?" [tag="WP"|tag="PNQ"|tag="CJT"]? [tag="RB.?"|tag="RB"|tag="VM"]0,5 [lemma="be" & tag="V.*"] "RB.?"0,2 [tag="DT.?"|tag="PP$"]0,1 "CD"0,2 [tag="JJ.?"|tag="RB.?"|word=","]0,3 "NN.?.?"0,2 2:"NN.?.?" [tag!="NN.?.?"]` [120, 195, 196].

$$p\_strength(w,p) = \frac{f_n}{f_u * f_t} \tag{4.13}$$

where f_n is the number of times p predicates w, f_u is the number of *unique* words predicated by p and f_t is the total number of times w is predicated. This formulation discourages pairs in which the predicate is common throughout the corpus. The intuition, similar to MI, is that if an event is observed which occurs more often than expected, it should be scored high. A unique aspect of predication, however, is that a predicate may be widely applicable, or conversely, a noun may be widely predicated – two conditions that are accounted for in this formalisation.

Root Word	Predicate	Frequency
person	member	22
person	party	16
person	solicitor	10
person	employee	9
person	director	7
person	victim	7
person	friend	5
person	resident	5
person	subject	5
person	chairman	4

Table 4.15: Word-predicate pairs for *person*-[noun] (f_{raw} = 28,705; f_{rel} = 0.000299%) in the BNC.

Predications from the enTenTen and BNC corpora were scored using the $p\_strength$ function above. Further, the scores were normalised for each root word to avoid biases from root or predicate frequency. Table 4.16 shows sample scores for *people*-[noun]. Three levels of increasing interest were assessed based on the normalised scores: if it was below the median, the mean or 1 SD + mean of all the scores for that root. This approach was necessary because raw $p\_strength$ score are not comparable across roots as they may not be normally distributed.

This heuristic is applied only if the unit of analysis is a predication. Like selectional violation, MetID uses a scaled bonus. For a given predication, points are accumulated if the normalised $p\_strength$ is less than the median, mean, 1 SD + mean or unobserved for the given root. The large bonus is scaled as follows for a root word, t, with $p\_strength$ scores S:

0 if $p\_strength(t) \geq SD(S) + mean(S)$
1 if $mean(S) \leq p\_strength(t) < SD(S) + mean(S)$
2 if $median(S) \leq p\_strength(t) < mean(S)$
3 if $p\_strength(t) < median(S)$
4 if t has never been observed

That is, if 4 is the case, a large bonus will be applied, and if 0 is the case, no bonus is applied. Like the selectional association scores, predication strength is distributed along an unbounded power-law curve in which the median is consistently below the mean. This heuristic operationalises a similar intuition to selection preference violation, here taking into account words' inherent like-

					$p\_strength$	**Normalised**
Root Word	**Predicate**	f_n	f_u	f_t	$\times 10^{-4}$	$p\_strength$
person	member	22	464	461	1.03	0.137
person	party	16	152	461	2.28	0.312
person	solicitor	10	56	461	3.87	0.533
person	employee	9	77	461	2.54	0.347
person	director	7	114	461	1.33	0.179
person	victim	7	242	461	0.63	0.081
person	friend	5	195	461	0.56	0.071
person	resident	5	63	461	1.72	0.233
person	subject	5	682	461	0.16	0.016
person	chairman	4	84	461	1.03	0.138
person	christian	4	47	461	1.85	0.251
person	customer	4	93	461	0.93	0.124
person	man	4	537	461	0.16	0.016
person	partner	4	114	461	0.76	0.100
person	tenant	4	55	461	1.58	0.213
person	year	4	646	461	0.13	0.013
person	client	3	70	461	0.93	0.123
person	driver	3	56	461	1.16	0.156
person	female	3	38	461	1.71	0.232
person	officer	3	77	461	0.85	0.111

Table 4.16: Word-predicate pairs for *person*-[noun] in the BNC. f_n is the number of times the root word is predicated by the predicate, f_u is the number of unique words the predicate predicates and f_t is the number of times the root word is predicated.

lihood to be predicated in general. However, because predications are considerably less frequent than any of the grammatical relationships measured in selectional violation, this heuristic often applies in the case where a predication has never been observed (case 4 above). The sparsity of previously observed predications could be reduced by using a larger reference corpus such those of the WaC collection[27], the use of which was computationally prohibitive in MetID.

4.5.3 Cluster Quality Heuristics

The quality of the clusters can have an effect on the associative quality between the clustered words and the seed nuclei. This can occur when a seed word is relatively infrequent in the lexical model. Two heuristics that account for cluster quality are applied in all circumstances (to all units, candidate metaphors, identified terms, etc.) unless MetID is configured to use WordNet as the lexical model. This is because when WordNet is configured, MetID measures word-similarity at execution time instead of using clusters. The system applies a weighted bonus for purity and a weighted penalty for entropy. Purity is applied as one-fifth of a normal bonus, times the geometric mean of the purity of the candidate topic and the candidate vehicle:

$$score := score + \frac{(1 - score)}{10}\sqrt{p_{topic} p_{vehicle}} \tag{4.14}$$

where p is a cluster's purity. Entropy, because it is a negative indicator, is implemented as a

[27] http://wacky.sslmit.unibo.it/doku.php?id=corpora; 16 January, 2014.

penalty. Again, the geometric mean of the topic and vehicle entropies is used to weight one-fifth of a normal penalty:

$$score := score / \frac{1 + \sqrt{e_{topic} e_{vehicle}}}{10} \tag{4.15}$$

where e is a cluster's entropy.

4.5.4 Coverage Heuristics

There are two heuristics that are not lexical or grammatical in nature. The first is a normal penalty if an identified topic or vehicle is not found in the lexical model, which can be the case for uncommon words. In practice, if this heuristic fires, the results will be dubious at best because the system has effectively been used to find a metaphor without a topic or vehicle. However, there are cases where a topic or vehicle is so well matched to a candidate metaphor and other bonus heuristics have been applied, that the absence of an identified term in the model can be overcome.

The last heuristic attempts to compensate for coverage in the lexical models. When an identified topic and vehicle pair are found, their average distance to each term in the candidate metaphor is minimised, providing the best score. But if either term is not found in *any* of the respective topic / vehicle clusters, the result is that no metaphor is found. When this is the case, MetID will re-run the same analysis without enforcing the pairing from *Metalude*. That is, the topics and vehicles are used as a single list, from which the best two are chosen as the candidate topic and vehicle. This effectively allows metaphors not given by *Metalude*. But without the imposed structure of the pairings, it allows spurious pairings like MAMMAL AS ANIMAL or THOUGHT AS IDEA. Thus, when this occurs, a large penalty is applied.

4.5.5 Reversing Heuristics

Because the heuristics sub-system applies bonuses and penalties to the score from the cluster search module, they can be reversed. MetID's output lists which heuristics fired and to what degree, so that users can reverse their effect on individual candidate's score. For example, the statement "My heart is a fire that burns with love" produces a candidate metaphor AFFECTION = WARMTH for the predication of *heart* and *fire* with a score of 0.92. In this example, two heuristics were applied: the full predication strength bonus (a large bonus) and the WordNet synonym bonus for *fire* and WARMTH (a normal bonus). To retrieve the initial score, first, the predication bonus can be reversed by solving Eq. 4.10 with $c = 1$:

$$0.92 = x + \frac{1-x}{2} + 1\frac{1-x}{4} \tag{4.16}$$

which yields $x = 0.68$. Next, to remove the synonym bonus, Eq. 4.10 can be solved with $c = 0$:

$$0.68 = x + \frac{1-x}{2} \tag{4.17}$$

which gives $x = 0.36$, the score for this candidate without any heuristics applied. Though the bonus / penalty system is a simplistic approach to accounting for metaphorical signals in text, its effects are easily analysed and can be removed altogether. The evaluations in the following chapter were performed using scores subjected to all the heuristics described above.

4.6 Summary

This chapter described how MetID was implemented. For a more technical description of the system, refer to appendix B, where the core algorithms are analysed and the architectural and design principles are described in more depth. As a rule, the aspects of MetID that relate to finding and interpreting metaphor were included in this chapter, whereas appendix B contains programmatic and computational considerations.

The system uses a number of NLP tools, like taggers and parsers, but implements a number of its own techniques, such as the text pre-processor, the cluster creation, the predication and the selectional association system. Each of these components were developed with respect to the goals of the overall project. Wherever possible, existing open-source solutions were used to address problems that are more or less solved (stemming and tagging) or beyond the scope of this research (parsing). There are two main components that comprise the core of MetID's behaviour: *Metalude* and the lexical models. *Metalude* provides the seed terms around which the clusters are built, as well as the mapping between topic and vehicle concepts. The lexical models, most of which are corpus-based, provide a means of associating observed words in the input with the terms in *Metalude*.

To test the various lexical models, they are designed to be interchangeable which, as we will see in the next chapter, offers a way to compare them to one another. Methodologically, the system's evaluation attempts to address the initial combinatoric complexity of corpus × model × distance function × input by broadly defining simple tasks with which to constrain configurations for more difficult tasks. The first results presented will be the quality of the clusters built with each configuration. Configurations that produced viable, broad-coverage clusters will be used in a binary decision task, where MetID is used to "pick the metaphor" from a set of statements. The models that perform best on the decision task will be used to validate the candidate metaphors using participant ratings. Finally, the method will be used in a corpus-analysis of how the metaphor of *contagion* has been adopted in finance, economics and politics.

Chapter 5

Evaluation & Results

5.1 Introduction

This chapter presents four evaluations of MetID: the results from the clustering task, two controlled experiments of the system's performance on the detection and interpretation of figurative language and a summary of a corpus-analysis case study. Because a number of variables contribute to the system's performance, each evaluation will motivate simplification of successive configurations[1]. The first evaluation reviews the clusters built with the lexical models, providing assurance that they are reliable and viable. The following three experiments evaluate MetID in increasingly complex situations and will each be presented with introduction, method, results and analysis sections. The first will test the system's ability to pick the more figurative of two sentences and will examine noun-based and verb-based statements. The second experiment is a human evaluation of the automatically generated interpretations (candidate metaphors). The last experiment reviews MetID's contribution to a terminological investigation of a *contagion* metaphor in finance, economics and politics.

5.2 Text Collections

Corpus construction is important for a number of NLP tasks [209, 214, 216] and in MetID, corpora provide the data with which to build the clusters[2]. In general, it has been found that larger corpora yield better results on many tasks, often compensating for sparsity and noise in smaller data-sets. The collections used and developed for MetID include different types, some of which were used in previous linguistic research (ANC and enTenTen), computational work (TASA, NIPS and BBY-FT-NYT) or were custom-built for this research (LexisNexis Finance). Table 5.1 summarises the collections referred to throughout in this chapter. Appendix C contains brief excerpts and descriptions of each corpus.

[1] A configuration is a choice of corpus, lexical model and similarity function.

[2] The role of corpora in semantic space modeling has been reviewed elsewhere and will not be covered here [14, 120, 136, 147, 149].

Corpus	Tokens	Documents	Open-Class Words	Sentences	Vocabulary	Size on Disk
ANC (MASC 1 & 2)	214,917	130	100,177	12,581	14,756	1.2MB
LexisNexis Finance	3,859,136	60	2,208,349	167,127	28,100	25MB
NIPS	5,280,609	1,738	2,703,695	258,236	30,133	35MB
TASA	10,766,809	38,972	5,199,253	687,736	70,089	154MB
BBC-FT-NYT	10,829,827	211	6,067,015	495,376	40,503	63MB
Partial enTenTen	72,146,944	92,327	31,225,288	3,122,227	69,745	365MB

Table 5.1: Text collections used in the development and testing of MetID. Tokens include all character sequences (numerals, names, punctuation, etc.). Open-class words are the valid English words after pre-processing. The vocabulary is the number of unique, open-class words. The two finance collections, LexisNexis and BBC-FT-NYT were developed specifically for the development of MetID and other preliminary research [67, 68, 69]. The remaining collections have been used in other computational and linguistic research (ANC [152], NIPS [189], TASA [136], enTenTen [120]).

5.3 Experiment 0: Cluster Evaluation

Because building word-clusters is a central task, every configuration of MetID was tested. There are three types of lexical models: WordNet, distributional semantic models (DSMs) and the co-occurrence likelihood estimation (COLE) model. A slightly different method was used to build clusters for each type. In WordNet, clusters are not stored but are instead computed during execution. This is because interacting with WordNet is relatively fast, and because it is not a corpus-based model, the clusters will were the same for every configuration. For DSMs, a semantic space was built using one of the similarity functions, the 200 nearest neighbours were retrieved and stored for later use (see section 4.4.4)[3]. The relative frequencies of each clustered word in the configured corpus were used to compute the cluster's purity and entropy (Eqs. 4.8 and 4.9). In the COLE-AMI model, the co-occurrence likelihood was computed between the seed terms and each word in a collection's vocabulary. Like the DSM clusters, COLE-AMI clusters were saved in a database with their purity and entropy. There was no difference in clustering topic-terms versus vehicle-terms and the cluster-pairing given by *Metalude* is not realised in the saved data.

5.3.1 Method 1: WordNet

As noted above, when configured to use WordNet, MetID computes word relatedness online instead of using pre-built clusters. By default, the system uses the Lin similarity [144] (see in section 4.4.3) but can be configured to use any relatedness measure available in the Perl WordNet::Similarity module [173][4]. During execution, the system computes every pair-wise distance between unique words in the current unit of analysis and every *Metalude* seed. Because noun and verb hierarchies lack a common root in WordNet, the abstract root-nodes were enabled in WordNet::Similarity to allow comparison between any entry. And because MetID does not perform word-sense disambiguation, the system chooses the best (ie. most similar) entry for each word in

[3] The resulting semantic spaces are available at http://www.scss.tcd.ie/~gerowa/semantic_spaces/.

[4] http://wn-similarity.sourceforge.net/; 28 February, 2013.

a pair[5]. Using WordNet provided a baseline for non-corpus-based semantic models and it will be used throughout the experiments regardless of its performance.

5.3.2 Method 2: Distributional Semantic Models

The distributional semantic models (DSMs) make up the majority of those evaluated in MetID. These include variants of LSA, HAL, BEAGLE and COALS. BEAGLE, COALS and HAL were run using their reference implementation in the S-Space[6] package and LSA was implemented in Ruby using an interface to SVDLIBC[7]. When building clusters with the DSMs, a semantic space was constructed from the configured corpus, after which each of the unique seeds from *Metalude* were used to build 200-word clusters. The relative frequency of each neighbor was used to calculate the cluster's purity and entropy[8].

Not every DSM is compatible with every corpus and not all similarity functions work with every model, which constrains valid configurations. Tables 5.2 and 5.3 show valid configurations corpus × model and model × similarity function respectively. The main constraint for corpus × model is that the number of documents in the corpus must be greater than the target representation's dimensionality. Each valid configuration were used to generate word clusters.

	ANC	**LexisNexis**	**BBC-FT-NYT**	**NIPS**	**TASA**	**enTenTen**
LSA-100	✓	✗	✓	✓	✓	✓
LSA-300	✗	✗	✗	✓	✓	✓
LSA-400	✗	✗	✗	✓	✓	✓
LSA-500	✗	✗	✗	✓	✓	✓
HAL	✓	✓	✓	✓	✓	✓
HAL-400	✗	✓	✓	✓	✓	✓
HAL-1400	✗	✓	✓	✓	✓	✓
BEAGLE-128	✓	✓	✓	✓	✓	✓
BEAGLE-256	✓	✓	✓	✓	✓	✓
BEAGLE-512	✓	✓	✓	✓	✓	✓
BEAGLE-1024	✓	✓	✓	✓	✓	✓
COALS-800	✗	✓	✓	✓	✓	✓
COALS-14k	✗	✓	✓	✓	✓	✓
COALS-SVD-100	✗	✗	✓	✓	✓	✓
COALS-SVD-200	✗	✗	✓	✓	✓	✓
COALS-SVD-800	✗	✗	✓	✓	✓	✓

Table 5.2: Compatible distributional models for each text collection.

[5] If MetID were to perform word-sense disambiguation, it could be counter-productive to the goal of testing its strategy to finding metaphor in text, because the disambiguation task might subsume the identification task.

[6] https://code.google.com/p/airhead-research/; 9 August, 2013.

[7] http://tedlab.mit.edu/~dr/SVDLIBC/; 6 March, 2013.

[8] This strategy avoids the need for an empirical cut-off for distance from the nucleus.

	Cosine	**Correlation**	**Euclidean**	**Spearman**	**Jaccard**
LSA-100	✓	✓	✓	✓	✗
LSA-300	✓	✓	✓	✓	✗
LSA-400	✓	✓	✓	✓	✗
LSA-500	✓	✓	✓	✓	✗
HAL	✓	✗	✓	✗	✓
HAL-400	✓	✗	✓	✗	✓
HAL-1400	✓	✗	✓	✗	✓
BEAGLE-128	✓	✓	✓	✓	✗
BEAGLE-256	✓	✓	✓	✓	✗
BEAGLE-512	✓	✓	✓	✓	✗
BEAGLE-1024	✓	✓	✓	✓	✗
COALS-800	✓	✓	✓	✗	✓
COALS-14k	✓	✓	✓	✗	✓
COALS-SVD-100	✓	✓	✓	✓	✗
COALS-SVD-200	✓	✓	✓	✓	✗
COALS-SVD-800	✓	✓	✓	✓	✗

Table 5.3: Compatible distributional models for each similarity function.

DSM Results

There are four variables that contribute to the clusters' composition: corpus, model, similarity function and the full / held-out distinction. This section briefly reviews the contribution of each variable with respect to purity and entropy. The full results, some of which are referred to here, can be found in appendix A.

Each of the collections in table 5.2 were used build clusters. The ANC and NIPS collections tended to have moderately higher purity and lower entropy than other collections. Between BBC-FT-NYT and LexisNexis Finance, which are both made-up of news texts, the former tended to have higher purity and lower entropy. The TASA collection, commonly used in semantic modeling applications, exhibited slightly lower purity and higher entropy than NIPS, BBC-FT-NYT and LexisNexis Finance. This may be due to the diversity of topics in TASA, not apparent in the others. The largest corpus, enTenTen, had scores similar to TASA, but usually with greater variance. This is perhaps because while enTenTen is topically diverse like TASA, it contains less formal documents from websites, blogs, news and social media outlets. Overall, no single corpus was a consistent outlier across model or similarity function.

Purity and entropy were used to ensure models provided consistent, viable clusters with each corpus and similarity function. Overall, BEAGLE clusters had higher purity and lower entropy than other models. However, HAL and COALS produced better clusters on the smaller corpora (ANC, LexisNexis and NIPS) than BEAGLE and LSA. This may be due to the frequency cut-offs employed by HAL and COALS. Also, HAL is designed to use Euclidean distance which tended to produce more variable scores than other functions. This is because Minkowski distances are unbounded which will inherently allow more outliers than a bounded metric like cosine and correlation.

LSA and COALS-SVD variants reduce the dimensionality of their representations using singular value decomposition (SVD). Across variants of LSA and COALS-SVD, the cluster qualities do not differ significantly, which points to SVD's preservation of the seed terms' influence on the resulting clusters. The non-SVD COALS variants use a frequency cut-off instead of dimension reduction, and in general, COALS-14k showed higher purity than COALS-800. On the other hand, BEAGLE variants, which refine word representations by permuting the co-occurrence matrix, show a drop in purity and a rise in entropy as the number of permutations is increased. The HAL models are relatively consistent across variant, though Euclidean distance (HAL's default similarity function) produces more variable scores than other functions.

Across similarity function, cosine and correlation yield similar results in each configuration. This appears to hold regardless of corpus, model and variant. Where compatible, the Spearman rank coefficient behaves similarly to cosine and correlation, except in LSA. This may be due to the row-entropy normalisation employed in pre-processing co-occurrence matrices for LSA. The Jaccard index, which is only compatible with HAL and non-SVD variants of COALS, is comparable to Euclidean scores, though it exhibits more variance in HAL than COALS. Excluding the Jaccard index, which is only applicable to five configurations, table 5.4 shows correlations for each function in LSA across all corpora.

Correlation	**Strength**	**Sig.**
Cosine × Correlation	$r = 0.999$	$p < 0.0001$
Cosine × Euclidean	$r = 0.835$	$p < 0.0001$
Cosine × Spearman	$r = 0.547$	$p < 0.01$
Correlation × Euclidean	$r = 0.830$	$p < 0.0001$
Correlation × Spearman	$r = 0.563$	$p < 0.01$
Euclidean × Spearman	$r = 0.182$	$p = 0.3942$

Table 5.4: Correlations of average similarity in LSA variants using different similarity functions.

The average purity and entropy should be similar for full and held-out corpora unless the held-out version was significantly different from the full corpus. Recall the hold-out procedure helps avoid representing predominantly figurative relations as if they were literal, when using a corpus-semantic model to analyse text from the same corpus. The only configuration in which the held-out clusters were significantly different from the full-corpus was with NIPS, using COALS-14k and Euclidean distance. All other scores between full and held-out clusters were within 1 SD of one-another.

5.3.3 Method 3: Co-occurrence Likelihood Estimation

The co-occurrence likelihood estimation (COLE) model, described in section 4.4.5, was based on document-ranking systems. COLE-AMI is a modified version of previous work with a visiting scholar, Dr. Hua Huo and Prof. Khurshid Ahmad [103, 104, 105]. For clustering, instead of ranking documents based on estimates of terms' co-occurrence likelihood, the algorithm computes every pair-wise estimation between a collection's vocabulary and the seed terms. The resulting score measures how likely two words are to be found near one-another and was used as a measure of their relatedness. COLE-AMI offers a corpus-based, statistical alternative to semantic space representations. For computational reasons, the enTenTen corpus could not be used in COLE-AMI model because it consists of a prohibitively large number of documents (92,327) which made the computation-time requirements intractable[9].

Table A.1 contains the average similarity, purity and entropy of the clusters built using the COLE-AMI model. Two other COLE-based models were explored, but COLE-AMI was the only one found to produce usable clusters. The other two, based respectively on multinomial and Bernoulli distribution language models, are described in appendix A, section A.1.1. In the COLE-AMI scores, TASA has the highest variance in similarity, as well as the highest average purity. The ANC and NIPS collections exhibit purities and entropies comparable to the DSMs reported above. Overall, the BBC-FT-NYT and LexisNexis collections have the largest variation for purity and entropy, whereas TASA has the most in terms of similarity. None of the held-out collections have significantly different quality scores than the full-corpus versions.

5.3.4 Summary of Cluster Evaluation

With the exception of WordNet, the word clusters constitute the lexical models used to associate observed words, MetID's input, with the figurative seeds from *Metalude*. Allowing a degree of elasticity in what constitutes relatedness among words is important. As we saw in the second chapter, the semantics of figurative language are complex and often include different lexical, semantic and pragmatic relationships. WordNet restricts the possible lexical-semantic relationships to those it explicitly encodes (done by experts). Alternatively, using the corpus-based models, word-word relations can 1) be separated from specific relationships like synonymy and antonymy and 2) allow a corpus to "dictate" its own associations. The experiment in the next section will test the models with their referenced similarity functions. For example, cosine similarity will be used for LSA and BEAGLE, Euclidean distance for HAL and the correlation function for COALS.

The following experiments are three increasingly difficult tasks: identification, interpretation and analysis. Each evaluation will motivate simplifying the design of the next by discarding poorly performing configurations. The first experiment is a "pick the figurative statement" task in which MetID will be used to examine pairs of literal / figurative statements. The best-performing models on that task will be tested further for their ability to generate interpretations for literal and non-literal statements – the results of which are evaluated by human participants. The final experiment will use MetID in a corpus-analysis of figurative terminology in financial and political texts.

[9] The TASA corpus, which could only processed using the hold-out method, took 90,579 CPU-hours.

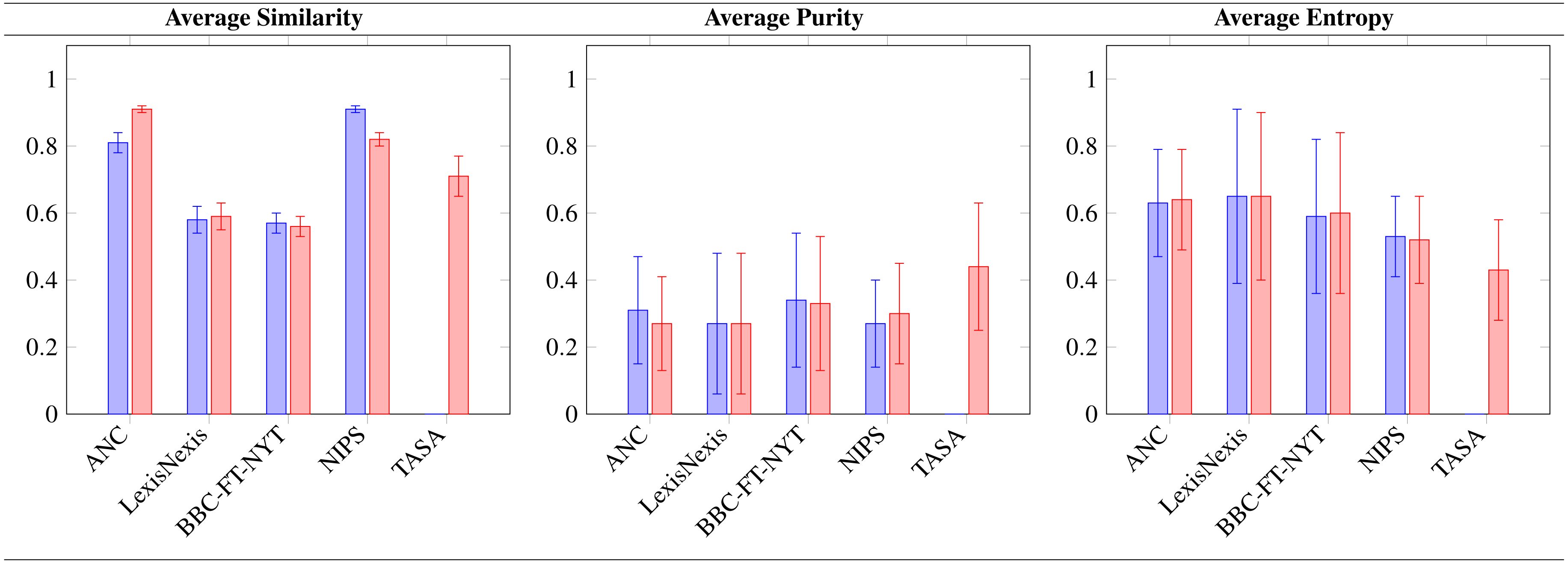


Figure 5.1: Average similarity (left), purity (Eq. 4.8; middle) and entropy (Eq. 4.9; right) of clusters built with COLE-AMI for the full-corpus (in blue) and 30% held-out versions (in red) and the TASA, ANC, NIPS, BBC-FT-NYT and LexisNexis collections. Errors bars are 1 SD of the mean.

5.4 Experiment 1: A Literal / Non-literal Choice Task

5.4.1 Introduction

This experiment was designed to test MetID's ability to discern literal from non-literal statements. To do this, two sets of test materials were used, both from previous in studies on metaphor comprehension [80, 212]. The design of the experiment is a binary decision task where two statements are submitted for analysis by MetID: a literal statement and a similar figurative statement. If it performs as intended, the system will score figurative statements higher than the literal counterparts as they should produce higher-scoring candidate metaphors. The results will be used to refine top-performing configurations for further evaluation on an interpretation task (section 5.5). This experiment exemplifies a strength of MetID: that by framing metaphor identification analogous to estimation across possible metaphors, the system can score statements on a cline of metaphoricity. This evaluation does not use the candidate metaphors produced by MetID, it only uses the highest score as an indication of how well-suited the best candidate metaphor is for a given statement.

In the best case, MetID will score all non-literal statements higher than literal statements – effectively "winning" each choice. Conversely, the worst case would be if MetID were to choose the more literal statement in each pair. As we will see, some models perform better on different kinds of statements. To reduce the experimental complexity, this experiment tested configurations with the TASA and enTenTen corpora given their large size (enTenTen) and their use in other NLP tasks (TASA) [122, 138, 184, 204, 214].

5.4.2 Method

Materials

The materials consisted of 75 literal / non-literal sentence pairs (tables 5.5 and 5.6). Of the 75 pairs, 30 were noun-based statements in canonical copular form and were based on the materials from a 1997 paper that explored how metaphors are processed [80]. For each metaphor in the original set, a literal statement was created that shared the same determiner and subject or object. The set of verb-based statements was a subset of materials used in another study examining the comprehension of such metaphors [212]. A subset of the materials in [212] contain statements with a particular verb used in alternatingly literal and non-literal statements. Each verb-based statement was in subject-noun or object-noun constructions (consistent for each pair). For DSMs, coverage in the corpus can cause MetID to overlook words, therefore, statement pairs from [80] and [212] were discarded if every open-class did not occur at least 40 times in the TASA corpus in any form. The final sets of noun- and verb-based statements are shown in tables 5.5 and 5.6 respectively.

Literal	**Non-literal**
Crime is a problem.	Crime is a disease.
A dancer is a person.	Dancers are butterflies.
Some surgeons are professionals.	Some surgeons are butchers.
A beggar is a person.	Beggars are parasites.
A brain is an organ.	The brain is a computer.
Ideas are thoughts.	Some ideas are diamonds.
A smile is a feature.	A smile is a magnet.
Experience is memory.	Experience is a fountain.
Beauty is a trait.	Beauty is a ticket.
Love is an emotion.	Love is a journey.
Malls are places.	Some malls are jungles.
Jobs are occupations.	Some jobs are prisons.
An education is a process.	An education is a doorway.
Angry words are communication.	Angry words are knives.
Faith is a belief.	Faith is a fortress.
Smallpox is a disease.	Crime is a disease.
Aspirin is a medicine.	A vacation is medicine.
Some performances are operas.	His life is an opera.
Some people are butchers.	Some surgeons are butchers.
Flees are parasites.	Beggars are parasites.
A calculator is a computer.	The mind is a computer.
Some rocks are diamonds.	Some ideas are diamonds.
Some monuments are fountains.	Experience is a fountain.
A trip is a journey.	Love is a journey.
Polio is a virus.	Rumors are viruses.
The Amazon is a jungle.	Some malls are jungles.
Some buildings are prisons.	Some jobs are prisons.
A castle is a fortress.	Faith is a fortress.
A gun is a weapon.	Humor is a weapon.
A novel is a book.	Some professors are books.

Table 5.5: Noun-based statement pairs used in experiment 1.

Literal	Non-literal
He restrained his tongue.	He bolstered his tongue.
He fastened a bandage.	He buckled a bandage.
He decorated the hero.	He garnished the hero.
He maneuvered his dance partner.	He piloted his dance partner.
She saved her money.	She filed her money.
She consumed the material.	She devoured the material.
He opened his old wound.	He unlocked his old wound.
She opened the gate.	She uncorked the gate.
She confined the spill.	She bridled the spill.
She delivered a message.	She shot a message.
The doctor mended the cut.	The doctor darned the cut.
The student stretched his string.	The student craned his string.
The man shot the cannonball.	The man evicted the cannonball.
The woman rejected the proposal.	The woman killed the proposal.
The woman repaired his suit.	The woman sutured his suit.
The man stole their solution.	The man kidnapped their solution.
They ended the alliance.	They melted the alliance.
They dropped the candy into the bag.	They parachuted the candy into the bag.
They withdrew the invitation.	They retreated the invitation.
They released the prisoner.	They unleashed the prisoner.
We excised the scene.	We amputated the scene.
The engine wore out.	The engine frayed out.
Confetti fell on the arena.	Confetti rained on the arena.
His waistline grew no matter what.	His waistline inflated no matter what.
She was trying to provide for her children.	She was floundering to provide for her children.
The ancient car fell apart.	The ancient car unravelled apart.
The trial displaced all other stories.	The trial pushed all other stories.
The station wagon travelled back home.	The station wagon limped back home.
The boats moved along shore.	The boats danced along shore.
The boy grabbed his bike and went home.	The boy grabbed his bike and flew home.
The building shook from the earthquake.	The building shivered from the earthquake.
The bulldozer travelled towards town.	The bulldozer lumbered towards town.
The child expressed his need for attention.	The child howled for attention.
The clouds gathered on the horizon.	The clouds swarmed on the horizon.
The garbage truck ate the debris.	The garbage truck consumed the debris.
The runners ran through the streets.	The runners streamed through the streets.
The bike moved along the trail.	The tiptoed along the trail.
The plants adapted to the constraints.	The plants obeyed the constraints.
The poster hung over the desk.	The poster hovered over the desk.
The van was idling on the road.	The van was slumbering on the road.
The truck climbed up the slope.	The truck crawled up the slope.
The bread rose to perfection	The bread climbed to perfection.
The waste contaminated the workers.	The waste infected the workers.
The troops forced their way through the defenses.	The troops stormed their way through the defenses.
The house decayed over time.	The house wilted over time.

Table 5.6: Verb-based pairs used in experiment 1.

Procedure

MetID was run in decision mode on each of the statement pairs. The higher scoring statement in each pair is the one for which MetID found a better candidate metaphor – implying it was more figurative than the alternative. This computation was done according to the procedure outlined in section B.4. For each input pair, the result is the difference between the non-literal score and the literal score. If MetID correctly judged the literal to be less figurative than the non-literal, the difference is positive.

Of the 135 possible configurations of model × corpus × similarity function, the 18 listed in table 5.7 were used with the TASA and enTenTen collections, yielding a total of 36 configurations. Two sets of results will be presented: those run with the TASA corpus and those with enTenTen corpus. Recall the TASA collection was developed to represent general knowledge among American, college-educated students, while the enTenTen corpus is an accumulation of web-based text with no restrictions on type, genre, topic or domain. In all cases, the clusters built with the held-out collections were used (see section 4.4.2).

Model	Variables	Similarity	Notes
WordNet		Lin Similarity [144]	v3.1; Not corpus-based [167]
HAL		Euclidean	Reference implementation [150]
HAL	Keep 400	Euclidean	
HAL	Keep 1400	Euclidean	
COALS	Keep 800	Correlation	
COALS	Keep 14000	Correlation	Reference implementation [188]
COALS	SVD-100	Correlation	Dimensionally reduced
COALS	SVD-200	Correlation	Dimensionally reduced
COALS	SVD-800	Correlation	Dimensionally reduced
BEAGLE	128 permutations	Cosine	
BEAGLE	256 permutations	Cosine	
BEAGLE	512 permutations	Cosine	Reference implementation [113]
BEAGLE	1024 permutations	Cosine	
LSA	100	Cosine	
LSA	300	Cosine	Reference implementation [136]
LSA	400	Cosine	
LSA	500	Cosine	
COLE		Augmented Mutual Information	

Table 5.7: Lexical models tested in experiment 1. This selection provides a cross-section of various types (WordNet, DSMs and COLE-AMI) and variants (dimensionality, frequency cut-offs, etc.).

5.4.3 Results

Results will be presented for the TASA corpus first and then for the enTenTen collection, after which a discussion will briefly compare performance across corpora. In each case, the effect of the model and the grammatical type (noun- or verb-based) will be examined. The scores will be used to rank the performance of each configuration, but the results will also be subjected to a two-way ANOVA to explore the contribution of each variable on the system's performance. While the ANOVA explores differences in the group-wise means and tests for interactions effects, the ranking results will be used to examine what configurations perform best.

TASA

To establish the role of model configuration and grammatical type, the scores were analysed with a two-way ANOVA. Statement-pairs that were not computable, or resulted in a tie (the same score for both statements) were not included. Intuitively, this test addresses the variables (lexical model and grammatical type) that influence the identification of non-literal statements. Keep in mind that ties and "misfires" usually occur when a potential topic and / or vehicle term were identified, but could not be found in the lexical model. Coverage is an important aspect affected by the corpus and lexical model, but it is not necessarily indicative of MetID's design.

The analysis of variance found that the grammatical type had a significant effect on the scores ($F_{1,675} = 146.42$, $p < 0.0001$) and accounted for 17.8% of the variance (partial η^2). This effect is shown in figure 5.2 (bottom right). Taken alone, the lexical model did not have a significant effect on the scores ($p = 0.356$) (figure 5.2; upper right). However, there was a significant interaction between model and grammatical type ($F_{17,675} = 3.55$, $p < 0.0001$) that accounted for 8.2% of the variance (partial η^2). This interaction is simply that the choice of model made a significant difference for noun-based statements but not for verbs. In fact, none of the models perform significantly above chance (0) on the verb statements, whereas as all but one model (HAL-400) scored above chance on the noun statements. Not considering coverage, LSA-400 and LSA-500 are the two top-performing models for noun-based statements.

Knowing the grammatical form of a statement plays a significant role in MetID's performance is important, however, a simple question remains: which model is best? Table 5.8 shows the raw results for each model tested. For noun statements, LSA-500, COLE-AMI and COALS-800 all had zero losses but a large number of misfires. Looking at the ratio of wins to losses over misfires and ties ($\frac{Wins/Losses}{Misfires+Ties+1}$), WordNet and COALS-SVD-200 are the best. This ranking accounts for coverage concerns. For nouns, WordNet and COALS-SVD-200 – both of which achieved full coverage – perform the best[10]. The ANOVA found that WordNet performed about 20% above chance and COALS-SVD-200 about 15% though their coverage was relatively high. For verbs, the best configuration is again WordNet, however, recall that none of the configurations performed significantly above chance in this case.

[10] Note the scores for this ranking have no interpretable analogue.

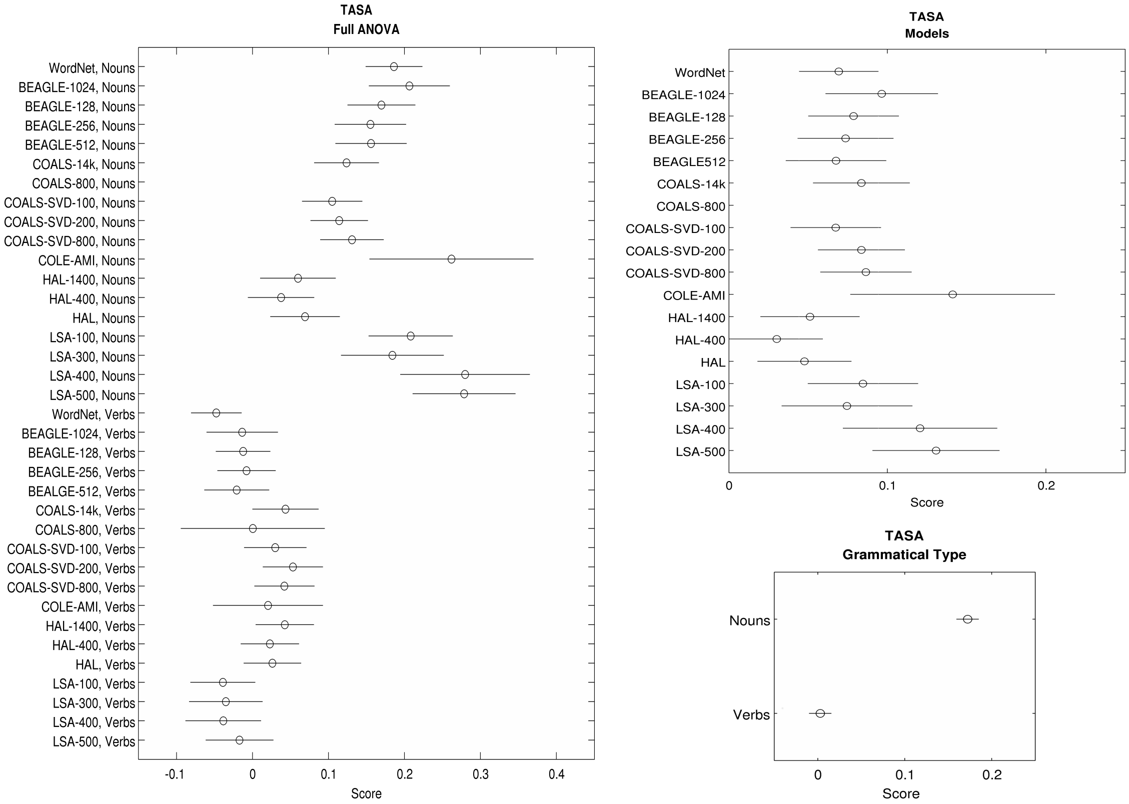


Figure 5.2: Comparison of mean scores in experiment 1 across model and grammatical type using the TASA corpus. The scores are the difference between MetID's scores for the non-literal statement and the literal statement (0 is the 50% chance baseline). These scores measures how much better a given configuration performed when trying to choose the more figurative statement from a pair. Shown here are the full analysis (left) and group-wise means for model (top right) and grammatical type (bottom right). Error-bars are symmetric confidence intervals measured using Tukey's HSD ($\alpha = 0.05$). Note that in the noun case, COALS-800 was only able to produce results for three pairs, and therefore did not meet the minimum statistical criteria for inclusion.

Model	**Type**	**Misfires**	**Ties**	**Wins**	**Losses**	$Wins - Losses$	$Wins/Losses$	$\frac{Wins/Losses}{Misfires+Ties+1}$
WordNet	Noun	0	0	23	7	16	3.29	3.29
COALS-SVD-200	Noun	0	1	22	7	15	3.14	1.57
BEAGLE-128	Noun	7	3	18	2	16	9	0.82
COALS-14k	Noun	5	3	19	3	16	6.33	0.70
COALS-SVD-800	Noun	3	4	19	4	15	4.75	0.59
COALS-SVD-100	Noun	2	2	18	8	10	2.25	0.45
HAL	Noun	11	0	16	3	13	5.33	0.44
BEAGLE-512	Noun	10	2	15	3	12	5	0.38
LSA-300	Noun	20	1	8	1	7	8	0.36
BEAGLE-1024	Noun	14	2	12	2	10	6	0.35
HAL-400	Noun	9	0	16	5	11	3.2	0.32
HAL-1400	Noun	12	2	12	4	8	3	0.20
BEAGLE-256	Noun	10	2	13	5	8	2.6	0.20
LSA-400	Noun	23	1	5	1	4	5	0.20
LSA-100	Noun	16	1	10	3	7	3.33	0.19
LSA-500	Noun	20	1	9	0	9	NaN	0.00
COLE-AMI	Noun	25	1	4	0	4	NaN	0.00
COALS-800	Noun	27	0	3	0	3	NaN	0.00
WordNet	Verb	0	5	15	25	-10	0.60	0.10
COLE-AMI	Verb	37	0	6	2	4	3.00	0.08
COALS-SVD-200	Verb	14	5	15	11	4	1.36	0.07
HAL-1400	Verb	10	7	15	13	2	1.15	0.06
HAL-400	Verb	10	7	15	13	2	1.15	0.06
COALS-SVD-100	Verb	14	7	14	10	4	1.4	0.06
HAL	Verb	13	3	15	14	1	1.07	0.06
COALS-SVD-800	Verb	13	6	14	12	2	1.17	0.06
COALS-14k	Verb	16	8	12	9	3	1.33	0.05
LSA-500	Verb	20	5	10	10	0	1	0.04
BEAGLE-128	Verb	4	8	11	22	-11	0.5	0.04
COALS-800	Verb	34	6	3	2	1	1.5	0.04
BEAGLE-256	Verb	10	7	10	18	-8	0.56	0.03
LSA-400	Verb	25	4	6	10	-4	0.6	0.02
LSA-100	Verb	15	8	7	15	-8	0.47	0.02
LSA-300	Verb	22	6	6	11	-5	0.55	0.02
BEAGLE-512	Verb	16	7	6	16	-10	0.38	0.02
BEAGLE-1024	Verb	20	7	5	13	-8	0.38	0.01

Table 5.8: Results on experiment 1 using the TASA corpus, ordered by the win-loss ratio over the number of misfires and ties (last column). Misfires imply that an identified term was not found in the lexical model. Note that disregarding coverage issues will provide a different ranking.

enTenTen

The results using the enTenTen collection were similar to the TASA results. Because WordNet is not a corpus-based lexical model, the scores are the same as they were for TASA; it is included here for comparison purposes. The analysis of variance (two-way ANOVA) showed that grammatical type contributed significantly to the variance in scores ($F_{1,604} = 109.72$, $p < 0.0001$) with an effect size of of 8.2% (partial η^2). As a main-effect, the choice of model did not contribute significantly ($p = 0.5624$). There was a significant interaction between model and grammatical type ($F_{16,604} = 2.525$, $p < 0.001$) with partial $\eta^2 = 0.063$. This interaction was the same as that found for the TASA corpus: in the noun case, the model made a significant difference, but in the verb case, it did not. Figure 5.3 shows the means comparisons.

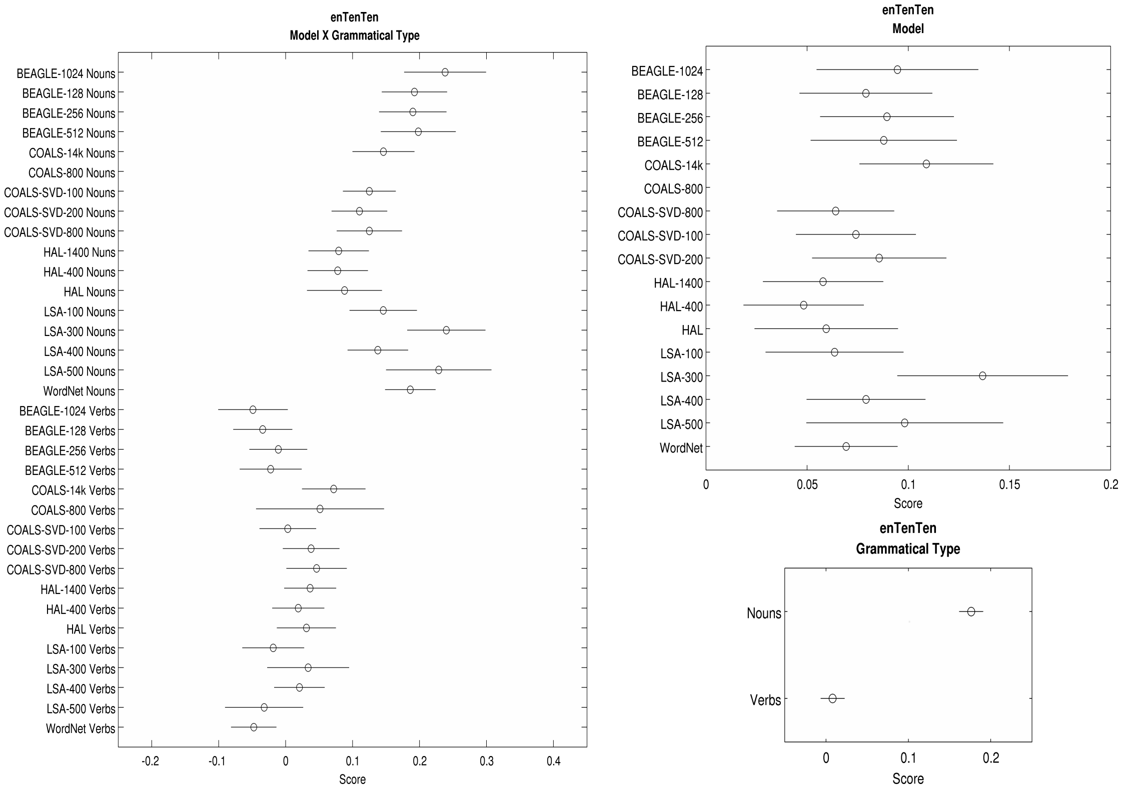


Figure 5.3: Comparison of means (two-way ANOVA) across model and grammatical type for the enTenTen corpus. Shown here are the full analysis (left) and group-wise means for model (top right) and grammatical type (bottom right). Error-bars are symmetric confidence intervals measured using Tukey's HSD (α = 0.05). COALS-800 did not produce any results and was not included in the ANOVA.

The ANOVA provides a comparison of group-wise means, but again, considering coverage, a different picture emerges of the best-performing models. Table 5.9 ranks the configurations, taking into account their coverage. For nouns, WordNet scored the highest, while COALS-14k and two COALS-SVD rank second, third and fourth. Note that the COALS models are the best-performing DSMs in the noun cases which is similar to the results for the TASA corpus. The best-performing models for verb-based statements are WordNet and HAL-400 – neither of which performed significantly above chance in the ANOVA.

Model	**Type**	**Misfires**	**Ties**	**Wins**	**Losses**	$Wins - Losses$	$Wins/Losses$	$\frac{Wins/Losses}{Misfires+Ties+1}$
WordNet	Noun	0	0	23	7	16	3.29	3.29
COALS-14k	Noun	9	2	17	2	15	8.5	0.71
COALS-SVD-100	Noun	2	1	19	8	11	2.375	0.59
COALS-SVD-800	Noun	9	4	15	2	13	7.5	0.54
BEAGLE-1024	Noun	16	3	10	1	9	10	0.50
COALS-SVD-200	Noun	5	1	18	6	12	3	0.43
HAL-1400	Noun	9	1	16	4	12	4	0.36
LSA-400	Noun	10	0	16	4	12	4	0.36
BEAGLE-128	Noun	12	1	14	3	11	4.67	0.33
BEAGLE-512	Noun	16	1	11	2	9	5.5	0.31
HAL	Noun	16	1	11	2	9	5.5	0.31
LSA-500	Noun	21	2	6	1	5	6	0.25
BEAGLE-256	Noun	12	2	12	4	8	3	0.20
HAL-400	Noun	9	1	13	7	6	1.86	0.17
LSA-300	Noun	17	1	9	3	6	3	0.16
LSA-100	Noun	13	1	10	6	4	1.67	0.11
COALS-800	Noun	28	1	1	0	1	NaN	0.00
WordNet	Verb	0	5	15	25	-10	0.6	0.10
HAL-400	Verb	9	7	18	11	7	1.64	0.10
COALS-800	Verb	17	10	13	5	8	2.6	0.09
COALS-SVD-800	Verb	15	7	14	9	5	1.56	0.07
HAL-1400	Verb	9	8	15	13	2	1.15	0.06
LSA-400	Verb	11	4	14	16	-2	0.88	0.05
COALS-SVD-100	Verb	15	10	11	9	-2	1.22	0.05
COALS-14k	Verb	15	10	11	9	2	1.22	0.05
HAL	Verb	14	10	11	10	1	1.1	0.04
LSA-300	Verb	26	7	7	5	2	1.4	0.04
BEAGLE-256	Verb	16	6	10	13	-3	0.77	0.03
COALS-SVD-200	Verb	14	8	10	13	-3	0.77	0.03
LSA-100	Verb	17	8	9	11	-2	0.82	0.03
BEAGLE-128	Verb	17	7	9	12	-3	0.75	0.03
BEAGLE-512	Verb	19	7	8	11	-3	0.73	0.03
LSA-500	Verb	26	7	5	7	-2	0.71	0.02
BEAGLE-1024	Verb	25	5	5	10	-5	0.5	0.02

Table 5.9: Results for experiment 1 using the enTenTen corpus with each model for each grammatical type, ordered by the win / loss ratio over the number of misfires and ties.

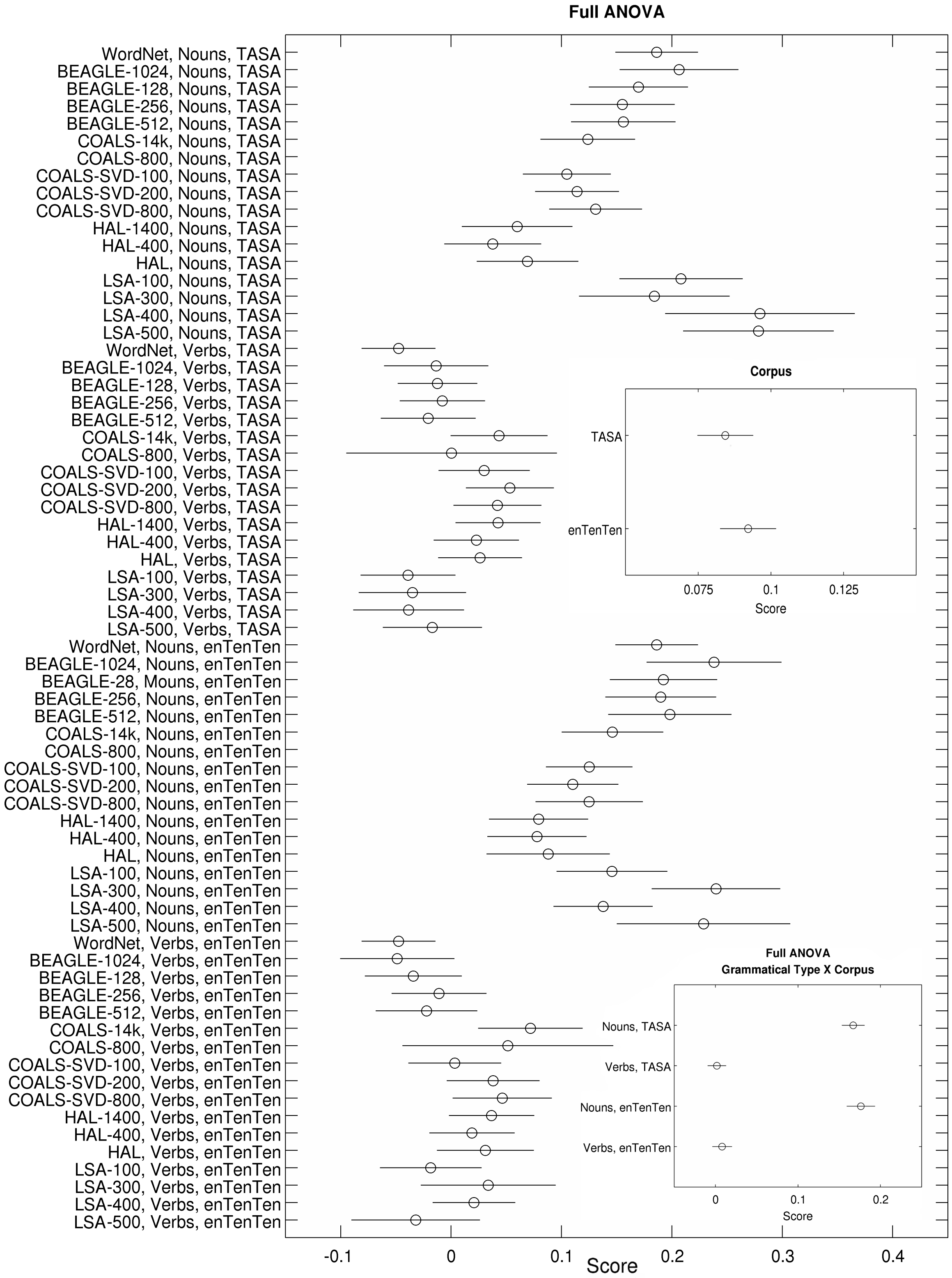


Figure 5.4: Means comparison for the three-way ANOVA between corpus, model and grammatical type for experiment 1. In-set comparisons are for the group-wise means for corpus (TASA vs. enTenTen; top) and grammatical type × corpus (bottom). All error-bars are symmetric confidence intervals measured using Tukey's HSD test ($\alpha = 0.05$).

Summary

To verify the choice of corpus did not play a significant role in this experiment, a three-way ANOVA was run on model × grammatical type × corpus. The results were similar to those found using the collections individually. There was a main-effect of grammatical type ($F_{1,1269} = 248.26$, $p < 0.0001$, partial $\eta^2 = 0.164$) and an interaction between grammatical type and model ($F_{16,1269} = 5.7$, $p < 0.0001$, partial $\eta^2 = 0.067$). Including the corpus, there were no other significant main, two- or three-way effects (all $p > 0.05$). The full comparison of means is shown in figure 5.4.

The analyses of variance for the individual corpora as well as the three-way analysis above, confirmed two pieces of information. First, that the grammatical type has an influence on MetID's performance. Second, that in the noun-case, the lexical model plays a significant role in how well the system performs. Regardless of corpus or lexical model, the grammatical type accounted for more variance than the other variables. This result, which will be explored more in the next experiment, is perhaps predicted by the motivations of the method implemented in MetID.

5.4.4 Analysis & Discussion

The results of this experiment provide evidence for two observations about MetID: that it is more attenuated to identifying noun-based metaphors and that when processing such metaphors the lexical model is important. This experiment was designed as a kind of easiest possible task; choosing the more figurative of two statements means a null model will achieve 50% accuracy – defined as the number of wins over the total number of trials. None of the configurations performed significantly better than the null model for verbs. For nouns, on the other hand, some configurations achieved nearly 75% accuracy – a compelling reason to further explore performance with these configurations.

Consider the noun / verb distinction. Recall our review in chapter 2 of the various types of metaphor and how they relate to conceptual theory of metaphor (CMT) and two theories of comprehension: category matching and structure mapping [80, 133, 235]. Both comprehension theories, as well as CMT, are grounded in the idea that a metaphor consists of one concept (a topic) in terms of another (a vehicle). By instantiating a metaphor as a verb-based statement, the grammatical structure can obscure the topic-vehicle structure. Take (xx), which was used in this experiment:

(xx) She devoured the material.

MetID will identify *devour* and *material* as potential topic and vehicle terms. Additionally, there is likely a degree of selectional violation for the relation `dobj(devour,material)`. However, matching *devour* to a term in *Metalude* will likely require overlooking (or at least proceeding without) an explicit lexical semantic relationship, because the seed terms consist mostly of noun-concepts. This requires the lexical model to explicitly associate a *Metalude* noun-concept to the observed verb-concept. Unfortunately, the more common (and more frequently explicated) relationships among nouns are likely to override their verb relationships. For (xx) in the best case, MetID would relate *devour* to EATING, FOOD or perhaps CONSUMING and *material* to INFORMATION or KNOWLEDGE. These associations would produce metaphors such as FOOD AS KNOWLEDGE or CONSUMING AS KNOWING. However, the results show that relating terms from the verb-based statements is considerably harder than from noun-based statements – so much so that the choice of model does not make a significant difference. This implies that the core strategy of MetID is better suited to noun-based analysis, a theme that will be discussed more in the following chapter.

In noun cases, performance varies between different lexical models. WordNet performs relatively well and is the best in terms of coverage (all statements were processed). However, WordNet's win / loss ratio is lower than some other models, the two best being BEAGLE-128 (18 wins, 2 losses) and COALS-14k (19 wins, 3 losses) for the TASA corpus. Using the enTenTen collection, BEAGLE-1024 is the best at 10 wins, 1 loss, seconded by COALS-14k (17 wins and 2 losses). Looking at the top five corpus-based models in the noun-case for both corpora, it turns out that variants of COALS comprise 8 of the top 10 and variants of BEAGLE are the remaining two. This ostensibly speaks to COALS being better at representing figurative associations between observed words and the *Metalude* terms. Further, the LSA variants are all in the bottom five DSMs. This is interesting because LSA-500 scored relatively high on the ANOVA, which implies that when LSA gets it right, it does so with a large margin of error (ie. the non-literal statement score *much* higher than the literal). Generally, the newer models (COALS and BEAGLE) perform better than the older ones (LSA and HAL).

This experiment showed that MetID is able to identify non-literal noun-based metaphors with about 75% accuracy when they are framed in a binary choice format. Identification is the first goal of the system, the second being interpretation. It could be the case that configurations that performed poorly here are nonetheless able to provide accurate interpretations for known metaphors. This is the topic of the next experiment. The results here were used to narrow the range of configurations tested in the next experiment. WordNet is included again, as it provided good coverage and performed comparably to the DSMs. COLE-AMI performed uniformly low and will not be tested further. Because identification is somewhat independent of the interpretation task, the best performing variants of each DSM will be included in the next task. Specifically, we will look at LSA-400 and LSA-500, the best of the HAL variants (HAL), the best COALS model (COALS-14k; taking coverage into account) and the two best BEAGLE variants (BEAGLE-128 and BEAGLE-1024).

5.5 Experiment 2: Interpreting Non-literal Statements

5.5.1 Introduction

The first experiment was a relatively simple task which tested MetID's ability to choose a figurative statement from pairs of literal / figurative sentences. The system's performance was significantly affected by the grammatical form of the statement, and the choice of lexical model was important when processing noun-based statements. The first experiment only used the score of the top candidate metaphor to rate how likely it was that a statement was a metaphor in general. The current experiment evaluates the actual metaphors MetID generates as interpretations to the input statements. Because the lexical model is perhaps the most interesting variable, this experiment continues to test the performance of different configurations across models. Only the TASA collection is used because it has been used in similar research and was shown in the previous experiment to perform comparably well to the enTenTen collection. The configurations evaluated are listed in table 5.10. Because this experiment explores MetID's ability to generate interpretations for *figurative* language, the system was also tested on literal statements. Conceivably, some candidate metaphors will be accurate interpretations for literal statements, but in general, MetID should provide better interpretations to figurative language. Thus, the experiment tests the contribution of three variables: the lexical model, the grammatical form of a statement and statements' literalness.

Model	Distance Metric	Corpus	Notes
WordNet	Lin Similarity [144]	N/A	Version 3.1 [167]
COALS-14k	Correlation	TASA	Reference implementation [188]
BEAGLE-1024	Cosine	TASA	Uses 1024 permutations
BEAGLE-128	Cosine	TASA	Uses 128 permutations
HAL	Euclidean	TASA	Reference implementation [150]
LSA-400	Cosine	TASA	Uses 400 dimensions
LSA-500	Cosine	TASA	Uses 500 dimensions

Table 5.10: Configurations of MetID tested in experiment 2.

The procedure for this experiment is based on [83, 198, 212] and uses human participants to qualify sentences and their potentially metaphorical paraphrases. The crux of the experiment is that instead of asking participants to generate the "correct" metaphor for a given sentence, the task is instead framed as a paraphrasing exercise. By allowing participants to rate the quality of a short paraphrase – irrespective of a literal / figurative distinction – the task remains consistent across materials. Further, it does not rely on peoples' intuition about figurative language, which as we have seen, can be deceptively complex. An online survey was used to gather ratings for the sensibility of sentences and the quality of related paraphrases – ie. the top-ranked candidate metaphor processed by MetID.

This experiment evaluates the interpretation function of MetID. While the first experiment used aggregate scores across the best candidate metaphors to build a score of "metaphoricity", here the candidate metaphors themselves are used as potential interpretations. In the best case, MetID would produce uniformly correct interpretations for all literal and non-literal sentences, but the literal interpretations would have lower scores than the non-literal (indicating their metaphoricity). Additionally, in the best case, verb-based statements would be interpreted as accurately as noun-based. In the worst outcome, not only would no configuration produce good interpretations, but they would do so without regard to the lexical model, the literal / non-literal distinction or the grammatical form of the statements. As we will see, the lexical model, grammatical form and the literal / non-literal distinction all contribute significantly to MetID's performance. The results also point to an interaction between grammatical form and literalness.

5.5.2 Method

Participants

A total of 291 people participated in the user-study. An initial group of 31 acquaintances were recruited by email, whom, upon completion, were asked to share the study via email, Facebook, Google+ and Twitter. The initial group was made up of 17 women, 14 men and consisted primarily of friends and colleagues in the Dublin area. There was no incentive to participate, nor any penalty for not completing the survey. Participants were briefed with an information page about how their data would be gathered, stored, analysed and potentially published after which the rating tasks were explained with two examples. Of the 291 participants who were presented with the instruction page (those who clicked "Continue" after informed consent), 147 were excluded from analysis for the following reasons:

- The participant did not complete the survey or opted not to submit their results upon completion. (98 participants)
- The participant failed two or more of six planted questions – implying they misunderstood the task, or were not completing the survey mindfully. (39 participants)
- The participant reported being under 18 years old. (5 participants)
- The participant reported not being fluent in English. (4 participants)
- There was evidence of technical problems. (1 participant)

This resulted in 144 participants, whom reported being fluent in English and over 18 years old (*M* = 34.4, SD = 14.1).

Materials

Sentences. 80 literal and non-literal sentences from the first experiment were used in this study. The statements' literal to non-literal pairings were discarded; they were processed individually. The sentences are listed in appendix C and consisted of 20 literal noun-based statements, 20 non-literal noun-based, 20 literal verb-based and 20 non-literal verb-based statements.

Paraphrases. Each sentence was processed by every configuration in table 5.10. The best candidate metaphor was taken by ranking *all* the candidates, produced by the given configuration, for a given input. Not every sentence could be processed by every configuration, given the requirement that the open-class words be represented in the lexical model. What would have been a total of 560 sentences (80 × 7) became 432, that were included in the survey. The full set of materials is shown in appendix C, table C.4.

The paraphrases are grammaticalised versions of the candidate metaphors from MetID. Grammaticalisation refers to making the resulting metaphor a valid phrase; if the candidate metaphor was LOVE = WARMTH, the paraphrase became "LOVE IS WARMTH". Grammaticalisation was done manually for each of the 432 instances, a sample of which is shown below in table 5.11. The scores for each of the top-ranked candidate metaphors were not used in the survey, but will be used in analysing the results. Following similar studies' method, participants were not told the paraphrases were potentially metaphorical [198]. This reduced participants' tendency to overthink figurative interpretations, and to instead rely on validating the synoptic accuracy of the paraphrases.

Model	**Type**	**Figurative?**	**Score**	**Sentence**	**Candidate Metaphor / Paraphrase**
WordNet	noun	yes	0.97	His life is an opera	ACTIVITY IS MUSIC
WordNet	noun	yes	0.90	Crime is a disease	A PROBLEM IS A DISEASE
WordNet	noun	yes	0.78	A vacation is medicine	MONEY IS FOOD
WordNet	noun	yes	0.75	Dancers are butterflies	A HUMAN IS AN INSECT
WordNet	noun	yes	0.54	Some surgeons are butchers	STEALING IS HITTING
WordNet	noun	no	0.89	Crime is a problem	A PROBLEM IS A DISEASE
WordNet	noun	no	0.87	Some urban schools are crowded	CONTROLLING IS PUSHING
WordNet	noun	no	0.81	That lost painting is a portrait	AN OPINION IS A VIEW
WordNet	noun	no	0.80	A snail is a pest	A HUMAN IS A PIG
WordNet	noun	no	0.61	A lion is an animal	AN ANIMAL IS A HUMAN
WordNet	noun	no	0.61	That creature in the net is a crab	AN ANIMAL IS A HUMAN
WordNet	noun	no	0.59	Some ideas are great	AN IDEA IS A COMMODITY
WordNet	noun	no	0.59	Some jobs are constraining	A JOB IS A POSITION
WordNet	noun	no	0.59	Some lectures are boring	SPEECH IS A GAME
WordNet	noun	no	0.54	My brother is a butcher	STEALING IS HITTING
WordNet	noun	no	0.51	A salmon is a fish	A HUMAN IS A FISH
WordNet	noun	no	0.50	Cereal is a food	A HUMAN IS FOOD
WordNet	noun	no	0.23	The Earth is a planet	A BODY IS THE EARTH
WordNet	noun	no	0.11	Sharks have sharp teeth	A HUMAN IS A FISH

Table 5.11: Sample materials for the user study generated using the WordNet model with Lin similarity. Similar materials were derived using the other configurations listed in table 5.10. The type may be noun or verb, depending on the form of the statement. The score refers to MetID's top-scoring candidate metaphor, which is presented in a valid grammatical form as a paraphrase of the sentence.

Procedure

The survey was administered anonymously as a web-based questionnaire. Full instructions and example screen-shots can be found in appendix C. Participants were asked for two ratings: first for the sensibility of a sentence (“How sensible is:”) and then for the quality of a related summary (“How well is it summarized by:”). Each rating was on a seven-point semantic differential Likert scale from *bad* to *excellent*. Paraphrases were presented in block caps to distinguish them from sentences, and each pair was grouped with alternating white and grey backgrounds to distinguish them visually. Each survey contained a total of 60 questions (sentence-paraphrase pairs) in six blocks of ten per page. After completing the final block, participants were allowed to exit without submitting their results. Six of the questions (one per page; 10%) were planted questions designed to verify participants’ understanding and mindfulness in completing the task.

Each participant answered 60 random question-pairs, of which 54 were the results of three-variable configurations: model (7 levels; see table 5.10) × grammatical type (2 levels; noun or verb) × literalness (2 levels; literal or non-literal). This yields a 7 × 2 × 2 design. Sensibility ratings (which pertain to a sentence independent of its paraphrase) were used to disqualify 24 sentences that elicited a mean rating below 4, across all participants[11]. The remaining sentences had the following mean sensibility ratings: 6.73 (SD=0.44) for literal nouns, 5.569 (SD=0.6) for non-literal nouns, 6.67 (SD=0.27) for literal verbs and 5.13 (SD=0.73) for literal verbs.

A small pilot study was conducted with five colleagues familiar with experimental design, from whom feedback was solicited. The feedback prompted minor changes to the layout of the questionnaire as well correcting two mistyped paraphrases. The results of the pilot study were not included in the analysis. Ethical approval was granted by the School of Computer Science & Statistics, Trinity College Dublin on 17 April, 2013.

5.5.3 Results

Ratings for sensibility and paraphrase quality were combined by weighting the paraphrase scores attenuated by the sensibility scores. The score for a given sentence-paraphrase pair was calculated as the product of the *paraphrase rating* and the *sensibility rating* divided by 7 (because the rating was on a seven-point scale). Henceforth, a ”score” for an input pair refers to this calculation, which ranges from 0.14 to 7. The scores for every question were averaged over all responses, in which there were more than 10 in every case. The rationale for scaling the quality scores down by the sensibility ratings was to avoid reporting false positives. While there is no reason to trust bad quality ratings over good ones if they both received lower sensibility ratings, this situation would result in falsely reporting the successful interpretation of a given input pair. Thus, the reported results have a conservative slant toward the negative. An alternative analysis, which was not performed, would be to simply not scale the quality scores by the sensibility ratings – a topic discussed more in the following section.

[11] These sentences are listed in appendix C, table C.3 with their mean sensibility rating.

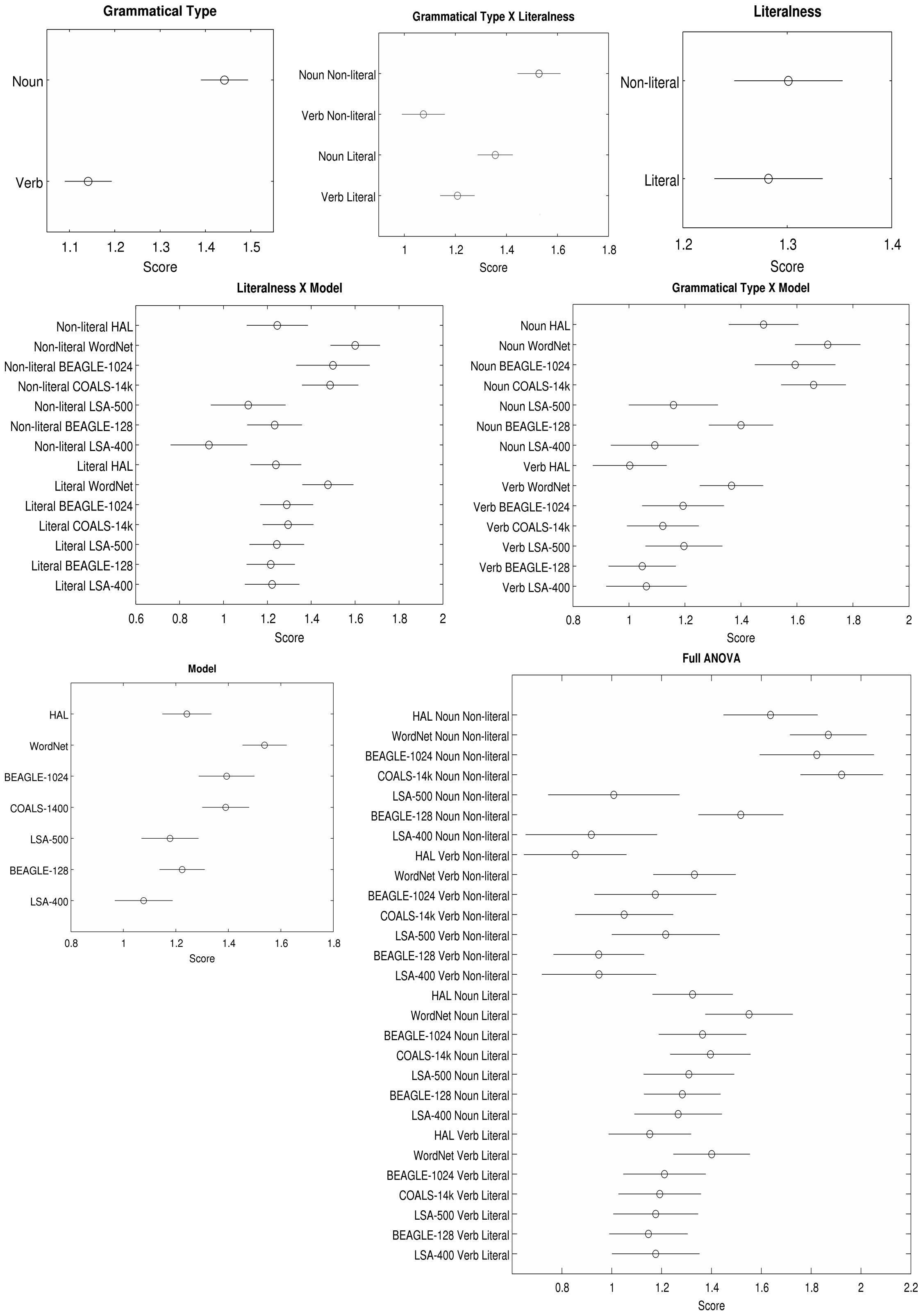


Figure 5.5: Comparison of mean scores for all combinations of model, grammatical type and literalness. Points represent within-group means and error-bars are symmetric confidence intervals calculated using Tukey's HSD test ($\alpha = 0.05$).

A three-way ANOVA was used to compare the effects of model, grammatical type and literalness. Model (7 levels) and grammatical type (2 levels) contributed significantly to variation in the scores (model $F_{6,378} = 4.52$, $p < 0.001$; grammatical type $F_{1,378} = 27.75$, $p < 0.001$). The effect-size (as partial η^2) show that choice of model accounted for 7% of the variance and grammatical type accounted for 8%. As a main effect, literalness did not contribute significantly. Figure 5.5 compares the means across all groups.

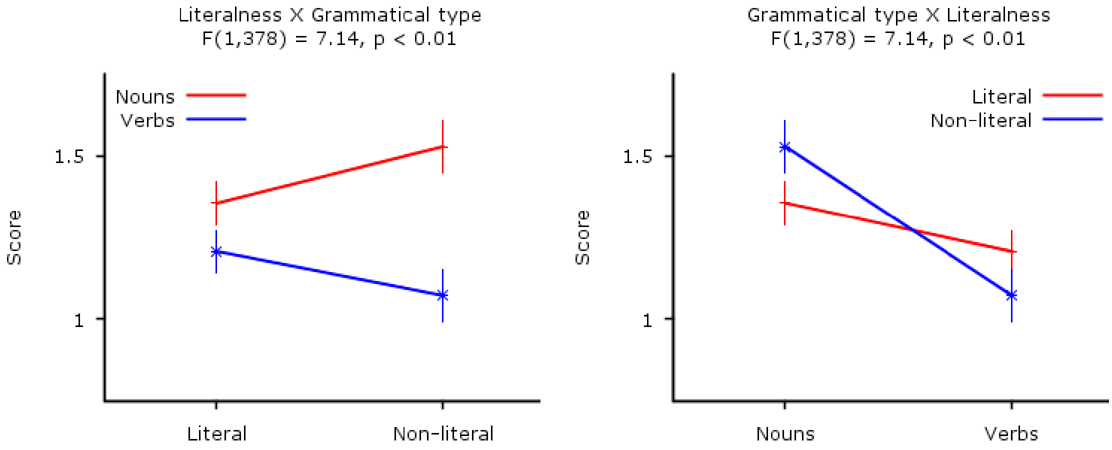


Figure 5.6: Interaction diagrams for the effect of literalness on scores for noun- and verb-based statements (left) and the same interaction for the effect of grammatical type on literal and non-literal statements (right). The vertical error-bars are 2 times the standard deviation of the within-group mean.

There was also a significant interaction: a two-way effect between grammatical type and literalness ($F_{1,378} = 7.14$, $p < 0.01$). The interaction is that for non-literal statements, grammatical type had significantly larger effect on scores than it did for literal statements (see figure 5.6; left). Paired samples t-tests for each case found significant differences with respect to literalness for nouns (t_{202} =2.771, $p < 0.01$; two-tailed) and verbs (t_{200} =2.669, $p < 0.01$; two-tailed). This means not only did literalness mediate the relevance of grammatical type, it did so differently for nouns and verbs: nouns achieve higher scores in non-literal statements while verbs are lower. The interaction can also be interpreted as grammatical type mediating the contribution of literalness. This interaction (figure 5.6; right) is significant across literalness in both cases: literal (t_{234} =2.701, $p < 0.01$; two-tailed) and non-literal (t_{168} =4.885; $p < 0.0001$; two-tailed). Interpreted this way, the scores were generally worse for verb-based statements compared to nouns, but the decrease was more pronounced for non-literal statements than literal. All other two- and three-way interactions were non-significant (all $p > 0.05$).

Variance in the mean scores is one way to examine which models perform well in different situations, but it disregards some important information from MetID – namely the score of the candidate metaphor. Because no configuration performed well enough to be considered an outright success in terms of its paraphrase ratings, which configurations provided good correlations between model scores and participants' ratings was also explored. This is important because though MetID will always generate a "best" interpretation, it may not be very good – a property evident in the resulting score. Optimally, a low score from MetID would be matched by low participant ratings; if people rate the paraphrase low, MetID's interpretation should also have a low score. Thus, a high correlation between the ratings and model scores would indicate that MetID was able to account for the relative quality of its interpretations. Table 5.12 shows correlations for each configuration of MetID between the model scores and the ratings from participants.

In table 5.12, first note that WordNet provides the strongest correlation overall, at 0.254. Looking at literalness and grammatical type, note that neither case produced particularly strong correlations, though nouns were generally stronger than verbs and non-literal statements stronger than literal. Over the individual models, WordNet performs best on literal statements overall, as well as on verb-based literal statements. A corpus-based model, however, outperforms WordNet in non-literal cases as well as for statements irrespective of literalness. LSA-500 is the best-performing model for non-literals, nouns and for verb non-literals (which perform relatively well at 0.585). These correlations are not indicative of the interpretations' quality, but instead point to the circumstances in which MetID's scores align with the quality of its output – an aspect of the system that will be discussed in the next section.

Model	**All**	**Literal**	**Non-literal**	**Noun**	**Verb**	**Noun literal**	**Noun non-literal**	**Verb literal**	**Verb non-literal**
Overall	*0.117*	*0.020*	0.102	0.314	*-0.041*	0.259	*-0.044*	0.137	0.187
WN	**0.254**	**0.283**	0.274	0.398	0.085	0.475	*-0.003*	**0.456**	*0.064*
BEAGLE-128	0.223	0.161	0.209	0.230	**0.186**	0.103	0.231	0.188	0.475
LSA-500	*0.210*	-0.238	**0.327**	**0.424**	*-0.057*	0.563	*-0.068*	0.141	**0.585**
LSA-400	*0.102*	*0.077*	*0.055*	0.373	*0.096*	-0.480	**0.286**	-0.148	0.560
COALS-14k	*0.088*	0.05	-0.223	*0.152*	-0.239	*-0.023*	-0.214	-0.148	*-0.059*
HAL	*0.001*	*0.043*	-0.167	0.195	-0.17	**0.61**	-0.239	*0.12*	-0.396
BEAGLE-1024	*-0.117*	-0.238	0.241	**0.424**	-0.188	-0.563	-0.299	0.348	*0.080*

Table 5.12: Correlations between MetID's top-ranked candidate metaphor scores and those elicited by participants in the paraphrase rating task. The top row is the overall correlation without regard to the lexical model, the remaining are ordered by their overall performance. The best model in each situation (column) is shown in bold and the better of each variable (vertical delimiters) is underlined. Non-significant results are shown in italics ($p > 0.01$) whereas all others are significant ($p < 0.01$) and $N = 406$ in all cases.

5.5.4 Analysis & Discussion

This experiment evaluated the quality of interpretations for literal and figurative statements generated by MetID. In the absence of a gold-standard for metaphor interpretation tasks (see [198] and [197]), participant ratings were used to measure the accuracy of the system. Two sets of results were presented: the participant scores for MetID's interpretations and correlations between participant ratings and scores from the system. The correlations highlight a strength unique to MetID: that the candidate metaphors are generated with a degree of confidence. As we saw, some lexical models produced scores that correlate well to participants' ratings. Taken with the first experiment, these data point the system's ability to identify certain types of figurative language – namely noun-based metaphors. The results show that the noun / verb distinction plays a significant role in performance. Literalness, on the other hand, is important when processing noun-based statements but not verbs. In fact, verb-based statements are uniformly hard for MetID to interpret accurately. Similar to the first experiment, the choice of lexical model contributed significantly to the system's accuracy of interpretation as well as the correlations between model and participant scores. Overall, this evaluation supports MetID's role as a useful way to analyse noun-based metaphors, but in other circumstances it has considerable limitations.

The means analysis (figure 5.5) explored the contribution of three factors in MetID's performance: grammatical type, literalness and the choice of lexical model. The best performing situation is interpreting figurative, noun-based statements with the COALS-14k model. Here, MetID achieves a score near 2.0[12], which is not particularly strong. Intuitively, this means MetID does well about 38% of the time. However, in the group-wise means it was found that noun-based statements were significantly easier than verbs to accurately interpret. Literalness alone, on the other hand, had little effect on performance, but appears to mediate the contribution of the noun / verb distinction. This interaction is a new finding, though as will be discussed, is perhaps due to theoretical underpinnings of the method. A second finding is that the choice of lexical model contributed significantly to variance in the scores, especially when analysing noun statements.

The interaction between grammatical type and literalness can be described in two ways: that grammaticality mediates the contribution of literalness, or that literalness mediates the contribution of the grammaticality (see figure 5.6). The interaction shows that the method is better suited to interpreting figurative language as it occurs in noun-based constructions, rather than verb-based. This points to a fundamental aspect of MetID that is perhaps grounded in the foundations of *Metalude*. *Metalude* encodes most of its figurative relations with noun-based, nominal concepts, which means depending on the metaphor, MetID's cluster analysis must usually relate observed terms to the nominal concepts. The dominant theories of metaphor comprehension (reviewed in chapter 2) are formulated as noun-concept processing procedures – that a topic concept is understood "as" or "using" a vehicle concept. These theories require a lexical transition from action concepts to nominal analogues, and it has been proposed that they address an idealised noun-based conception of metaphor [121, 122, 206]. While the nominalisation of concepts and procedures for explanation and theory-building is not rare, with metaphor, communicative efficacy may also play a role in lexicalisation. Take the example (xxi):

(xxi) She devoured the material.

The metaphors implied by (xxi), IDEAS ARE MATERIAL and perhaps IDEAS ARE FOOD, can also be used to instantiate noun-based metaphor:

(xxii) The new material is food for thought.

Note how the copular construction in (xxii) used to equate *material* to *food*. In the noun-based example, the underlying metaphor is hard to avoid because it is made apparent in the surface structure of the statement. On the other hand, interpreting (xxi) requires more outside knowledge:

[12] Recall scores range from 0 to 7.

that people eat or *devour* and that we do not generally refer to what people eat as "material". These additional steps beckon the figurative interpretation of (abstract) *material* as something edible. While MetID attempts to make use of this kind of information with selectional preference violations, this technique is either too weak or too narrowly defined to interpret (xxi). Conversely, (xxii) makes the topic-vehicle mapping lexically clear, as nouns, reducing the system's need for stereotypical knowledge.

The interaction between grammatical type and the literal / figurative distinction does not account for all the variance in scores; the lexical model also makes a significant difference. Overall, WordNet performs best, but in many circumstances the DSMs produce comparable results. For example, the highest scoring configuration (though not significantly higher than WordNet) is COALS-14k on figurative, noun-based statements. In the same case, BEAGLE-1024 and HAL also produce comparable scores. WordNet was used on this task to provide a baseline alternative to the corpus-based semantic models. Because WordNet is developed by lexicographic and psycholinguistic research, it represents generally applicable lexical semantic relationships. Its structure and granularity (especially with nouns) means that word-relatedness can be measured with techniques that use explicit semantics and information content (the hyponym tree, synsets, glosses, etc.). In contrast, the DSMs rely on a semantic space representation in which word vectors realise their semantics as a statistical combination of co-occurrence patters. Observing that COALS-14k performs comparably to WordNet on noun-based non-literal interpretations supports COALS' ability to represent semantics similar to WordNet. However, the question of exactly *how* a COALS word vector equates to an explicit model like WordNet is essentially unanswerable [97, 136, 137, 217]. The semantics of DSM representations are abstract and only emerge with the use of vector similarity measurements. Entries in a semantic space are effectively points in a hyper-space where the dimensions themselves do not represent anything. DSMs are helpful in computational tasks, not only because they reduce reliance on external resources, but also because they provide *unified* representations. This allows the construction and re-use of a single data structure to use on a number of tasks, such as interpreting figurative statements in text.

The scores for DSMs' within-group means show some stratification in the non-literal cases (figure 5.5, left-middle). In literal statements, the only significant difference is that WordNet, outperforms HAL and BEAGLE-128. This is true overall and for literal verb-based statements. Alternatively, the noun-cases have a number of significant differences among the lexical models. Here, LSA performs relatively poorly compared to WordNet, COALS-14k and BEAGLE-1024. Given that COALS and BEAGLE were developed, in part, to address weaknesses in earlier, less linguistically informed models like LSA, it is perhaps not surprising that they perform better. The differences in the non-literal noun case, support the overall MetID approach as one that is tuned to figurative language, rather than literal. Further, the lack of significant variation in the verb-cases (both literal and non-literal) among the DSMs mimics WordNet's findings: none of the DSMs compensate for WordNet's inability to interpret verb-based statements. In short, MetID as a whole fails to interpret verb-based statements – a failure of the method overall, not individual lexical models. Potential strategies to accurately identify and interpret verb-based metaphors will be discussed in the final chapter.

It is important to remember that scores on this experiment are ratings from people, and as such, may include any number of outside factors. Perhaps the most confounding of these is the fact that if a sentence was itself rated less than perfectly sensible (7 on the sensibility scale), it will diminish the paraphrase's score, no matter how high. This means that sentence sensibility ratings never *raised* the score of the paraphrases, forcing the scores to err on the low side[13]. That being the case, the best models only performed at about 38% of optimum (2 in a range from 0 to 7). If this is a conservative estimate, the best configurations may in fact be generating interpretations at about 50% accuracy. However, 38% (or 50%) do not imply that MetID "got it right" 38 times out of 100, instead, it means that on average, the paraphrase were rated near the mid-point between "bad" and "excellent".

The way sensibility scores were used to scale down the quality ratings is perhaps too conservative. The rationale was that trusting a falsely positive judgement would yield falsely positive results. However, the opposite is equally true: the analysis should also avoid trusting negative results. This experiment is effectively slanted toward a negative result. A less stringent design would be to use the sensibility scores to discard sentences below a threshold and consider all quality ratings. Given the average sensibility scores, we can guess that, on average, the scores for literal sentences would go up slightly in this regard, while the scores for non-literal statements (which had significantly lower sensibility ratings) would be moderately higher than reported. Future work on MetID could adjust evaluation to compensate for the conservative nature of the results reported here.

The correlation analysis showed a slightly different picture than the ANOVA (table 5.12). Similar to the means, the noun statements score higher overall, as do non-literal statements, showing that MetID is better in these cases. However, the interaction between grammatical type and literalness does not appear in the overall correlations. In fact, it does not manifest in any of the individual models. Instead, note the success for verb-based figurative language in BEAGLE-128 and the LSA variants. Keep in the mind that good correlations here do not mean MetID generated good interpretations. Instead, it shows that the system "knew" it was doing as poorly as people judged it to. For verbs, this confirms that the model scores were generally lower – inline with participant ratings (hence the stronger correlation). For literal noun-statements, HAL does relatively well at 0.61. For non-literal statements, the best two models, LSA-400 and BEAGLE-128, achieve 0.286 and 0.231 respectively. While these findings underscore how MetID enables testing and refinement, without looking at results for individual statements, the system's success remains unclear. To explore how, when and why MetID succeeded and failed, the next section examines individual examples from this experiment.

[13] For example, if a sentence was rated at 4 out of 7 for sensibility, and its paraphrase was a 7 out of 7, the resulting score will be 4.0.

Types of Success & Failure

This section reviews some specific ways the system commonly succeeds and fails; table 5.13 shows five such cases. In addition to these situations, an outstanding source of noise is peoples' intuitions. In the absence of a gold-standard for a metaphor interpretation task, this experiment used ratings elicited from people. This strategy obscures the fact that the metaphors are often idealised abstractions, and sometimes not independently interpretable. One example of this is that when interpreting "The runners streamed through the street." with the WordNet model, MetID produced the metaphor CROWD AS LIQUID. In terms of root analogies it is hard to think of a more accurate metaphor. However, its average quality was rated at only 2.2. This low score is likely because the sentence is not obviously figurative, obfuscating the relationship between the verb *streaming* and the concept LIQUID. Such examples are perhaps discouraging for this evaluation, but inevitable.

	Example	
Type	Sentence	Candidate Metaphor
1. Unpaired Root Analogy	*The truck soared down the slope.*	ROAD AS RISE
2. Misidentified Term(s)	*He piloted his dance partner.*	RELATIONSHIP AS MUSIC
3. Lexical Semantic Failure	*A vacation is medicine.*	ELEMENTARY AS DISEASE
4. Category or Feature Mapped	*The mind is a computer.*	THINKING AS CALCULATING
5. Selectional Violation	*He piloted his dance partner.*	CONTROLLING AS LEADING

Table 5.13: Six examples of common mistaken and successful interpretations.

1. Unpaired Root Analogies. After MetID identifies two terms in a statement as a possible topic-vehicle pair, it tries to find the best pair in *Metalude* by minimising within-cluster distances between topics and vehicles. It is possible, especially when using lower dimension semantic space models, that no topic or vehicle is found in any of the clusters. In this case, MetID will "decouple" *Metalude*'s pairs and use them as a bag words (applying a large penalty). This allows *new* metaphors not given by *Metalude* to be identified, which is conceivably a good idea. However, it often results in uninterpretable candidate metaphors such as ROAD AS RISE, ENGINE AS WHITE or PASTA AS STAGNATION[14]. Because MetID applies a large penalty in this situation, this type of failure seldom has an adverse effect on identification tasks, but it often means that the best interpretation is wildly inaccurate.

2. Misidentified Terms. By looking for candidate topic and vehicle terms separately, MetID operationalises a nominal view of metaphor: that metaphors consist of two objects. As we will discuss in the next chapter, this strategy influences the system's ability to interpret verb-based statements. It can also fail in other ways. Because the system ranks candidate metaphors that may have different observed topic and vehicle terms, the "correct" interpretation may have been found, but scored lower than others. Take (xxiii):

[14] Perhaps this could be a metaphor for a chef's culinary skills, but surely not for the statement for which it was generated: *That creature in the net is a crab*.

(xxiii) The runners streamed through the streets.

For (xxiii), MetID (with LSA-500) provided the interpretation A BODY IS A LANDSCAPE. In that case the system identified *runners* as the topic and *streets* as the vehicle, which minimised the respective distances to BODY and LANDSCAPE. Cases like this are precisely the motivation for not only analysing a sentence as a whole, but also certain sub-units like dependencies and predications. Analysing these constituent units allows for multiple metaphors per sentence and enables well-suited topics and vehicles to score higher in a figurative relation than they might outside a relation. This separation is only as good as the difference between top-scoring candidates for the sentence and sub-units. In (xxiii), *runners*→BODY and *streets*→LANDSCAPE are mutually closer than any other pairings, however, the fourth-best candidate was indeed CROWD AS LIQUID. This situation raises the question of how many better interpretations exists *near* the top of MetID's rankings – a question that will be discussed in the next chapter.

3. Lexical Semantic Failures. The focus of this experiment and the previous one, was to test MetID on identification and interpretation respectively, as well as narrow the list of good lexical models. By testing the models on the same materials, with the same configuration for the heuristics, these evaluations found significant differences in performance from one model to another. These differences are defined by the models providing good or bad associations between observed terms and the seeds from *Metalude*. However, this association can simply be wrong, making it unlikely for MetID to generate an accurate interpretation. Take (xxiv) for example:

(xxiv) The boy grabbed his bike and flew home.

Using LSA-500, MetID found the best candidate to be ORGANISATION AS SHIP. This metaphor might make sense when a government "steers the course" or a CEO "weathers the storm", but these are not the case for (xxiv). What happened here is that LSA wrongfully associated *bike* with ORGANISATION, and *flew* with SHIP (which is more reasonable). Nonetheless, MetID relies centrally on the lexical model to provide associations that account for the figurative use of observed terms. In this example, the failure is the misassociation of *bike* with ORGANISATION.

4. Category Matching & Feature Mapping. One situation in which MetID succeeds as it was designed to, is when the candidate terms match their observed counterparts via a categorical or featural association. In the first case, matching an observed term to its super-ordinate category as a *Metalude* term, is analogous to category matching (cf. Glucksberg) which has been simulated in WordNet [122] as well as DSMs [121, 184, 215, 221]. In the other case, feature mapping (cf. Gentner) occurs when the lexical model associates terms by salient features of the observed term to those of the nucleus. This has also been simulated computationally and using DSMs [76, 222]. In this experiment, the interpretations generated for (xxv) and (xxvi) exemplify MetID's use of categorical and featural information respectively.

(xxv)	*The mind is a computer.*	$\Longrightarrow$	ORGAN AS MACHINE
(xxvi)	*An education is a doorway.*	$\Longrightarrow$	OPPORTUNITY AS PATH

The underlying success in these situations rests on the lexical model making associations along categorical or featural relations. For the corpus-based models (ie. not WordNet), this is done using word-vectors – representations without explicit relations. This abstract relation between a semantic space representation and actual categories or features, makes it impossible to elaborate on how, or precisely *what* properties were used / represented leading to the association [137, 147, 149]. The strengths and weaknesses of this strategies will be explored further in the concluding chapter, but it is worth noting that WordNet, an explicitly coded model, performs comparably to the best-performing distributional models.

5. Selectional Violation. Another situation where MetID performs well is when it uses selectional preferences to find violations in object / subject-verb constructions. One example is the sentence "He piloted his dance partner" for which WordNet produced the interpretation CONTROLLING AS LEADING. In this case, MetID's selectional strength heuristic applied a large bonus, having observed that `dobj(piloted,partner)` constituted a high degree of selectional preference violation. That is, "partners" are seldom observed to be "piloted". Because the heuristics are applied individually to each candidate interpretation, such cases still rely on the lexical model to associate the topic and vehicle terms. Thus, while the heuristics promote interpretations that exhibit certain properties, such as selectional preference violation, the interpretations themselves are still the product of the lexical model.

5.5.5 Summary

The goal of this experiment was to evaluate MetID's ability to generate interpretations to non-literal statements. Overall, it was found that MetID performed better on non-literal statements than literal – a fundamental goal. It was also found that noun-based statements were considerably easier for the system to interpret than verb-based statements. Moreover, MetID can use interchangeable lexical models, the choice of which significantly affects performance on noun-based statements. While the best configurations do significantly better than chance in terms of human ratings, this is only the case for noun statements. This finding is supported by other research about how figurative statements are identified and processed [25, 82, 121, 212] and it supports the lexical models' ability to adequately represent noun-based concepts. An interaction was also found, in which the noun / verb distinction significantly mediated the effect of the literal / figurative distinction.

These evaluations highlight a strength of MetID's design with regard to providing scored rank-orders of candidate metaphors, as opposed to a single result. This design enabled a correlation analysis between the system's scores and people's ratings, which explored significant differences among models' ability to "know" how well they were doing. Interestingly, WordNet is among the top-performing models for both evaluations (ratings and correlations). It remains to be seen if MetID can be used to aid the simultaneous identification and interpretation of figurative language. To explore this idea, the next experiment uses MetID with WordNet to examine the term *contagion* as it came to be used figuratively in finance and economics.

5.6 Experiment 3: Using MetID in Terminological Research

The third experiment is a case-study involving a corpus-based analysis of the term *contagion* as it is used figuratively in finance and politics. The first two experiments addressed metaphor identification and interpretation, while this evaluation tests MetID's usefulness in broader research setting. The results presented here are part of a more comprehensive analysis in [71] and are presented with permission of my co-authors. This study uses MetID to find potential metaphors instantiated by the term *contagion* in a corpus of US congressional documents. The full study was presented at the European Symposium on Language for Specific Purposes, in July of 2013 [71].

5.6.1 Background

The term *contagion*, which was initially used in a religious context to describe something as morally defiling[15], is used in biology and medicine to refer to diseases that are passed among organisms. Recently, it has come into use in finance and politics where it refers to adverse financial phenomena that propagate between institutions (see Figure 5.7). Although the semantic features of *contagion* that account for movement and contraction make it apropos to describing spreading financial problems, many semantic features from its literal use in biology are not found in the new domain. Additionally, other feature are highlighted and exaggerated in finance that are found relatively infrequently medicine and biology. Indeed, a coherent definition of *contagion* in finance remains illusive [71].

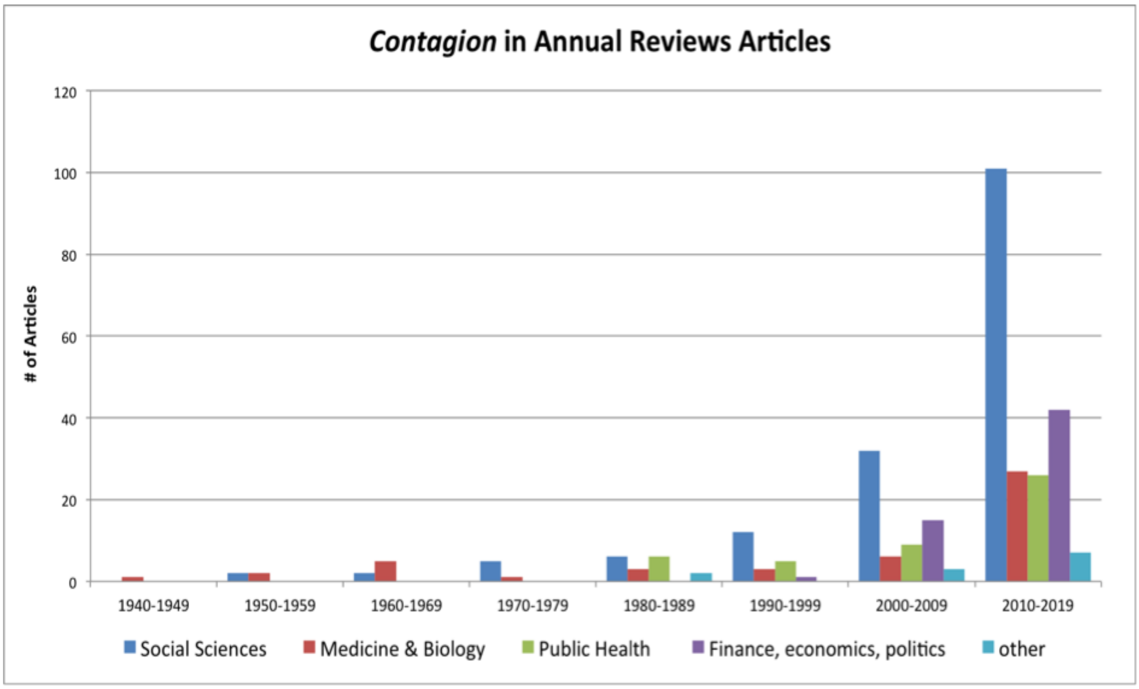


Figure 5.7: Types of *Annual Reviews* articles in which *contagion* occurred. Derived from www.annualreviews.org; adapted here from [71].

[15] Oxford English Dictionary: www.oed.com.

Contagion is used to talk about institutional behaviours that spread between institutions. The mechanism of this movement, the conditions for contraction and even the symptoms themselves, are a complex apparatus of financial policies, circumstances and behaviours [49, 50, 131]. Thus, as a term, *contagion* fills a gap in economic vocabulary, allowing news outlets, financial executives, researchers and politicians to simplify complex (and necessarily negative) financial workings. In [71], we undertake an historical, lexical, semantic and conceptual analysis of the term. MetID was used in the conceptual analysis, to find root analogies that undergird the figurative term's behaviour at this level. By using MetID in this capacity, the research serves a kind of case-study from which we can glean some of MetID's strengths and weaknesses in an applied setting.

5.6.2 Data, Methods & Results

The data used in [71] for the conceptual analysis of *contagion* consisted of a set of documents from the US congress: hearings, testimonies, reports and press releases. Hearings and testimonies usually constitute legal commitments on the part of the authors and are made up of deliberate, formal language. Reports and press releases are typically commissioned research and public relations announcements, respectively. The documents analysed ranged from 2001 to 2012 and were downloaded from www.senate.gov and www.congress.gov. The collection contained a total of 267,256 tokens in which *contagion* was found 96 times in 87 different sentences. The term was found only in the singular noun form; *contagious*, *contagiousness* and *contagions* were not present. Moreover, every use of the term referred to finance and economics, never biology or medicine.

Though the metaphor of contagion in finance – that spreading economic problems are like diseases – the word *contagion* may not always be a vehicle, despite the underlying metaphor using it as such. For example, in the phrase "to defended against contagion", *contagion* is actually the topic of the metaphor DISEASE AS WAR. To take such cases into account, we examined both metaphors where the term was found as a topic term and those where it was a vehicle. Although MetID performed only moderately above chance in the first two experiments, in this study the output was analysed manually. Specifically, we reviewed candidate interpretations that were not just the top-scoring metaphors. MetID was run in rank mode, using WordNet as the semantic model, to provide the 20 best-scoring candidate interpretations for every sentence containing *contagion*. After this, all interpretations that did not use *contagion* as a topic or vehicle term were discarded[16]. Table 5.14 shows some example interpretations generated by MetID.

Though the concept of contagion in finance instantiates a metaphor of PROBLEM AS DISEASE, the term itself has more diverse uses. The sample results in Table 5.14 exemplify some of this diversity. These samples also highlight an important weakness of MetID. Take the first two sentences where MetID provided the interpretation DISEASE AS INVASION, which is perhaps plausible looking at the identified topic and vehicle terms. However, in the first sentence, MetID selected *contagion* as the topic and *entering* as the vehicle. Without examining the sentence, this is certainly a reasonable pair of words for which to provide an interpretation, but in the sentence

[16] Though these metaphors might be interesting to a broader analysis of financial and political language, the goal of this experiment was to analyse the term *contagion* specifically.

Sentence	Candidate Metaphor	Topic Term	Vehicle Term	Score
[...] entering a critical phase as policy initiatives undertaken so far have not prevented systemic contagion.	DISEASE AS INVASION	contagion	entering	0.88
[...] contagion may spread further in the very short term.	DISEASE AS INVASION	contagion	spread	0.74
[...] a material impact in addressing market contagion.	DISEASE AS WAR	contagion	impact	0.71
The contagion is driven primarily by what other securities are owned [...]	DISEASE AS WAR	contagion	need	0.60
[...] has come a new strain of global contagion [...]	DISEASE AS IDEA	contagion	strain	0.83
[...] as part of its operations can extend the contagion risk [...]	DISEASE AS IDEA	contagion	part	0.79
Banks have solvency regulation to protect depositors and to defend the banking system from contagion risk.	DISEASE AS IDEA	contagion	regulation	0.71
Anticipating future sources of contagion is difficult [...]	DISEASE AS IDEA	contagion	source	0.70
[...] a real contagion risk to the financial system [...]	DISEASE AS IDEA	contagion	system	0.70
General investor panic is the final reason for contagion.	DISEASE AS EMOTION	contagion	panic	0.78
The contagion is driven primarily by what other securities are owned [...]	DISEASE AS EMOTION	contagion	security	0.69
Financial contagion to the US from further deterioration [...]	DISEASE AS EMOTION	contagion	deterioration	0.59
Contagion from the Greek debt crisis and [...], which too have solvency problems.	PREVENTION AS OBSTACLE	contagion	problem	0.82
[...] Fueled Contagion Ultimately: private-label mortgage securitization turned out to be an edifice [...]	PREVENTION AS OBSTACLE	contagion	edifice	0.65
[...] as part of its operations can extend the contagion risk [...]	FAILURE AS DIVISION	contagion	part	0.84

Table 5.14: A sample set of sentences from the congressional texts analysed with MetID; adapted from [71].

itself, they are completely unrelated. This is not the case in the second sentence where MetID identified *contagion* and *spread*, which are directly (grammatically) related. This problem appears considerably more often in natural language, like the congressional documents analysed here, than in the materials for the first two experiments. Also, consider the five sentences which produced the metaphor DISEASE AS IDEA in table 5.14 for which there is range of vehicle terms. Here, it is not clear why WordNet related the vehicle concept IDEA to the terms *strain*, *part* and *source*.

The candidate metaphors MetID generated were grouped into their root analogies to provide a high-level analysis of *contagion*'s conceptual behaviour (see Table 5.15). The root analogies are categorised into the topic and vehicle concepts on the map of root analogies [92]. For example, the metaphor DISEASE AS INVASION occurs in the sector relating the topic concept *Human, Senses, & Society* to the vehicle concept *Space & Place*. In the corpus, metaphors instantiated with the term *contagion* were most commonly found relating *Human, Senses, & Society* to *Human / Animal, Body & Senses*, which serve personify institutions, equating them to human senses. This use of *contagion* likely owes to the term's biological and medical origins. Other common kinds of metaphors found are those that relate *Living Things & Substances* to *Human / Animal, Body & Senses* as well as those relating *Values, Qualities & Quantities* to *Activity & Movement* and *Space & Place*. These are metaphors where changes in quantities and qualities are thought of as movement and other material changes such as "boiling" or "solidifying". These metaphors help imbue the concept of contagion with its abilities to move, spread and grow, as seen in the example "a disturbing level of contagion has already been evident around the hemisphere." Here, contagion is thought of as a quantity, using metaphors like CHANGE IN QUANTITY AS CHANGE IN ELEVATION.

	Topic					
Vehicle	Activity & Movement	Human, Senses, & Society	(Living) Things & Substances	Value, Qualities, & Quantities	Emotions, Experiences &, Relationships	Thinking & Communications
Things & Substances	15 (10, 5)	39 (27, 12)	7 (5, 2)	0 (0, 0)	3 (2, 1)	9 (0, 9)
Human / Animal Body, & Senses	13 (6, 7)	208 (93, 115)	155 (58, 97)	69 (8, 61)	12 (6, 6)	5 (2, 3)
Activity & Movement	41 (22, 21)	11 (2, 9)	7 (5, 2)	99 (41, 58)	0 (0, 0)	9 (0, 9)
Space & Place	16 (8, 8)	23 (21, 2)	46 (31, 15)	98 (67, 31)	45 (39, 6)	0 (0, 0)

Table 5.15: Root analogies found in the congressional corpus; adapted from [71]. The total is given and in parentheses are the number of times *contagion* was selected as a topic term and as a vehicle term, respectively.

Looking at the use of *contagion* specifically as a topic term in the metaphors provided by MetID, we find there is more uniformity in the range of topic concepts of the root analogies (see Table 5.15; left numbers in parentheses). Note that when *contagion* is found as a topic, it is more apparent in root analogies relating to *Values, Qualities, & Quantities* and *Emotions, Experience, & Relationships* than anywhere else. This implies that *contagion* is not restricted to figurative use as a disease, but that it can be measured in a technical (ie. financial) sense. These features are new to finance, as they were not found in the biological domain [71]. On the other hand, instances where *contagion* is observed as a vehicle term are where it is used to make sense of another concept. In these cases, common root analogies range from rather general vehicle concepts like SPACE, TIME and MOVEMENT to more specific ones, like WAR, BUSINESS and OBSTACLE (see Table 5.15; right numbers in parentheses). Metaphors relating to *Human / Animal, Body, & Senses* liken institutions (banks, markets, countries, etc.) to living beings, presumably to enable them to have problems like diseases – that is, to contract and spread contagion. *Contagion* instantiates a range of metaphors that relate the term's literal meaning in biology to the complex mechanics of international economics. Metaphors about movement, space and place are particularly apt because they provide financial problems the ability to move – a defining feature of disease.

5.6.3 Discussion

The motivation for this experiment was to assess how MetID can aid terminological research. As mentioned above, the results reviewed here are part of a larger analysis carried out in [71], that examined the semantic, grammatical and conceptual behaviour of *contagion* in financial discourse. In this study, MetID proved to be useful and helped to augment an otherwise manual corpus-based analysis. Unlike the first two experiments, which used only the top-scoring interpretation produced by MetID, in this evaluation, a wider range of output was analysed (the 20 top-scoring candidates for each sentence). While this placed more importance on the researchers' intuition, similar to traditional corpus-studies [34, 35, 219, 234], it resulted in a more qualitative analysis of language.

This experiment also highlights a crucial weaknesses of MetID that did not arise in previous experiments: the system will sometimes select distant topic and vehicle terms that are unrelated. Because the first item MetID analyses for a given input is the sentence as a bag of words (all possible word-pairs), it can produce a number of candidate interpretations that fit well with a pair of terms, despite them having no relationship. There are three ways MetID could be changed to help avoid providing spurious interpretations resulting from poorly chosen term-pairs. First, a heuristic could be added to penalise candidate interpretations that were generated for distant or grammatically unrelated terms. Alternatively, the system could apply a bonus to selected terms if they occur in a relationship. Thirdly, and perhaps the most robust solution, would be to use semantic parsing techniques to retrieve higher-level relationships [42, 57]. These relationships would add to the existing grammatical relations from the dependency parses – perhaps rendering the bag of words analysis superfluous altogether. The first two alterations would not be difficult to add to MetID with its current architecture, but the third would require augmenting the structural processing component to include semantic parsers.

Looking generally at the onset of the use of *contagion* in financial texts provided what may be a typical case of metaphorical term-borrowing from one domain to another. The idea of contagion, long since borrowed from ethics, is used extensively in biology and medicine. In the 20th century, it was increasingly used in other domains and was recently adopted in finance and economics. Perhaps this adoption was enabled by the increasing interconnectedness of global finance where the term is an apt description of new and complex problems. Adapting the concept came with some constraints, intentional or not. Though there appear to be different types of financial contagion, little attention is given to what a contagion *is*, the circumstances in which it can emerge, or what contracts it. These aspects are addressed readily in biology and medicine, where the focus is usually on something else – that is, contagion is typically a property of other objects. This is evident in how the term is commonly used as an adjective in biology and medicine, whereas it is only used as a noun in finance and politics.

5.7 Summary of Results

The results of the word clustering task, which is central to the method, showed that across most lexical models, the method produced viable clusters. The identification task (the first experiment) used the reference implementations for the DSMs, WordNet with Lin similarity and the COLE-AMI model. Metaphor identification was tested by analysing pairs of literal / figurative statements and observing how often MetID ranked the figurative higher than the literal. Overall, noun-based statements were easier than verb-based statements and the lexical model made a significant difference but the choice of corpus did not. Without taking coverage into account, the best configurations (LSA-400 and LSA-500) performed at about 75% accuracy (50% baseline), but WordNet and COALS-SVD-200 were the top when considering coverage. The best-performing configurations on the identification task were used to evaluate MetID's ability to generate accurate interpretations to figurative statements (second experiment). In this task, participants rated candidate interpretations of figurative and literal statements. In the best case (figurative, noun-based statements) the top models were COALS-14k and WordNet, at about 38% accuracy (0% baseline). However, correlating MetID's scores with people's ratings, produced different results (table 5.12). The interpretation experiment also found an interaction where grammaticality mediated the contribution of metaphoricity on the system's performance, which is further discussed in the next chapter. The final experiment was a case study from previously published work [71] where MetID was used to aid a lexicographic analysis of *contagion* in biomedical and financial texts. The system was used to extract a number of potential metaphors instantiated by the term. The case study highlighted the role of computational techniques, and specifically MetID, in corpus analysis and terminological research.

5.7.1 Role of Grammatical Structure in MetID

One thing not tested in the first two experiments, and glossed over in the third, is MetID's ability to process long sentences. Looking for topic and vehicle terms in a statement means that unrelated words from a particularly long sentence may be found for a candidate metaphor. Take (xxvii), from the congressional corpus discussed in the third experiment:

> (xxvii) The European crisis is entering a critical phase as policy initiatives undertaken so far have not prevented systemic contagion.

MetID will analyse 8 items for (xxvii):

1. Full Sentence (all unique ordered word pairs)
2. `amod(european,crisis)`
3. predication: *crisis-phase*
4. `nsubj(crisis,entering)`
5. `dobj(phase,entering)`
6. `nn(policy,initiatives)`
7. `nsubj(initiatives,prevented)`
8. `dobj(contagion,prevented)`
9. `amod(contagion,systemic)`

In (xxvii), the top-scoring candidate involving *contagion* is extracted from the first item (the bag-of-words analysis). MetID isolates *contagion* and *entering* as instantiating the metaphor DISEASE = INVASION (*score* $\approx$ 0.88), which is a plausible interpretation: that disease invades living beings similarly to how economic contagion enters a geo-political region. However, examining the dependency structure of (xxvii) (figure 5.7.1) we see that *entering* does not refer to *contagion*, but instead to a *phase*^{DOBJ} of *the European crisis*^{NSUBJ}. This example typifies a situation that arises, particularly when analysing longer sentences, where MetID produces interpretable output, but is not informed enough by grammatical structure to point to a linguistic metaphor. That is, though it may underly the spirit of (xxvii), the metaphor DISEASE = INVASION is not directly instantiated.

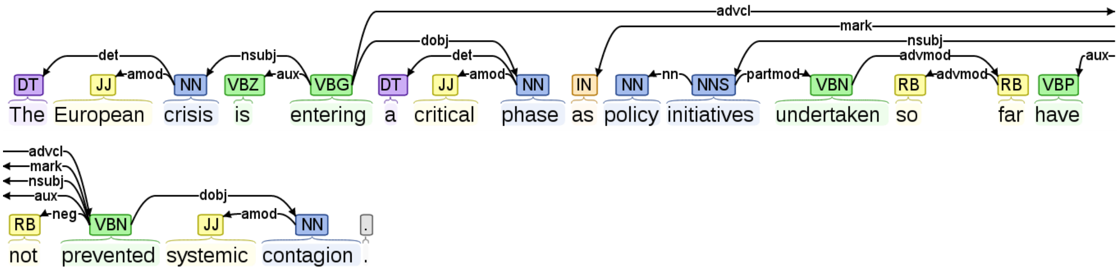


Figure 5.8: Dependency structure of (xxvii), from which a number of relations are extracted for analysis by MetID. Parsed using the Stanford Parser and visualised with CoreNLP.

The third experiment, unlike the first two, explored real-world text: a corpus of US congressional documents. Note that in (xxvii), seven out of nine items of analysis are dependencies, however, given the targeted analysis of *contagion*, only items 1, 8 and 9 were analysed. Indeed, neither 8 nor 9 are figurative – the top candidate between both scoring 0.55. The bag-of-words analysis (item 1) allows any pair of words to instantiate a candidate metaphor, which provides the largest range of analysis.

As exemplified in the third experiment, MetID provides a systematic way of analysing *potential* metaphors in language. Specifically, the system was used to extract commonly co-occurring topic and vehicle concepts. Particularly in a targeted analysis, like the *contagion* case study, MetID can provide a synoptic analysis of commonly co-observed topic and vehicle domains. In the *contagion* analysis, we analysed sentences, but that input could be pared down to individual clauses or phrases. This could be helpful, because when analysing long sentences, further inspection is often needed to find metaphors. This process, however, highlights the strength of MetID as a tool to aid terminological analysis, but it also points to a limitation: grammatical structure does not always inform the extraction of paired topic and vehicle concepts. There are some ways MetID could be changed to better accommodate long sentences.

One option to reduce spurious results on long sentences is to penalise distant pairings. In (xxvii) the surface-level, linear distance between *entering* and *contagion* (used to generate the top-scoring candidate metaphor) is 14. Penalising long-distance pairs could be done using linear distance in the surface structure (which is known to be proportionally minimal to parse complexity [32]) or by using distance in the dependency structure. This would discourage distant and / or unrelated words from constituting a term-pair for which to generate candidate interpretations. An alternative to the distance penalty would be to increase the scope of grammatical analysis. Currently, the system only looks for `nsubj`, `dobj`, `nn` and `amod` relations to analyse, because subject-/object-verb and modification relations were originally proposed to evince selectional preference. Conceivably, any dependency could be used in the spirit of selectional association or colligation analysis. A third option to mitigate the distant-term problem is to perform some kind of semantic parsing [5, 42, 79]. Similar to semantic-level selectional preference induction [41, 52, 183], semantic parsing could annotate a statement with information such as lexical semantic operators (negation, combination, juxtaposition), frame-based information (roles, agentising, abstraction) or even pragmatic information (intent, polarity, affect). Such semantic information could be used similarly to the dependency structures in promoting semantically related pairs and penalising unrelated pairs. In the absence of an implemented solution to the long-sentence problem, MetID was used as a means to detect topic and vehicle domains that commonly co-occur. When the sentence is short, like those analysed in the first two experiments, this often results in extracting an instantiated metaphor, but in long sentences, the extraction is often implicit or even accidental. All the same, this use of MetID is unique in lexicographic and terminological research and can provide insight into figurative concepts apparent in naturally occurring text [71, 73].

The concluding chapter will further discuss the results presented here and show how they relate to the technical and theoretical motivations presented in chapters 2 and 3. It will also discuss how MetID fits into a larger theme of research in statistical semantics, NLP and figurative language. The system's limitations will be more thoroughly explored in light of the research goals and MetID's design.

Chapter 6

Discussion

This project explored the use of linguistic resources and computational techniques in an effort to address a complex NLP task: automatically identify and interpret certain kinds of figurative language. The goal was to extend existing NLP methods by combining statistical semantic models and linguistic resources to address noun- and verb-based metaphors in a unified way. The goal was to help examine the efficacy of statistical semantic models, existing NLP tools and linguistic resources to help identify metaphor in naturally occurring text. By comparing three types of lexical models (WordNet, semantic spaces and co-occurrence likelihood estimation) it was found that the choice of model is responsible for significant portion of the system's performance. Below is a discussion of the findings presented in the previous chapter, their implications and a review of ideas for future work.

In metaphor, there are two key theories of comprehension: category matching and feature mapping (cf. Glucksberg and Gentner, respectively). These theories provide a backdrop against which to analyse if and how computational models can represent the necessary lexical and conceptual information required to process metaphor. Metaphor theory provides a guide to the structural and semantic requirements of the system and help in its evaluation. By exploring the performance of MetID, both technical and fundamental weaknesses became apparent. One weakness, for example, is the need to represent both categorical *and* featural properties to process a broader range of linguistic metaphor. Linguistic research has produced resources that provide a starting point for metaphor processing [10, 92, 133]. While the information in *Metalude* is central to MetID, the only functional requirement is a list of mapped topic-vehicle terms. *Metalude* fits this requirement, but there are other options, such as Lakoff's *Master Metaphor List*. There are also two new projects on automated metaphor analysis – the *METAL*[1] [231] and *MetaNet*[2] – that may yield more resources for systems like MetID. These resources could code more detailed information such as relationships between metaphors, translations from verb-based expressions or annotations for different types of figurative language like metonymy, ellipsis and synecdoche.

[1] http://www.theatlantic.com/technology/archive/2011/05/why-are-spy-researchers-building-a-metaphor-program/239402/; 6 August, 2013.

[2] https://metanet.icsi.berkeley.edu/metanet/; 6 August, 2103.

The choice of semantic model is the other central component and was the primary focus in the system's evaluation. The design of MetID allows any model that provides relatedness scores between words to be used. WordNet takes a lexicographic approach where experts are tasked with coding entries and relationships [167]. Alternatively, corpus-based semantic models are perhaps better suited to information extraction tasks as they reduce reliance on outside resources. Such models (LSA, for example) require a corpus to construct a semantic space in which word relatedness is measured as proximity. The third type of model explored in MetID was the use of co-occurrence likelihood estimation as an analogue to word relatedness. The results suggest that some semantic space models perform comparably well to WordNet, but that WordNet usually provides better coverage (see sections 5.4 and 5.5). This highlights the strength of resources like WordNet while underscoring some of the advances in corpus-based semantics. The remainder of this chapter reviews the implications of the findings in chapter 5, including the architecture of MetID, the semantic models and the heuristics. Linguistic implications will also be discussed, particularly how the findings relate to metaphor theory, meaning in language and the effects of grammatical structure on metaphor processing.

6.1 Computational Implications

MetID combined a word-clustering strategy [199, 201] with three additional features: a series of heuristics, intrinsic cluster-quality measurements and an interchangeable lexical semantic model. The architecture places the search for root analogies at the center of metaphor identification and interpretation, combining them into a single task. By ranking potential interpretations (root analogies), identification is addressed by calculating scores over all possible interpretations. This places the database of root analogies (*Metalude*) at the core of metaphor processing, though, it can be replaced by any similarly structured resource. The heuristics are auxiliary to the main algorithm and operate on the score of individual candidate interpretations. These include violations of selectional preference, predications, cluster quality and lexical cues. The synthesis of these methods is unique to MetID and, the implications are discussed here as they relate to computational aspects of the research.

MetID uses the cluster-quality heuristic to compensate for intrinsic abstractness and vagueness in natural language represented in the corpus-based semantic models. No matter how technically sound these models may be, word meaning is produced by patterns of use in the text. Building clusters allows MetID to temper word associations with a cluster's purity and entropy (Eqs. 4.8 and 4.9). Using these measures attempts to address a fundamental aspect of meaning in language: words often represent ambiguous concepts. For example, if the word *true* is associated in a cluster PASSION, the association should be weaker if PASSION is disparately defined (an entropic cluster) and should be stronger if it is uniformly defined (a pure cluster). This means the clusters serve two purposes: to *relate* words to the seeds and to *qualify* the seeds themselves.

Perhaps the biggest finding from the development and testing of MetID is that the selectional preference violation heuristic is not enough to account for verb-based metaphors. By using a database of paired topic and vehicle terms, MetID operationalises a mostly *nominal* conception of metaphor – apparent in Goatly's theory of root analogies as well as conceptual metaphor theory. The selectional violation heuristic is designed to raise the score of observed word-pairs found in relationships that violate normative argument classes (cf. Wilks). That is, when a noun selects a verb from an abnormal class, MetID applies a bonus to the score proportional to how abnormal the class is, given pre-calculated observations from the BNC and enTenTen collections. However, candidate interpretations (the root analogies) are themselves not informed by the selectional violations. Instead, the word-pairs are matched like any other metaphor, whether or not they violates selectional preferences, using the paired-cluster search algorithm. This does not mean that a verb-based metaphor cannot, in theory, be correctly identified. The lexical models can associate observed verbs with topics or vehicles as they would any other word, regardless of POS-class. Associating verbs is more complicated because they usually have to cross POS-class to the predominantly nominal set of seeds in *Metalude*.

The heuristics are designed to augment the core algorithm. The results suggest some of these heuristics are actually more important than was assumed in designing MetID. Perhaps the most important is selectional preference violation, which has been reported to accurately account for a range of novel, verb-based metaphors as Ekaterina Shutova, et al. suggest [201]. In this context, selectional preference induction is done at the semantic level, making judgments on the semantic type of word selected by a root-word, as opposed to the lexical-level implementation in MetID, where words simply select words. However, it is not clear to me how selectional preferences can be used to provide candidate interpretations the way the cluster-search algorithm does [230]. That is, selectional violations may *constitute* linguistic metaphor, but it remains to be seen how they could provide interpretations. On the other hand, the lexical heuristics presented by Goatly may not be as accurate as originally proposed [92]. Relying exclusively on these lexical cues was tested by Shutova, et al. in [200] where the authors found the cues often do not signal a metaphor (table 6.1). Unlike selectional violation, which should be more central to the system, the lexical cues are good heuristics to apply conditionally to augment the score.

In evaluating MetID it was apparent that some heuristics were less relevant than predicted. For example, the bonus applied when synonymy is detected between an identified topic or vehicle and its candidate metaphor's counterpart almost never occurs unless using WordNet as the lexical model. This is because in the distributional models synonymy is defined as vector similarity being 1.0. For this to occur, not only must words share frequency distribution vectors, but they have to share the same relative frequency. On the other hand, when MetID uses WordNet, it chooses the senses of a pair of words that maximises their Lin similarity, which means all synonymous senses will yield a score of 1.0, thus applying the synonym bonus. This heuristic could be changed to allow any score above a threshold to be considered a synonym. Alternatively, the heuristic could be applied more fluidly as a scaled bonus, relative to how ”much” synonymy is detected between a pair of words (ie. how close their similarity score is 1.0).

Cue	Sample	Metaphors	Precision
metaphorically speaking	7	5	0.71
so to speak	49	35	0.71
utterly	50	16	0.32
literally	50	13	0.26
completely	50	13	0.26
figurative	50	9	0.18

Table 6.1: Precision of using Goatly's cues to identify metaphors in the BNC. Adapted from [200].

Though MetID seeks to combine metaphor identification and interpretation into one task, a subtle distinctions persists. The interpretation of a statement is generated by the cluster search algorithm and the identification task is effectively a judgment based on the resulting scores. That is, the cluster search *finds* potential metaphors, attributing them a score based on the strength of the within-cluster associations. The score is then augmented with a series of conditional bonuses and penalties applied by each heuristic. Separating the identification task from interpretation could be helpful from a functional point of view – especially if the identification task generated typed output that designated the interpretation method (see figure 6.2). This architecture would allow different interpretation mechanisms determined by the type of metaphor, allowing a selectional preference-based analysis to take precedent for verb-based metaphors, or a sense-tagging approach to process single-word metaphors. While such a system might simplify (or at least modularise) metaphor processing, it would be a departure from metaphor theory [75, 80, 133].

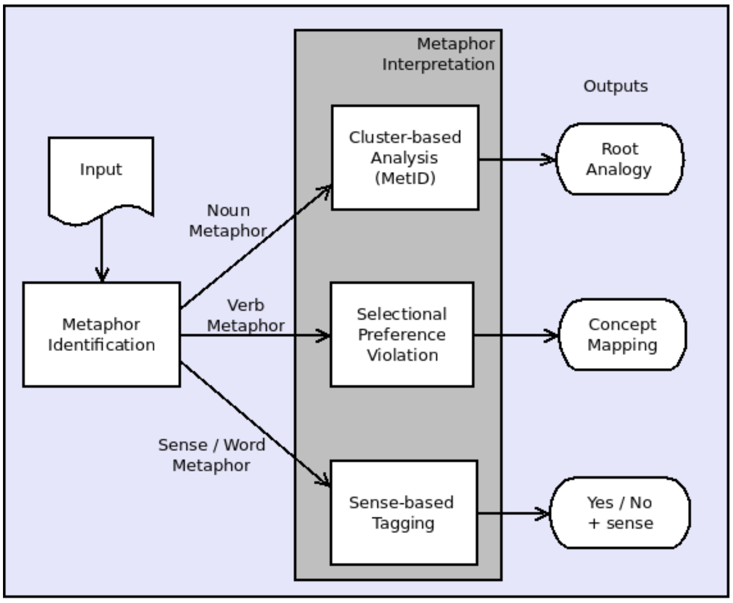


Figure 6.1: An alternative architecture for metaphor processing in which the method of interpretation is dictated by the output from the identification task. The interpretation's output depends on what type of metaphor was processed.

Using an external set of possible interpretations is a strength and weakness in MetID: if a statement fits a candidate metaphor, the statement 1) is identified as a metaphor and 2) is interpretable by that candidate. This design provides a helpful separation between linguistic work, exploring the conceptual structure of metaphor, and computational research on the machinery of processing metaphor. However, the biggest weakness of this design is that metaphors not provided by *Metalude* will not be identified. Two systems, CorMet and CMI, attempt to address this issue by finding classes of common mappings, using them to identify new or novel mappings [15, 158]. The classes in CorMet and CMI are analogous to MetID's use of topic and vehicle terminology from *Metalude*, but are built from text analysis. While this strategy can potentially find new metaphors, it will overlook many that have become lexicalised – especially verb-based phrases like "rising stocks" and "falling crime" [69].

Another way in which a bifurcated processing architecture could accommodate a wider range of metaphors is by taking into account theories of competing processing metaphors [25, 84]. For example, were a system to implement category matching and feature mapping procedures specifically, different kinds of linguistic metaphors could be processed in accordance to different processes (see table 2.2. As it is currently implemented, MetID is better suited to addressing categorical metaphors given the strictly lexical modelling of similarity. This is because categorical information (especially in WordNet) can be encoded more succinctly than featural information [147]. That is, category memberships require less information than unbounded sets of features needed to implement feature mapping.

6.1.1 Role of Lexical Models & Corpora

MetID was designed, in part, to evaluate the ability of semantic space models to help process metaphor. WordNet was included to provide a kind of base-line: an explicitly coded model as an alternative to vector-space models. In the evaluation, WordNet was consistently among the best-performing in almost all circumstances (see sections 5.4 and 5.5). Disregarding coverage, which as noted previously is mostly an issue of frequency as opposed to method, LSA-500 and COALS-14k performed comparably well to WordNet on identification and interpretation tasks respectively, though scores from LSA-400 were more strongly correlated with participant data (experiment 2). This can be interpreted as a success for these models, as they were able to (automatically) construct a semantic model comparable to the hand-built WordNet. One goal of corpus-based models is to rely exclusively on text to build viable representations without appeal to an external source or authority. Using text alone allows a corpus to define words by their use and association with other words. WordNet, on the other hand, which is primarily coded by psycholinguists and lexicographers, is like a highly-structured thesaurus or dictionary. So what makes WordNet a consistently good model for metaphor processing?

Figurative language can be defined as a transfer between two concepts that disregards literal aspects of one to make sense of another. The complex part is that making sense of figurative relations requires knowledge about the topic and vehicle. This knowledge is largely conceptual, though there appear to be linguistic constraints on how the relationships are instantiated [68, 69, 92, 212]. For example, while it is lexically normative (and common) to say that numbers rise, it is not literally true: rising is a concept that relates to elevation and numbers are abstract representations of quantity. This kind of conceptual information is precisely what gives metaphor its communicative efficacy and explanatory power [71]. The results of testing WordNet against the corpus-based, semantic space models points to the importance of the information in WordNet. That is, there is conceptual knowledge in WordNet that has been used to develop entries and relations. In metaphor processing, the information coded in distributional models' vectors is analogous to the explicit information coded in WordNet. Though WordNet can be limited, its explicit representation of this type of knowledge enables it to perform competitively with the best corpus-based models.

The choice of corpora has a significant effect on the performance of semantic space models [13, 184, 217]. A number of collections were used in MetID to build word clusters with the distributional models (see section 5.2), but only the two largest (TASA and enTenTen) were used to evaluate the system. This was mainly due to coverage considerations: a number of seed words were not frequent enough in the smaller corpora to be represented in some models. Even the enTenTen corpus, which has the largest vocabulary at 69,745 words, exhibited coverage problems in certain models like COALS-800. With some models, this is due to dimensional reduction thresholds (LSA and COALS-SVD) or to minimum frequency thresholds (HAL and COALS). These constraints are designed to assure better representations and the simplest way to compensate for this is to use larger collections. Another way to compensate for this would be to enlarge the clusters beyond 200-words. Doing so would increase the inclusiveness of the seeds, but would lower the average quality of clusters but was beyond the scope of this thesis.

6.2 Linguistic Implications

6.2.1 Effects of Grammatical Structure & Literalness on Metaphor Processing

In the second experiment, an interaction was found between the grammatical structure and the literal / non-literal distinction of a statement. The results show that the grammatical structure has significantly more effect on MetID's ability to process figurative statements than literal statements (section 5.5; figure 5.6). The effect underscores the role of linguistic structure on metaphor processing more generally: for figurative statements, the grammatical structure plays a stronger role than for literal statements. This interaction was found in the interpretation task, where MetID's interpretations were rated by people. Participants' ratings indicate that MetID was better at interpreting figurative statements than literal – which is perhaps a product of the root analogy approach. However, the interaction shows that literal verb-statements are more accurately interpreted than non-literal verb-statements, without considering the grammatical structure. These findings indicate that grammatical structure has a unique (or at least exaggerated) role in mediating the interpretation of figurative language specifically.

It could also be that verb-based metaphors are harder for *people* to interpret. In a post-hoc analysis of participant data from the second experiment, it was found that the average sensibility score[3] for figurative statements was generally lower than for literal statements (figure 6.2). Because the sensibility ratings were used to down-weight the paraphrase ratings, figurative statements

[3] The rating for the statement, not of the interpretation / paraphrase produced by MetID.

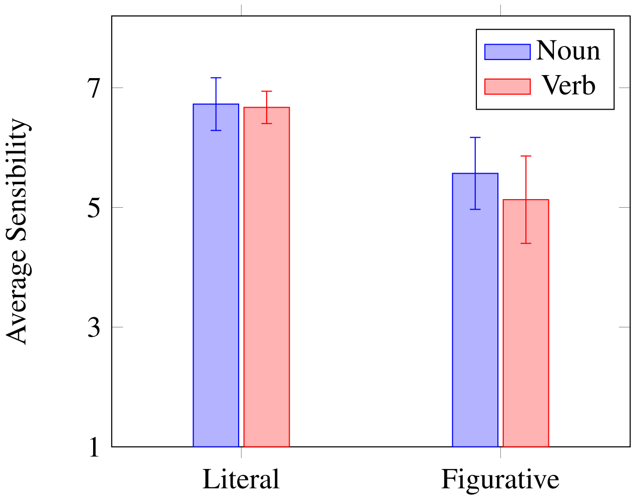


Figure 6.2: Average sensibility from 1 (bad) to 7 (excellent) for the four classes of sentences in experiment 2 (section 5.5). Recall that the sensibility scores were used to down-weight the paraphrase ratings – any score less than 7 will lower the score of the corresponding paraphrase rating. Error-bars are 1 SD of the mean and *N* varies in each class due to exclusion criteria.

tended to be lower than literal statements, on average. Taken with the interaction effect described above, this implies there is something uniquely difficult about figurative verb statements; the absence of an interaction in the sensibility scores implies that verb metaphors are different from noun metaphors.

6.3 Conclusions & Future Work

MetID underscores how linguistic theory and description can help define computational problems. Linguistic metaphor is a complex and broadly defined phenomenon closely related to reasoning about concepts. The system operationalises a view of how metaphors appear in text: a related pair of words associated with a set of paired concepts. This conception neglects some of the creativity inherent in the use of language and reduces the range of metaphorical expression to those represented in *Metalude*. Assuming that a linguistic metaphor is a paired association to one or more root analogies limits the observable phenomena, but it provides constraints under which to approach the problem computationally. On one hand, MetID's strategy has programmatic advantages because it provides a clear goal for identification and a mechanism for interpretation. On the other hand, the list of potential interpretations may be incomplete, redundant or wrong. MetID attempts to alleviate this potential problem by ranking candidate metaphors, instead of choosing a single metaphor for a statement. Ranking the results lets users examine the results for potentially helpful interpretations, a process highlighted by the *contagion* case-study (section 5.6).

One of the problems in the evaluated models, was that the corpus-based models tended not to encode all the information necessary to process metaphors, whereas the explicitly coded model, WordNet, performed consistently and was best in terms of coverage. This implies that WordNet codes some of conceptual knowledge needed to address metaphor. Optimally, a metaphor processing system could extract word associations *and* conceptual information from a corpus of text. Though semantic space models offer an approach to conceptual representation, these models appear not to adequately capture it to the extent needed to process figurative language. Some of the corpus-based models do indeed appear to represent information like features and categories (see section 5.5.4), but none of them are a comprehensive solution to word-concept associations.

MetID was designed to capture the fact that metaphor is firstly a conceptual phenomenon (the within-cluster search) and secondly, a linguistic phenomenon signalled by various surface and statistical cues (the heuristics) [92, 182, 230]. Previous metaphor processing systems did not preserve this separation. For example, figurative sense-tagging only marks an instance of a single word, providing no analogue to the topic-vehicle structure apparent at the conceptual level [22, 218]. Other systems use selectional preference violations, where deviations from normative subject- / object-verb constructions constitutes metaphor [199, 200]. There are, however, two systems that attempt to address the conceptual structure: Mason's CorMet and Baumer's CMI. Both of these methods find word-pairs in certain relations signalling concept mappings that were not commonly found elsewhere [15, 158]. MetID embodies a kind of hybrid strategy, combining the identification and interpretation tasks into a single problem. The system effectively measures the likelihood that an observed statement is an instance of every possible root analogy. This preserves a cline of metaphoricity [46] and allows the system to use other cues like co-text markers, selectional violations and predication. MetID also attempts to preserve the conceptual structure that is fundamental to metaphor: the relationship between topic and vehicle concepts. While it does not perform well enough to be considered an outright solution, its architecture is unique and exemplifies a step toward comprehensive metaphor processing.

To better address verb metaphors, selectional preference violations should be more central to the algorithm. Instead of the core algorithm looking for term-pairs that maximise paired associations with the root analogies, it could instead use selectional violations to select and prioritise observed word-pairs. Further, by conditionally applying a feature-selection process to extract nominal-like concepts from verbs in the input, a system could "translate" verb-statements to a nominal form, making them more compatible with the representations in *Metalude*. In addition to promoting the use of selectional preferences violation, and addressing the nominal and action concept disconnect, there are some specific areas for future work. The first is further research on metaphor in language, with the aim of developing more particular definitions of how linguistic metaphor relates to conceptual structure. The second is in computational semantics, where a number of new approaches have been proposed as alternatives to the semantic space models. Last, work in NLP, which has grown considerably with the adoption of machine learning techniques, can offer new insights into the automation and validation of linguistic and conceptual processes.

Metaphor in Language

The method presented in chapter 3 relies on a structured definition of metaphors – namely using a vehicle to make sense of a topic concept. This definition is apparent in many linguistic metaphors but is perhaps most easily observed in the structure of noun-noun predications. Alternatively, for action metaphors, the conceptual structure is not necessarily evident in the linguistic structure. Take (xxviii), for example:

(xxviii) The boy got on his bike and flew home.

Here, the verb *flew* selects an exemplary or prototypical feature, perhaps FAST, SPEED or EXPEDIENCE, yielding an interpretation FLYING = FAST MOVEMENT. In (xxviii), the structure of the metaphor is a verb (*flew*) as a nominal concept (FAST MOVEMENT). Because action metaphors are more common than noun-based metaphors [133], it is important for the structure to afford equivalent expressions. One way to do this is, instead of asserting a figurative equality, to assert an *aspect* of the verb as in FLYING AS SPEED. The conceptual structure of verb-based metaphors is similar to the topic-vehicle structure typical of root analogies, but it emphasises the selective role the topic has on the vehicle. For example, in (xxviii), the vehicle concept FAST is selected by the topic concept BOY ON BIKE. Developing verb-based representations like FLY = MOVE QUICKLY is one way to better address verb-based metaphors in the style of root analogies. It is likely, however, that noun-concepts are easier and more simple to deal with because they provide stronger prototype and exemplar information [74] and may have finer resolution on scales like imageability and concreteness [181]. Overcoming these obstacles is a matter of tenacity for corpus-linguistic research as these concerns are not likely to be fatal in developing structures of equations like those relating noun-concepts.

Much of the literature in metaphor research relies on somewhat intuitive concept derivations ([19] and [135], for example). Even corpus-driven analyses tend to triangulate common categories, types or domains of concepts in figurative language [34, 35, 45, 128]. Though a number of internal and contextual properties of linguistic metaphors have been shown to bare on metaphors' interpretation, it is not clear how a verb provides a property for use with the topic [92]. This selection process has been studied for noun metaphors, where features and categories are more clear [135] in which topics place constraints on vehicles [212]. It could be that this process is quite different for nouns and verbs [47], perhaps due to more articulated differences among nouns [53]. Indeed, nominal concepts behave differently than action concepts [74] and are more readily accessed as prototypes than verbs [232], perhaps making them better suited to define metaphor.

The distinction between noun- and verb-based metaphors is one of many in figurative language. This research attempted to address noun and verb metaphors in a unified way, but generally failed in the verb case. While this may be due to predominantly nominal definitions of metaphor, it could also be due to less frequent explication of verbs. Though text often evinces the properties of nominal concepts (ie. talking about *things*), verbs are often more contextual, abstract, vague and implicit. Other types of metaphor may prove similarly difficult. For example, adjective metaphors like "the urban brain" or "a hot temper" are perhaps more similar to action metaphors given their use of vehicle properties. Systems like MetID could benefit from further analysis of how words relate, at the lexical, grammatical and semantic levels, to the concepts they employ.

Computational Semantics

The metaphor interpretation task relies centrally on word associations provided by a lexical model. Statistical methods of associating words have made considerable progress over the last two decades. Many of the semantic space models used in this research were designed to provide analogues to mental representations and cognitive processes [111, 146, 149]. Three new approaches have become popular focal points for cognitive and computational research: probabilistic, tensor and graph-based models. Each of these strategies attempt to address some cognitive phenomena like semantic growth, memory and priming, as well as computational tasks like word association and clustering. Because some of MetID's weaknesses are due to the semantic models mis-associating words, these new approaches could provide better alternatives to semantic space models.

Probabilistic, non-parametric models have emerged as a way to model lexical semantics [23, 60, 63, 90, 204]. Probabilistic models begin by assuming a prior-probability distribution over possible observations and a set of latent variables that contribute to posterior observations. Using Bayes' theorem and a sampling procedure, the latent variables are tuned to develop the posterior into the observed distribution. In statistical semantics, the Bayesian approach has two main advantages over semantic space representations. First, it can account for unseen and unknown variables, such as authorship, topic and domain [89, 189, 204]. Second, and perhaps more importantly, probabilistic models relax assumptions of completeness in the input, allowing structure to be gleaned from low-density data. MetID could use probabilistic semantic models to allow less complete and noisier text collections to develop word representations. The Bayesian paradigm has strengths independent of individual model results, like rule-learning from sparse data [51, 60, 210] and it could potentially provide more accurate word associations than the models tested in MetID.

So-called tensor models are an extension to semantic space models like LSA [13, 41, 127]. Instead of using high-dimensional matrices to represent words, two or more such representations are combined to form an n-way tensor. Tensor-based representations have been built to combine word-document, word-word, word-POS and word-dependency matrices into one structure. In *distributional memory*, a three-way tensor is constructed from POS, windowed co-occurrence and dependency matrices and performs comparably well to best-in-task semantic space models on a range of tasks [13]. The strength of the tensor-based models is not necessarily that they perform better than semantic space models, but that they offer a *unified representation* for different tasks. Tensor models embody a largely mathematical advancement in statistical semantic models, and in applications like MetID, they could combine different semantic space representations.

Graph-based semantic models were proposed in linguistics long before they were implemented computationally. The theory of a mental semantic network accounts for memory, recall, priming and association tasks [11, 56]. Implementations of semantic networks have also been around for some time, WordNet being perhaps the most iconic [167]. Recently, empirical findings in the emergent structure of naturally occurring networks have sparked renewed interest in their application to linguistics and cognitive science [6, 11, 166]. Computational, graph-based models can be used to construct directed networks using word-context, word-document or dependency information. These models capitalise on long-standing graph-theoretic procedures, using them to simulate semantic operations. For example, word-word associations can be drawn by minimising path- or node-traversals from one entry to another [31, 100]. Concepts can be extracted by finding clusters of highly connected nodes [205] and such connectivity mimics the development of lexica in children [204]. Unlike semantic space models, graph-based representations can be annotated with explicit relations (word class, dependency or co-occurring distance information). In systems like MetID, graph-based methods could provide concept extraction, word associations and explicit representations, all of which would contribute to addressing a wider range of figurative language.

NLP & Machine Learning

Significant progress has been made in NLP by applying machine learning methods to outstanding problems. The decision to avoid machine learning methods in MetID was motivated by three considerations. First, machine learning algorithms tend to rely on structured, curated data-sets [125, 209]. Because figurative language is ubiquitous and diverse in communication, finding or creating viable training data has proved difficult [198][4]. Second, machine learning techniques can obfuscate the relationship between a trained model, the representations used to create it and those it is used to analyse. For example, in connectionist systems where nodes in a network are connected with weighted paths such that the weights correctly map training inputs to outputs, there is effectively no representational analogue to what the weights *mean*, despite the fact that they constitute the solution. In representational algorithms like WordNet, LSA and MetID as a whole, there is an answer to how or "why" the system produced the output – even in vector-space models where the vectors themselves represent "meaningless" dimensions. Though machine learning is uniquely positioned to build scalable systems for complex data-mining and pattern-matching tasks, such techniques would shed little light on how metaphors are used, identified or interpreted. Finally, machine learning algorithms typically employ a feature selection process where a set of features are chosen (sometimes automatically) and validated by a ranking process [242]. This means that features chosen to establish a model may have no principled connection to the task or the data. In this work, given the amount of experimental and linguistic research on the properties of metaphor, feature selection would risk discarding established features of metaphor use, in turn, making it difficult to relate the system's performance and theories of metaphor comprehension.

This research is situated half-way between representationalism and non-representationalism: the distributional models (and some of the heuristics) adopt a corpus-centric / use-based conception of lexical semantics. One problem with this approach is that concepts are represented as words themselves. There are some models that extract categories, frames and other kinds of conceptual information using co-occurrence patterns [12, 14, 57, 184]. Similar representational methods in NLP could be used to augment or replace resources like *Metalude* in MetID.

In addition to the lexical models, MetID also implemented some corpus-based heuristics for measuring selectional preference violation and predicative strength (see section 4.5). These two heuristics, the first of which has been introduced elsewhere [182, 199, 200], rely on a simplification of figurative language: that it violates otherwise "normal" language. The conception that

[4] Personal communication with E. Shutova; 29 March, 2012.

figurative language is in some way abnormal, or violates normative structures, is not supported by cognitive linguistic findings [24, 85, 133, 211]. One example is how the noun *stock* selects verbs like *rise*, *fall*, *crash* and *climb* [69], all of which are figurative. What selectional violation and predication strength actually measure is a degree of deviation from typical language. Selectional violation appears to be a viable way to address novel, verb-based metaphors [200, 201], but lexicalised metaphors, like rising prices or falling temperatures, require extrinsic conceptual knowledge about quantities and movement. For this reason, NLP techniques may continue to need outside knowledge to make sense of that which is seldom explained in text.

6.4 Concluding Remarks

Metaphor processing continues to present a unique challenge for computational fields. It is perhaps one of the most complex and creative conceptual phenomena evident in language and there is seemingly limitless potential in the use of metaphor as an expository device. Though many become lexicalised over time, metaphors engender creativity that is both common and complex which is precisely what imbues it with such communicative efficacy. Building on cognitive and linguistic theories of metaphor comprehension and use, this thesis contributes to a paradigm that uses computational modeling to explore, express and test such theories. Designing and implementing MetID allowed various aspects of how metaphors are used in natural language to be operationalised and tested. The results highlight the need for better definitions of figurative language and improved technical apparatus for relating textual information to conceptual information. Further, the dominant theories of metaphor are largely nominal in nature and have an important role in shaping systems to automatically identify and interpret metaphor. Data-driven approaches, such as corpus-based semantic models, can help address metaphor processing tasks like those explored in this research. Though the breadth of figurative language may never be fully accounted for with data-driven strategies, computer science and NLP can leverage existing resources and theories in the cognitive sciences, to build more accurate and flexible metaphor processing systems. Systems like MetID contribute to an increasingly comprehensive understanding of language and communication, and the processes that make them such powerful carriers of knowledge and meaning.

Bibliography

[1] A. Alter and Y. S. Schüler. Credit spread interdependencies of European states and banks during the financial crisis. *Journal of Banking & Finance*, 2012.

[2] K-H. Bae, A. G. Karolyi, and R. M. Stulz. A new approach to measuring financial contagion. *Review of Financial Studies*, 16 (3): 717-763, 2003.

[3] D. L. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR 1998*, pages 96-103, 1998.

[4] S. Banergee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pages 805-810, 2003.

[5] S. Bangalore and A. Joshi. Supertagging: an approach to almost parsing. *Computational Linguistics*, 25 (2): 237-265, 1999.

[6] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439): 509-512, 1994.

[7] J. Barnden. Consciousness and common-sense metaphors of mind. In S. O'Nuallain, P. McKevitt, and E. Mac Aogain (Eds.) *Two Sciences of Mind: Readings in Cognitive Science And Consciousness*, pages 331-340. John Benjamins Publishing Company, 1997.

[8] J. Barnden. Uncertainty and Conflict Handling in the ATT-Meta Context-Based System for Metaphorical Reasoning. In *Proceedings of Third International Conference on Modeling and Using Context (CONTEXT 2001)*, Lecture Notes in Artificial Intelligence, 2116. Springer-Verlag, 2001.

[9] J. Barnden, S. Glasbey, M. Lee, and A. Wallington. Reasoning in metaphor understanding: The ATT-Meta approach and system. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1188-1193, 2002.

[10] J. Barnden. Metaphor and metonymy: Making their connections more slippery. *Cognitive Linguistics*, 21 (1): 1-34, 2010.

[11] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen. Networks in Cognitive Science. *Trends in Cognitive Science*, 17 (7): 348-360, 2013.

[12] M. Baroni, A. Lenci, and L. Omnis. ISA meets lara: A fully incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 49-56, 2007.

[13] M. Baroni and A. Lenci. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36 (4): 673-721, 2010.

[14] M. Baroni, B. Murphy, E. Barbu, and M. Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34: 222-254, 2009.

[15] E. P. S. Baumer, D. Hubin, and B. Tomlinson. Computational Metaphor Identification. Technical Report: LUCI-2010-002, University of California, Irvine, 2010.

[16] D. Baur. Financial contagion and the real economy. *Journal of Banking & Finance*, 36: 2680-2692, 2012.

[17] M. P. Beardsley. Metaphor. In P. Edwards (Ed.) *Encyclopedia of Philosophy*, vol. 5, Macmillan: New York, USA, 1967.

[18] R. A. Bentley, M. J. O'Brien, and W. A. Brock. Mapping collective behavior in the big-data era. *Behavioral & Brain Science*, in press.

[19] D. Berggren. The use and abuse of metaphor. *Review of Metaphysics*, 16 (2): 237-258, 1962.

[20] R. C. Berwick, P. Pietroski, B. Yankama, and N. Chomsky. Poverty of the Stimulus Revisited. *Cognitive Science*, 35 (7): 1207-1242, 2011.

[21] R. C. Berwick, A. D. Friederici, N. Chomsky, J. J. Bolhuis. Evolution, brain, and the nature of language. *Trends in Cognitive Science*, 17 (2): 89-98, 2013.

[22] J. Birke and A. Sarkar. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy*, pages 329-336, 2006.

[23] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (4-5): 993-1022, 2003.

[24] L. Boroditsky. Metaphoric structuring: understanding time through spatial metaphors. *Cognition* 75: 1-28, 1999.

[25] B. F. Bowdle and D. Gentner. The career of metaphor. *Psychological Review*, 112: 193-216, 2005.

[26] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra and J.C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18: 467-479, 1992.

[27] C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12: 177-210, 1997.

[28] K. Burke. *A Grammar of Motives*. Prentice Hall: New York, 1945.

[29] L. Burnard. *Reference Guide for the British National Corpus (XML Edition)*, 2007.

[30] L. Cammeron and A. Deignan. The emergence of metaphor in discourse. *Applied Linguistics*, 27 (4), 2006.

[31] R. F.-i-Cancho, R. V. Solé, and R. Köhler. Patterns in syntactic dependency networks. *Physical Review E*, 69 (5): 051915, 2004.

[32] R. F.-i-Cancho, R. V. Solé, and R. Köhler. Euclidean distance between syntacticaally linkeed words. *Physical Review E*, 70 (5): 056135, 2004.

[33] D. Cer, M-C. de Marneffe, D. Jurafsky, and C. D. Manning. Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1628-1632, 2010.

[34] J. Charteris-Black and T. Ennis. A comparative study of metaphor in Spanish and English financial reporting. *English for Specific Purposes* 20: 249-266, 2001.

[35] J. Charteris-Black and A. Musolff. 'Battered hero' or 'innocent victim'? A comparative study of metaphors for euro trading in British and German financial reporting. *English for Specific Purposes* 22: 153-176, 2003.

[36] N. Chomsky. *Cartesian Linguistics*. Cambridge University Press: Cambridge, UK, 1983.

[37] N. Chomsky. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger Publishers: Westport, CT, USA, 1986.

[38] N. Chomsky. *The Science of Language*. Cambridge University Press: Cambridge, UK, 2012.

[39] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16: 22-29, 1990.

[40] D. Cruse. *Lexical Semantics*. Cambridge University Press: Cambridge, MA, 1986.

[41] T. Van de Cruys. A Non-negative Tensor Factorization Model for Selectional Preference Induction. In *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, pages 83-90, 2009.

[42] D. Das, N. Schneider, D. Chen, and N. A. Smith. Probabilistic frame-semantic parsing. In *Proceedings of human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (HLT-NAACL 2010).*, 2010.

[43] M. Davies. Corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25 (4): 447-464, 2010.

[44] C. J. Davis. Contagion as Metaphor. *American Literary History*, 14 (4): 828-836, 2002.

[45] A. Deignan. *Metaphor and Corpus Linguistics*. John Benjamins Publishing Company, 2005.

[46] A. Deignan. A corpus linguistic perspective on the relationship between metonymy and metaphor. *Style*, 39 (1): 72-91, 2005.

[47] A. Deignan. Corpus linguistics and metaphor. In R. W. Gibbs, (Ed.) *The Cambridge Handbook of Metaphor and Thought*Cambridge University Press: NY, USA, 2008.

[48] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the Semantic Web. *The VLDB Journal*, 12 (4): 303-319, 2003.

[49] R. Dodd and P. Mills. Outbreak: U.S. Subprime Contagion. *Finance & Development*, 45 (2): 14-19, 2008.

[50] M. Dungey G., Milunovich, and S. Thorp. Unobservable shocks as carriers of contagion. *Journal of Banking & Finance*, 34 (5): 1008-1021, 2010.

[51] A. D. Endress. Bayesian learning and the psychology of rule induction. *Cognition* 127 (2): 159-176, 2013.

[52] K. Erk. A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 216-223, 2007.

[53] V. Evans. Lexical concepts, cognitive models and meaning-construction. *Cognitive Linguistics*, 18 (1): 491-534, 2007.

[54] D. Fass. Collative Semantics: a semantics for natural language processing. *Doctoral Dissertation*, New Mexico State University, NM, USA, 1988.

[55] D. Fass. met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics*, 17 (1): 49-90, 1991.

[56] C. Fellbaum (Ed.) *WordNet: An Electronic Lexical Database*. MIT Press: Cambridge, MA, USA, 1998.

[57] C. Fillmore. Frame semantics. *Linguistics in the morning calm*, 111-137, 1982.

[58] W. Forbes. *Behavioural Finance*. Chichester: Wiley, 2009.

[59] N. Francis and H. Kucera. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University, Department of Linguistics, Providence, RI, USA, 1964.

[60] M. C. Frank, and J. B. Tenenbaum. Three ideal observer models for rule learning in simple languages. *Cognition*, 120 (3): 360-371, 2011.

[61] R. M. French. *The Subtlety of Sameness*. MIT Press: MA, 1995.

[62] C. D. Frith and U. Frith. The Neural Basis of Mentalizing. *Neuron*, 50 (4): 531-534, 2006.

[63] N. Fuhr. Probabilistic Models in Information Retrieval. *The Computer Journal*, 35 (3): 243-255, 1992.

[64] R. Gabbard, M. Marcus, and S. Kullick. Fully Parsing the Penn Treebank. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (ACL 2006)*, pages 184-191, 2006.

[65] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1606-1611, 2007.

[66] W. A. Gale and G. Sampson. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2 (3): 217-237, 1995.

[67] A. Gerow and M. T. Keane. Mining the Web for the "Voice of the Herd" to Track Stock Market Bubbles. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 2244-2249, 2011.

[68] A. Gerow and M. T. Keane. Identifying Metaphor Hierarchies in a Corpus Analysis of Finance Articles. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci 2011)*, pages 2049-2054, 2011.

[69] A. Gerow and M. T. Keane. Identifying Metaphoric Antonyms in a Corpus Analysis of Finance Articles. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci 2011)*, pages 3288-3292, 2011.

[70] A. Gerow and K. Ahmad. Diachronic Variation in Grammatical Relations. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, 2012.

[71] A. Gerow, K. Ahmad, and S. Glucksberg. The Concept of Contagion in Finance: A Computational Corpus-based Approach. To appear in *Proceedings of the 19th European Symposium on Languages for Special Purposes (LSP 2013)*, 2013.

[72] A. Gerow and M. T. Keane. It's Distributions All The Way Down! *Behavioral & Brain Sciences*, in press, 2013.

[73] A. Gerow, K. Ahmad, and S. Glucksberg. Contagion in Finance. *Working Paper*, under review.

[74] D. Gentner. Some interesting differences between verbs and nouns. *Cognition and brain theory*, 4 (2): 161-178, 1981.

[75] D. Gentner. Structure-Mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2): 155-170, 1983.

[76] D. Gentner. Metaphor as Structure Mapping: The Relational Shift. *Child Development*, 59 (1): 47-59, 1988.

[77] D. Gentner and P. Wolff. Metaphor and knowledge change. In E. Dietrich and A. Markman (Eds.) *Cognitive Dynamics: Conceptual change in humans and machines*. Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2000.

[78] R. W. Gibbs. *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press: New York, USA, 2008.

[79] D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28 (3): 245-288, 2002.

[80] S. Glucksberg, S. M. McGlone, and D. Manfredi. Property attribution in metaphor comprehension. *Journal of Memory and Language*, 36: 50-67, 1997.

[81] S. Glucksberg. *Understanding Figurative Language*. Oxford University Press: Oxford, UK, 2001.

[82] S. Glucksberg. The psycholinguistics of metaphor. *Trends in Cognitive Science*, 7: 92-96, 2003.

[83] S. Glucksberg and C. Haught. Can Florida Become Like the Next Florida? When Metaphoric Comparisons Fail. *Psychological Science*, 17 (11): 935-938, 2006.

[84] S. Glucksberg and C. Haught. On the Relation Between Metaphor and Simile: When Comparison Fails. *Mind & Language*, 21, 360-378, 2006.

[85] J. Grady, S. Taub, and P. Morgan. Primitive and Compound Metaphors. In A. E. Goldberg (Ed.) *Conceptual structure, discourse and language*, pages 177-187, 1996.

[86] S. Greco. Metaphorical headlines in business, finance and economic magazines. *Linguistica e Fililogia*, 28: 193-211, 2009.

[87] S. Th. Gries. Corpus-based methods and cognitive semantics: The many senses of *to run*. In S. Th. Gries and A. Stefanowitsch (Eds.) *Corpora in Cognitive Linguistics*, pages 57-100, 2006.

[88] S. Th. Gries and A. Stefanowitsch (Eds.). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Mouton de Gruyter: Berlin, Germany, 2006.

[89] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (1): 5228-5235, 2004.

[90] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, 17: 537-544, 2004.

[91] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in Semantic Representation. *Psychological Review* 114 (2): 211-244, 2007.

[92] A. Goatly. *The Language of Metaphors*. Routledge: London, UK, 1997.

[93] A. Goatly. Metaphors as Resource for the Conceptualisation and Expression of Emotion. In K. Ahmad (Ed.) *Affective Computing and Sentiment Analysis*, pages 13-26, 2012.

[94] H. Haker, et al. Mirror neuron activity during contagious yawning – an fMRI study. *Brain Imaging and Behavior*, 7 (1): 28-34, 2012.

[95] Y. Hao. PhD Thesis: *Stereotype, Simile and Metaphor*. University College Dublin, School of Computer Science & Informatics, 2010.

[96] I. Hardie and D. MacKenzie. Assembling an Economic Actor: The Agencement of a Hedge Fund. *The Sociological Review*, 55 (1): 57-80, 2007.

[97] S. Harnad. The Symbol Grounding Problem. *Physica D*, 42: 335-346, 1990.

[98] S. Harnad. To cognize is to categorize: cognition is categorization. In H. Cohen and C. Lefebvre (Eds.) *Handbook of Categorization*, Elsevier: Amsterdam, Netherlands, 2003.

[99] Z. Harris. Distributional structure. In *Papers in structural and transformational linguistics*, pages 775-794: D. Reidel Publishing Company, 1970.

[100] A. Herdağdelen, K. Erk, and M. Baroni. Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 50-53, 2009.

[101] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.) *WordNet: An electronic lexical database*, pages 305-332. The MIT Press: Cambridge, MA, 1998.

[102] G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, 1994.

[103] F. Huo, J. Liu and B. Feng. A new method for retrieval based on relative entropy with smoothing. *Lecture Notes in Computer Science: Algorithmic Applications in Management*, vol. 3521: 821-824, 2005.

[104] H. Huo, J. Liu and B. Feng. Multinomial Approach and Multiple-Bernoulli Approach for Information Retrieval Based on Language Modeling. *Lecture Notes in Artificial Intelligence*, 3613: 580-583, 2005.

[105] H. Huo, A. Gerow, K. Ahmad. Measuring Multi-Term Co-occurrence with Multinomial and Bernoulli Language Models. TCD Working Paper, 2012.

[106] International Monetary Fund. Press Release, April 1999, No. 99/14.

[107] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37: 547-579, 1901.

[108] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19-33, 1997.

[109] K. S. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28 (11): 11-21, 1972.

[110] L. L. Jones and Z. Estes. Metaphor comprehension as attributive categorization. *Journal of Memory and Language*, 53: 110-124, 2005.

[111] M. N. Jones, W. Kintsch, and D. J. K. Mewhort. High-dimensional semantic space accounts priming. *Journal of Memory and language*, 55: 534-552, 2006.

[112] L. L. Jones and Z. Estes. Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55: 18-32, 2006.

[113] M. N. Jones and D. J. K. Mewhort. Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, 114 (1): 1-37, 2007.

[114] D. Jurafsky and J. H. Martin. *Speech and Language Processing (2nd edition)*. Prentice Hall: Upper Saddle River, NJ, USA, 2008.

[115] D. Jurgens and K. Stevens. The s-space package: An open source package for word space models. In *System Papers of the Association of Computational Linguistics*, 2010.

[116] G. Karypis. CLUTO: A Clustering Toolkit. Technical Report: #02-017, University of Minnesota, 28 November, 2003.

[117] M. T. Keane and A. Gerow. It's Distributions All The Way Down!. *Behavioral & Brain Sciences*, in press, 2013.

[118] B. Keysar, Y. Shen, S. Glucksberg, and W. S. Horton. Conventional language: How metaphorical is it? *Journal of Memory and Language*, 43: 576-593, 2000.

[119] A. Kilgarriff and M. Palmer (Eds.). *Computers and the Humanities (special issue based on SENSEVAL-1)*, 34 (1-2), 1999.

[120] A. Kilgarriff, P. Rychlý, P. Smrz, and D. Tugwell. The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105-116, 2004.

[121] W. Kintsch. Metaphor comprehension: a computational theory. *Psychonomic Bulletin Review*, 7: 257-266, 2000.

[122] W. Kintsch. How the mind computes the meaning of metaphor: A simulation based on LSA. In R. W. Gibbs (Ed.) *The Cambridge Handbook of Metaphor and Thought*, Cambridge University Press: New York, NY, USA, 2008.

[123] W. Kintsch and P. Mangalath. The Construction of Meaning. *Topics in Cognitive Science*, 3 (2): 346-370, 2011.

[124] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (Eds.), *The Computer and Literary Studies*. Edinburgh University Press: Edinburgh, UK, 1973.

[125] R. Kneser. Statistical language modeling using a variable context length. In *Proceedings of ICSLP 1996*, pages 494-497, 1996.

[126] S. Knudsen. Scientific metaphors going public. *Journal of Pragmatics*, 35 (8): 1247-1263, 2003.

[127] T. G. Kolda and V. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 10 June, 2008.

[128] Z. Kővecses. *Metaphor, 2nd edition*. Oxford University Press: Oxford, UK, 2010.

[129] H. Kucera, W. N. Francis, and J. Carroll. *Computational Analysis of Present-Day American English*. Brown University Press: Providence, RI, USA. 1967.

[130] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals Mathematical Statistics*, 22: 79-86, 1951.

[131] A. Kyle and W. Xiong. Contagion as a Wealth Effect. *Journal of Finance*, 56 (4): 1401-1440, 2001.

[132] G. Lakoff. The contemporary theory of metaphor. In A. Ortony (Ed.) *Metaphor and Thought, 2nd Edition*. Cambridge University Press: Cambridge, UK, 1992.

[133] G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press: Chicago, IL, USA, 1980.

[134] G. Lakoff and A. Schwartz. The master metaphor list. University of California Berkeley Technical Report: Cognitive Linguistics Group, 1991.

[135] G. Lakoff and M. Turner. *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press: Chicago, IL, USA, 2009.

[136] T. K. Landauer and S. T. Dumais. A Solution to Plato's Problem: The Latent Semantic Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104: 211-240, 1997.

[137] T. K. Landauer, P. E. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Analysis*, 25 (2-3): 259-284, 1998.

[138] T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum, 2004.

[139] V. Lavrenko. Localized smoothing for multinomial language models. CIIR Technical Report: IR-222, 2000.

[140] C. Leacock and M. Chodrow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.) *WordNet: An electronic lexical database*, pages 305-332. MIT Press: Cambridge, MA, 1998.

[141] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39 (4): 1-32, 2013.

[142] H. Lee, Y. Peirsman, A. Chang, N Chambers, M. Surdeanu, and D. Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the CoNLL-2011 Shared Task*, 2011.

[143] D. Lin. Using syntactic dependency as a local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, pages 64-71, 1997.

[144] D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th Intentional Conference on Machine Learning (ICML 1998)*, pages 296-304, 1998.

[145] B. Lönneker. Is there a way to represent metaphors in WordNet? Insights from the Hamburg Metaphor Database. In *Proceedings of the ACL 2003 Workshop on the Lexicon and Figurative Language*, pages 18-26, 2003.

[146] M. M. Louwerse. Symbolic or embodied representations: A case for symbol interdependency. In T. K. Landauer, et al. (Eds.) *Handbook of latent semantic analysis*, pages 107-120: Erlbaum: Mahwah, NJ, USA, 2007.

[147] M. M. Louwerse. Embodied relations are encoded in language. In *Psychonomic Bulletin & Review*, 15 (4): 838-844, 2008.

[148] M. M. Louwerse and W. Van Peer. How cognitive is cognitive poetics? The interaction between symbolic and embodied cognition. In G. Brône and J. Vandaele (Eds.) *Cognitive Poetics*, De Gruyter: Germany, 2009.

[149] M. M. Louwerse. Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, 3 (2): 273-302, 2011.

[150] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28: 203-208, 1996.

[151] D. MacKenzie. Long-Term Capital Management and the sociology of arbitrage. *Economy and Society*, 32 (3): 349-380, 2003.

[152] C. Macleod, N. Ide, and R. Grishman. The American National Corpus: Standardized Resources for American English. In *Proceedings of the Second International Language Resources and Evaluation Conference (LREC 2000)*, pages 831-836, 2000.

[153] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press: Cambridge, MA, USA. 1999.

[154] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth Language Resource and Evaluation Conference (LREC 2006)*, 2006.

[155] M.-C. De Marneffe and C. D. Manning. *Stanford typed dependencies manual, 2008*, http://nlp. stanford. edu/software/dependencies manual.pdf; 6 August, 2013.

[156] J. Martin. A computational theory of metaphor. University of California, Berkeley Technical Report: 88/465, Computer Science Division, 1988.

[157] J. Martin. MetaBank: A Knowledge-Base of Metaphoric Language Conventions. University of Colorado Technical Report #CU-CS-523-91, 1991.

[158] Z. J. Mason. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30 (1): 23-44, 2004.

[159] P. H. Matthews. *Oxford Concise Dictionary of Linguistics*. Oxford, UK & New York, NY, USA: Oxford University Press, 1997.

[160] R. McDonald and J. Nivre. Analyzing and Integrating Dependency Parsers. *Computational Linguistics*, 37 (1): 198-230, 2011.

[161] A. M. McMahon. *Understanding Language Change*. Cambridge University Press: Cambridge, UK, 1994.

[162] D. S. McNamara. Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition. *Topics in Cognitive Science*, 3 (1): 3-17, 2011.

[163] K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, & Computers* 37 (4): 547-559, 2005.

[164] D. Metzler, V. Lavrenko, and B. W. Croft. Formal multiple-Bernoulli models for language modeling. In *Proceedings of SIGIR 2004*, pages 540-541, 2004.

[165] J.-B. Michel, et al. Quantitative analysis of culture using millions of digitized books. *ScienceExpress* 10.1126, 2011.

[166] R. Mihalcea and D. Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press: Cambridge, UK, 2011.

[167] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38 (11): 39-41, 1995.

[168] J. J. A. Mooij. *A Study of Metaphor*. Longman: Harlow, UK, 1976.

[169] J. H. Neely. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106 (3): 226-254, 1997.

[170] L. Nummenmaa, et al. Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences*, 109 (24): 9599-9604, 2012.

[171] A. Ortony (Ed.). *Metaphor and Thought, 2nd edition*. Cambridge University Press: Cambridge, UK, 1992.

[172] S. Patwardhan. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, University of Minnesota, Duluth, USA, 2003.

[173] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity – Measuring the Relatedness of Concepts. In *Proceedings of HLT-NAACL: Demonstration Papers at HLT-NAACL 2004*, pages 38-41, 2004.

[174] M. S. Pernick. Contagion and Culture. *American Literary History*, 14 (4): 858-865, 2002.

[175] S. M. Platek, et al. Contagious yawning and the brain. *Cognitive Brain Research*, 23 (2): 448-452, 2005.

[176] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14 (3): 130-137, 1980.

[177] Pragglejaz Group. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22 (1): 1-39, 2012.

[178] S. Prasada, et al. Conceptual distinctions amongst generics. *Cognition*, 126 (3): 405-422, 2013.

[179] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of EMNLP-2010, Boston, USA*, 2010.

[180] M. Ramscar and D. Yarlett. Semantic grounding in models of analogy: an environmental approach. *Cognitive Science*, 27: 41-71, 2001.

[181] J. Reilly and J. Kean. Formal Distinctiveness of High?and Low?Imageability Nouns: Analyses and Theoretical Implications. *Cognitive science*, 31 (1): 157-168, 2007.

[182] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448-453, 1995.

[183] P. Resnik. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How (pages 52-57)*, 1997.

[184] B. Riordan and M. N. Jones. Redundancy in Perceptual and Linguistic Experience: Comparing Feature-Based and Distributional Models of Semantic Representation. *Topics in Cognitive Science*, 3 (2): 303-345, 2011.

[185] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC 1994), Gaithersburg, USA*, 1994.

[186] M. A. Rodriguez. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15 (2): 442-456, 2003.

[187] P. M. Roget [1852]. In S. M. Lloyd (Ed.) *Roget's Thesaurus*. Longman Group Limited: Essex, UK, (1962, 1982).

[188] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Cognitive Science*, submitted, 2009.

[189] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In *20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[190] W. Ruts et al. Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36 (3): 506-515, 2004

[191] M. Sahlgren, A. Holst, and P. Kanerva. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci 2008)*, 2008.

[192] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18 (11): 613-620, 1975.

[193] M. Schürmann, et al. Yearning to yawn: the neural basis of contagious yawning. *Neuroimage*, 24 (4): 1260-1264, 2005.

[194] J. R. Searle. Can Computers Think? In D. J. Chalmers (Ed.) *Philosophy of Mind*, pages 669-675. Oxford University Press: Oxford, UK, (2002), 1983.

[195] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decisions Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[196] H. Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, 1995.

[197] E. Shutova. Models of Metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688-697, 2010.

[198] E. Shutova and S. Teufel. Metaphor Corpus Annotated for Source – Target Domain Mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[199] E. Shutova, L. Sun, and A. Korhonen. Metaphor Identification Using Verb and Noun Clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1002-1010, 2011

[200] E. Shutova, S. Teufel, and A. Korhonen. Statistical Metaphor Processing. *Computational Linguistics*, 39 (2), 1-52, 2012.

[201] E. Shutova, T. Van de Cruys, and A. Korhonen. Unsupervised Metaphor Paraphrasing Using a Vector Space Model. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, 2012.

[202] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19 (1): 143-177, 1993.

[203] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Blackwell: Oxford, UK, 1986.

[204] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2004.

[205] M. Steyvers and J. B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive science*, 29 (1): 41-78, 2005.

[206] P. Stockwell. *Cognitive Poetics: An Introduction*. Routledge: London, UK, 2002.

[207] B. Stone, S. Dennis, and P. J. Kwantes. Comparing Methods for Single Paragraph Similarity Analysis. *Topics in Cognitive Science*, 3 (1): 92-122, 2011.

[208] L. Sun and A. Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1023-1033, 2009.

[209] M. Tan, W. Zhou, L. Zheng, and S. Wang. A Scalable Distributed Syntactic, Semantic, and Lexical Language Model. *Computational Linguistics*, 38 (3): 631-671, 2011.

[210] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331 (6022): 1279-1285, 2011

[211] P. H. Thibodeau and L. Boroditsky. Metaphors we think with: The role of metaphor in reasoning. *PloS ONE*, 6 (2): 2011.

[212] L. A. Torreano, C. Cacciari, and S. Glucksberg. When Dogs Can Fly: Level of Abstraction as a Cue to Metaphorical Use of Verbs. *Metaphor and Symbol*, 20 (4), 259-274, 2012.

[213] K. Toutanova, A. Haghighi, and C. D. Manning. Joint Learning Improves Semantic Role Labeling. In *Proceedings of Association for Computational Linguistics (ACL 2005)*, 2005.

[214] P. D. Turney. Mining the Web for Synonyms: PMI-IR verses LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML 2001)*, 2001.

[215] P. D. Turney and M. L. Littman. Corpus-based Learning of Analogies and Semantic Relations. *Machine Learning*, 60 (1-3): 251-278, 2005.

[216] P. D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(2): 379-416, 2006.

[217] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37: 141-188, 2010.

[218] P. D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 680-690, 2011.

[219] J. Ureña and P. Faber. Reviewing imagery in resemblance and non-resemblance metaphors. *Cognitive Linguistics*, 12 (1): 123-149, 2010.

[220] A. Utsumi. The Role of Feature Emergence in Metaphor Appreciation. *Metaphor and Symbol*, 20 (3): 151-172, 2005.

[221] A. Utsumi. Interpretative Diversity Explains Metaphor-Simile Distinction. *Metaphor and Symbol*, 22 (4): 291-312, 2007.

[222] A. Utsumi. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35: 251-296, 2011.

[223] K. M. Uyeda and G. Mandler. Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, 12 (6): 587-595, 1980.

[224] T. Veale. The Analogical Thesaurus. In *Proceedings of the 15th Innovative Applications of Artificial Intelligence Conference*, pages 606-612, 2003.

[225] T. Veale and Y. Hao. Comprehending and Generating Apt Metaphors: A Web-driven, Case-based Approach to Figurative Language. In *Proceedings of the 2007 Meeting of the Association for the Advancement of Artificial Intelligence (AAAI 2009)*, pages 1471-1476, 2007.

[226] T. Veale. A Context-sensitive, Multi-faceted model of Lexico-Conceptual Affect. In *Proceeding the 50th Annual Conference of the Association for Computational Linguistics (ACL 2012)*, 2012.

[227] D. P. Vinson and G. Vigliocco. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40 (1): 183-190, 2008.

[228] P. Wiemer-Hastings and I. Zipitria. Rules for Syntax, Vectors for Semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society (CogSci 2001)*, 2001.

[229] Y. Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6: 53-74, 1975.

[230] Y. Wilks. Making preferences more active. *Artificial Intelligence*, 11 (3): 197-223, 1978.

[231] Y. Wilks, A. Dalton, J. Allen, and L. Galescu. Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. In *Proceedings of The First Workshop on Metaphor in NLP*, pages 36-44, 2013.

[232] M. D. Wilson. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, 20 (1): 6-11, 1988.

[233] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pages 192-196, 1999.

[234] M. White. Metaphor and economics: the case of growth. *English for Specific Purposes*, 22 (2): 131-151, 2003.

[235] P. Wolff and D. Gentner. Structure-Mapping in Metaphor Comprehension. *Cognitive Science*, 35 (8): 1456-1488, 2011.

[236] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133-138, 1994.

[237] D. Yarowsky. Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual meeting of the ACL*, pages 189-196, 1995.

[238] S. M. Zeno, S. H. Ivens, R. T. Millard, and R. Duvvuri. *The educator's word frequency guide*. Touchstone Applied Sciences Associates: Brewster, NY, 1995.

[239] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of SIGIR 2001*, 2001.

[240] Y. Zhao and G. Karypis. Criterion Functions for Document Clustering. University of Minnesota Technical Report: #01-40, 21 February, 2002.

[241] W. Zhang, T. Yoshida, T. B. Ho and X. Tang. Augmented mutual information for multi-word extraction. *International Journal of Innovative Computing, Information and Control*, 5: 543-554, 2009.

[242] M. Zhitomirsky-Geffet and Ido Dagan. Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35 (3): 435-461, 2008.

[243] G. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley: Reading, MA, 1949.

Appendix A

Additional Results

A.1 Cluster Quality

The cluster analysis presented in section 5.3 summarised the findings without reference to the specific (or exhaustive) scores. Here, the clusters are reviewed in more detail and all scores are presented for each of configuration. The COLE-AMI model was presented in chapter 4, however, two more COLE models, based on language models, are described and evaluated here. Because neither of these COLE variants produced viable clusters (see section A.1.1), they were not evaluated in MetID. This section also contains a review of the clusters from each of the DSMs (LSA, COALS, BEAGLE and HAL), with each corpus (section 5.2) and each similarity (section 4.4.4).

A.1.1 COLE-based Language Models

Method

Language modeling is technique used in NLP for language generation and simulation tasks. The most basic example of a language model (LM) is called a *unigram* model where the relative frequency of every observed word is used to construct a probability distribution for all words. Using this distribution, language can be "modeled" by sampling from the distribution. This will generate a randomly selected sequence of words with a similar distribution to the observed text. A unigram LM will generate unintelligible strings because it does not account for phrasal, grammatical or morphological constraints. A more advanced LM is an *n*-gram model, where the proceeding n words, $(w_1, w_2, ..., w_{n-1})$, are used to calculate the conditional probability $p(w_n|w_1, w_2, ..., w_{n-1})$ with which the next word will occur. The intuition with *n*-gram models is that given a preceding n words, what is the most likely next word? By making use of punctuation and capitalisation information, a 5-gram model can generate nearly intelligible sentences which are considerably more sensible than a unigram model [153]. By constructing probability distributions over observed words, language models are a powerful tool for the analysis and generation of natural language and are ostensibly well suited to estimate co-occurrence [139, 164].

A good language model, however, will have to account for never-before-seen observations. Statistically, the intuition for this requirement is that some probability mass must be preserved for unobserved so the model can generate *new* terms in a sequence. One option is to assume that every word has been observed precisely once, which works in-practice, but will exaggerate the probability of low-frequency words. Other ways of addressing this have been proposed, such as Laplace smoothing [114], Good-Turing smoothing [66] and Kneser-Ney smoothing [125]. These smoothing techniques use the observed prior distribution to preserve some probability mass for new and low-frequency words. In recent literature, Dirichlet smoothing has been found to be an efficient, accurate method of local smoothing for n-gram models [139, 164].

The methods described here are based on [105], an extension of previous work, [103, 104, 139, 164], which use smoothed language models to estimate co-occurrence likelihood. Here, two models based on multinomial and Bernoulli distributions are introduced which are presented together to highlight their similarities. In the multinomial model, the sequence $(t_1, t_2, ..., t_n)$ is treated as a sequence of independent events occurring with independently random priors. The n-length sequence consists of as many random variables making probability of its observation obtainable by taking the product of probabilities for each term. This probability is generated by model M of document by D by multiplying each term's probability:

$$p(t_1, t_2, ..., t_n | M_D)) = \prod_{i=1}^{n} p(t_i | M_D) \tag{A.1}$$

In the Bernoulli model, the sequence is represented as a vector of binary attributes for each term in the vocabulary, V, indicating its presence in the sequence [164]. The terms are again assumed to occur with independent randomness. The likelihood $p(t_1, t_2, ..., t_n | M_D)$ of the sequence is the product of two probabilities with respect to M_D: that of producing the sequence and that of *not* producing another:

$$p(t_1, t_2, ..., t_n | M_D)) = \prod_{t_i \in seq} p(t_i | M_D) \prod_{t_i \notin seq} 1 - p(t_i | M_D) \tag{A.2}$$

Both models are built by observations of document D, comprised of a vocabulary, V, where each term $t_1, t_2, ..., t_n$ occurs with frequencies $f_1, f_2, ..., f_{|V|}$. The model for each document is parameterised by the vector $M_D = \langle M_{f_1}, M_{f_2}, ..., M_{f_{|V|}} \rangle \in [0, 1]^V$, which indicates the probability of omission or inclusion of $t \in V$, where $M_{f_i} = p(t_i | M_D)$ and the model frequencies, M_f, sum to 1. The length of a document, $|D|$, is defined as the sum of its terms' frequencies f_i that are used to compute the MLE of Eq. A.3.

$$p(M_D | D) \approx p(D | M_D) p(M_D) \tag{A.3}$$

Which gives

$$\widehat{M_D} \approx argmax_{M_D} p(D | M_D) p(M_D) \tag{A.4}$$

where $p(D|M_D)$ is the likelihood of the document under M_D and $p(M_D)$ is the prior of the model itself.

The multinomial model samples a multinomial distribution for each word in M_D, hence its name. When M_D parameterises the distribution and the model prior is a Dirichlet distribution, the conjugate prior, is defined as

$$\widehat{M_D} \approx argmax_{M_D} \frac{\Gamma(|D| + \sum_{i=1}^{|V|} \alpha_i)}{\prod_{i=1}^{|V|} \Gamma(f_i + \alpha_i) \prod_{i=1}^{|V|} (M_{f_i})^{f_i + \alpha_i + 1}} \tag{A.5}$$

where Γ is $s = (s-1)!$ and α_i is the hyper-parameter characterising the Dirichlet distribution. Eq. A.5 can be solved by expectation maximisation, which yields:

$$\widehat{M_{f_i}} = \frac{f_i + \alpha_i - 1}{|D| + \sum_{i=1}^{|V|} \alpha_i - |V|} \tag{A.6}$$

One choice of hyper-parameters α_i is to attribute equal probability to all terms $t \in seq$. However, this allows zero probabilities if the collection is small or sparse (ie. $\frac{|V|}{|C|}$ is low). Extending previous work to improve individual document models, a corpus-wide model is used to inform the smoothing of each document. Specifically, $\alpha_i = \mu \frac{f_{ci}}{|C|} + 1$ which provides Dirichlet smoothing will be used [139]. Here, μ is a smoothing factor, f_{ci} is the ith term's frequency in collection C. The probability of observing t_i given $\widehat{M_D}$ is

$$p(t_i|\widehat{M_D}) = \frac{f_i + \mu \frac{f_{ci}}{|C|}}{|D| + \sum_{i=1}^{|V|} \mu \frac{f_{ci}}{|C|}} \tag{A.7}$$

Sampling t in M_D from a Bernoulli distribution gives

$$\widehat{M_D} \approx argmax_{M_D} \prod_{i=1}^{|V|} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i) + \Gamma(\beta_i)} (M_{f_i})^{f_i + \alpha_i + 1} (1 - M_{f_i})^{|D| - f_i + \beta_i - 1} \tag{A.8}$$

to which the solution by expectation maximisation is

$$\widehat{M_f} = \frac{f_i + \alpha_i - 1}{|D| + \alpha_i + \beta_i - 2} \tag{A.9}$$

In the Bernoulli model, setting $\alpha_i = \frac{\mu_{fci}}{|C|} + 1$; $\beta_i = \frac{|C|}{f_{ci}} + \mu(1 - \frac{f_{ci}}{|C|}) - 1$ yields the form of Dirichlet smoothing used in the multinomial model, leading to term probability defined as

$$p(t_i|\widehat{M_D}) = \frac{f_i + \mu \frac{f_{ci}}{|C|}}{|D| + \frac{|C|}{f_{ci}} + \mu - 2} \tag{A.10}$$

To summarise, for the multinomial and Bernoulli models respectively, how likely it is a sequence of terms will co-occur in a document can be calculated by equations A.11 and A.12 respectively:

$$p_{multinomial}(t_1,t_2,...,t_n|D) = \prod_{i=1}^{n} \frac{f_i + \mu \frac{f_{ci}}{|C|}}{|D|} + \sum_{i=1}^{|V|} \mu \frac{f_{ci}}{|C|} \tag{A.11}$$

$$p_{Bernoulli}(t_1,t_2,...,t_n|D) = \prod_{t_i \in seq} \frac{f_i + \mu \frac{f_{ci}}{|C|}}{|D| + \frac{|C|}{f_{ci}} + \mu - 2} \prod_{t_i \notin seq} \frac{f_i + \mu \frac{f_{ci}}{|C|}}{|D| + \frac{|C|}{f_{ci}} + \mu - 2} \tag{A.12}$$

COLE-based Clusters

The language model methods appear to have generated nearly uniform clusters, evidenced in the low variances in each measure. The language models perform more accurately than AMI when ranking documents [105], but when using the same scores to cluster words, the estimations are too uniform to differentiate related words. AMI, on the other hand, is more affected by word-document frequency, which perhaps provides word-pairs the "space" to be re-ordered, yielding less uniform clusters. Recall that similarity, purity and entropy are computed using the 200 closest words to a given nucleus. This means that low variances across all clusters implies that the top-200 closest words change very little from one nucleus to another. Looking at purity and entropy, aside from the implications of the low variance, the language model clusters have and low purities very high entropies for every corpus. Without regard to how the word-ranking occurred, these clusters are of particularly low quality, especially compared to the DSM clusters. These results show that AMI is the only viable method with which to use the COLE models in MetID.

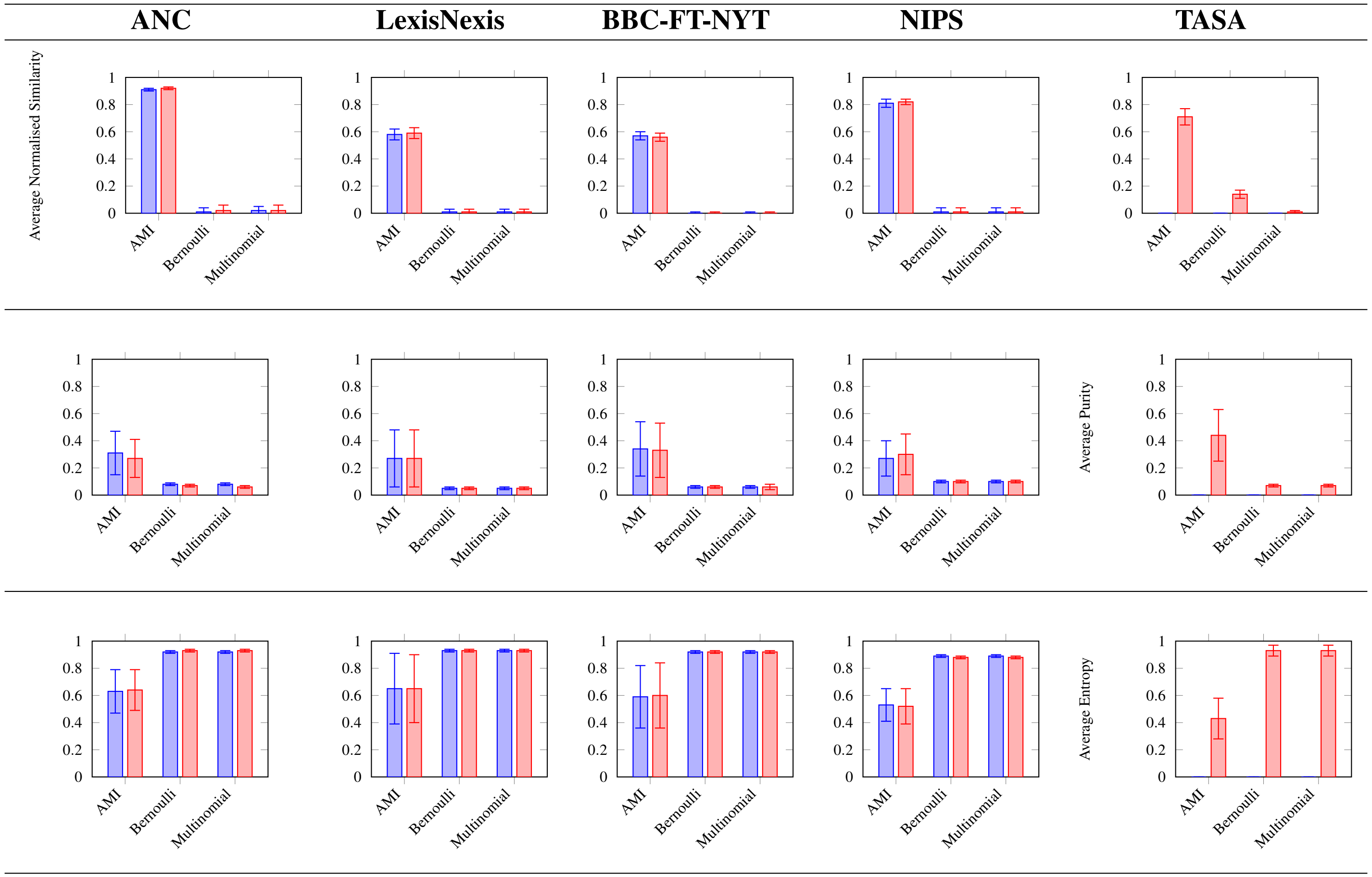


Figure A.1: Average similarity (top), purity (Eq. 4.8; middle) and entropy (Eq. 4.9; bottom) of clusters built with COLE models with each method (AMI, multinomial language models and Bernoulli language models) for the full-corpus (in blue) and 30% held-out versions (in red) and the TASA, ANC, NIPS, BBC-FT-NYT and LexisNexis collections. Errors bars are 1 SD of the mean.

A.1.2 Distributional Semantic Models

LSA

LSA generates reduced dimensionality semantic spaces (100, 300, 400 and 500) from log-entropy normalised document-word co-occurrence matrices. LSA is designed to work with cosine similarity, but its vectors are compatible with correlation, Euclidean and Spearman functions. When analysing LSA spaces, lower dimensional representations effectively "force" words into more efficient representations. Intuitively, this leads to a loss of granularity with a gain in latent contextual relations. Lower-dimension spaces are likely to generate more uniform clusters with respect to average distance, purity and entropy. Also, it is expected that higher average similarity in smaller spaces, where a feature-vector has less opportunity to encode distance. Lastly, spaces built with larger collections (higher document-vocabulary ratio) will likely be less pure and more entropic, given the range of observations for a single word.

First consider the average distance among words for LSA shown in table A.1. The scores confirm of our intuitions about dimensional reduction and corpus size – namely that smaller spaces produce more uniform clusters, and that larger corpora (enTenTen and TASA) yield higher average similarity. This holds between collections, where the smallest (BBC-NYT-FT[1]) shows the lowest average similarity. It also holds between full-corpus and held-out versions. Looking at specific collections, note NIPS is consistently lower in each variant. This may be because NIPS is made up of specialist language, as opposed to the other corpora which contain news (BBC-NYT-FT, LexisNexis) or general language (TASA, ANC and enTenTen). The fact that words in the NIPS collections exhibit lower mean similarity points to 1) the use of more specialised, terms and 2) a higher document / vocabulary ratio. Conversely, note that enTenTen is usually the highest, which may be due to the less structured text.

Looking at the average similarity in the clusters shows similarities between the methods (model variant, corpus, similarity function, etc.) but it does not necessarily a measure the overall quality. The cluster purities of the LSA variants (table A.2) give an impression of cluster quality in each configuration. Here note that cosine and correlation functions provide similar scores and that Euclidean and Spearman see increases of a little more than half for every corpus / model configuration. Over the corpora, note that TASA and enTenTen score lower than the other collections in general. This is likely due to the size of the vocabulary and the diversity of topics apparent in those collections. Also note that ANC (in LSA-100) has the highest purity, which may be because of the corpus' relatively small size. Unlike average similarity, purity does not appear to be correlated with the dimensionality of the semantic spaces. Last, the variance across full and held-out versions of each corpus is never more than 1 SD apart for any configuration (not shown in table A.2). This supports the role of average purity as a reliable quality measure for corpora of varying size.

For the average entropy in LSA clusters, note the variance from full to held-out versions are all still within 1 SD of one another. Also, TASA and enTenTen are the highest, which further supports their diversity of vocabulary and topics, especially when compared to the specialist language in the

[1] In terms of documents.

Model	Full Corpus				30% Held-out			
Corpus	Cosine	Correlation	Euclidean	Spearman	Cosine	Correlation	Euclidean	Spearman
LSA-100:								
ANC	0.71	0.71	0.43	0.67	0.71	0.71	0.44	0.56
BBC-NYC-FT	0.75	0.75	0.47	0.41	0.76	0.76	0.49	0.51
NIPS	0.64	0.64	0.48	0.37	0.62	0.62	0.46	0.40
TASA	0.80	0.80	0.61	0.35	0.76	0.76	0.61	0.34
enTenTen	0.90	0.90	0.57	0.78	0.86	0.88	0.57	0.76
LSA-300:								
NIPS	0.44	0.44	0.30	0.31	0.43	0.43	0.30	0.31
TASA	0.65	0.65	0.50	0.20	0.60	0.60	0.50	0.20
enTenTen	0.81	0.81	0.45	0.62	0.78	0.79	0.45	0.57
LSA-400:								
NIPS	0.41	0.40	0.28	0.28	0.39	0.40	0.28	0.28
TASA	0.60	0.60	0.47	0.17	0.56	0.55	0.47	0.17
enTenTen	0.78	0.78	0.42	0.56	0.51	0.51	0.18	0.34
LSA-500:								
NIPS	0.38	0.38	0.27	0.26	0.37	0.37	0.26	0.26
TASA	0.56	0.56	0.45	0.16	0.51	0.51	0.45	0.16
enTenTen	0.76	0.76	0.40	0.52	0.72	0.72	0.40	0.47

Table A.1: Average similarity (inverse distance; see section 4.4.4) of words in clusters built with LSA for each similarity function, for both the full-corpus (in blue) and 30% held-out versions (in red). The standard error for each sample was less than 0.001 ($N \approx$ 95,800). In each case, the similarity ranges from 0 (completely dissimilar) to 1 (completely similar), even for unbounded functions like Euclidean, which are expert-normalised within each space.

Model	Full Corpus				30% Held-out			
Corpus	Cosine	Correlation	Euclidean	Spearman	Cosine	Correlation	Euclidean	Spearman
LSA-100:								
ANC	0.27	0.27	0.30	0.24	0.23	0.24	0.27	0.31
BBC-NYC-FT	0.22	0.22	0.27	0.34	0.25	0.26	0.25	0.36
NIPS	0.23	0.24	023	0.39	0.22	0.22	0.28	0.34
TASA	0.13	0.14	0.23	0.33	0.15	0.15	0.25	0.33
enTenTen	0.15	0.15	0.22	0.17	0.15	0.14	0.27	0.16
LSA-300:								
NIPS	0.26	0.26	0.30	0.37	0.26	0.26	0.31	0.35
TASA	0.15	0.15	0.27	0.33	0.14	0.14	0.30	0.35
enTenTen	0.12	0.11	0.21	0.13	0.12	0.12	0.23	0.14
LSA-400:								
NIPS	0.26	0.26	0.31	0.37	0.25	0.25	0.32	0.35
TASA	0.14	0.14	0.27	0.33	0.14	0.14	0.28	0.35
enTenTen	0.12	0.12	0.23	0.14	0.19	0.19	0.25	0.34
LSA-500:								
NIPS	0.27	0.27	0.32	0.37	0.26	0.26	0.31	0.35
TASA	0.12	0.12	0.29	0.32	0.13	0.13	0.32	0.34
enTenTen	0.11	0.11	0.22	0.13	0.12	0.11	0.23	0.14

Table A.2: Average purity (Eq. 4.8) of clusters built with LSA with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red). Note Cosine similarity is the function for the reference implementation of LSA.

NIPS corpus. Again, cosine and correlation scores are similar, whereas Euclidean and Spearman scores are less consistent across corpora and configuration of dimensionality.

Model	**Full Corpus**				**30% Held-out**			
Corpus	**Cosine**	**Correlation**	**Euclidean**	**Spearman**	**Cosine**	**Correlation**	**Euclidean**	**Spearman**
LSA-100:								
ANC	0.65	0.65	0.65	0.67	0.70	0.69	0.67	0.62
BBC-NYC-FT	0.71	0.70	0.72	0.46	0.70	0.70	0.70	0.59
NIPS	0.66	0.66	0.55	0.49	0.65	0.65	0.54	0.49
TASA	0.84	0.84	0.75	0.52	0.82	0.82	0.72	0.51
enTenTen	0.82	0.82	0.76	0.79	0.82	0.82	0.70	0.79
LSA-300:								
NIPS	0.60	0.60	0.49	0.47	0.60	0.59	0.48	0.48
TASA	0.84	0.84	0.69	0.50	0.83	0.83	0.66	0.49
enTenTen	0.85	0.85	0.77	0.83	0.85	0.85	0.74	0.81
LSA-400:								
NIPS	0.60	0.60	0.49	0.46	0.60	0.60	0.48	0.48
TASA	0.84	0.84	0.67	0.51	0.83	0.83	0.65	0.50
enTenTen	0.82	0.82	0.76	0.79	0.71	0.70	0.68	0.49
LSA-500:								
NIPS	0.60	0.60	0.47	0.46	0.60	0.60	0.47	0.47
TASA	0.85	0.85	0.65	0.51	0.84	0.84	0.62	0.50
enTenTen	0.85	0.85	0.75	0.82	0.85	0.85	0.73	0.81

Table A.3: Average entropy (Eq. 4.9) of clusters built with LSA with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red). Note Cosine similarity is the function for the reference implementation of LSA.

COALS

The next DSM reviewed is COALS, with five variants: 800, 14k, SVD-100, SVD-200 and SVD-400. The first two, COALS-800 and COALS-14k, use the most frequent 800 and 14,000 words to build the semantic space. The SVD variants use all observed words and reduce the resulting space to the configured dimensionality using singular value decomposition. In the COALS models, correlation is designed to be used for similarity measurements, though cosine, Euclidean and Spearman functions can also be used. Because average similarity is less indicative of a cluster sets' quality than relative purity and entropy, in the following models, only purity and entropy are reported.

Figures A.2 and A.3 show the average relative purity and entropy for clusters built with the COALS-SVD variants for each compatible corpus / similarity function configuration. Looking at the purities, cosine and correlation again provide similar results and both variances are similar. The Spearman scores are also relatively similar to the cosine and correlation. On the other hand, the scores using Euclidean distance (which is unbounded) have generally higher purity and greater variance. Recall that purity is a measure of uniformity among a cluster's contributing types, which implies Euclidean distance generally chose more uniform members for a given nucleus, whereas the correlation function tended to choose more varied constituents. Looking across SVD variants (100, 200 and 800) there does not appear to be a general trend. This supports SVD's ability to preserve the cluster make-up across dimensionalities. Looking at each corpus, note the higher scores for the NIPS and BBC-FT-NYT collection compared to enTenTen and TASA. This supports the intuition that the topically diverse texts in the TASA and enTenTen collections beget slightly lower purities. Last of all, note that all the full corpora are within the 1 SD of the scores for their held-out counter-parts.

The entropies for the COALS-SVD variants (figure A.3) point to similar findings as the purities with one notable exception: the average entropies exhibit less variance. It remains that entropy scores under the Euclidean function show more variation, especially compared to correlation. Also note that NIPS and BBS-FT-NYT have slightly lower entropy than other collections. Again, the variation between SVD variants and full / held-out versions is negligible in most cases.

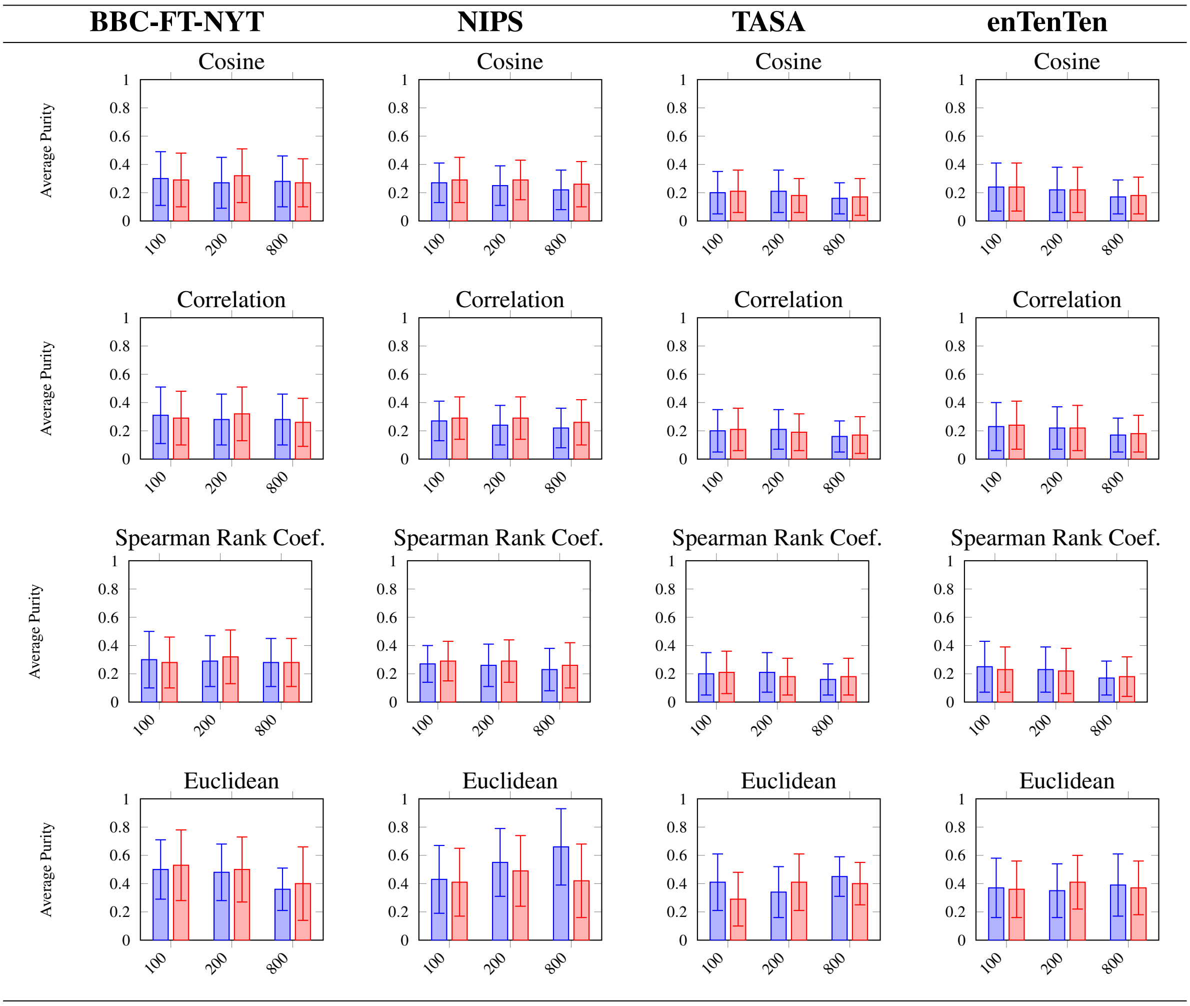


Figure A.2: Average purity (Eq. 4.8) of clusters built with the COALS-SVD variants (100, 200 and 800) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, ANC, NIPS and BBC-FT-NYT and enTenTen collections. Errors bars are 1 SD of the mean.

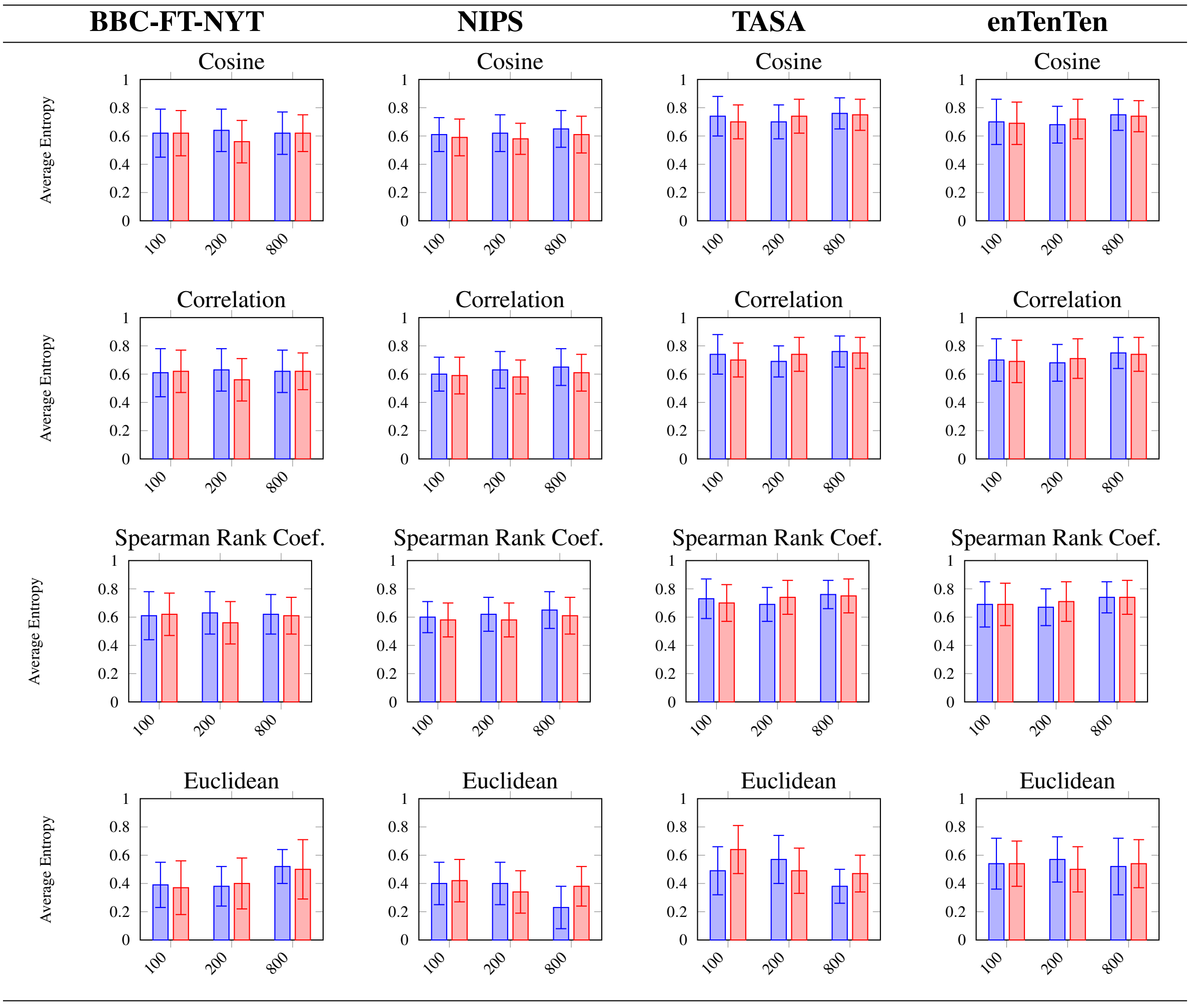


Figure A.3: Average entropy (Eq. 4.9) of clusters built with the COALS-SVD variants (100, 200 and 800) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, ANC, NIPS and BBC-FT-NYT and enTenTen collections. Errors bars are 1 SD of the mean.

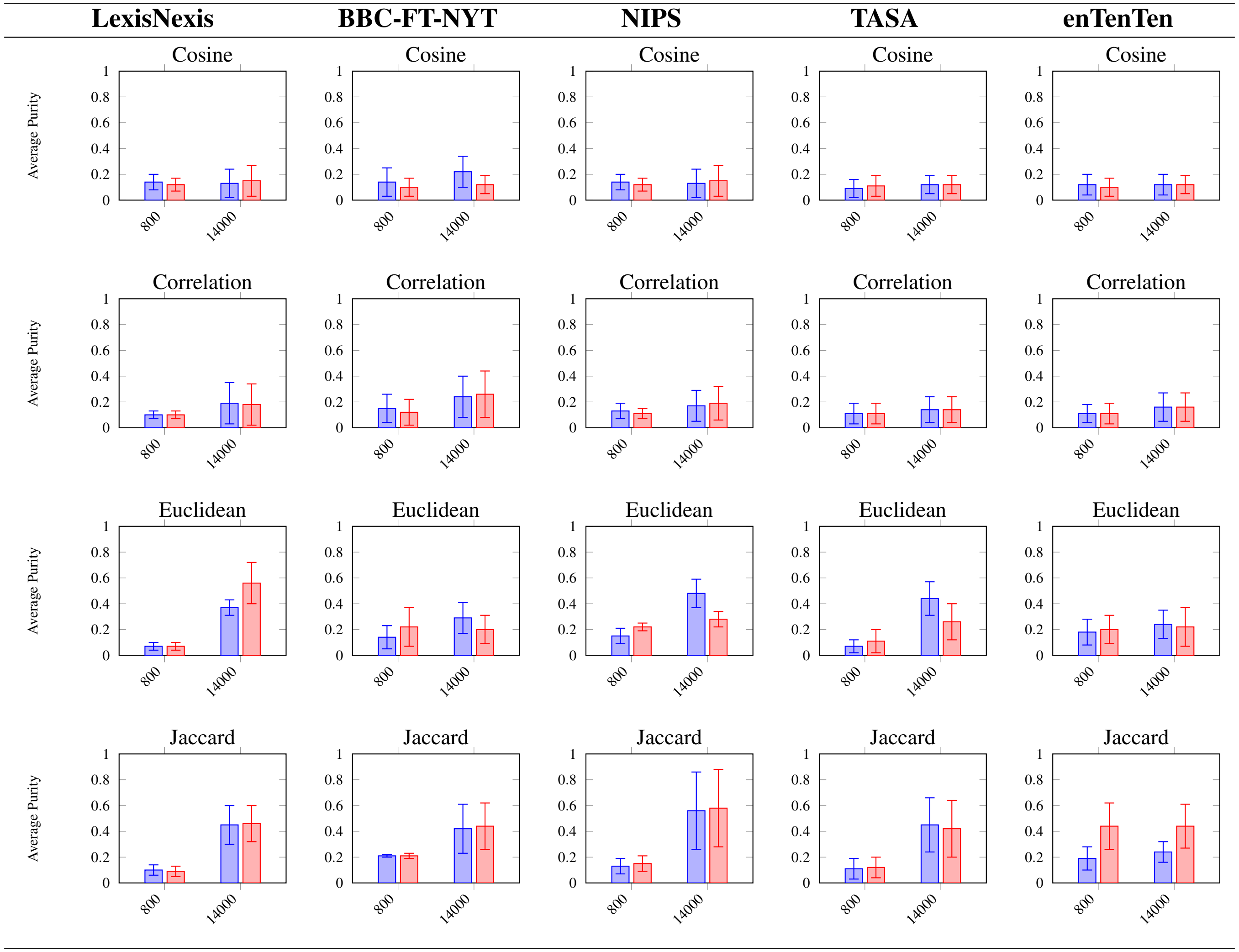


Figure A.4: Average purity (Eq. 4.8) of clusters built with the COALS variants (800 and 14000) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, NIPS, BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

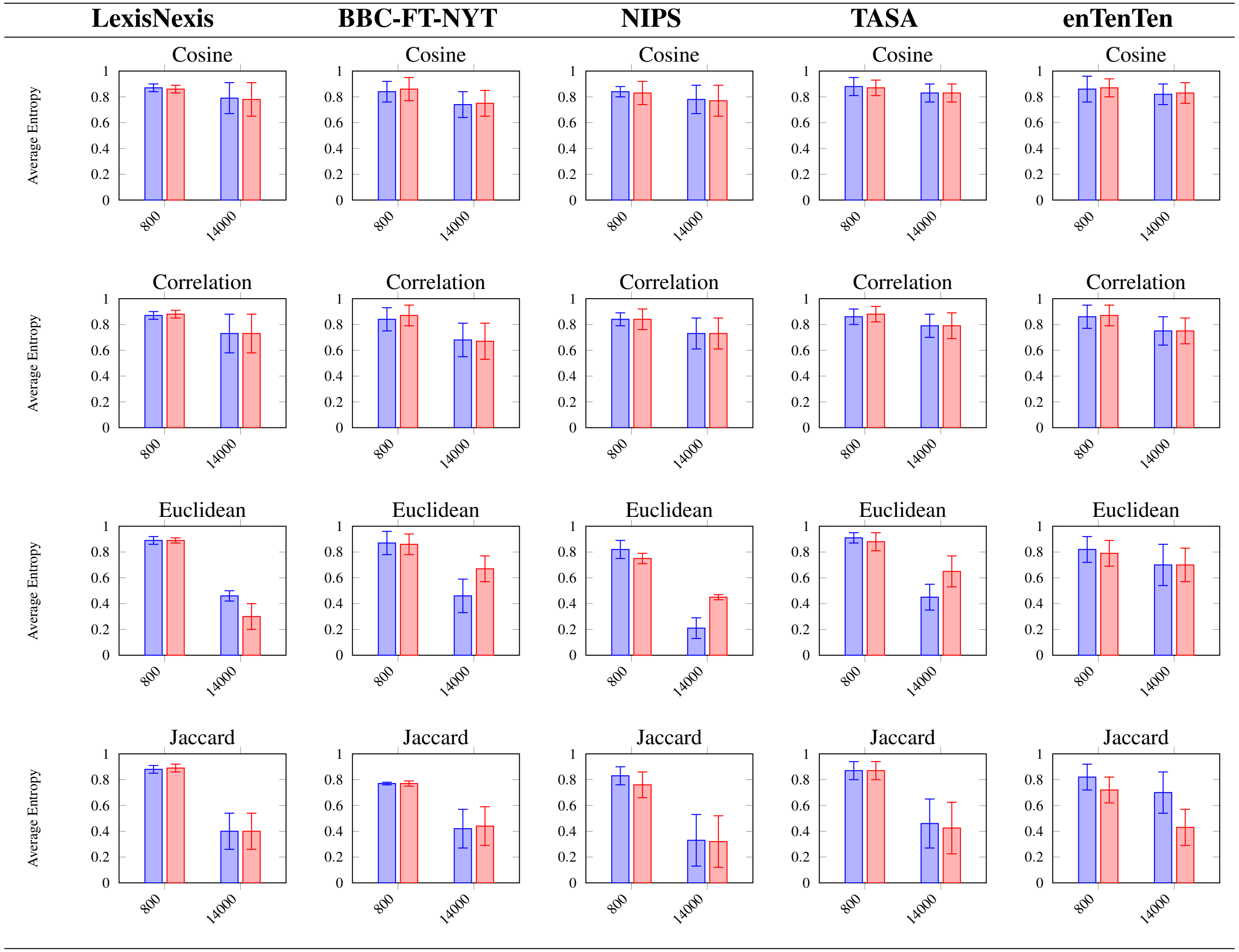


Figure A.5: Average entropy (Eq. 4.9) of clusters built with the COALS variants (800 and 14000) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, NIPS, BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

The second group of COALS models are the variants that do not reduce the dimensionality of the semantic space, but instead only represent the 800 or 14,000 most frequent words. The purities for the 800 and 14k variants are shown in figures A.4 and A.5 respectively. Note that in each, the variances with COALS-14k are generally larger than COALS-800. This may be due to increased contribution of less frequent words, which are less likely to be found in consistent contexts. Again note how Euclidean distance yields scores with greater variation across collections, as well as across 800 / 14k variants. This also seems to have an affect on the variance, which is greater in most of the COALS-14k scores than COALS-800. Lastly, note the differences between variants when measured with the Jaccard function. Overall, the average purities here are lower than the clusters built with the COALS-SVD models.

The average purities are generally higher than the COALS-SVD variants. They also tend to be slightly lower for COALS-14k than COALS-800. Again, the differences for Euclidean and Jaccard scores across variants are more pronounced than correlation or cosine functions. TASA is the most similar across variant and is actually not significantly different for cosine or correlation functions. For the correlation scores (the default similarity function for COALS), none of the full collection scores are significantly different than their held-out version – which is not the case with Euclidean or Jaccard functions.

BEAGLE

BEAGLE is designed to use cosine similarity to measure the distance between word-vectors. There is no size restriction on applicable corpora – not for vocabulary, documents or document-vocabulary ratio. Unlike SVD-based models like LSA and some of the COALS variants, BEAGLE does not implement dimensional reduction. Instead, the permutation process is a kind of re-encoding routine that gradually refines a word's contextual information into a representation. Looking at the purities, there is a downward trend as the number of permutations increases (from 128 to 1024). Variance in the scores also appears to have a concomitant decrease with the number of permutations. However, these trends are almost non-existent with clusters built using Euclidean distance, where the variance in purity is greatest.

The entropy scores generally increase with the number of permutations BEAGLE applies, again with the exception of those using Euclidean distance. The BBC-FT-NYT and LexisNexis collections have the highest overall entropy with the cosine function (BEAGLE's default similarity metric). This is perhaps due to their lower document-vocabulary ratios, especially compared to TASA and enTenTen. Alternatively, TASA has the highest entropies, which may be a reflection of its topical diversity.

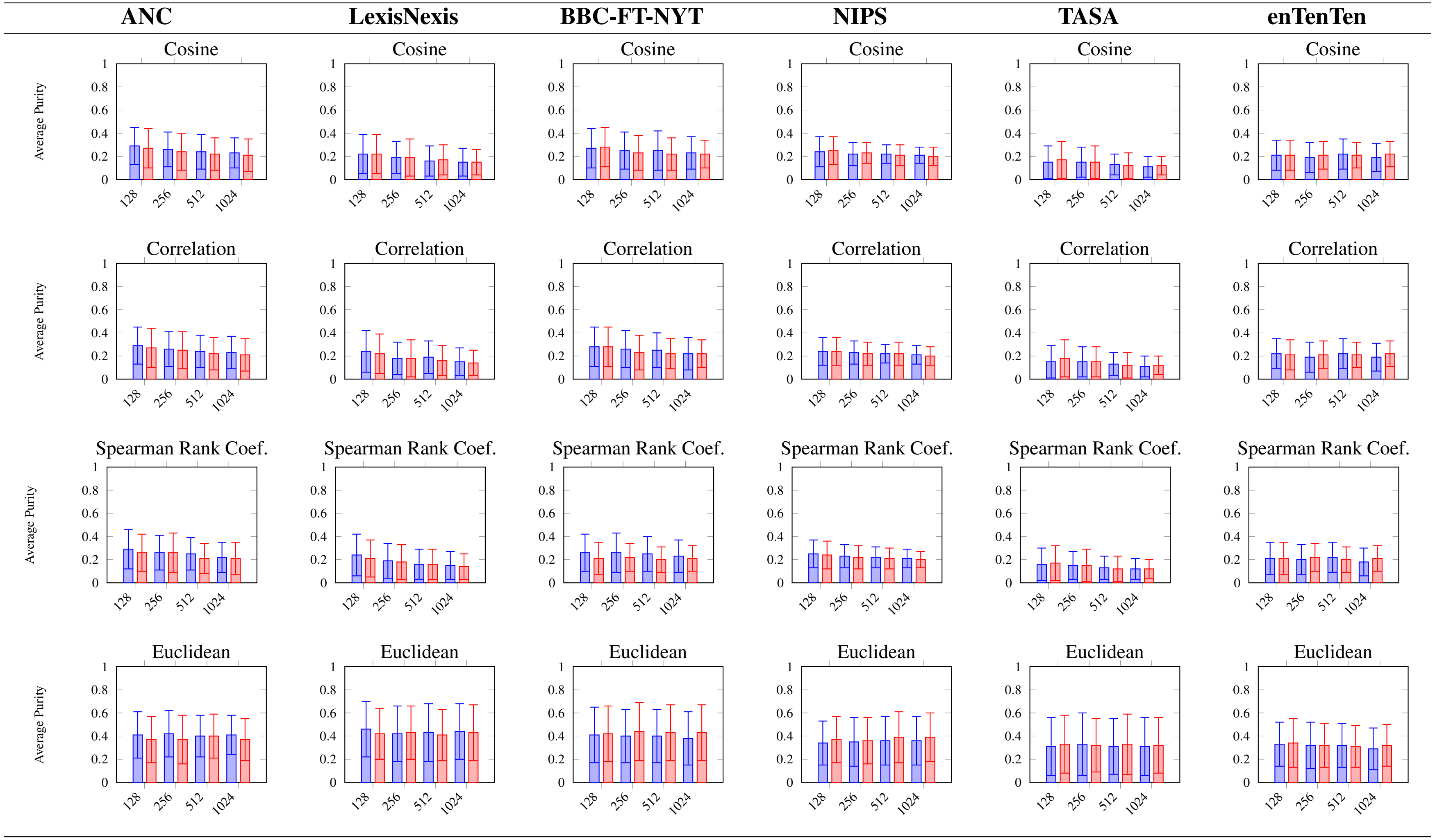


Figure A.6: Average purity (Eq. 4.8) of clusters built with the four BEAGLE variants (128, 256, 512 and 1024) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, ANC, NIPS and BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

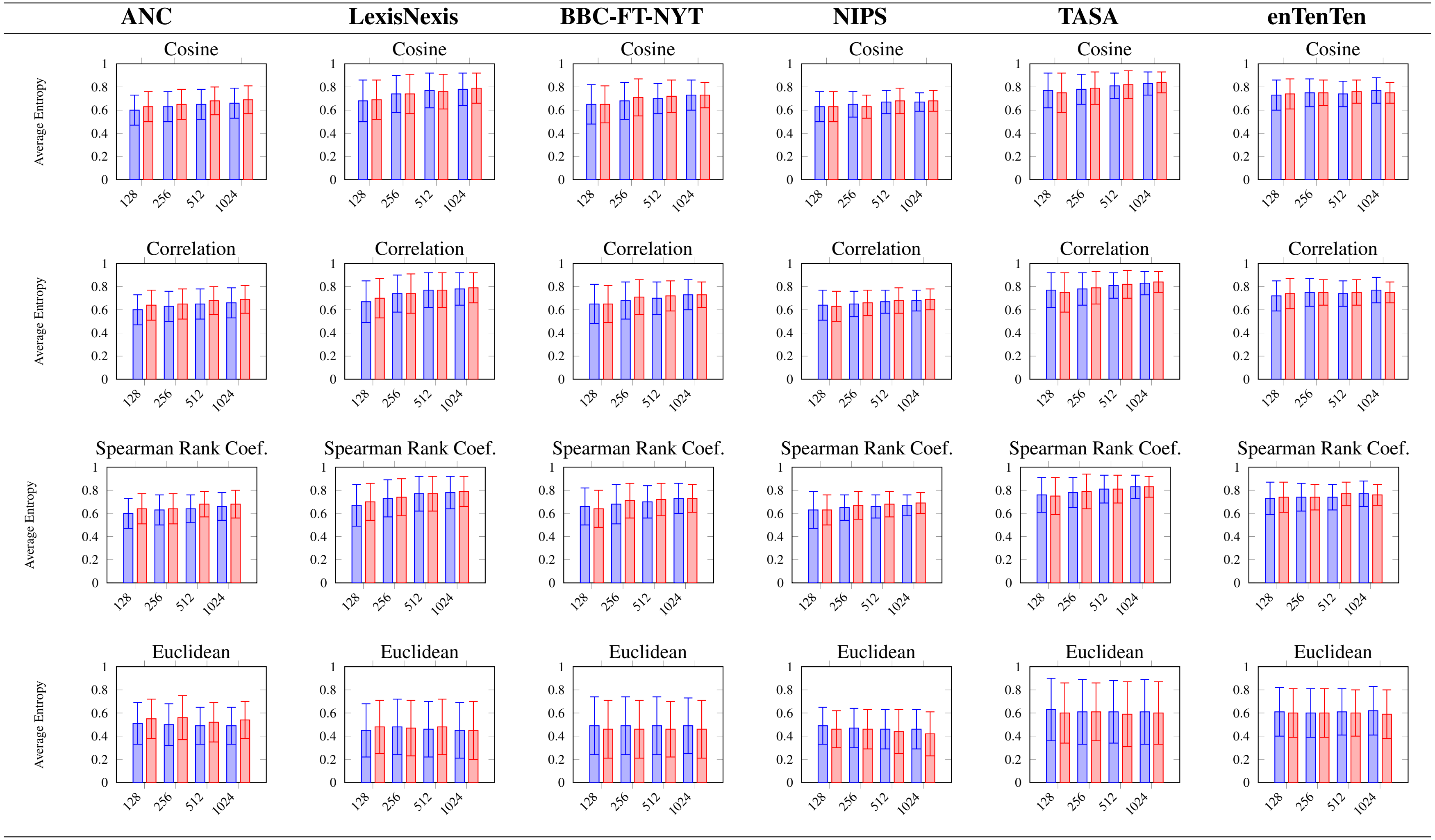


Figure A.7: Average entropy (Eq. 4.9) of clusters built with the four BEAGLE variants (128, 256, 512 and 1024) with each similarity function for both the full-corpus (in blue) and 30% held-out versions (in red) of the TASA, ANC, NIPS, BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

HAL

The clusters built with HAL, which is designed to use Euclidean distance and does not use SVD, are similar to the previous DSMs. There are three variants of HAL in which all words are represented or the top 400 or 1400 are kept. The TASA and enTenTen clusters have slightly lower purities than the other collections, especially under cosine similarity. Euclidean distance, again, exhibits higher variance than the cosine or correlation functions. Model-wise, there are not significant differences between the variants. Moreover, none of the configurations produce significantly different purities between full and held-out versions of the collections. The average scores for entropy in TASA and enTenTen are comparable and cosine produces relatively stable scores compared to Euclidean and Jaccard functions – a finding mirrored by the other DSMs.

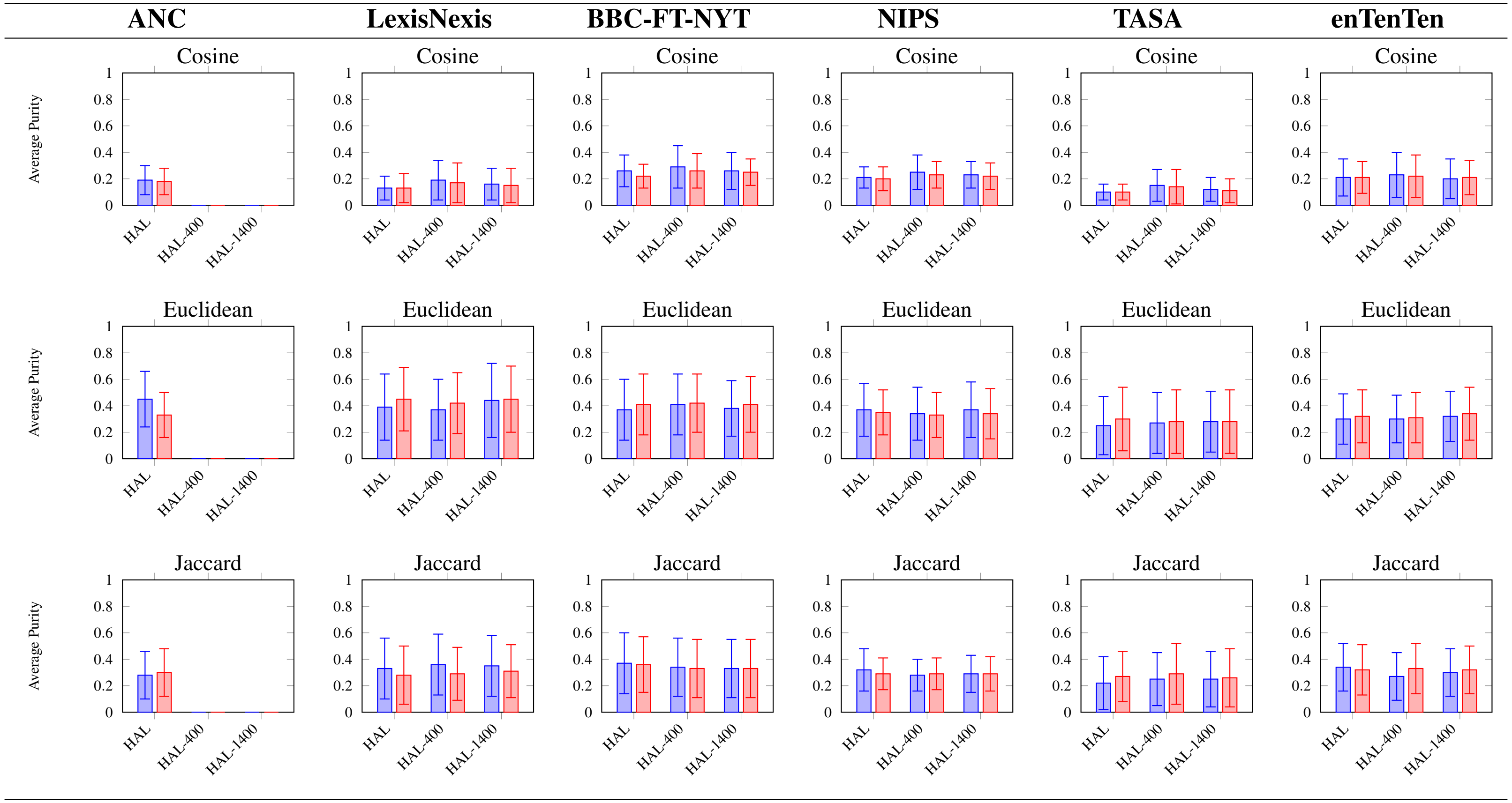


Figure A.8: Average purity (Eq. 4.8) of clusters built with HAL variants (HAL, HAL-400, HAL-1400) with each similarity function for both the full-corpus and 30% held-out versions of the TASA, ANC, NIPS, BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

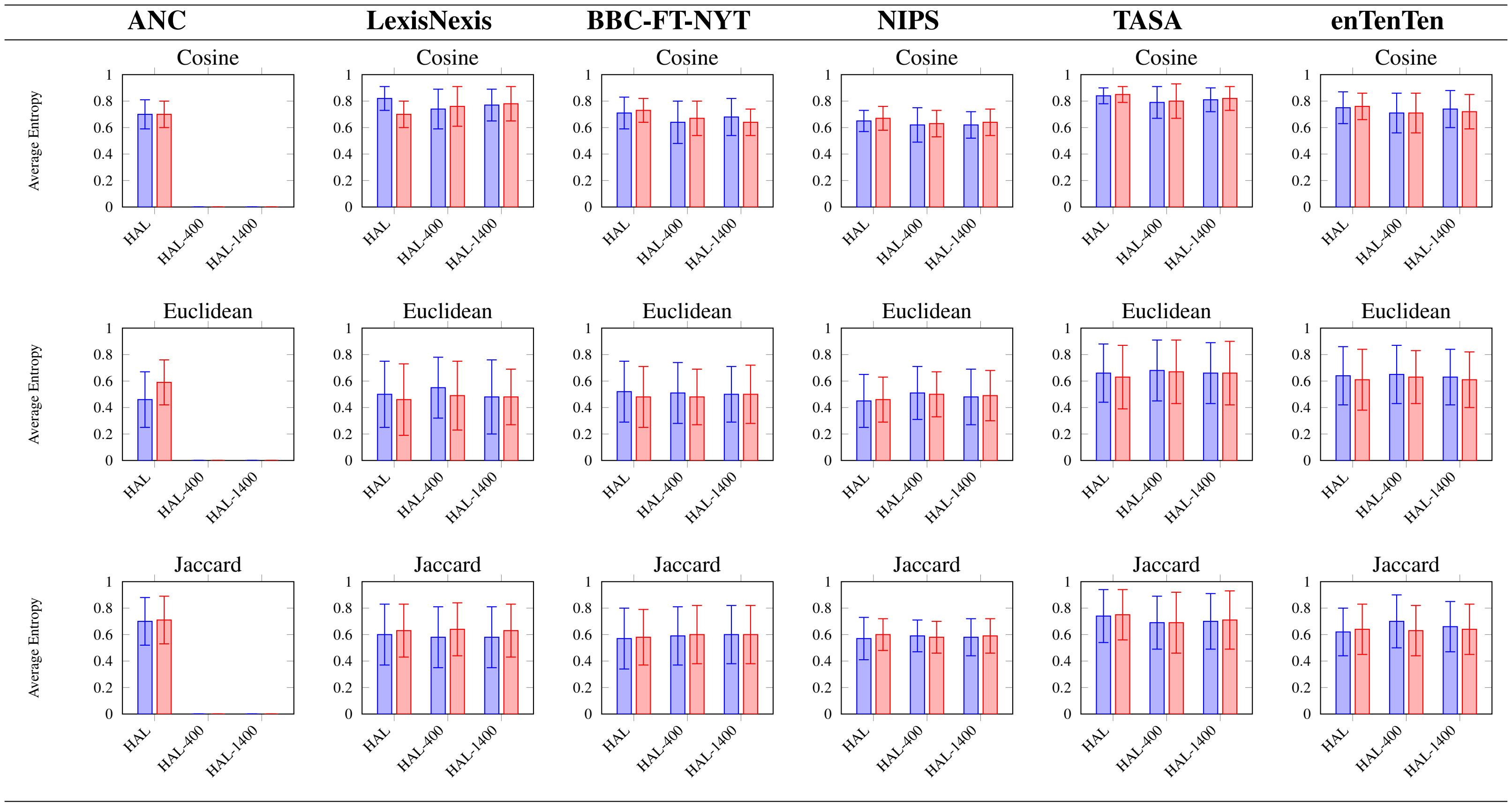


Figure A.9: Average entropy (Eq. 4.9) of clusters built with HAL variants (HAL, HAL-400, HAL-1400) with each similarity function for both the full-corpus and 30% held-out versions of the TASA, ANC, NIPS, BBC-FT-NYT, LexisNexis and enTenTen collections. Errors bars are 1 SD of the mean.

Appendix B

MetID Technical Overview

MetID was designed, developed and tested with a constant integration model. The system consists of a series of independent scripts for interacting with a transactional database. Input and output are typically semi-structured, text-files like CSVs. While the specification was modular, the design deviated from this to accommodate agility in the testing phase, as well as non-threaded parallelisation and syncronisation of tasks (ie. more than one instance of the same script working on different data). The result was a database, containing most of the required elements for analysing instances of figurative language. Interaction with the system usually amounts to interacting with the database is some manifold way. This appendix first outlines the design principles which lead to the implementation of the method specified in chapter 3. Second, I review the system as implemented, discussing various technicalities as they relate to methods and algorithms discussed in chapter 4. Third, the database design is discussed in detail, as it affords the system with most of its extensibility and evaluative potential. The concluding section presents a short computation and time analysis of the two central algorithms – word clustering and metaphor identification.

MetID is available online[1], as is the software for the COLE-based models discussed in section 4.4.5[2]. The former is a Java application with an interface for typical users. MetID, on the other hand, is a series of scripts, with some embedded documentation and pointers to their respective use. However, the MetID database (MySQL 6) embodies a lot of the system's functionality. As such, MetID is not in a state to be run by typical users. The following sections explain various parts of the system, including the design of the database.

B.1 Design & Development Principals

From the start of the research phase for this project, the design principles were dictated by two on-going necessities: pace and agility. To address the research goals, three initial decisions were made with regard to the process of implementing MetID: 1) use of central persistent data-store, 2) the use of scripts, rather than an application and 3) the pipe-lining of various development and use-cases. While each of these decisions potentially detracts from the system's maintainability

[1] http://www.scss.tcd.ie/˜gerowa/metid

[2] http://www.scss.tcd.ie/˜gerowa/co-occurrence

and interoperability, they enabled agility in exploring aspects of the project to better address the research goals.

The central data-store is a relational SQL database, implemented in MySQL 6.0 on Debian GNU/Linux. This allowed the development of scripts and programs in multiple languages (Ruby, Perl and Java) with common access to authoritative data. The database contains metaphor examples (from *Metalude*), the text collections and the word clusters (the result of the corpus-based semantic models). The database is available online[3] and is SQL-compliant with any foreign-key capable engine[4].

There was not a uniform approach to developing scripts to interact with data. However, a general paradigm of development employed Unix-like principles with heavy use of file-structure and the data-store, minimising in-memory operations. While this detracted from the speed and agility of some development tasks (such as cleaning and parsing corpora) it allowed longer-term flexibility in developing various experimental tasks, such as those presented in chapter 5. One example was the extraction of selectional preferences using the BNC and TASA text collections. This involved determining representative grammatical relations from Sketch Engine [120][5] with its Python API, piping that output with POSIX I/O redirection through a series of Perl scripts to parse and clean the output, and finally to a Ruby script which calculated the selectional metrics in relation to those already present in the database. This process typified the development and reuse of scripts, like the Perl scripts implementing the text cleaning pipeline, which was also used to clean and parse the corpora and examples from *Metalude*.

B.2 Modules

Programmatically, MetID is organised into sub-systems correlating to the those in figure 4.2. These sub-systems consist of a text stream, semantic stream, database and analysis. The implementation presented in chapter 4 can be viewed as a procedural organisation of how the various tasks are related, as opposed to the technical specifications presented.

B.2.1 Scaffolding & Text Stream

Some prerequisite tasks had to be addressed before MetID's modular development began. These included scraping various resources, like *Metalude*, the BNC *n*-gram frequencies and predications, as well as creating custom corpora. The web-scrapers were written in Ruby and used of HTTP requests to pull content directly from the web after which it was cleaned and stored. In constructing the custom corpora (BBC-FT-NYT and LexisNexis collections), Perl scripts were used to clean and parse the raw text from the web.

[3] http://www.scss.tcd.ie/~gerowa/metid

[4] InnoDB was used by default.

[5] http://www.sketchengine.co.uk/; 11 March, 2013

In the absence of an ideal text pre-processing tool, a series of independent Perl scripts were developed. These scripts can be piped together to perform various cleaning tasks. For those tasks listed in section 4.3, and for cleaning other resources like *Metalude*, the same series of scripts was used. Table B.1 shows the chain of Perl scripts used to clean text prior to parsing and storage. In addition to the Perl standard functions, a number of Unix tools were used such as `sed`, `grep`, `tr`, and `uniq`, making these scripts dependent on a Unix-like environment.

Step	Script	Role
1	delete_duplicate_docs.pl	Removes duplicate text files.
2	delete_non_english_docs.pl	Removes documents which contain less than 80% valid English words.
3	chunk_sentences.pl	Separates sentences, leaving one per line.
4	expand_contractions.pl	Expands contractions to separate words.
5	lexicalise_patterns.pl	Patterns like URLs are converted to tokens like <URL>.
6	remove_weird_punctuation.pl	Removes typographic punctuation and special characters.
7	separate_punctuation.pl	Wraps all punctuation in spaces.
8	split_invalid_hyphenations.pl	Those hyphenated words not in a dictionary are split into separate words.
9	lexicalise_symbols.pl	Symbols like $ are replaced with tokens like <DOLLAR>.
10	embed_postags.pl	Appends every word with its POS tag (eg. 'risk|VB').
11	remove_long_words.pl	Removes long words.
12	remove_stopwords.pl	Removes common closed-class words.
13	remove_nonwords.pl	Removes any word not found in an English dictionary.
14	convert_to_uppercase.pl	Converts all characters to uppercase.

Table B.1: Scripts used in the text pre-processing sub-system. Each script is technically optional, however, the order in which they operate is fixed. Some steps rely on previous steps, such as separating punctuation before removing non-words. Steps 10 and 11 are particularly optional and were not used in preparing corpora for use in the lexical models. Step 3 uses the OpenNLP Toolkit and step 10 uses TreeTagger.

The text stream takes the output from the pre-processing routines, to persist useful data in the database. This module centered around the use-case where a user wants to prepare a corpus for use with a corpus-based model such as LSA. The main script takes a cleaned corpus (a directory of pre-processed plain-text files, representing documents) and inserts them into the database. During the insertion, the texts are parsed with the Stanford Parser [33, 155][6]. Within-sentence phrases were also separated by traversing the resulting parse-trees and nouns (single- and multi-word) are also stored.

This module includes some maintenance routines. The first of which was for stemming all words in the database. Though stemming a single word is comparable in processing time to a database lookup, various implementations differ across languages and libraries. To assure that stemming was consistent when using different programming platforms, a single Java program was written to extract, stem and save all unique word-stem pairs in the database. This table later replaces the actual stemmer in subsequent routines. Another set of scripts provided some one-off reporting tools, which informed a summary comparison of the text collections. By writing scripts to report various aspects like size, vocabulary, average document length, the database, later analysis was simplified.

[6] http://nlp.stanford.edu/software/lex-parser.shtml; 3 March, 2013.

Output 1: Candidate Assertions

The output of the text stream, described in section B.4, is a set of candidate assertions. Technically these are relational triples of the form $\langle w_1, w_2, gramel \rangle$. However, *gramel* may be null, though that will exclude the application of some of the grammatical heuristics mentioned above, like predication or selectional preference violation. This output is unordered, but preserves a trace to the text collection, document, sentence and optionally, the embedded phrase in which it was found. Each layer of this trace, which is stored in the database, can be used in the other modules of the system: the semantic stream and metaphor identification. These candidates are then passed to the semantic stream where each item will undergo the cluster analysis.

B.2.2 Semantic Stream

The semantic stream is responsible for building a model with which to associate words, and to use the model to build clusters around terms from *Metalude*. There are two types of corpus-based models – DSMs and COLE models. The corpus-based models are described in section 4.4.4. This section describes how each class of model was implemented to build and persist word-clusters.

The DSMs were implemented using the S-Space package[7] (see section 4.4.4), with the exception of LSA, which was implemented in Ruby. S-Space provides binary Java programs which take a corpus as input and produce a semantic space as a MATLAB sparse matrix, compatible with the same package's *Semantic Space Explorer*. The binaries for HAL, COALS and BEAGLE were used to generate the semantic spaces for each combination of model × similarity function × corpus combination, which resulted in 172 spaces in total[8].

The COLE models were implemented in a Java program based on one by Dr. Hua Huo described and tested in [105]. There were two modifications to the original application. First, it was made headless – that is the GUI was converted to non-interactive CLI, so that it could be called by other scripts. The second change was the implementation of threading. Given a single estimation in the TASA corpus takes about 2 seconds and the search must exhaust $|V \times R|$, for the corpus' vocabulary V and the *Metalude* terms R, which for TASA was a total of 17,657,876 pairs. Java's variable thread-pool system was used to create new threads at runtime as they become available given system load. Ad-hoc tests showed a speed increase of approximately an order of magnitude. The Java programs for COLE modeling are available online[9].

The WordNet model used Lin similarity to get distances between term-pairs [144, 167]. WordNet 3.1 was used as was an implementation of Lin similarity in the Perl modules Lingua::WordNet and WordNet::Similarity [173]. Because WordNet similarity calls are relatively inexpensive, an exhaustive set of a clusters was not built (or saved) for this model. This results in slightly slower execution in this configuration. Given that clusters were not built for the WordNet model, and that WordNet is not corpus-based, the cluster quality metrics, purity and entropy, are not applicable in this configuration.

[7] https://code.google.com/p/airhead-research/; 9 August, 2013.

[8] Available at http://www.scss.tcd.ie/˜gerowa/metid/.

[9] http://www.scss.tcd.ie/˜gerowa/co-occurrence

S-Space

S-Space is an open-source software package which provides reference implementations for a number of distributional semantic models [115]. As of this writing S-Space includes implementations of document-based models (LSA and Salton's VSM), co-occurrence models (HAL and COALS) and approximation models (random indexing, BEAGLE and ISA). It also provides libraries for common tasks like vector comparison and matrix reduction. S-Space was used for the HAL, COALS and BEAGLE models. LSA was re-implemented in Ruby, which proved faster, but S-Space was used to interact with the sparse-matrices of the other DSMs.

Output 2: Word Clusters

After a model is constructed with a corpus, the result is a semantic space is a binary sparse matrix file. The space is then loaded by S-Space's *Semantic Space Explorer* to get the nearest neighbours for each of seed word. The algorithm for creating the models and clusters is described in section 4.4.4. Each cluster consisted of the 200 nearest neighbors to the nucleus, which is a term in *Metalude*. For each cluster, a number of similarity functions were used: cosine, Euclidean, Pearson correlation, Jaccard coefficient and Spearman's rank coefficient. The same method was used for the COLE models, except that S-Space was not used, and the similarity function is dictated by the COLE method. Here, a Ruby script was used to analyse the output of the COLE Java program, to find the top 200 most likely co-occurring terms for each seed in *Metalude*. Clusters from the DSMs and COLE models were stored in the database. The schema for this storage is given in B.1 (rightward branch of the Collections table). The functional aspects of cluster verification are described in 5.3. This analysis was done entirely with the database, without respect to the models or the resulting spaces. The evaluation of the clusters used the purity and entropy calculations, which were derived from the clustered words' relative frequency in the configured corpus.

B.2.3 Core Module: Analysis / Metaphor Identification

The analysis module is responsible for taking relational triples or sentences from the text stream, and using the word-clusters from semantic stream to derive a ranked list of candidate metaphors. The algorithm is describe more formally later in this appendix, but the intuition behind it is that if MetID observes two related words, which respectively cluster around a terms from *Metalude*, then it may be an instance of that metaphor. This process is described in section 4.2. The analysis module consists of Ruby script that takes arguments to set parameters for the model, similarity function, corpus and of course, the text to analyse. This module is also where the heuristics are implemented (see section 4.5). Some of these heuristics use information from the parsed sentence (output from the text stream). The cluster quality heuristics also implemented here. This module has options for special use-cases. These include forcing a choice of topic or vehicle term in a sentence. This feature is specific useful in investigating a particular metaphor, for example the use of the word *contagion* in financial texts. Another helpful feature is the ability to group candidate metaphors into Goatly's *Map of Root Analogies* [92] which allows a broad corpus analysis of various patterns of metaphor-use.

Output 3: Potential Metaphors

The final output from MetID is a rank-ordered list of candidate metaphors for the each item analysed from the text stream. The output is a CSV file, which is typically piped from the final script and written iteratively. This output, given its format and size, is usually analysed further to present the results like those in chapter 5. Often, the results were limited to the top-scoring candidate metaphor[10]. In practice, the results were limited to the top 20 best candidates. An example of these final results is given at the end of this appendix.

B.3 Database Design

The MetID database began as a persistence schema for various long-term data like the corpora and *Metalude* examples. For technical considerations of time and efficiency, the database (and its schema) grew to accommodate other data like the word clusters, stems, dependency parses and phrases extracted from text. By using a central data-store, various development and use-cases were able to be made parallel. For example, at a given point, a corpus could be being cleaned and parsed while another was being used to build word clusters, while a third program could be using the database for analysis. Figure B.1 contains the database schema.

With regard to figure B.1, there are two main trees descending from the Collections table: Clusters and Documents. The Documents tree organises the text into collections, containing tables for chunked sentences, phrases, parses and an unused table of extracted noun-terms. The Clusters tree contains the word clusters generated from a given collection. Within the Clusters tree there is a table of methods which corresponds to the lexical model (type, variant and full / held-out distinction) as well as a table of all clustered words. Separate from the Collections table are three resources, only one of which was used. In the top left of figure B.1 is a table containing Shutova's raw results [198][11], the veracity of which she expressed concern[12]. In the upper right is a scraped copy of Lakoff's Master Metaphor List [134]. This resource, while derived from Lakoff's extensive work on metaphor, is less structured and less internally consistent that *Metalude*. Lastly, is a schema for Goatly's *Map of Root Analogies*, which was scraped with permission from the *Metalude* website[13]. The root table, Root_Analogies, references a set of examples, which in turn has a table of dependency parses, MetRelations.

[10] This score appears to follow a type-2 power-law with a Pareto distribution – the farther down the rank-order, the less different the scores.

[11] Thanks to Ekaterina Shutova for making these results available in their raw form.

[12] Personal communication; 29 March, 2012. I, however, am not clear whether she is concerned with her reported results or the reliability of her participants' or perhaps both.

[13] http://www.ln.edu.hk/lle/cwd/project01/web/home.html; 12 February, 2013.

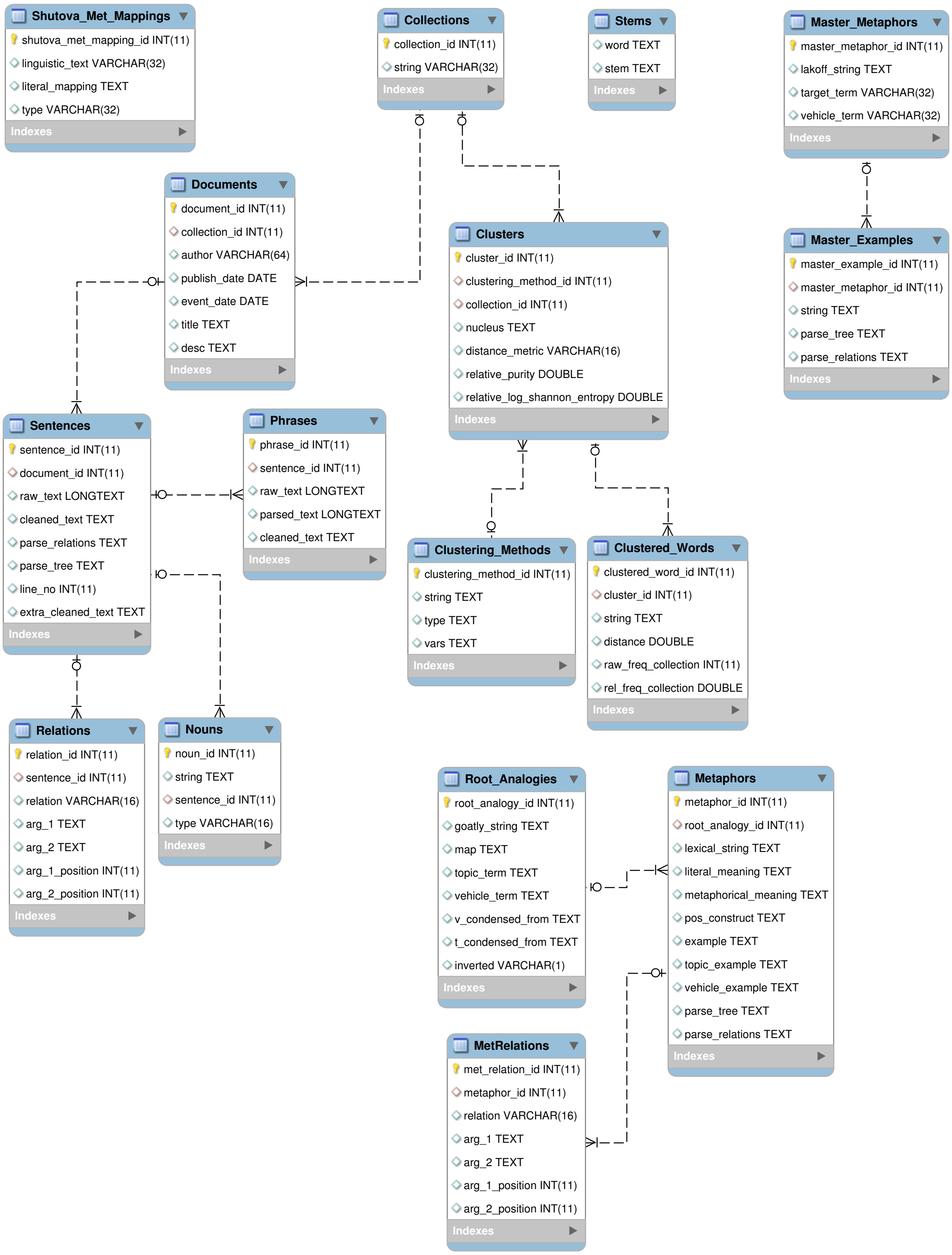


Figure B.1: The MetID database entity / relation schema. At the final stages of development, the database was 18GB on disk.

B.4 Sample Run

MetID can be run in two modes: *rank* or *decide*. *Rank* mode generates an ordered list of candidate metaphors in the way described in section 4.2. The input for this mode is a text file of line-delimited sentences in UTF-7/8 or ASCII encoding. The output, then, is an ASCII-encoded CSV spreadsheet of the results. Each unit extracted from the input will have a list of 20 best candidate metaphors, which can be combined manually or independently analysed. MetID runs at the command line in a Unix-like environment. The output is written to STDOUT, which can be redirected to an output file. STDERR is used for progress information, shown in table B.2, as well as any run-time errors.

```
Loading candidate mapping...
Processing "My heart is a fire that burns with love.":
Parsing and cleaning sentence...
Looking for probable targets and vehicles...
Terms were guessed:  (heart, fire)
Getting additional triples to analyse...
Building similarity calls...
sentence("heart love fire"):  Finding best topic and vehicle terms...
sentence("heart love fire"):  Applying bonus and penalty heuristics...
pred(heart,fire):  Finding best topic and vehicle terms...
pred(heart,fire):  Applying bonus and penalty heuristics...
nsubj(fire,burn):  Finding best topic and vehicle terms...
nsubj(fire,burn):  Applying bonus and penalty heuristics...
All done!
```

Table B.2: Sample progress output from MetID.

Rank-mode

If MetID is run in *decide* mode, it will try to choose the more figurative of a pair of sentences. These sentences are input as a text file, with each pair on a line, separated by a bar |. Decide mode works similarly to the rank mode, except that for each sentence the score is calculated as the mean of maxima over all units of analysis. That is, the result is the sentence with the higher average score across all of its extracted units. The output in *decide* mode includes the scores for each sentence and their difference (figure B.2).

```
Crime is a problem.|Crime is a disease.
Dancers are people.|Dancers are butterflies.
The bike moved along the trail.|The bike tiptoed along the trail.
```

⇓

Sentence$_1$	**Score$_1$**	**Sentence$_2$**	**Score$_2$**	**Score$_2$-Score$_1$**
Crime is a problem	0.92	Crime is a disease	0.98	0.06
Dancers are people	0.59	Dancers are butterflies	0.87	0.28
The bike tiptoed along the trail	0.84	bike moved along the trail	0.90	0.06

Figure B.2: Example input and output data for MetID when run in decide mode.

Decide Mode

In addition to the input files, options are passed at execution time to select the model-type (WordNet, DSM or COLE), the specific model (LSA-400, COALS-14000, etc.), the distance function (cosine, Euclidean, Lin similarity, etc.) and corpus with which the seed-clusters were built. The output is a plain-text CSV file with fields for the input, unit of analysis, identified topic term, distance from topic-term to the candidate topic, identified vehicle term, its distance to the candidate vehicle, the location of the candidate metaphor on the *map of root analogies*, the topic cluster's purity and entropy, the vehicle cluster's purity and entropy, a the list of heuristics applied and the score. Table B.3 lists the heuristics referenced in the system's output. Table B.4 contains output from the sentence "My heart is on fire with love." in rank mode using the WordNet model. Because WordNet does not generate clusters, the for purity and entropy are omitted. Tables B.5 and B.6 show the results for the same sentence run with the LSA-500 and BEAGLE-512 models respectively.

#	Heuristic	Type	Description
1	Non-word	Penalty	The identified topic or vehicle, are not valid words.
2a	WN Synonyms	Bonus	The identified topic is a synonym of the candidate topic.
2b	WN Synonyms	Bonus	The identified vehicle is a synonym of the candidate vehicle.
3a	Marker	Bonus	The sentence contains a strong marker.
3b	Cue	Bonus	The sentence contains a marker.
4	Unpaired Metaphor	Large Penalty	Could not find a metaphor with a pairing given by *Metalude*.
5	Predication	Large Bonus*	If the unit is a predication, and the identified vehicle predicates the topic.
6	Selectional Violation	Large Bonus**	If the identified topic and vehicle are in a relationship which violates the selectional association of the root word.
7	Hypernym	Penalty	If the identified vehicle and topic are nouns, and the vehicle is a hypernym of the the topic.

Table B.3: The heuristics described in section 4.5 as referenced in the example output in tables B.4, B.5 and B.6.

*The predication bonus is scaled by .25, .5, .75 or 1 depending on whether the observed predication strength is stronger than the median, mean or 1 SD + mean of all other predications with the same root.

**The selectional violation bonus is scaled similarly to the predication bonus, but with added respect to the grammatical relationship.

Unit	Candidate Metaphor	Topic	Dist$_t$	Vehicle	Dist$_v$	RA Sector	Heuristics	Score
sentence	AFFECTION=WARMTH	heart	1.0	love	0.80	2B	2a	0.97
sentence	PASSION=HEAT	love	1.0	fire	0.71	2B	2a	0.96
sentence	LOVE=HEAT	love	1.0	fire	0.71	2B	2a	0.96
sentence	KNOWN=OPEN	love	1.0	heart	0.59	3B B1	2a	0.94
sentence	KNOW=SEE	love	1.0	fire	0.50	3B	2a	0.93
sentence	EMOTION=IMPRESSION	love	0.82	heart	0.92	2B B3		0.93
sentence	AFFECTION=WEALTH	heart	1.0	love	0.36	2A	2a	0.92
sentence	IMPROVEMENT=RAISE	love	0.36	fire	1.0	1C D1	2b	0.92
sentence	AFFECTION=MONEY	heart	1.0	fire	0.32	2A	2a	0.91
sentence	SUBSTANCE=HUMAN	heart	1.0	love	0.21	6B	2a	0.90
sentence	HUMAN=MEAT	love	0.21	heart	1.0	5A	2b	0.90
sentence	ACTIVITY=PLACE	fire	0.84	heart	0.70	4D		0.88
sentence	EMOTION=HEAT	love	0.82	fire	0.71	2B		0.88
sentence	INTEREST=PROXIMITY	fire	0.83	heart	0.66	3D		0.87
sentence	EMOTION=LIGHT	love	0.82	heart	0.66	2B		0.87
sentence	EFFECT=IMPRESSION	fire	0.55	heart	0.92	2B		0.87
sentence	QUANTITY=WATER	heart	0.54	fire	0.93	1A		0.86
sentence	CRITICISM=HEAT	fire	0.87	heart	0.57	2B		0.86
sentence	EXPERIENCE=IMPRESSION	fire	0.52	heart	0.92	2B		0.86
sentence	ACTIVITY=HEAT	fire	0.84	heart	0.57	2B		0.85
pred(heart,fire)	AFFECTION=WARMTH	heart	1.0	fire	0.42	2B	2a 5	0.92
pred(heart,fire)	IMPROVEMENT=RAISE	heart	0.33	fire	1.0	1C D1	2b 5	0.91
pred(heart,fire)	AFFECTION=WEALTH	heart	1.0	fire	0.32	2A	2a 5	0.91
pred(heart,fire)	AFFECTION=MONEY	heart	1.0	fire	0.32	2A	2a 5	0.91
pred(heart,fire)	SUBSTANCE=HUMAN	heart	1.0	fire	0.15	6B	2a 5	0.89
pred(heart,fire)	HUMAN=MEAT	fire	0.15	heart	1.0	5A	2b 5	0.89
pred(heart,fire)	ACTIVITY=PLACE	fire	0.84	heart	0.70	4D	5	0.88
pred(heart,fire)	PASSION=HEAT	fire	0.95	heart	0.57	2B	5	0.88
pred(heart,fire)	EMOTION=IMPRESSION	fire	0.59	heart	0.92	2B B3	5	0.87
pred(heart,fire)	INTEREST=PROXIMITY	fire	0.83	heart	0.66	3D	5	0.87
pred(heart,fire)	EFFECT=IMPRESSION	fire	0.55	heart	0.92	2B	5	0.87
pred(heart,fire)	QUANTITY=WATER	heart	0.54	fire	0.93	1A	5	0.86
pred(heart,fire)	CRITICISM=HEAT	fire	0.87	heart	0.57	2B	5	0.86
pred(heart,fire)	EXPERIENCE=IMPRESSION	fire	0.52	heart	0.92	2B	5	0.86
pred(heart,fire)	ACTIVITY=HEAT	fire	0.84	heart	0.57	2B	5	0.85
pred(heart,fire)	OPINION=CURRENT	heart	0.92	fire	0.46	3A C3	5	0.84
pred(heart,fire)	INTELLIGENCE=LIGHT	fire	0.72	heart	0.66	3B	5	0.84
pred(heart,fire)	FEELING=EATING	heart	0.92	fire	0.46	2A	5	0.84
pred(heart,fire)	ANGER=HEAT	fire	0.79	heart	0.57	2B	5	0.84
pred(heart,fire)	FEELING=CONTROL	heart	0.92	fire	0.44	3B	5	0.84
nsubj(fire,burns)	INTELLIGENCE=LIGHT	fire	0.72	burn	0.98	3B	6	0.91
nsubj(fire,burns)	IMPROVEMENT=RAISE	burn	0.40	fire	1.00	1C D1	6 2b	0.91
nsubj(fire,burns)	EMOTION=LIGHT	fire	0.60	burn	0.98	2B	6	0.87
nsubj(fire,burns)	ACTIVITY=SHOOTING	fire	0.85	burn	0.67	4C	6	0.85
nsubj(fire,burns)	ACTIVITY=SAILING	fire	0.85	burn	0.67	4C	6	0.85
nsubj(fire,burns)	ACTIVITY=PLACE	fire	0.85	burn	0.67	4D	6	0.85
nsubj(fire,burns)	ACTIVITY=SWIMMING	fire	0.85	burn	0.67	4C	6	0.85
nsubj(fire,burns)	HAPPINESS=LIGHT	fire	0.51	burn	0.98	2B	6	0.84
nsubj(fire,burns)	HOPE=LIGHT	fire	0.49	burn	0.98	2B	6	0.84
nsubj(fire,burns)	EXCITEMENT=LIGHT	fire	0.49	burn	0.98	2B	6	0.83
nsubj(fire,burns)	PASSION=HEAT	fire	0.96	burn	0.49	2B	6	0.83
nsubj(fire,burns)	LIGHT=LIQUID	burn	0.98	fire	0.47	5A	6	0.83
nsubj(fire,burns)	EMOTION=COLOUR	fire	0.60	burn	0.80	2B	6	0.81
nsubj(fire,burns)	DISEASE=INVASION	burn	0.63	fire	0.76	5D C5	6	0.81
nsubj(fire,burns)	CRITICISM=HEAT	fire	0.87	burn	0.49	2B	6	0.80
nsubj(fire,burns)	EMOTION=SOUND	fire	0.60	burn	0.76	2B B3	6	0.80
nsubj(fire,burns)	ACTIVITY=HEAT	fire	0.85	burn	0.49	2B	6	0.79
nsubj(fire,burns)	COMMUNICATION=CONTACT	burn	0.80	fire	0.52	3D C3 6	0.79	
nsubj(fire,burns)	REPUTED=LIGHT	fire	0.33	burn	0.98	1B	6	0.79
nsubj(fire,burns)	IMPRESSIVE=LIGHT	fire	0.33	burn	0.98	1B	6	0.79

Table B.4: Output from running MetID in rank mode on the single sentence “My heart is a fire that burns with love.”, using WordNet as the lexical model. Note that three units are analysed: the sentence as a whole, the predication (*heart*, *fire*) and the nominal subject (*fire*, *burn*). RA Sector refers to the *map of root analogies* (figure 3.4). Referenced heuristics are listed in table B.3.

Unit	Candidate Metaphor	Topic	Dist$_t$	Vehicle	Dist$_v$	RA Sector	Purity$_t$	Entropy$_t$	Purity$_v$	Entropy$_t$	Heuristics	Score
sentence	AFFECTION=WARMTH	love	0.41	fire	0.52	2B	0.17	0.81	0.07	0.91		0.72
sentence	EMOTION=COLOUR	heart	0.41	fire	0.30	2B	0.07	0.90	0.08	0.88		0.67
sentence	AWARENESS=PROXIMITY	heart	0.48	love	0.20	3D	0.07	0.90	0.04	0.85		0.66
sentence	SADNESS=DISCOMFORT	heart	0.35	burn	0.30	2B	0.09	0.87	0.08	0.88		0.65
sentence	ARGUING=ATTACKING	love	0.33	heart	0.30	3C	0.18	0.81	0.03	0.92		0.65
sentence	ARGUING=WOUNDING	love	0.33	fired	0.21	3C	0.18	0.81	0.12	0.64		0.64
pred(heart,fire)	EMOTION=COLOUR	heart	0.41	fire	0.30	2B	0.07	0.90	0.08	0.88	5	0.67
pred(heart,fire)	SADNESS=DISCOMFORT	heart	0.35	fire	0.28	2B	0.09	0.87	0.08	0.88	5	0.65
nsubj(fire,burn)	DISCOMFORT=MEAT	burn	0.30	fire	0.66	unpaired	0.08	0.88	0.06	0.92	4 6	0.51
nsubj(fire,burn)	DISCOMFORT=SUBSTANCE	burn	0.30	fire	0.63	unpaired	0.08	0.88	0.07	0.92	4 6	0.51
nsubj(fire,burn)	DISCOMFORT=CROWD	burn	0.30	fire	0.63	unpaired	0.08	0.88	0.08	0.89	4 6	0.51
nsubj(fire,burn)	TYING=MEAT	burn	0.25	fire	0.66	unpaired	0.10	0.84	0.06	0.92	4 6	0.51
nsubj(fire,burn)	DISCOMFORT=SOLUTION	burn	0.30	fire	0.61	unpaired	0.08	0.88	0.07	0.91	4 6	0.51
nsubj(fire,burn)	DISCOMFORT=HAPPENING	burn	0.30	fire	0.60	unpaired	0.08	0.88	0.07	0.92	4 6	0.51
nsubj(fire,burn)	TYING=SUBSTANCE	burn	0.25	fire	0.63	unpaired	0.10	0.84	0.07	0.92	4 6	0.50
nsubj(fire,burn)	DISCOMFORT=COOKING	burn	0.30	fire	0.58	unpaired	0.08	0.88	0.07	0.91	4 6	0.50
nsubj(fire,burn)	UNCLEAR=MEAT	burn	0.21	fire	0.66	unpaired	0.18	0.77	0.06	0.92	4 6	0.50
nsubj(fire,burn)	TYING=CROWD	burn	0.25	fire	0.63	unpaired	0.10	0.84	0.08	0.89	4 6	0.50
nsubj(fire,burn)	RELINQUISH=MEAT	burn	0.16	fire	0.66	unpaired	0.48	0.40	0.06	0.92	4 6	0.50
nsubj(fire,burn)	TYING=SOLUTION	burn	0.25	fire	0.61	unpaired	0.10	0.84	0.07	0.91	4 6	0.50
nsubj(fire,burn)	UNCLEAR=SUBSTANCE	burn	0.21	fire	0.63	unpaired	0.18	0.77	0.07	0.92	4 6	0.50
nsubj(fire,burn)	DISCOMFORT=ANGER	burn	0.30	fire	0.56	unpaired	0.08	0.88	0.07	0.91	4 6	0.50
nsubj(fire,burn)	UNCLEAR=CROWD	burn	0.21	fire	0.63	unpaired	0.18	0.77	0.08	0.89	4 6	0.50
nsubj(fire,burn)	TYING=HAPPENING	burn	0.25	fire	0.60	unpaired	0.10	0.84	0.07	0.92	4 6	0.50
nsubj(fire,burn)	RELINQUISH=SUBSTANCE	burn	0.16	fire	0.63	unpaired	0.48	0.40	0.07	0.92	4 6	0.50
nsubj(fire,burn)	RELINQUISH=CROWD	burn	0.16	fire	0.63	unpaired	0.48	0.40	0.08	0.89	4 6	0.50
nsubj(fire,burn)	TYING=COOKING	burn	0.25	fire	0.58	unpaired	0.10	0.84	0.07	0.91	4 6	0.50
nsubj(fire,burn)	UNCLEAR=SOLUTION	burn	0.21	fire	0.61	unpaired	0.18	0.77	0.07	0.91	4 6	0.50

Table B.5: Example run of MetID on the sentence “My heart is a fire that burns with love”, using the LSA-500 model.

Unit	Candidate Metaphor	Topic	Dist$_t$	Vehicle	Dist$_v$	RA Sector	Purity$_t$	Entropy$_t$	Purity$_v$	Entropy$_t$	Heuristics	Score
sentence	LOVE=HEAT	love	0.62	burn	0.51	2B	0.12	0.83	0.07	0.83	2a	0.88
sentence	GOOD=CLEAN	love	0.66	fire	0.53	1B	0.21	0.75	0.12	0.85		0.78
sentence	UNDERSTAND=HOLD	love	0.58	fire	0.57	3B	0.10	0.87	0.10	0.88		0.77
sentence	FEAR=COLD	love	0.55	fire	0.57	2B	0.10	0.87	0.05	0.89		0.76
sentence	IDEA=PLACE	love	0.54	fire	0.56	3D	0.09	0.89	0.09	0.90		0.76
sentence	TRAFFIC=BLOOD	fire	0.39	heart	0.68	6A	0.10	0.87	0.13	0.73		0.75
sentence	IDEA=CLOTHES	love	0.54	fire	0.53	3A	0.09	0.89	0.11	0.85		0.75
sentence	PLACE=BODY	fire	0.56	heart	0.51	6B	0.09	0.90	0.04	0.89		0.75
sentence	IDEA=SOUND	love	0.54	fire	0.50	2B B3	0.09	0.89	0.11	0.87		0.74
sentence	IDEA=DISEASE	love	0.54	heart	0.49	3B	0.09	0.89	0.07	0.86		0.74
sentence	DISEASE=IDEA	heart	0.49	love	0.54	3B	0.07	0.86	0.09	0.89		0.74
sentence	ANGER=HEAT	love	0.51	burn	0.51	2B	0.11	0.86	0.07	0.83		0.74
sentence	SIGHT=SOUND	love	0.50	fire	0.50	5B	0.10	0.87	0.11	0.87		0.73
sentence	SOUND=TOUCH	fire	0.50	love	0.48	5B	0.11	0.87	0.11	0.85		0.73
sentence	TOUCH=SOUND	love	0.48	fire	0.50	5B	0.11	0.85	0.11	0.87		0.73
sentence	EXCITEMENT=HEAT	love	0.45	burn	0.51	2B	0.10	0.87	0.07	0.83		0.72
sentence	EVIL=DARK	love	0.38	fire	0.55	1B	0.13	0.80	0.11	0.85		0.72
sentence	IDEA=SMELL	love	0.54	fire	0.39	2B B3	0.09	0.89	0.11	0.86		0.72
sentence	HAPPINESS=LIGHT	love	0.46	fire	0.46	2B	0.12	0.87	0.13	0.81		0.72
sentence	EXCITEMENT=LIGHT	love	0.45	fire	0.46	2B	0.10	0.87	0.13	0.81		0.71
pred(heart,fire)	TRAFFIC=BLOOD	fire	0.39	heart	0.68	6A	0.10	0.87	0.13	0.73	5	0.75
pred(heart,fire)	PLACE=BODY	fire	0.56	heart	0.51	6B	0.09	0.90	0.04	0.89	5	0.75
pred(heart,fire)	EMOTION=BODY	fire	0.27	heart	0.51	3B	0.13	0.84	0.04	0.89	5	0.68
pred(heart,fire)	DISEASE=EMOTION	heart	0.49	fire	0.27	3B	0.07	0.86	0.13	0.84	5	0.68
pred(heart,fire)	EMOTION=DISEASE	fire	0.27	heart	0.49	3B	0.13	0.84	0.07	0.86	5	0.68
pred(heart,fire)	INACTIVITY=SLOW	heart	0.20	fire	0.51	4C	0.06	0.77	0.10	0.88	5	0.67
pred(heart,fire)	EXTREMITY=FEAR	heart	0.18	fire	0.45	1B	0.09	0.75	0.10	0.87	5	0.65
pred(heart,fire)	EMOTION=FLUID	fire	0.27	heart	0.29	3B	0.13	0.84	0.08	0.82	5	0.64
nsubj(fire,burn)	PASSION=HEAT	fire	0.28	burn	0.51	2B	0.11	0.84	0.07	0.83	6	0.61
nsubj(fire,burn)	EMOTION=HEAT	fire	0.27	burn	0.51	2B	0.13	0.84	0.07	0.83	6	0.61
nsubj(fire,burn)	EMOTION=GAS	fire	0.27	burn	0.42	2A	0.13	0.84	0.07	0.85	6	0.58
nsubj(fire,burn)	EXTREMITY=FEAR	burn	0.20	fire	0.45	1B	0.09	0.75	0.10	0.87	6	0.57
nsubj(fire,burn)	EMOTION=ELECTRICITY	fire	0.27	burn	0.34	2B	0.13	0.84	0.05	0.86	6	0.56

Table B.6: Example run of MetID on the sentence “My heart is a fire that burns with love”, using the BEAGLE-512 model.

B.5 Analysis of Algorithms

There are two main algorithms at work in MetID: word clustering and metaphor identification. This section will define these algorithms more formally and provide a brief complexity analysis. The space requirements of each algorithm are not considered because neither appears to exhibit abnormal behaviour in this regard. Because the text stream algorithms (cleaning, parsing, etc.) are not unique to MetID, they are not reviewed here. Likewise, because the individual DSM algorithms are explained and analysed in their respective literature, they are not covered here.

Word Clustering

The word clustering algorithm is responsible for the main output of the semantic stream described above. The algorithm described here is exhaustive over collections, models and seed terms (from *Metalude*). In the implementation of this procedure, the collection and model may be given individually, instead of iterated through. Figure B.3 describes the algorithm.

```
in: Database of collections and clusters, $DB$.
in: Collection of document-segmented texts (corpus), $C$.
in: Set of target-vehicle pairs $\langle target, vehicle\rangle$, $R$.
in: Set of semantic models $\{WordNet, LSA300, LSA100, etc.\}$, $M$.
in: Set of vector distance functions $\{cosine, euclidean, etc.\rangle\}$, $\Phi$.
in: Number of nearest neighbors with which to populate clusters, $n$.
set: $Clusters = \{\varnothing \mapsto \{\varnothing\}\}$
Dump all documents in $C$ from $DB$ for use in $M$.
for $m \in M$:
    if: $m \notin DB_{methods}$
        Insert $m$ into $DB_{methods}$.
    end if
    Generate semantic space $\mathbf{S}$ using $m$.
    for $\varphi \in \Phi$:
        for Term $t \in (R_{targets} \cup R_{vehicles})$:
            $Cluster_t \mapsto \{\varnothing\}$
            for $i \in 0...n$:
                push: $i^{\text{th}}$ nearest neighbour of $t$ in $\mathbf{S}$: $Cluster_t \mapsto \{..., \langle t_i, \varphi(t, t_i)\rangle\}$.
            end for
            for Neighbour $w \in Cluster_t$:
                Calculate relative frequency, $frel_w$, and raw frequency, $fraw_w$, of $w$ in $C$.
                Insert $\langle w, frel_w, fraw_w\rangle$ into $DB$.
            end for
            Calculate entropy, $entropy_{Cluster}$, and purity, $purity_{Cluster}$, using $w \in Cluster$.
            Amend $Cluster$ in $DB$ with $entropy_{Cluster}$ and $purity_{Cluster}$.
        end for
    end for
    Save $\mathbf{S}$ to disk.
end for
```

Figure B.3: The word clustering algorithm implemented in the semantic stream module. Note that in practice, M and Φ may be specified (as m and φ) similar to how C is given.

In the exhaustive case, the outer most loops (lines 9 and 14) contribute exponentially to the complexity. In this case, the algorithm is in multi-polynomial space where complexity is $M^{\Phi^{2R^n}} \in O(n^m)$ where m is factorial. However, if M and Φ are specified (as c and φ beginning at line 15) it runs in regular polynomial space with $2R^n \in O(n^m)$ where m is non-factorial. This puts time-complexity for each cluster (of which there were 479 in every case) at $|M| * |\Phi| * |R| * n$ which for all DSMs is $16 * 6 * 2 * n = 64n$, plus the COLE models $1 * 3 * 2 * n = 6n$ plus the WordNet-based model $1 * 1 * 2 * n = 2n$ resulting in a total of $72n * 479 = 34488n$. And because n was set to 200, this resulted in a grand total of $6,897,600$ operations. A clustering operation, however, is not entirely atomic. For instance, Ruby implements a standard $O(n)$ HASH_INSERT algorithm which is used in line 18, as well as a loop using Ruby's $O(\log n)$ HASH_READ. There are also three calculations (lines 18 and 24) and two database insertions (lines 22 and 25) for each cluster. Additionally, there is a conditional insert (line 11) and a persistence operation for each model (line 28).

Metaphor Identification

The metaphor identification algorithm is the last and perhaps most important procedure in MetID. It takes a set of word-relation-word triples from the text stream and a set of word-clusters generated by the semantic stream; it returns a rank-ordered list of candidate metaphors for each item in the input. The algorithm presented here is in its exhaustive form, where it examines every triple with every set of clusters. Another use-case, perhaps yielding more tractable results, is to examine a given sentence using a particular set of clusters. For example, given the results of chapter 5, it may be enticing to simply use the TASA clusters built using the COALS-14k model, to identify metaphors in a single newspaper article. This use-case was typical of the evaluation of the system.

The algorithm begins with $O(m^{s^{r^2}})$ for the control loops (lines 7-9), $O(kn)$ for calculating the heuristics, purity and entropy, $O(2n)$ for normalising the results (line 26), and finally $O(n \log n)$ for sorting them (line 27). This puts it in log-polynomial with $n * m^{s^{r^2}} * 2n * n \log n \in O(n^m \log n)$ where n is linear and m is factorial. Implemented, there was $|M| = 64$, $|R| = 479$, S is of varying size depending on the use-case, but each element has 2 words. This makes our final number of operations $4608 * 2|S| \log |S|$. Like the preceding section, this complexity analysis is not exhaustive.

```
1   **in:** Set of relational triples $\langle w_1, w_2, rel \rangle$ *or* set of word-pairs $\langle w_1, w_2, rel = \varnothing \rangle$, $S$.
    **in:** Set of target-vehicle pairs $\langle target, vehicle \rangle$, $R$.
    **in:** Set of semantic spaces (models) $\{WordNet\text{-}Lin, LSA300\text{-}cos, LSA300\text{-}euc, etc.\}$, $M$.
    **in:** Number of results to return (ordered best to worst), $n$.
5   **set:** $\hat{S}$ = $\{\varnothing\}$ (to hold results).
    **for** $m \in M$:
        **for** $s \in S$:
            **for** $r \in R$:
                **for** each word $w \in s$:
10                  $d_1$ := the distance from $w$ to $r_{target}$ under $m$.
                    $d_2$ := the distance from $w$ to $r_{vehicle}$ under $m$.
                    $Score := \frac{d_1+d_2}{2}$. ($Score \sim$ "$s$ is in an instance of $r$")
                    Apply grammatical bonuses for $s_{rel}$ as $Score$ += $\frac{1-Score}{2}$.
                    Apply grammatical penalties for $s_{rel}$ as $Score$ /= 2.
15                  Apply lexical bonuses for $s_{w_1}$ and $s_{w_2}$ as $Score$ += $\frac{1-Score}{2}$.
                    Apply lexical penalties for $s_{w_1}$ and $s_{w_2}$ as $Score$ /= 2.
                    **if:** $m \neq WordNet$
                        $Score$ /= $m_{purity}$.
                        $Score$ *= $m_{entropy}$.
20                  **end if**
                    **push:** $\langle s, r_{target}, r_{vehicle}, Score \rangle$ on to $\hat{S}$.
                **end for**
            **end for**
        **end for**
25  **end for**
    Normalise $\hat{S}_{Score}$ to range $[0,1]$.
    Sort $\hat{S}$ descending by $Score$.
    **Return:** $\{\hat{S}_{0...n}\}$.
```

Figure B.4: The metaphor identification algorithm. M and S may by specified at execution time as m and s (lines 6 and 7).

Appendix C

Data, Text Collections & Test Materials

C.1 *Metalude* Seed Terminology

Table C.1 contains all the topic-vehicle pairs from *Metalude*. Topics and vehicles were manually lexicalised to condense multi-word terms into single words. Usually, this involved removing adjective modifiers, resulting in more general, singular nouns.

Topic	Original Topic	Vehicle	Original Vehicle	Sector
ACCEPTANCE		VITALITY		3B
ACHIEVEMENT		HIGH		1D
ACTIVITY		GAME	BALL GAME	4C
ACTIVITY		GAME	BOARD GAME	4C
ACTIVITY		TRAVEL	BOAT TRAVEL	4C
ACTIVITY		BUILDING		4C
ACTIVITY		GAME	CARD GAME	4C
ACTIVITY		DANCING		4C
ACTIVITY		FIGHTING		4C
ACTIVITY		FISHING		4C
ACTIVITY		HUNTING		4C
ACTIVITY		GAME	GAMBLING GAME	4C
ACTIVITY		GAME		4C
ACTIVITY		MUSIC		4C
ACTIVITY		PERFORMANCE		4C
ACTIVITY		PLACE		4D
ACTIVITY		SAILING		4C
ACTIVITY		SHOOTING		4C
ACTIVITY		SWIMMING		4C
ACTIVITY		THEATRE		4C
ACTIVITY		RACE		4C
ACTIVITY		PATH		4D
ACTIVITY		WRITING		4C
ACTIVITY		HIGH		4D
ACTIVITY		ABOVE		4D
ACTIVITY		BODY	HUMAN BODY	4B
ACTIVITY		LIVING		4B
ACTIVITY		MOVEMENT	MOVEMENT (FORWARD)	4C
ACTIVITY		AGRICULTURE		4C
ACTIVITY		LIQUID		2A
ACTIVITY	EXCITED ACTIVITY	ELECTRICITY		2B
ACTIVITY	EXCITED ACTIVITY	HEAT		2B

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
ACTIVITY	INTENSE ACTIVITY	SPEED		4C
AFFECTION		WARMTH		2B
AFFECTION		MONEY		2A
AFFECTION		WEALTH		2A
AGREEMENT		PROXIMITY		3D,D2
ANGER		HEAT		2B
ANIMAL		HUMAN		6B
ANNOYANCE		FRICTION		2B
ANTAGONISM		FRICTION		2B
ARGUING		ATTACKING		3C
ARGUING		FIGHTING		3C
ARGUING		HITTING		3C
ARGUING		PUNCHING		3C
ARGUING		WOUNDING		3C
ARGUING		CUTTING		3C
ARGUMENT		BUILDING		3A
ARGUMENTS		WEAPONS		3C
ARGUMENTS		AMMUNITION		3C
AWARENESS		HIGH		3D
AWARENESS		OUT		3D,A2
AWARENESS		FIXING		3D
AWARENESS		CAPTURE		3D
AWARENESS		PROXIMITY		3D
BAD		LOW		1D
BAD		SMELLY		2A,B2
BAD		POOR		1A
BAD		CHEAP		1A
BEGIN		MOVEMENT	START MOVING	4C
BELIEVING		WALKING		3C
BELIEVING		TRAVELLING		3C
BETTER		RISE		1C,D1
BODY		HUMAN		5B
BODY	HUMAN BODY	EARTH		5D,A5
BROKEN		LOW		4D
BUILDING		BODY		6B
CALM		BALANCE		2D
CATEGORY		SECTION	DIVIDED AREA	3D
CAUSE		FORCE		4C
CAUSE		LOW		4D
CAUSE		PATH		4C
CAUSE		LINK		4C
CAUSE		CONNECTION		4C
CEASE		STOP		4C
CERTAINTY		LOW		3D
CERTAINTY		SOLIDITY		1C
CERTAINTY		FIRMNESS		1C
CESSATION		DEATH		4B
CHANGE		MOVEMENT		4C,C1
CHANGE	CHANGE BEHAVIOUR	BEND		1C,C4
CHANGEABLE		FLEXIBLE		1C,D3
CHANGEABLE		SOFT		1C,D3
CHARACTER		BODY	BODY PART	3B
CHARACTER		FLUID		3B
CHARACTER		METAL		1A
CHOICE		SPACE	SPACE TO MOVE	4D,C4
CHOOSE		SEPARATE		3D

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
COLOUR		MINERAL		1A
COLOUR		PLANT		1A
COMMUNICATION		CONTACT		3D,C3
COMMUNICATING		COOKING		3A,C3
COMMUNICATING		SERVING		3A,C3
COMMUNICATION		FLOW		3A,C3
COMMUNICATION		MOVEMENT		3C
COMMUNICATION		TRAVEL		3C
COMPETITION		RACE		4C
COMPETITION		WAR		4C
COMPETITION		VIOLENCE		4C
COMPETITION	COMPETITIVE EQUALITY	SPEED	EQUALITY OF SPEED	4C
COMPREHENSIBILITY		LOW		3D
COMPREHENSIBILITY		STRAIGHTNESS		3D
CONCEPTION	CREATE AN IDEA	CLOTH	MAKE CLOTH	3A
CONCEPTION	START OF AN IDEA	BIRTH		3B
CONFLICT	CONFLICTING PURPOSE	DIRECTION	OPPOSITE DIRECTION	4C,D4
CONSIDER		LOOK		3B
CONSIDER		TRAVEL	TRAVEL OVER	3C,D3
CONSIDER		INTO		3C,D3
CONSIDERING		CALCULATING		3A
CONTINUATION		DISTANCE		4C,D4
CONTINUE		MOVEMENT	GO ON	4C
CONTROL		HANDLE		4B,C4
CONTROL		OWN		4B,C4
CONTROL		PUSH		1C,D1
CONTROL		DOWNWARD	PUT DOWN	1C,D1
CONTROL		LEAD		4C
CONTROL		GUIDE		4C
CONTROL		ABOVE		1D
CONTROL		HANDLING		4B,C4
CONTROL		DESCEND		1C,D1
CONTROLLED		BELOW		1D
CORRECTNESS		POINT	POSITION AT A POINT	3D
CORRECTNESS		STRAIGHTNESS		3D
CRITICISING		ATTACKING		3C
CRITICISING		FIGHTING		3C
CRITICISING		HITTING		3C
CRITICISING		PUNCHING		3C
CRITICISING		WOUNDING		3C
CRITICISING		CUTTING		3C
CRITICISM		HEAT		2B
CROWD		LIQUID		5A
DEAD		LOW		1D,D4
DECEIT		DOUBLENESS		3B
DECREASE		CONTRACT		1C,D1
DECREASE		CUT		3B,E4
DECREASE		FALL		1C,D1
DESIRE		BENDING		2D
DESIRE		ATTRACTION		2D
DESIRE		APPETITE		2A
DETACHMENT	NO RELATIONSHIP	DISTANCE		2D
DETACHMENT	NO RELATIONSHIP	SEPARATION		2D
DETERIORATE		FALL		1C,D1
DETERIORATE		LOWER		1C,D1
DEVELOP		GROW		4B

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
DEVELOPING		MOVEMENT	MOVING FORWARD	4B
DEVELOPMENT	DEVELOPMENT OF AN IDEA	GROWTH		3B
DIFFERENCE		DISTANCE		3D
DIFFICULT		SLOW		4C
DIFFICULTY		HARDNESS		1B
DIFFICULTY		MUD	MUDDY GROUND	4C
DIFFICULTY		OBSTACLE		4C
DIFFICULTY		DISEASE		4B
DISEASE		EMOTION		3B
DISEASE		IDEA		3B
DISEASE		WAR		5D,C5
DISEASE		INVASION		5D,C5
DISINTEREST		DISTANCE		3D
DISTRACTION	MENTAL DISTURBANCE	DIVISION		2D
DISTRACTION	MENTAL DISTURBANCE	INCOMPLETENESS		2D
EASE		SPEED		4C
EFFECT		COOKING	FOOD PREPARATION	2A
EFFECT		IMPACT		4C
EFFECT		MARK		4C
EFFECT		PRESSURE		4C
EFFECT		IMPRESSION	SENSE IMPRESSION	2B
ELECTRICITY		LIQUID		5A
ELEMENTARY		LOW		3D
EMOTE	CAUSE BAD EMOTIONS	HURT		2B,C2
EMOTE	CAUSE BAD EMOTIONS	INJURE		2B,C2
EMOTE	CAUSE EMOTION	STIR		2A,C2
EMOTION		CURRENT		2A
EMOTION		WAVE		2A
EMOTION		EXPLOSION		2B
EMOTION		FOOD		2A
EMOTION		EATING		2A
EMOTION		GAS		2A
EMOTION		HEAT		2B
EMOTION		HIGH		2D
EMOTION		LIGHT		2B
EMOTION		COLOUR		2B
EMOTION		LIQUID		2A,D3
EMOTION		MOVEMENT		2C,C4
EMOTION		TOUCH		2B
EMOTION		IMPACT		2B
EMOTION		WEATHER		2B
EMOTION		DISEASE		3B
EMOTION		IMPRESSION	SENSE IMPRESSION	2B,B3
EMOTION		SMELL		2B,B3
EMOTION		SOUND		2B,B3
EMOTION		ANIMAL		3B
EMOTION		HUMAN		3B
EMOTION		CONTROL	PERSON CONTROLLED	3B
EMOTION		PLANT		3A
EMOTION		MINERAL		3A
EMOTION		ELECTRICITY		2B
EMOTION		BODY	BODY PART	3B
EMOTION		FLUID		3B
EMPOWER	GAIN POWER	RISE		1C,D1
ENCOURAGE		SUPPORT		4C,D4
ENGINE		ANIMAL		6B

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
ENGINE		HUMAN		6B
EVENT	EVENTS IN TIME	LINE		4D
EVENT	EVENTS IN TIME	ROW		4D
EVENT		LIQUID		2A
EVIL		DARK		1B
EVIL		BLACK		1B
EVIL		DIRT		1B
EVIL		WASTE		1A
EXCITEMENT		LIGHT		2B
EXCITEMENT		COLOUR		2B
EXCITEMENT		HEAT		2B
EXISTENCE		PROXIMITY		3D,A3
EXISTENCE		HIGH		1D,D4
EXPERIENCE		IMPRESSION	SENSE IMPRESSION	2B
EXPERIENCE		EATING		2A
EXPERIENCE		FOOD		2A
EXPERIENCE		WEATHER		2B
EXPERIENCE		RELATIONSHIP		3B
EXPERIENCE		LIQUID		2A
EXPRESSION	EMOTIONAL EXPRESSION	OUTFLOW		2A,C2
EXPRESSION		OUTFLOW		3A,C3
EXPRESSION		HIGH		3D
EXPRESSION		OUT		3D,A2
EXTREMITY		FEAR		1B
FAILURE		DIVISION		4D
FAILURE		FALLING		1C,D1
FAILURE		SHIPWRECK		4C
FAILURE		SINKING		4C,D4
FAILURE		BACKWARDS		4C
FAMOUS		LIGHT		1B
FASHION		CURRENT		3A,C3
FEAR		COLD		2B
FEELING	FEELING EMOTION	EATING	BEING EATEN	2A
FEELING		CONTROL	CONTROLLING PEOPLE	3B
FEW		SMALL		1D
FOOD		HUMAN		6B
FREEDOM		RELEASE		4D,C4
FREEDOM		SPACE	SPACE TO MOVE	4D,C4
FUNDAMENTAL		LOW		3D
FUTURE		AHEAD		4D
FUTURE		FORWARDS		4D
GOOD		CLEAN		1B
GOOD		WHITE		1B
GOOD		FIRST		4C
GOODNESS		PURITY		1A
GROUP	SOCIAL ORGANISATION	BODY		5B
GROUP	SOCIAL ORGANISATION	BUILDING		5D,A5
HAPPENING		ARRIVING		4D,C4
HAPPENING		TRAVELLING		4D,C4
HAPPINESS		LIGHT		2B
HAPPY		HIGH		2D
HEALTH		HIGH		1D,D4
HELP		SUPPORT		4C,D4
HONESTY		STRAIGHTNESS		1D
HOPE		LIGHT		2B
HOSTILITY	EMOTIONALLY HOSTILE	HARD		2B

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
HUMAN		ANIMAL		5B
HUMAN		ARMY		5B
HUMAN		BIRD		5B
HUMAN		BODY		5B
HUMAN		BREAD		5A
HUMAN		DOUGH		5A
HUMAN		PASTA		5A
HUMAN		BUILDING		5D,A5
HUMAN		CAT		5B
HUMAN		CHICKEN		5B
HUMAN		CLOTH		5A
HUMAN		MATERIAL		5A
HUMAN		COW		5B
HUMAN		DOG		5B
HUMAN		FISH		5B
HUMAN		FLOWER		5A
HUMAN		FOOD		5A
HUMAN		FRUIT		5A
HUMAN		GRASS		5A
HUMAN		CORN		5A
HUMAN		HORSE		5B
HUMAN		IMPLEMENT		5A
HUMAN		UTENSIL		5A
HUMAN		INSECT		5B
HUMAN		MACHINE		5A
HUMAN		APPLIANCE		5A
HUMAN		MAMMAL		5B
HUMAN		MEAT		5A
HUMAN		MILK		5A
HUMAN		MONKEY		5B
HUMAN		PIG		5B
HUMAN		PLANT		5A
HUMAN		REPTILE		5B
HUMAN		RODENT		5B
HUMAN		SHEEP		5B
HUMAN		SHIP		5A
HUMAN		SUPERNATURAL		5B
HUMAN		MYTH	MYTHICAL BEING	5B
HUMAN		SWEET		5A
HUMAN		DESSERT		5A
HUMAN		TREE		5A
HUMAN		OBJECT	VALUABLE OBJECT	5A
HUMAN		COMMODITY		5A
HUMAN		VEGETABLE		5A
HUMAN		VEHICLE		5A
HUMAN		WATERBIRD		5B
HUMAN		SEABIRD		5B
HUMANS		LIQUID		5A
IDEA		DISEASE		3B
IDEA		IMPRESSION	SENSE IMPRESSION	2B,B3
IDEA		SMELL		2B,B3
IDEA		SOUND		2B,B3
IDEA		ANIMAL		3B
IDEA		HUMAN		3B
IDEA		CONTROL	PERSON CONTROLLED	3B
IDEA		PLANT		3A

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
IDEA		BUILDING		3A
IDEA		COMMODITY		3A
IDEA		PLACE		3D
IDEA		POSITION		3D
IDEA		CLOTH		3A
IDEA		CLOTHES		3A
IGNORED		INVISIBLE		3B,B1
IMITATE		FOLLOW		4C
IMPOLITE		ROUGH		2B
IMPORTANCE		CENTRALITY		1D
IMPORTANCE		HIGH		1D
IMPORTANCE		WEIGHT		1C
IMPORTANT		FIRST		4C
IMPORTANT		BIG		1D
IMPRESSIVE		LIGHT		1B
IMPROVEMENT	IMPROVE STATUS	RAISE		1C,D1
INACTIVE		LOW		4D
INACTIVITY	LESS ACTIVE	SLOW		4C
INACTIVITY		ABSENCE		4D
INACTIVITY		IMMOBILITY		4C
INCOMPREHENSIBLE		UNCLEAR	NOT CLEAR	3B,B1
INCOMPREHENSIBLE		CROOKED	NOT STRAIGHT	3D
INCREASE		RISE		1C,D1
INCREASE		EXPAND		1C,D1
INFLUENCE		MAGIC		4C
INFLUENCE		PRESSURE		4C
INFLUENCE		LEAD		4C
INFLUENCE		GUIDE		4C
INFORMATION		PREY		3A
INFORMATION		MINERAL		3A
INFORMATION		COMMODITY		3A
INTELLIGENCE		LIGHT		3B
INTEREST		FIXING		3D
INTEREST		CAPTURE		3D
INTEREST		PROXIMITY		3D
INVOLVEMENT		PRESENCE		4D
IRRELEVANCE		WANDERING		3C
JOB		POSITION		4D
JUSTICE		STRAIGHT		1D
KNOW		SEE		3B
KNOWLEDGE		VIEW		3B
KNOWLEDGE		FLUID		3A
KNOWLEDGE		FOOD	FOOD AND DRINK	3A
KNOWN		UNCOVERED		3B,B1
KNOWN		OPEN		3B,B1
LANDSCAPE		BODY		6B
LANGUAGE		PLANT		3A
LANGUAGE		HUMAN		3B
LANGUAGE	LANGUAGE QUALITY	TASTE		3A,A2
LAW		STRAIGHT		1D
LESS		LOW		1D
LESS		SMALL		1D
LIFE		DAY		4D
LIFE		PATH		4D
LIFE		WRITING		4C
LIFE		HIGH		1D,D4

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
LIGHT		LIQUID		5A
LIQUID		CROWD		6B
LIQUID		HUMANS		6B
LISTENING		EATING		3A,C3
LISTENING		DRINKING		3A,C3
LOUD		HIGH		5D
LOVE		HEAT		2B
MACHINE		HUMAN		6B
MACHINE		ANIMAL		6B
MEANS		ROAD		4C,D4
MEANS		TRACK		4C,D4
MEANS		TRANSPORT		4C
MIND		BUILDING		5D,A5
MIND		CONTAINER		5D
MONEY		FOOD		6A
MONEY		LIQUID		6A
MONEY		BLOOD		6A
MORALITY		HIGH		1D
MORE		HIGH		1D
MORE		BIG		1D
NERVOUSNESS		TENSION		2B
NORMALITY		STRAIGHTNESS		2D
NUMEROUS		BIG		1D
OBEY		FOLLOW		4C
OBJECT		HUMAN	HUMAN ?	6B
OBVIOUS		CLEAR		3B,B1
OCCURRENCE		HIGH		1D,D4
OPERATION		HIGH		4D
OPERATION		ABOVE		4D
OPINION		VIEW		3B
OPINION		PERSPECTIVE		3B
OPINION		ORIENTATION		3B
OPINION		CURRENT		3A,C3
OPINION		PLACE		3D
OPINION		BEND		1C,C4
OPINION		POSITION		3D
OPPORTUNITY		TRANSPORT		4C
OPPORTUNITY		OPENING		4C
ORGANISATION		SHIP		4C,A5
ORGANISATION		MACHINE		5A
ORGANISATION		PLANT		5A,C4
ORGANISATION	ORGANISATION PART	BODY	BODY PART	5B
PASSION		HEAT		2B
PAST		BEHIND		4D
PAST		BACKWARDS		4D
PERIOD		DAY		4D
PERIOD		LENGTH		4D
PERIOD		DISTANCE		4D
PERIOD		SPACE		4D
PESSIMISM		DARK		2B
PITCH	SOUND FREQUENCY	HEIGHT		5D
PLACE		BODY		6B
PLANT		HUMAN		6B
PLANT		ANIMAL		6B
POINT	POINT IN TIME	POSITION		4D
POSSIBILITY		OPENING		4C

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
POWER		ABOVE		1D
POWER		CENTRALITY		1D
POWERLESS		BELOW		1D
PREVENTION		OBSTACLE		4C
PROBABILITY		PROXIMITY		4D
PROBLEM		DISEASE		4B
PROBLEM		WEIGHT		2B
PROCESS		BODY	HUMAN BODY	4B
PROCESS		LIVING		4B
PROCESS		MOVEMENT	MOVEMENT (FORWARD)	4C
PURPOSE		DIRECTION		4C,D4
PURPOSELESS		DIRECTIONLESS		4C,D4
QUALITY		MONEY		1A
QUALITY		WEALTH		1A
QUALITY		SHAPE		1D
QUALITY		SIZE		1D
QUALITY		TASTE		2A,A1
QUALITY		TEXTURE		2A,A1
QUALITY		HIGH		1D
QUANTITY		LENGTH		1D
QUANTITY		SIZE		1D
QUANTITY		WATER		1A
RACE		COLOUR		1B
RANK		METAL		1A
READING		EATING		3A,C3
READING		DRINKING		3A,C3
REBUTTAL		DEFENSE		3C
REDUCE	REDUCE STATUS	LOWER		1C,D1
RELATIONSHIP		MONEY		2A
RELATIONSHIP		WEALTH		2A
RELATIONSHIP		MUSIC		2B
RELATIONSHIP		PROXIMITY		2D
RELATIONSHIP		COHESION		2D
RELIABILITY		SOLIDITY		1C
RELIABILITY		FIRMNESS		1C
RELINQUISH	LOSE POWER	DESCEND		1C,D1
RELINQUISH	GIVING UP	BACKWARDS		4C
REPRESSION	NO FREEDOM	ENCLOSURE		4D,C4
REPRESSION	NO FREEDOM	SPACE	LIMIT TO SPACE	4D,C4
REPRESSION	NO FREEDOM	TYING		4D,C4
REPRESSION	NO FREEDOM	BINDING		4D,C4
REPURPOSE	CHANGE PURPOSE	DIRECTION	CHANGE DIRECTION	4C,D4
REPUTED		LIGHT		1B
RESPONSIBILITY		WEIGHT		2B
REVEAL	MAKE KNOWN	DIG	DIG UP	3B,C3
REVEAL	MAKE KNOWN	OPEN		3B,C3
REVEAL	MAKE KNOWN	SHOW		3B,C3
REVEAL	MAKE KNOWN	DRAW		3B,C3
REVEAL	MAKE KNOWN	UNCOVER		3B,C3
SAD		LOW		2D
SADNESS		DARK		2B
SADNESS	UNPLEASANT EMOTION	COLD		2B
SADNESS	BAD EMOTION	DISCOMFORT		2B
SADNESS	BAD EMOTION	PAIN		2B
SANITY		STRAIGHTNESS		2D
SANITY		BALANCE		2D

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
SERIOUSNESS		DEPTH		3D,D1
SERIOUSNESS		WEIGHT		1C
SEX		VIOLENCE		4C
SHARE	SHARE PURPOSE	ALIGN		4C,D4
SIGHT		SOUND		5B
SIMILARITY		PROXIMITY		3D
SITUATION		PLACE		4D
SITUATION		WEATHER		2B
SOLUTION		WAY	WAY ROUND	4C
SOLUTION		OVER		4C
SOLUTION		THROUGH		4C
SOUND		LIQUID		5A
SOUND		TASTE		5B
SOUND		TOUCH		5B
SPEECH	VERBAL COMMUNICATION	GAME	(BALL)GAME	3C
SPEECH	AWKWARD SPEECH	WALKING	AWKWARD WALKING	3C
SPEECH	VERBAL COMMUNICATION	TRAVEL		3C
SPEECH	VERBAL COMMUNICATION	MOVEMENT		3C
STAGNATION	NO DEVELOPMENT	IMMOBILITY		4C
STAGNATION	NO DEVELOPMENT	CIRCULARITY		4C
STATE		PLACE		4D
STATUS	STATE OF AN ORGANISATION	HEALTH		5B
STATUS		HIGH		1D
STEAL		HIT		1C,D1
STEAL		CUT		1C,D1
STEAL		LIFT		4C,D4
SUBJECT		PLACE		3D
SUBORDINATE		LOW		1D
SUBSTANCE		HUMAN	HUMAN ?	6B
SUCCEEDING		MOVEMENT	MOVING FORWARD	4B
SUCCESS	COMPETITIVE SUCCESS	LEADING		4C
SUCCESS	COMPETITIVE SUCCESS	RACE	WINNING A RACE	4C
SUCCESS		DISTANCE		4C,D4
SUCCESS		SPEED		4C
SUCCESS		HIGH		1D
SUCCESS	SUCCESS IN ARGUMENT	VICTORY		3C
SUFFICIENCY	BE GOOD ENOUGH	RISE		1C,D1
SYSTEM		MACHINE		5A
SYSTEM		PLANT		5A,C4
TEXT		CLOTH	MAKE CLOTH	3A
TEXT		BUILDING		3A
TEXT		CONTAINER		3A
TEXT		PATH		3C,D3
TEXT		STRUCTURE		3A
TEXT		CLOTH		3A
TEXT		CLOTHES		3A
THINKING		CALCULATING		3A
THINKING		CONTROL	CONTROLLING PEOPLE	3B
THINKING		WALKING		3C
THINKING		TRAVELLING		3C
THOUGHT		RELATIONSHIP		3B
TIME		MONEY		1A
TIME		COMMODITY		1A
TIME		SPACE		4B
TIME	TIME ELAPSING	TRAVEL		4D
TOPIC		PLACE		3D

Continued on next page

Table C.1 continued

Topic	Original Topic	Vehicle	Original Vehicle	Sector
TOUCH		SOUND		5B
TRAFFIC		LIQUID		6A
TRAFFIC		BLOOD		6A
TRUTH		STRAIGHTNESS		3D
UNAWARENESS		DISTANCE		3D
UNAWARENESS		LOW		3D
UNCERTAINTY		INSTABILITY		1C
UNCHANGING		HARD		1C,D3
UNCHANGING		RIGID		1C,D3
UNCHANGING		STATIC		1C,D3
UNCONSCIOUSNESS		LOW		3D
UNDERSTAND		SEE		3B
UNDERSTAND		GRASP		3B
UNDERSTAND		HOLD		3B
UNDERSTANDING		EYESIGHT		3B
UNDERSTANDING		PENETRATION		3B
UNDERSTANDING	UNDERSTANDING	PENETRATION	PENETRATION	3B
UNDERSTANDING		SHARPNESS		3B
UNEMOTIONAL		COLD		2B
UNFEELING		HARD		2B
UNFRIENDLY		COLD		2B
UNHEALTHY		LOW		1D,D4
UNIMPORTANCE		PERIPHERY		1D
UNIMPORTANCE		EDGE		1D
UNIMPORTANT		LOW		1D
UNIMPORTANT		POOR		1A
UNIMPORTANT		CHEAP		1A
UNKNOWN		COVERED		3B,B1
UNKNOWN		INVISIBLE		3B,B1
UNPLEASANT		ROUGH		2B
UNRELIABILITY		INSTABILITY		1C
UNSUCCESSFUL		SLOW		4C
VALIDITY	IDEA'S VALIDITY	VITALITY		3B
VALUE		METAL		1A
VEHICLE		ANIMAL		6B
VEHICLE		HUMAN		6B
WEATHER		ACTIVITY	HUMAN ACTIVITY	4B
WEATHER		QUALITY		4B
WORDS		PREY		3A
WORDS		FLUID		3A
WORDS		FOOD	FOOD AND DRINK	3A
WORDS		HUMAN		3B
WORK		AGRICULTURE		4C
WORRY		WEIGHT		2B
WORTHLESS		EMPTY		1D
WORTHLESSNESS		WASTE		1A

Table C.1: All topic-vehicle pairs from *Metalude*. If a term was changed, its original is listed. Sector refers to where on the *map of root analogies* the entry is located (see figure 3.4.

C.2 Text Collections

These collections represent a range of corpus types. The ANC and TASA corpora were developed to be representative collections of general text [136, 152]. NIPS is a collection of peer-reviewed academic papers, which was used in a previous publication on diachronic changes in grammatical relations [70]. The two finance-related collections were purpose-built to explore how the language of finance relates to outside financial metrics [67, 68, 69]. Lastly, the enTenTen corpus is large collection built by automatically crawling portions of the internet.

American National Corpus

The American National Corpus (ANC) is a collection of American English texts and verbal transcripts spanning genres comparable to the British National Corpus (BNC)[1]. The full ANC contains a comparable number of words (100 million) to the BNC, but representative portions have been annotated by hand. These "manually annotated sub-corpora" (MASC) are designed to be representative of the full ANC, but manageable for human annotation. In MetID, MASC 1 and 2 were used as a whole. The vocabulary coverage of the combined MASC 1 and MASC 2 (130 documents) is about 95% of the full ANC and the genre breakdown is similarly proportioned. The annotations were not used in MetID because the pre-processing scheme was kept uniform across corpus. Instead, MASC 1 and 2 were used to pare the corpus down to a smaller size without relinquishing its representativeness. A sample document (119CWL041) is given here, note the partially redacted style:

> March 29, 1999
>
> Name Address City, ST Zip
>
> Dear Name:
>
> The 1999 Invest in Youth Campaign is in full swing. As a former board member, the success of the YMCA is still important to me. We must be able to reach all youth and families interested in values-based programs. The Invest in Youth campaign helps insure this. New initiatives in the inner-city are taking hold. The Urban Mission Branch is reaching out to middle school youth with programs based on caring, honesty, respect and responsibility for themselves and others. For some, these are very different messages from the ones heard in the street. They are learning to make positive choices concerning alcohol, tobacco and other drugs and to support each other when those choices are challenged. You have shared in the vision and the leadership of the YMCA of Greater Indianapolis. I now invite you to continue to support the mission and the message that is so important to building strong kids, strong families and strong communities. Please consider joining the Chairman's Roundtable with a gift of $1,000. I have enclosed a pledge card for your convenience. We would like to announce the success of this year's campaign at the Annual Meeting on April 27, so please return you pledge or gift within the next 30 days.

[1] http://www.americannationalcorpus.org/; 11 February, 2013.

> Thank you.
> Sincerely,
> Richard H. Gilbert, Jr. Past Board Member YMCA of Greater
> Indianapolis

LexisNexis Finance Articles

This purpose-built corpus was constructed using LexisNexis News[2] to retrieve the top 100 finance-related articles for every month from 2005 through 2009 were collected. Only English articles from major world newspapers were collected. This resulted in 6,000 documents over 60 months with an average of 64,319 tokens per month. Like the BBC, Financial Times, New York Times fiance corpus, these articles typify ones found in newspapers. Also like the BBC-FT-NYT collection, this corpus was diachronically organised, by month, making for 60 documents over the five-year period. This offers the broadest temporal grouping among the corpora. The following is an excerpt from one such article:

> A top think-tank wants the central government to set up a body to develop fiscal policy.
>
> It says the measure would accelerate the reform of decision-making on public finances and improve the government's ability to implement macroeconomic controls.
>
> The proposed fiscal policy committee would complement the central bank's Monetary Policy Committee, said the Institute of Finance and Trade Economics under the Chinese Academy of Social Sciences.
>
> In its annual report on the nation's financial policy, titled "Scientific Development Concept - New Thinking to Guide China's Fiscal Policy", the institute said it had become vital for the government to consider setting up a fiscal policy body.

NIPS Proceedings

The NIPS corpus consists of papers from thirteen volumes of Neural Information Processing Systems proceedings[3]. It contains about 5.2 million words in 1,738 documents published over 13 years from 1987 to 1999, with an average of 3,034 tokens per document. The corpus has been used previously in work on diachronic analysis, topic modeling, and relevance scoring [23, 70, 89, 189, 204]. The NIPS corpus has two unique features: it consists of advanced, academic language and is uniformly diachronic. Academic language, and in particular research papers, can

[2] http://www.lexisnexis.com/; 17 February, 2013.

[3] http://www.cs.nyu.edu/˜roweis/data.html; 11 February, 2013

be very technical and explanatory, and often gloss over defining terms that are common (and often unique) to the research community. One such example from the NIPS corpus is the term "connectivity" which refers to the density of connections in connectionist networks – an abstract machine which simulates undirected learning. The term is widely used, without definition, though it refers to something specific and only loosely related to the non-technical notion of a connection. Of course, there are a number of a other examples, including borrowed terms from mathematics, cognitive psychology and neurology – terms like *complexity*, *learning*, and *synapse*. Below is an excerpt from the first paper in the corpus; notice the neurological explanation of connectivity in the first line of the abstract:

> CONNECTIVITY VERSUS ENTROPY
> Yaser S. Abu-Mostafa
> California Institute of Technology
> Pasadena, CA 91125
> ABSTRACT
> How does the connectivity of a neural network (number of synapses per neuron) relate to the complexity of the problems it can handle (measured by the entropy)? Switching theory would suggest no relation at all, since all Boolean functions can be implemented using a circuit with very low connectivity (e.g., using two-input NAND gates). However, for a network that learns a problem from examples using a local learning rule, we prove that the entropy of the problem becomes a lower bound for the connectivity of the network.

TASA Corpus

The TASA corpus was created by Touchstone Applied Science Associates and used to develop *The Educator's Word Frequency Guide* [238][4]. It was compiled to be a representative set of texts to which typical American students would be exposed at varying stages of education. It contains almost 39,000 documents (paragraphs) of American English text on the subjects of language arts, health, home economics, industrial arts, science, social studies and business. An example document (#1728) is given here:

> A liquid is a strange substance. The principles that govern the behavior of solids and gases are much better understood than those that govern the behavior of liquids. The marvel is not that liquids behave as they do, but that they exist at all. In theory, it might seem more reasonable for a crystalline solid to melt to a fluid having molecules initially touching one another, and for further heating to cause the molecules to move faster and farther apart until something like a gas is produced, without any sharp transition in fluid properties along the way. This theoretical possibility is diagrammed in Figure 17-1 in a plot of volume per mole against temperature. Real substances

[4] Thanks to Professor Thomas Landauer for making this resource available.

> actually do behave this way above what is called their critical pressure, PC, which is 218 atm for H2O and 72 atm for CO2. But at lower pressures their behavior is more like that shown in Figure 17-2. The molar volume usually increases slightly upon melting (point E to point D), and then makes a sudden jump at the boiling point, TB, where the liquid changes to a gas (point B to point C).

The TASA corpus has been used in a variety of research applications. Particularly relevant here is its use as the data on which LSA and other distributional semantic models were developed [122, 136, 137, 188, 222]. TASA is a designed corpus, like the ANC, which gives us a degree of assurance that models built on it are somewhat representative. However, the TASA corpus consists of "documents" which are effectively paragraph from text-books, which while they may be helpful in investigating language-use as it is learned by school students, it is not necessarily representative of naturally occurring document segmentation like articles. TASA is perhaps the most widely used of the collections reported on here.

BBC, Financial Times & New York Times Finance Articles

This custom-built collection was made using automated web searches to select articles related to the Dow Jones, the FTSE 100 and the NIKKEI 225 from the New York Times (nyt.com), the Financial Times (ft.com) and the British Broadcasting Corporation (bbc.co.uk/news). The resulting corpus contained 17,713 articles with 10,418,266 tokens from 2006 to the beginning of 2010, with an average of 2,604,567 tokens per year and 4,428 tokens per article. After they were downloaded, the articles were stripped of HTML, converted to UTF-8 and their uniqueness ensured by keying them on the first 50 characters. The articles in this corpus were grouped in weekly documents. This collection was previously used to track changes in verb-distributions as they relate to markets fluctuations [67] and served as the basis for two studies on the structure of linguistic metaphors [68, 69]. The following paragraph is taken from a typical article:

> THE first week of the year may begin with traders following their hearts and end with them following their heads. John K. Lynch, the chief market analyst at Evergreen Investments, predicted that the stock market would have a subdued start to the year as traders react first to the inability of the Dow Jones industrial average to hang on to a gain for 2005. The loss last week, resulting in a decline of 0.6 percent for the Dow last year, "will affect sentiment going into the week," Mr. Lynch said. But come Friday, he added, the market is likely to be guided by sober analysis after the release that morning of the December employment report. Mr. Lynch estimated that 180,000 net new jobs were created last month, below the 200,000 consensus forecast in a Bloomberg News survey of economists but still greater than the monthly average of the last couple of years.

enTenTen Corpus

The TenTen corpora series, developed by Lexical Computing Ltd. for their Sketch Engine service[5], are large (10^{10} tokens) web-based corpora, crawled automatically from text-heavy portions of the web [120]. Version 1.0 of the enTenTen corpus contained 3.3 billion tokens, of which the first 92 thousand tokens were used here. Web corpora have become increasingly popular, as access to the internet becomes more widespread. Web corpora allow a large amount of data to be gathered quickly and without regard to source, genre, register or authorship. The size of such corpora has overcome some weaknesses of smaller, designed corpora in NLP tasks – particularly those which adopt a machine learning approach [209, 239]. The following is an excerpt taken from a blog about anthropology:

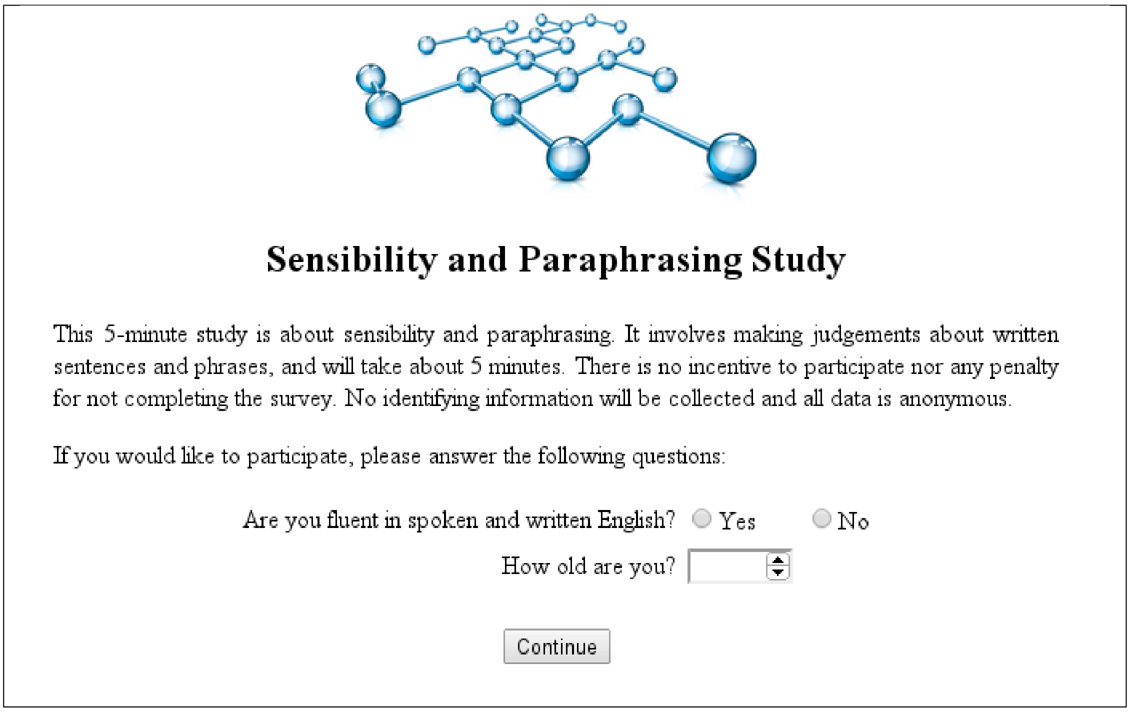
> Snakeman and the Ancient Mayan Medicine by: Jean-Philippe Soule and Luke Shullenberger.
>
> Snakes evoke fear and repulsion in western cultures. Yet the same animal that represents the devil in the bible was the symbol of medicine in Ancient Greece and is still found today on ambulances and pharmacies in many countries.
>
> The Ancient Mayan people revered snakes.
>
> Rattle snake representations and drawings have been found on numerous pieces of pottery and murals and is the most prominent feature on the head dress of their goddess of the herbs Ix Chel. Like the Chinese, they believed in the healing properties of certain species. The legacy they have left in Central America is a supposed cure-all snake bone medicine called Cascabel which is still used by traditional healers.

[5] http://www.sketchengine.co.uk/; 14 February, 2013.

C.3 Experiment 2: Materials

The user-based evaluation presented in the second experiment (see section 5.5) consisted of an online survey. Consenting participants were introduced and presented with the survey on web-pages, of which figures C.1, C.2 and C.3 are screen-shots.

The design of theses materials was based on similar studies in metaphor and sensibility [83, 198, 212]. However, one concern was brought to light about these materials: using the term ”sensible” is a bit vague with regards to what the sensibility ratings were used to. In experiment 2, these ratings were used to down-weight quality rating for the interpretations produced by MetID. That is, ”sensibility” is being used as a proxy for ease of understanding. There are certainly some sentences that are sensible but not easy to understand, or vice versa.



Sensibility and Paraphrasing Study

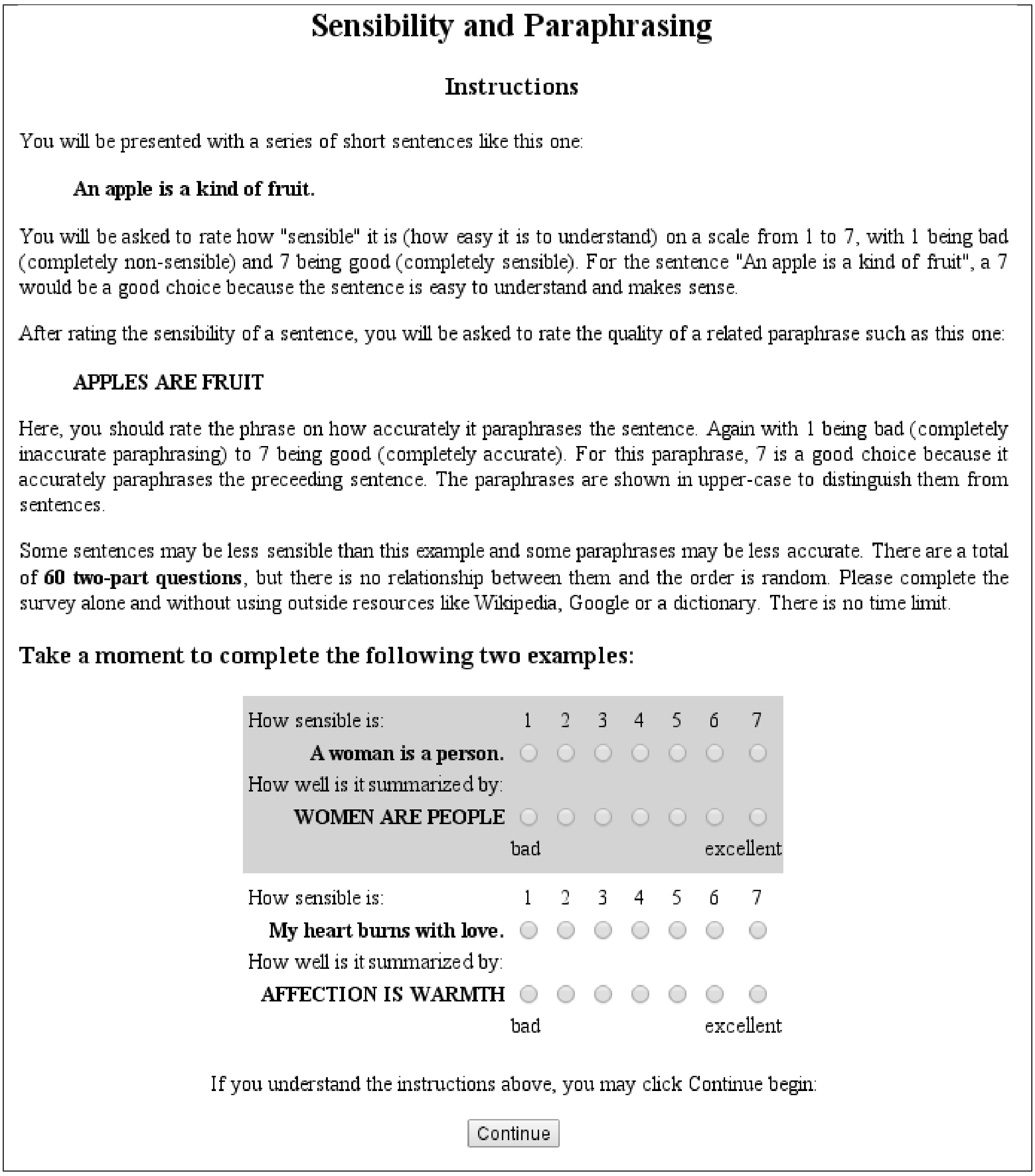
This 5-minute study is about sensibility and paraphrasing. It involves making judgements about written sentences and phrases, and will take about 5 minutes. There is no incentive to participate nor any penalty for not completing the survey. No identifying information will be collected and all data is anonymous.

If you would like to participate, please answer the following questions:

Are you fluent in spoken and written English? ○ Yes ○ No

How old are you? [ ]

Figure C.1: After informed consent, this introductory page was shown. Participants were asked to submit their age and fluency in English.



Sensibility and Paraphrasing

Instructions

You will be presented with a series of short sentences like this one:

An apple is a kind of fruit.

You will be asked to rate how "sensible" it is (how easy it is to understand) on a scale from 1 to 7, with 1 being bad (completely non-sensible) and 7 being good (completely sensible). For the sentence "An apple is a kind of fruit", a 7 would be a good choice because the sentence is easy to understand and makes sense.

After rating the sensibility of a sentence, you will be asked to rate the quality of a related paraphrase such as this one:

APPLES ARE FRUIT

Here, you should rate the phrase on how accurately it paraphrases the sentence. Again with 1 being bad (completely inaccurate paraphrasing) to 7 being good (completely accurate). For this paraphrase, 7 is a good choice because it accurately paraphrases the preceeding sentence. The paraphrases are shown in upper-case to distinguish them from sentences.

Some sentences may be less sensible than this example and some paraphrases may be less accurate. There are a total of **60 two-part questions**, but there is no relationship between them and the order is random. Please complete the survey alone and without using outside resources like Wikipedia, Google or a dictionary. There is no time limit.

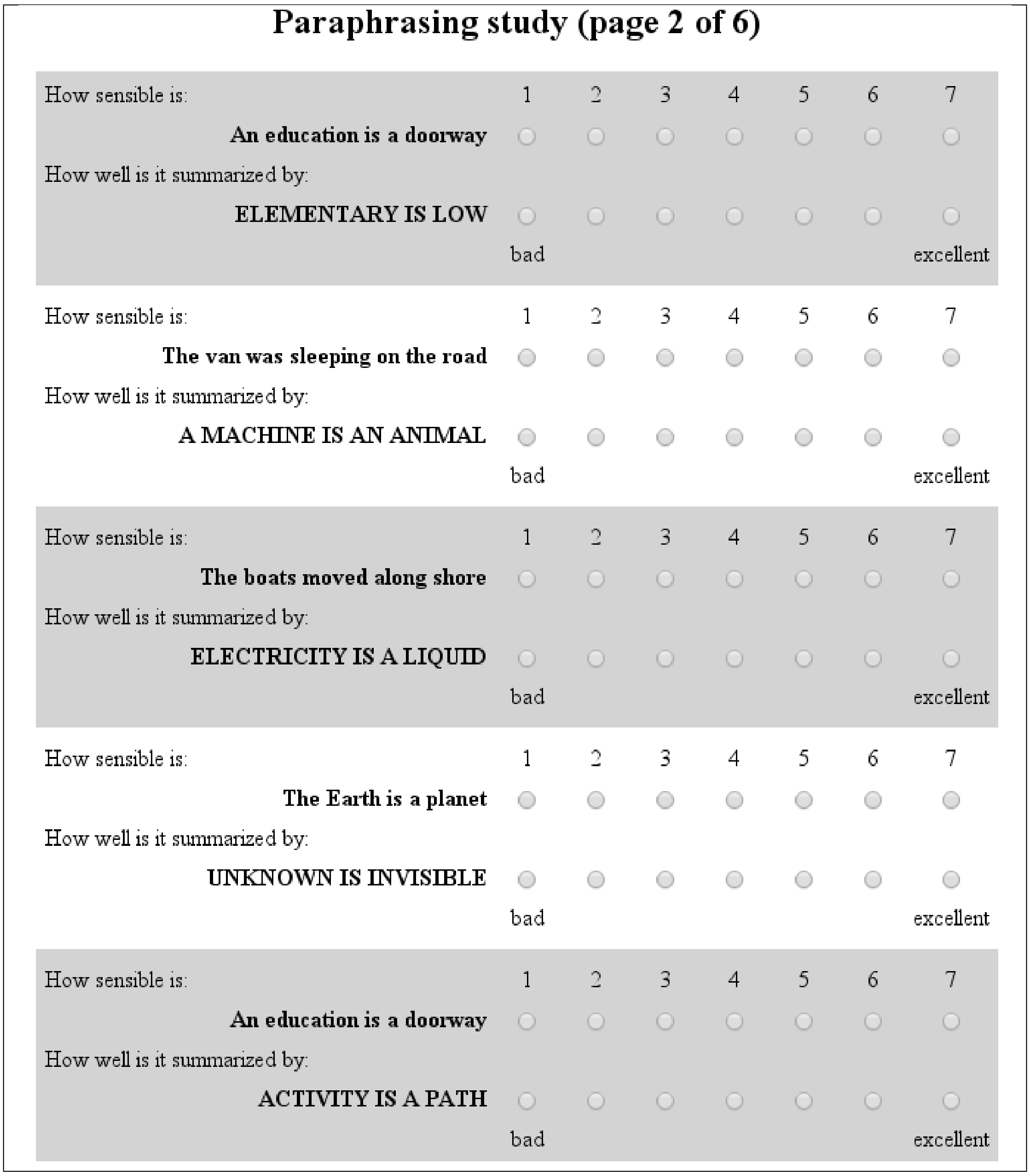
Take a moment to complete the following two examples:

How sensible is:	1	2	3	4	5	6	7
A woman is a person.	○	○	○	○	○	○	○
How well is it summarized by:							
WOMEN ARE PEOPLE	○	○	○	○	○	○	○
	bad					excellent	

How sensible is:	1	2	3	4	5	6	7
My heart burns with love.	○	○	○	○	○	○	○
How well is it summarized by:							
AFFECTION IS WARMTH	○	○	○	○	○	○	○
	bad					excellent	

If you understand the instructions above, you may click Continue begin:

Figure C.2: Instructions explaining the rating task with two demonstrative examples. Neither example appears in the actual survey.



Paraphrasing study (page 2 of 6)

How sensible is:	1	2	3	4	5	6	7
An education is a doorway	○	○	○	○	○	○	○
How well is it summarized by:							
ELEMENTARY IS LOW	○	○	○	○	○	○	○
	bad						excellent
How sensible is:	1	2	3	4	5	6	7
The van was sleeping on the road	○	○	○	○	○	○	○
How well is it summarized by:							
A MACHINE IS AN ANIMAL	○	○	○	○	○	○	○
	bad						excellent
How sensible is:	1	2	3	4	5	6	7
The boats moved along shore	○	○	○	○	○	○	○
How well is it summarized by:							
ELECTRICITY IS A LIQUID	○	○	○	○	○	○	○
	bad						excellent
How sensible is:	1	2	3	4	5	6	7
The Earth is a planet	○	○	○	○	○	○	○
How well is it summarized by:							
UNKNOWN IS INVISIBLE	○	○	○	○	○	○	○
	bad						excellent
How sensible is:	1	2	3	4	5	6	7
An education is a doorway	○	○	○	○	○	○	○
How well is it summarized by:							
ACTIVITY IS A PATH	○	○	○	○	○	○	○
	bad						excellent

Figure C.3: Example set of five sentence-paraphrase pairs as they were presented to participants. There were ten pairs per page and 60 in total.

Table C.2 lists the six planted questions used to verify participants' understanding and mindfulness in completing the survey. Each configuration listed in table 5.10 was used to generate interpretations for the statements in tables 5.5 and 5.6. This resulted in the 432 sentence-paraphrase pairs listed in table C.4. Recall that not every sentence could be processed by every model. Table C.3 lists those sentences that elicited a mean sensibility rating less than 4, and were thus not included in the analysis of their paraphrases.

Sentence	**Paraphrase**	**Intended Rating**
A desk is a kind of furniture	DESKS ARE FURNITURE	7
The team paraded across the field	PARADING IS MOVEMENT	7
He hurried to the bus stop	HURRYING IS MOVEMENT	7
A donkey is a flying door	DONKEYS ARE DOORS	0
The wall blazed the iron	BLAZING IS IRON	0
Melissa unbooked her pencils	BOOKING IS WRITING	0

Table C.2: Planted questions that all participants received to verify they were completing the survey correctly.

Sentence	**Mean Sensibility**
The van was sleeping on the road	2.11
They melted the alliance	2.13
The van was sleeping on the road	2.36
He buckled a bandage	2.6
The van was sleeping on the road	2.82
The van was sleeping on the road	2.85
The man kidnapped their solution	3.14
The engine frayed out	3.17
He buckled a bandage	3.18
He unlocked her old wound	3.23
The man kidnapped their solution	3.29
The engine frayed out	3.5
The plants obeyed the constraints	3.39
He unlocked his old wound	3.41
The van was sleeping on the road	3.44
The man kidnapped their solution	3.45
Beauty is a ticket	3.53
The plants obeyed the constraints	3.53
He unlocked his old wound	3.56
The engine frayed out	3.68
He unlocked her old wound	3.71
He unlocked her old wound	3.8
The man kidnapped their solution	3.8
He unlocked her old wound	3.87
He unlocked his old wound	3.93

Table C.3: Sentences in experiment 2 that elicited a mean sensibility rating less than 4, excluding them from analysis.

Model	Type	Fig?	Score	Sentence	Candidate Metaphor / Paraphrase
WordNet	noun	no	0.61	A lion is an animal	AN ANIMAL IS A HUMAN
WordNet	noun	no	0.54	My brother is a butcher	STEALING IS HITTING
WordNet	noun	no	0.80	A snail is a pest	A HUMAN IS A PIG
WordNet	noun	no	0.51	A salmon is a fish	A HUMAN IS A FISH
WordNet	noun	no	0.11	Sharks have sharp teeth	A HUMAN IS A FISH
WordNet	noun	no	0.50	Cereal is a food	A HUMAN IS FOOD
WordNet	noun	no	0.61	That creature in the net is a crab	AN ANIMAL IS A HUMAN
WordNet	noun	no	0.81	That lost painting is a portrait	AN OPINION IS A VIEW
WordNet	noun	no	0.89	Crime is a problem	A PROBLEM IS A DISEASE
WordNet	noun	no	0.23	The Earth is a planet	A BODY IS THE EARTH
WordNet	noun	no	0.87	Some urban schools are crowded	CONTROLING IS PUSHING
WordNet	noun	no	1.00	Some computer programs are complex	A SYSTEM IS A PLANT
WordNet	noun	no	0.59	Some ideas are great	AN IDEA IS A COMMODITY
WordNet	noun	no	0.59	Some jobs are constraining	A JOB IS A POSITION
WordNet	noun	no	0.59	Some lectures are boring	SPEECH IS A GAME
WordNet	noun	yes	0.90	Crime is a disease	A PROBLEM IS A DISEASE
WordNet	noun	yes	0.78	A vacation is medicine	MONEY IS FOOD
WordNet	noun	yes	0.75	Dancers are butterflies	A HUMAN IS AN INSECT
WordNet	noun	yes	0.97	His life is an opera	ACTIVITY IS MUSIC
WordNet	noun	yes	0.54	Some surgeons are butchers	STEALING IS HITTING
WordNet	noun	yes	0.61	Beggars are parasites	A HUMAN IS AN ANIMAL
WordNet	noun	yes	0.90	The mind is a computer	A MIND IS A BUILDING
WordNet	noun	yes	0.91	Some ideas are diamonds	AN IDEA IS A COMMODITY
WordNet	noun	yes	0.89	A smile is a magnet	DESIRE IS ATTRACTION
WordNet	noun	yes	0.88	Experience is a fountain	AN EXPERIENCE IS A LIQUID
WordNet	noun	yes	0.92	Beauty is a ticket	AN EFFECT IS A MARK
WordNet	noun	yes	0.93	Love is a journey	AN EMOTION IS MOVEMENT
WordNet	noun	yes	0.71	Rumors are viruses	SPEECH IS A GAME
WordNet	noun	yes	0.53	Some malls are jungles	A BUILDING IS A BODY
WordNet	noun	yes	0.92	Some jobs are prisons	A JOB IS A POSITION
WordNet	noun	yes	0.82	Alcohol is a crutch	A SOLUTION IS A WAY
WordNet	noun	yes	0.82	An education is a doorway	A POSSIBILITY IS AN OPENING
WordNet	noun	yes	0.92	Angry words are knives	ARGUMENTS ARE WEAPONS
WordNet	noun	yes	0.80	Faith is a fortress	A MIND IS A BUILDING
WordNet	noun	yes	0.91	Humor is a weapon	ARGUMENTS ARE WEAPONS
WordNet	verbs	no	0.77	The engine wore out	TO DETERIORATE IS TO FALL
WordNet	verbs	no	0.59	The ancient car fell apart	A MACHINE IS A HUMAN
WordNet	verbs	no	0.83	The boats moved along shore	ACTIVITY IS TRAVELLING
WordNet	verbs	no	0.74	The boy grabbed his bike and went home	A HUMAN IS A FISH
WordNet	verbs	no	0.59	The building shook from the earthquake	AN ARGUMENT IS A BUILDING
WordNet	verbs	no	0.80	The clouds gathered on the horizon	A LIQUID IS A CROWD
WordNet	verbs	no	0.90	The runners ran through the streets	COMPETITION IS SPEED
WordNet	verbs	no	0.80	The bike moved along the trail	ACTIVITY IS TRAVELLING
WordNet	verbs	no	0.91	The poster hung over the desk	COMMUNICATION IS FLOW
WordNet	verbs	no	0.75	The van was idling on the road	AN ORGANISATION IS A MACHINE
WordNet	verbs	no	0.90	The bread rose to perfection	A HUMAN IS DOUGH
WordNet	verbs	no	0.82	The house decayed over time	TO DETERIORATE IS TO FALL
WordNet	verbs	no	0.68	She opened the gate	SUFFICIENCY IS RISING
WordNet	verbs	no	0.88	She cleaned up the spill	DECREASING IS FALLING
WordNet	verbs	no	0.87	She delivered a message	A CONCEPTION IS A BIRTH
WordNet	verbs	no	0.81	The doctor mended the cut	MAKING BETTER IS RISING
WordNet	verbs	no	0.74	The woman rejected the proposal	A MIND IS A BUILDING
WordNet	verbs	no	0.29	The man stole their solution	A HUMAN IS MILK
WordNet	verbs	no	0.81	They withdrew the invitation	SPEECH IS TRAVEL
WordNet	verbs	no	0.70	They released the prisoner	A HUMAN IS A CAT
WordNet	verbs	yes	0.73	He buckled a bandage	COLOUR IS A PLANT

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
WordNet	verbs	yes	0.77	He piloted his dance partner	CONTROLING IS LEADING
WordNet	verbs	yes	0.76	She devoured the new material	LANGUAGE IS A TASTE
WordNet	verbs	yes	0.82	He unlocked his old wound	FREEDOM IS A RELEASE
WordNet	verbs	yes	0.91	She shot him a message	COMMUNICATION IS MOVEMENT
WordNet	verbs	yes	0.82	The woman killed the proposal	OBVIOUS IS CLEAR
WordNet	verbs	yes	0.75	The man kidnapped their solution	INTEREST IS CAPTURING
WordNet	verbs	yes	0.78	They melted the alliance	DECEIT IS DOUBLENESS
WordNet	verbs	yes	0.74	The engine frayed out	UNKNOWN IS COVERED
WordNet	verbs	yes	0.76	The boats danced on the shore	A HUMAN IS A SHIP
WordNet	verbs	yes	0.37	The boy grabbed his bike and flew home	COMPETITION IS SPEED
WordNet	verbs	yes	0.76	The clouds swarmed on the horizon	A GROUP IS A BODY
WordNet	verbs	yes	0.76	The runners streamed through the streets	A CROWD IS A LIQUID
WordNet	verbs	yes	0.69	The bike crawled along the trail	CONTROLING IS PUSHING
WordNet	verbs	yes	0.93	The plants obeyed the constraints	A SYSTEM IS A PLANT
WordNet	verbs	yes	0.75	The van was sleeping on the road	A MACHINE IS AN ANIMAL
WordNet	verbs	yes	0.67	The truck soared down the slope	COMPETITION IS SPEED
WordNet	verbs	yes	0.85	The house wilted over time	AN IDEA IS A BUILDING
WordNet	verbs	yes	0.82	He unlocked her old wound	FREEDOM IS A RELEASE
WordNet	verbs	yes	0.66	The poster hovered over the desk	CHANGE IS MOVEMENT
BEAGLE-1024	noun	no	0.12	A lion is an animal	A HUMAN IS A CAT
BEAGLE-1024	noun	no	0.09	A salmon is a fish	A FISH IS FISHING
BEAGLE-1024	noun	no	0.62	A smile is a attractive	ANNOYANCE IS FRICTION
BEAGLE-1024	noun	no	0.60	A snail is a pest	A FISH IS AN INSECT
BEAGLE-1024	noun	no	0.10	Cereal is a food	BREAD IS FOOD
BEAGLE-1024	noun	no	0.75	Crime is a problem	AN EXPERIENCE IS A RELATIONSHIP
BEAGLE-1024	noun	no	0.60	My brother is a butcher	BEING BIG IS RELINQUISHING
BEAGLE-1024	noun	no	0.79	Some apartments are big	A MIND IS A BUILDING
BEAGLE-1024	noun	no	0.73	Some computer programs are complex	A SYSTEM IS A MACHINE
BEAGLE-1024	noun	no	0.74	Some countries are unsafe	DEVELOPMENT IS MOVEMENT
BEAGLE-1024	noun	no	0.89	Some ideas are great	AN IDEA IS A PLACE
BEAGLE-1024	noun	no	0.78	Some jobs are constraining	HELP IS A SUPPORT
BEAGLE-1024	noun	no	0.45	Some urban schools are crowded	STATUS IS HEIGHT
BEAGLE-1024	noun	no	0.31	That lost painting is a portrait	A FLOWER IS SERIOUSNESS
BEAGLE-1024	noun	no	0.07	The Earth is a planet	SPACE IS THE EARTH
BEAGLE-1024	noun	yes	0.69	Crime is a disease	A LAW IS A DISEASE
BEAGLE-1024	noun	yes	0.65	The mind is a computer	THINKING IS CALCULATING
BEAGLE-1024	noun	yes	0.87	Some ideas are diamonds	AN IDEA IS AN IMPRESSION
BEAGLE-1024	noun	yes	0.64	A smile is a magnet	THOUGHT IS ELECTRICITY
BEAGLE-1024	noun	yes	0.71	Experience is a fountain	LISTENING IS DRINKING
BEAGLE-1024	noun	yes	0.61	Beauty is a ticket	BEING BIG IS DISINTEREST
BEAGLE-1024	noun	yes	0.69	Love is a journey	BAD IS SMELLY
BEAGLE-1024	noun	yes	0.77	Some jobs are prisons	AN IDEA IS A BUILDING
BEAGLE-1024	noun	yes	0.79	An education is a doorway	UNKNOWN IS OPEN
BEAGLE-1024	noun	yes	0.70	Angry words are knives	AN EMOTION IS A SOUND
BEAGLE-1024	verb	no	0.69	The engine wore out	AN ENGINE IS WHITE
BEAGLE-1024	verb	no	0.72	The ancient car fell apart	STEALING IS HITTING
BEAGLE-1024	verb	no	0.69	The boats moved along shore	TIME IS TRAVEL
BEAGLE-1024	verb	no	0.71	The boy grabbed his bike and went home	THINKING IS WALKING
BEAGLE-1024	verb	no	0.43	The clouds gathered on the horizon	COLDNESS IS A CROWD
BEAGLE-1024	verb	no	0.62	The runners ran through the streets	BEING BIG IS BEING UNAWARE
BEAGLE-1024	verb	no	0.64	The bike moved along the trail	TIME IS TRAVEL
BEAGLE-1024	verb	no	0.49	The poster hung over the desk	CORRECTNESS IS HOLDING
BEAGLE-1024	verb	no	0.63	The bread rose to perfection	GOING THROUGH IS MILK
BEAGLE-1024	verb	no	0.64	The house decayed over time	A PLANT IS BIG
BEAGLE-1024	verb	no	0.62	She opened the gate	A ROAD IS AN OPENING
BEAGLE-1024	verb	no	0.83	She cleaned up the spill	GOOD IS CLEAN

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
BEAGLE-1024	verb	no	0.46	She delivered a message	DETERIORATION IS COMMUNICATION
BEAGLE-1024	verb	no	0.78	The doctor mended the cut	GOOD IS WHITE
BEAGLE-1024	verb	no	0.46	The woman rejected the proposal	LAW IS REPRESSION
BEAGLE-1024	verb	no	0.32	The man stole their solution	UNCHANGING IS THINKING
BEAGLE-1024	verb	no	0.58	They released the prisoner	INJURY IS RELEASE
BEAGLE-1024	verb	yes	0.55	He piloted his dance partner	UNAWARENESS IS SHARPNESS
BEAGLE-1024	verb	yes	0.76	She devoured the new material	A PROCESS IS A BODY
BEAGLE-1024	verb	yes	0.59	He unlocked his old wound	AN OPENING IS PREY
BEAGLE-1024	verb	yes	0.61	The woman killed the proposal	DEAD IS UNSUCCESSFUL
BEAGLE-1024	verb	yes	0.90	The man kidnapped their solution	A SOLUTION IS A WAY
BEAGLE-1024	verb	yes	0.68	The boats danced on the shore	SADNESS IS DARK
BEAGLE-1024	verb	yes	0.70	The boy grabbed his bike and flew home	KNOWING IS SEEING
BEAGLE-1024	verb	yes	0.81	The runners streamed through the streets	MEANS ARE A ROAD
BEAGLE-1024	verb	yes	0.66	The plants obeyed the constraints	CALCULATING IS OBEYING
BEAGLE-1024	verb	yes	0.66	The van was sleeping on the road	INACTIVITY IS SLOWNESS
BEAGLE-1024	verb	yes	0.84	The house wilted over time	A LIFE IS A DAY
BEAGLE-1024	verb	yes	0.59	He unlocked her old wound	AN OPENING IS PREY
BEAGLE-128	noun	no	0.56	A lion is an animal	A HUMAN IS MEAT
BEAGLE-128	noun	no	0.64	My brother is a butcher	BEING BIG IS SWEETNESS
BEAGLE-128	noun	no	0.61	A snail is a pest	A FISH IS AN INSECT
BEAGLE-128	noun	no	0.09	A salmon is a fish	A FISH IS FISHING
BEAGLE-128	noun	no	0.26	Sharks have sharp teeth	POWERLESS IS TIME
BEAGLE-128	noun	no	0.60	An earthquake is a disaster	BALANCE IS A CATEGORY
BEAGLE-128	noun	no	0.13	Cereal is a food	A HUMAN IS BREAD
BEAGLE-128	noun	no	0.53	That creature in the net is a crab	ANTAGONISM IS A FISH
BEAGLE-128	noun	no	0.70	That lost painting is a portrait	REPRESSION IS A SPACE
BEAGLE-128	noun	no	0.88	Crime is a problem	A PROBLEM IS A DISEASE
BEAGLE-128	noun	no	0.81	Some apartments are big	A MIND IS A BUILDING
BEAGLE-128	noun	no	0.47	The Earth is a planet	A BODY IS THE EARTH
BEAGLE-128	noun	no	0.77	Some countries are unsafe	A STATE IS A PLACE
BEAGLE-128	noun	no	0.52	Some urban schools are crowded	QUALITY IS HIGH
BEAGLE-128	noun	no	0.55	Some computer programs are complex	A SYSTEM IS A MACHINE
BEAGLE-128	noun	no	0.91	Some ideas are great	AN IDEA IS A PLACE
BEAGLE-128	noun	no	0.74	A smile is a attractive	BAD IS SMELLY
BEAGLE-128	noun	no	0.81	Some colleges are pretty	HAPPINESS IS HIGH
BEAGLE-128	noun	no	0.75	Some jobs are constraining	A GROUP IS A BUILDING
BEAGLE-128	noun	no	0.61	Some lectures are boring	A FRUIT IS A SHOOTING
BEAGLE-128	noun	yes	0.70	Crime is a disease	AN EFFECT IS A DISEASE
BEAGLE-128	noun	yes	0.60	A vacation is medicine	PREVENTION IS COOKING
BEAGLE-128	noun	yes	0.59	Dancers are butterflies	A CONCEPTION IS A FLOWER
BEAGLE-128	noun	yes	0.72	His life is an opera	TO REDUCE IS TO LOWER
BEAGLE-128	noun	yes	0.68	Beggars are parasites	COLOUR IS A PLANT
BEAGLE-128	noun	yes	0.70	The mind is a computer	THINKING IS CALCULATING
BEAGLE-128	noun	yes	0.88	Some ideas are diamonds	AN IDEA IS AN IMPRESSION
BEAGLE-128	noun	yes	0.71	A smile is a magnet	AN EMOTION IS ELECTRICITY
BEAGLE-128	noun	yes	0.89	Experience is a fountain	AN EXPERIENCE IS A RELATIONSHIP
BEAGLE-128	noun	yes	0.75	Beauty is a ticket	SADNESS IS DARK
BEAGLE-128	noun	yes	0.75	Love is a journey	A LIFE IS A PATH
BEAGLE-128	noun	yes	0.60	Some malls are jungles	PREVENTION IS EXISTENCE
BEAGLE-128	noun	yes	0.75	Some jobs are prisons	A GROUP IS A BUILDING
BEAGLE-128	noun	yes	0.61	Alcohol is a crutch	DRINKING IS DISINTEREST
BEAGLE-128	noun	yes	0.79	An education is a doorway	KNOWLEDGE IS OPEN
BEAGLE-128	noun	yes	0.69	Angry words are knives	WORDS ARE CUTS
BEAGLE-128	noun	yes	0.51	Humor is a weapon	AN APPETITE IS A MACHINE
BEAGLE-128	verb	no	0.69	The engine wore out	WHITENESS IS AN ENGINE
BEAGLE-128	verb	no	0.72	The ancient car fell apart	SADNESS IS DARK

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
BEAGLE-128	verb	no	0.69	The boats moved along shore	TIME IS SPACE
BEAGLE-128	verb	no	0.57	The boy grabbed his bike and went home	AN OPENING IS BIG
BEAGLE-128	verb	no	0.87	The building shook from the earthquake	A TEXT IS A BUILDING
BEAGLE-128	verb	no	0.60	The clouds gathered on the horizon	INFORMATION IS A MINERAL
BEAGLE-128	verb	no	0.62	The runners ran through the streets	A ROAD IS A PITCH
BEAGLE-128	verb	no	0.64	The bike moved along the trail	TIME IS SPACE
BEAGLE-128	verb	no	0.51	The poster hung over the desk	HOLDING IS UNCOVERING
BEAGLE-128	verb	no	0.65	The bread rose to perfection	SEEING IS MILK
BEAGLE-128	verb	no	0.64	The house decayed over time	BIRTH IS BEING BIG
BEAGLE-128	verb	no	0.65	She opened the gate	AN EMOTION IS A TOUCH
BEAGLE-128	verb	no	0.46	She cleaned up the spill	CLEANLINESS IS MILK
BEAGLE-128	verb	no	0.70	She delivered a message	HELP IS A SUPPORT
BEAGLE-128	verb	no	0.69	The doctor mended the cut	EYESIGHT IS A CUT
BEAGLE-128	verb	no	0.59	The woman rejected the proposal	SADNESS IS DISCOMFORT
BEAGLE-128	verb	no	0.46	The man stole their solution	RACE IS A COLOUR
BEAGLE-128	verb	no	0.60	They withdrew the invitation	SADNESS IS AN EXPANSE
BEAGLE-128	verb	no	0.70	They released the prisoner	PASSION IS HEAT
BEAGLE-128	verb	yes	0.58	He buckled a bandage	PURITY IS CEASING
BEAGLE-128	verb	yes	0.68	He piloted his dance partner	FREEDOM IS A DANCE
BEAGLE-128	verb	yes	0.76	She devoured the new material	A PROCESS IS A BODY
BEAGLE-128	verb	yes	0.64	He unlocked his old wound	STRAIGHTNESS IS A THOUGHT
BEAGLE-128	verb	yes	0.70	She shot him a message	A TEXT IS A PATH
BEAGLE-128	verb	yes	0.74	The woman killed the proposal	EXISTENCE IS HEIGHT
BEAGLE-128	verb	yes	0.63	The man kidnapped their solution	HONESTY IS BIG
BEAGLE-128	verb	yes	0.61	They melted the alliance	BIRTH IS BEING HOT
BEAGLE-128	verb	yes	0.69	The engine frayed out	AN ORGANISATION IS AN ENGINE
BEAGLE-128	verb	yes	0.70	The boats danced on the shore	JUSTICE IS STRAIGHT
BEAGLE-128	verb	yes	0.69	The boy grabbed his bike and flew home	KNOWING IS SEEING
BEAGLE-128	verb	yes	0.52	The clouds swarmed on the horizon	A LIQUID IS A PERSPECTIVE
BEAGLE-128	verb	yes	0.81	The runners streamed through the streets	MEANS ARE A ROAD
BEAGLE-128	verb	yes	0.60	The bike crawled along the trail	SWEETNESS IS AN OPENING
BEAGLE-128	verb	yes	0.67	The plants obeyed the constraints	OBEYING IS ORIENTATION
BEAGLE-128	verb	yes	0.80	The van was sleeping on the road	THINKING IS WALKING
BEAGLE-128	verb	yes	0.59	The truck soared down the slope	FORWARDS IS UNRELIABLE
BEAGLE-128	verb	yes	0.63	The house wilted over time	A PLANT IS A PLACE
BEAGLE-128	verb	yes	0.60	He unlocked her old wound	DOUGH IS STRAIGHT
BEAGLE-128	verb	yes	0.59	The poster hovered over the desk	BEING UNCOVERED IS BEING PERIPHERAL
COALS-14000	noun	no	0.17	A lion is an animal	A PLANT IS AN ANIMAL
COALS-14000	noun	no	0.57	A snail is a pest	A FISH IS AN INSECT
COALS-14000	noun	no	0.08	A salmon is a fish	SWIMMING IS A FISH
COALS-14000	noun	no	0.07	Sharks have sharp teeth	CUTTING IS BELIEVING
COALS-14000	noun	no	0.54	An earthquake is a disaster	EARTH IS UNSUCCESSFUL
COALS-14000	noun	no	0.09	Cereal is a food	BREAD IS FOOD
COALS-14000	noun	no	0.55	That creature in the net is a crab	PREY IS A FISH
COALS-14000	noun	no	0.27	That lost painting is a portrait	CLOTHES ARE A LANDSCAPE
COALS-14000	noun	no	0.59	Crime is a problem	AN EXPERIENCE IS A RELATIONSHIP
COALS-14000	noun	no	0.57	Some apartments are big	ATTRACTION IS A DOG
COALS-14000	noun	no	0.03	The Earth is a planet	UNKNOWN IS INVISIBLE
COALS-14000	noun	no	0.55	Some countries are unsafe	DRINKING IS WAR
COALS-14000	noun	no	0.19	Some urban schools are crowded	WORK IS AGRICULTURE
COALS-14000	noun	no	0.33	Some computer programs are complex	A SYSTEM IS A MACHINE
COALS-14000	noun	no	0.80	Some ideas are great	AN IDEA IS AN IMPRESSION
COALS-14000	noun	no	0.62	A smile is a attractive	KNOWING IS SEEING
COALS-14000	noun	no	0.60	Some colleges are pretty	HAPPINESS IS HIGH
COALS-14000	noun	no	0.55	Some lectures are boring	A THEATRE IS A FEELING
COALS-14000	noun	yes	0.57	Crime is a disease	A HUMAN IS AN INSECT

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
COALS-14000	noun	yes	0.55	A vacation is medicine	A DAY IS HEALTH
COALS-14000	noun	yes	0.55	Dancers are butterflies	A DANCE IS AN INSECT
COALS-14000	noun	yes	0.56	His life is an opera	LIVING IS THEATRE
COALS-14000	noun	yes	0.60	Beggars are parasites	A PLANT IS AN ANIMAL
COALS-14000	noun	yes	0.52	The mind is a computer	A SYSTEM IS A MACHINE
COALS-14000	noun	yes	0.80	Some ideas are diamonds	AN IDEA IS AN IMPRESSION
COALS-14000	noun	yes	0.62	A smile is a magnet	KNOWING IS SEEING
COALS-14000	noun	yes	0.80	Experience is a fountain	AN EXPERIENCE IS AN IMPRESSION
COALS-14000	noun	yes	0.55	Beauty is a ticket	HAPPINESS IS PUNCHING
COALS-14000	noun	yes	0.58	Love is a journey	TIME IS TRAVEL
COALS-14000	noun	yes	0.65	Some jobs are prisons	A JOB IS DEATH
COALS-14000	noun	yes	0.59	Alcohol is a crutch	A PROBLEM IS A DISEASE
COALS-14000	noun	yes	0.60	An education is a doorway	AN OPPORTUNITY IS AN OPENING
COALS-14000	noun	yes	0.57	Angry words are knives	ARGUING IS CUTTING
COALS-14000	noun	yes	0.54	Faith is a fortress	TRUTH IS ATTACKING
COALS-14000	noun	yes	0.65	Humor is a weapon	AN EXPRESSION IS A WEAPON
COALS-14000	verb	no	0.56	The engine wore out	ELEMENTARY IS LOW
COALS-14000	verb	no	0.57	The ancient car fell apart	THE PAST IS BACKWARDS
COALS-14000	verb	no	0.40	The boats moved along shore	ELECTRICITY IS A LIQUID
COALS-14000	verb	no	0.34	The boy grabbed his bike and went home	KNOWING IS SEEING
COALS-14000	verb	no	0.56	The building shook from the earthquake	THE PAST IS BACKWARDS
COALS-14000	verb	no	0.32	The clouds gathered on the horizon	WATER IS HUNTING
COALS-14000	verb	no	0.56	The runners ran through the streets	BEING BIG IS A RACE
COALS-14000	verb	no	0.30	The bike moved along the trail	ELECTRICITY IS A LIQUID
COALS-14000	verb	no	0.44	The poster hung over the desk	BEING BIG IS THEATRE
COALS-14000	verb	no	0.66	The bread rose to perfection	BREAD IS FRUIT
COALS-14000	verb	no	0.56	The house decayed over time	GROWING IS BIG
COALS-14000	verb	no	0.56	She opened the gate	AN OPENING IS A HORSE
COALS-14000	verb	no	0.38	She cleaned up the spill	CLEANLINESS IS UNCOVERED
COALS-14000	verb	no	0.55	The doctor mended the cut	PREVENTION IS AN OBSTACLE
COALS-14000	verb	no	0.43	The woman rejected the proposal	A STATE IS A SUPPORT
COALS-14000	verb	no	0.23	The man stole their solution	STEALING IS THINKING
COALS-14000	verb	no	0.56	They released the prisoner	RELEASE IS DEATH
COALS-14000	verb	yes	0.61	He piloted his dance partner	A RELATIONSHIP IS MUSIC
COALS-14000	verb	yes	0.61	She devoured the new material	ELECTRICITY IS A LIQUID
COALS-14000	verb	yes	0.55	He unlocked his old wound	WORTHLESS IS EMPTINESS
COALS-14000	verb	yes	0.54	He unlocked his old wound	AN OPENING IS A CUT
COALS-14000	verb	yes	0.59	She shot him a message	SPEECH IS A GAME
COALS-14000	verb	yes	0.55	The woman killed the proposal	FIGHTING IS A STATE
COALS-14000	verb	yes	0.62	The man kidnapped their solution	A CROWD IS A LIQUID
COALS-14000	verb	yes	0.56	They melted the alliance	BEING BROKEN IS BEING LOW
COALS-14000	verb	yes	0.56	The engine frayed out	ELEMENTARY IS LOW
COALS-14000	verb	yes	0.66	The boats danced on the shore	A SHIP IS A DANCE
COALS-14000	verb	yes	0.32	The boy grabbed his bike and flew home	KNOWING IS SEEING
COALS-14000	verb	yes	0.61	The runners streamed through the streets	TRAFFIC IS BLOOD
COALS-14000	verb	yes	0.54	The bike crawled along the trail	A HORSE IS A REPTILE
COALS-14000	verb	yes	0.58	The plants obeyed the constraints	FREEDOM IS A RELEASE
COALS-14000	verb	yes	0.60	The van was sleeping on the road	THINKING IS WALKING
COALS-14000	verb	yes	0.54	The truck soared down the slope	A ROAD IS A RISE
COALS-14000	verb	yes	0.54	He unlocked her old wound	AN OPENING IS AMMUNITION
COALS-14000	verb	yes	0.59	The house wilted over time	THINKING IS WALKING
HAL	noun	no	0.01	A lion is an animal	A HUMAN IS MEAT
HAL	noun	no	0.51	My brother is a butcher	BEING BIG IS FLOWERING
HAL	noun	no	0.51	A snail is a pest	SWIMMING IS AN INSECT
HAL	noun	no	0.07	A salmon is a fish	MEAT IS A FISH
HAL	noun	no	0.53	An earthquake is a disaster	HEIGHT IS A DEFENSE

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
HAL	noun	no	0.06	Cereal is a food	BREAD IS FOOD
HAL	noun	no	0.50	That creature in the net is a crab	AN INSECT IS MEAT
HAL	noun	no	0.23	That lost painting is a portrait	MUSIC IS AN EXPRESSION
HAL	noun	no	0.51	Crime is a problem	A GROUP IS A BODY
HAL	noun	no	0.51	Some apartments are big	TIME IS SPACE
HAL	noun	no	0.01	The Earth is a planet	TIME IS SPACE
HAL	noun	no	0.51	Some countries are unsafe	TRANSPORTING IS THOUGHT
HAL	noun	no	0.02	Some urban schools are crowded	TIME IS TRAVEL
HAL	noun	no	0.16	Some computer programs are complex	CONTROL IS HANDLING
HAL	noun	no	0.52	Some ideas are great	KNOWING IS SEEING
HAL	noun	no	0.52	A smile is a attractive	THINKING IS WALKING
HAL	noun	no	0.51	Some colleges are pretty	HAPPINESS IS HIGH
HAL	noun	no	0.51	Some jobs are constraining	IMPORTANT IS BIG
HAL	noun	yes	0.51	A vacation is medicine	ELEMENTARY IS A DISEASE
HAL	noun	yes	0.51	Dancers are butterflies	A DANCE IS AN INSECT
HAL	noun	yes	0.52	Beggars are parasites	A MONKEY IS AN INSECT
HAL	noun	yes	0.41	The mind is a computer	THOUGHT IS TEXT
HAL	noun	yes	0.51	Some ideas are diamonds	IMPORTANT IS BIG
HAL	noun	yes	0.55	A smile is a magnet	THINKING IS CALCULATING
HAL	noun	yes	0.51	Experience is a fountain	A PROCESS IS MOVEMENT
HAL	noun	yes	0.53	Beauty is a ticket	BEING BIG IS INJURING
HAL	noun	yes	0.53	Love is a journey	A POSSIBILITY IS AN OPENING
HAL	noun	yes	0.51	Some jobs are prisons	IMPORTANT IS BIG
HAL	noun	yes	0.51	An education is a doorway	ACTIVITY IS A PATH
HAL	noun	yes	0.63	Angry words are knives	WORDS ARE WEAPONS
HAL	noun	yes	0.54	Humor is a weapon	IMPRESSIONS ARE WEAPONS
HAL	verb	no	0.63	The engine wore out	AN ENGINE IS CLOTHING
HAL	verb	no	0.51	The ancient car fell apart	A HUMAN IS A FISH
HAL	verb	no	0.28	The boats moved along shore	A SOUND IS A LIQUID
HAL	verb	no	0.16	The boy grabbed his bike and went home	KNOWING IS SEEING
HAL	verb	no	0.28	The clouds gathered on the horizon	A LIQUID IS HAPPINESS
HAL	verb	no	0.52	The runners ran through the streets	KNOWING IS SEEING
HAL	verb	no	0.15	The bike moved along the trail	IMPORTANT IS BIG
HAL	verb	no	0.53	The bread rose to perfection	UNIMPORTANT IS LOW
HAL	verb	no	0.55	The house decayed over time	NUMEROUS IS BIG
HAL	verb	no	0.54	She opened the gate	AN OPENING IS A HORSE
HAL	verb	no	0.14	She cleaned up the spill	A MIND IS A CONTAINER
HAL	verb	no	0.40	She delivered a message	READING IS A FASHION
HAL	verb	no	0.52	The doctor mended the cut	UNCHANGING IS HARD
HAL	verb	no	0.38	The woman rejected the proposal	MAKING BETTER IS RISING
HAL	verb	no	0.17	The man stole their solution	MAGIC IS GOODNESS
HAL	verb	no	0.55	They withdrew the invitation	DEFENSE IS FASHION
HAL	verb	no	0.39	They released the prisoner	CRITICISING IS FIGHTING
HAL	verb	yes	0.51	She devoured the new material	A HUMAN IS A PLANT
HAL	verb	yes	0.51	He unlocked his old wound	A CUT IS A DAY
HAL	verb	yes	0.21	The woman killed the proposal	WEAPONS ARE A DESCENT
HAL	verb	yes	0.77	The man kidnapped their solution	A SOLUTION IS A WAY
HAL	verb	yes	0.52	They melted the alliance	METAL IS A DEFENSE
HAL	verb	yes	0.51	The boats danced on the shore	LISTENING IS DRINKING
HAL	verb	yes	0.14	The boy grabbed his bike and flew home	A LIFE IS A PATH
HAL	verb	yes	0.50	The clouds swarmed on the horizon	A LIQUID IS AN INSECT
HAL	verb	yes	0.51	The runners streamed through the streets	CURRENT IS FRICTION
HAL	verb	yes	0.63	The plants obeyed the constraints	OBEYING IS GRASS
HAL	verb	yes	0.51	The van was sleeping on the road	A MIND IS A BUILDING
HAL	verb	yes	0.53	The truck soared down the slope	DOWNWARD IS FASHION
HAL	verb	yes	0.51	The house wilted over time	A GROUP IS A BUILDING

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
LSA-400	noun	no	0.18	A lion is an animal	A HUMAN IS MEAT
LSA-400	noun	no	0.28	Sharks have sharp teeth	PERIPHERY IS A WAY
LSA-400	noun	no	0.14	Cereal is a food	UNHEALTHY IS LOW
LSA-400	noun	no	0.52	That creature in the net is a crab	PASTA IS STAGNATION
LSA-400	noun	no	0.93	Crime is a problem	A PROBLEM IS A DISEASE
LSA-400	noun	no	0.79	Some apartments are big	PASSION IS HEAT
LSA-400	noun	no	0.13	The Earth is a planet	CORRECTNESS IS POINTING
LSA-400	noun	no	0.84	Some countries are unsafe	DEVELOPMENT IS MOVEMENT
LSA-400	noun	no	0.56	Some urban schools are crowded	TO IMITATE IS TO FOLLOW
LSA-400	noun	no	0.59	Some computer programs are complex	UNSUCCESSFUL IS SIMILARITY
LSA-400	noun	no	0.92	Some ideas are great	TO UNDERSTAND IS TO SEE
LSA-400	noun	no	0.58	A smile is a attractive	FIXING IS CALCULATING
LSA-400	noun	no	0.59	Some colleges are pretty	SUPERNATURAL IS FIXING
LSA-400	noun	no	0.65	Some jobs are constraining	ARGUING IS PUNCHING
LSA-400	noun	no	0.56	Some lectures are boring	WORTHLESSNESS IS IMMOBILITY
LSA-400	noun	yes	0.57	A vacation is medicine	PROXIMITY IS NORMALITY
LSA-400	noun	yes	0.55	The mind is a computer	TIME IS SOLIDNESS
LSA-400	noun	yes	0.85	Some ideas are diamonds	A SUBSTANCE IS A HUMAN
LSA-400	noun	yes	0.57	A smile is a magnet	EYESIGHT IS REPRESSION
LSA-400	noun	yes	0.65	Some jobs are prisons	ARGUING IS PUNCHING
LSA-400	noun	yes	0.64	An education is a doorway	SIMILARITY IS PROXIMITY
LSA-400	noun	yes	0.88	Angry words are knives	LANGUAGE IS A PERSON
LSA-400	verb	no	0.57	The engine wore out	AMMUNITION IS A SHIPWRECK
LSA-400	verb	no	0.65	The ancient car fell apart	UNCHANGING IS STATIC
LSA-400	verb	no	0.87	The boats moved along shore	IMPORTANT IS BIG
LSA-400	verb	no	0.54	The boy grabbed his bike and went home	FIRMNESS IS A MIND
LSA-400	verb	no	0.66	The building shook from the earthquake	REPRESSION IS BINDING
LSA-400	verb	no	0.38	The clouds gathered on the horizon	A SHIPWRECK IS INVOLVEMENT
LSA-400	verb	no	0.92	The runners ran through the streets	IMPORTANT IS BIG
LSA-400	verb	no	0.85	The bike moved along the trail	IMPORTANT IS BIG
LSA-400	verb	no	0.58	The bread rose to perfection	DOUGH IS A TEXTURE
LSA-400	verb	no	0.83	The house decayed over time	AN ARGUMENT IS A BUILDING
LSA-400	verb	no	0.83	She opened the gate	AN EXPRESSION IS HIGH
LSA-400	verb	no	0.70	She cleaned up the spill	A HUMAN IS A VEGETABLE
LSA-400	verb	no	0.47	She delivered a message	REPRESSION IS PESSIMISM
LSA-400	verb	no	0.62	The woman rejected the proposal	PESSIMISM IS FAME
LSA-400	verb	no	0.94	The man stole their solution	A SOLUTION IS A WAY
LSA-400	verb	yes	0.57	He piloted his dance partner	PERIPHERY IS A SHIPWRECK
LSA-400	verb	yes	0.95	She devoured the new material	A HUMAN IS A MATERIAL
LSA-400	verb	yes	0.65	He unlocked his old wound	UNCERTAINTY IS INSTABILITY
LSA-400	verb	yes	0.74	The woman killed the proposal	EXCITEMENT IS A COLOUR
LSA-400	verb	yes	0.94	The man kidnapped their solution	A SOLUTION IS A WAY
LSA-400	verb	yes	0.27	The boy grabbed his bike and flew home	DETERIORATION IS MINDING
LSA-400	verb	yes	0.88	The runners streamed through the streets	THINKING IS WALKING
LSA-400	verb	yes	0.55	The bike crawled along the trail	INCOMPREHENSIBLE IS UNCONSCIOUSNESS
LSA-400	verb	yes	0.69	The plants obeyed the constraints	IMPOLITENESS IS A PLANT
LSA-400	verb	yes	0.83	The house wilted over time	AN ARGUMENT IS A BUILDING
LSA-400	verb	yes	0.65	He unlocked her old wound	UNCERTAINTY IS INSTABILITY
LSA-500	noun	no	0.15	A lion is an animal	ACTIVITY IS HUNTING
LSA-500	noun	no	0.58	My brother is a butcher	A PIG IS STAGNATION
LSA-500	noun	no	0.24	Sharks have sharp teeth	INCOMPREHENSIBLE IS UNCLEAR
LSA-500	noun	no	0.17	Cereal is a food	A HUMAN IS MEAT
LSA-500	noun	no	0.35	That lost painting is a portrait	STRAIGHTNESS IS CHANGEABLE
LSA-500	noun	no	0.93	Crime is a problem	A PROBLEM IS A WEIGHT
LSA-500	noun	no	0.74	Some apartments are big	DESIRE IS ATTRACTION
LSA-500	noun	no	0.12	The Earth is a planet	UNCONSCIOUSNESS IS LOW

Continued on next page

Table C.4 continued

Model	Type	Fig?	Score	Sentence	Candidate Metaphor
LSA-500	noun	no	0.83	Some countries are unsafe	TO REDUCE IS TO LOWER
LSA-500	noun	no	0.25	Some computer programs are complex	HOSTILITY IS DETERIORATION
LSA-500	noun	no	0.93	Some ideas are great	KNOWING IS SEEING
LSA-500	noun	no	0.80	Some jobs are constraining	A CAUSE IS A CONNECTION
LSA-500	noun	no	0.56	Some lectures are boring	STAGNATION IS ALIGNMENT
LSA-500	noun	no	0.62	Some urban schools are crowded	ELEMENTARY IS LOW
LSA-500	noun	yes	0.67	The mind is a computer	TO DETERIORATE IS TO LOWER
LSA-500	noun	yes	0.87	Some ideas are diamonds	A LIFE IS WRITING
LSA-500	noun	yes	0.58	Beauty is a ticket	ARGUMENTS ARE INACTIVITY
LSA-500	noun	yes	0.64	Love is a journey	INCOMPREHENSIBLE IS CROOKED
LSA-500	noun	yes	0.80	Some jobs are prisons	A CAUSE IS A CONNECTION
LSA-500	noun	yes	0.81	An education is a doorway	ELEMENTARY IS LOW
LSA-500	noun	yes	0.77	Angry words are knives	A TEXT IS A STRUCTURE
LSA-500	verb	no	0.68	The engine wore out	AN ENGINE IS A CALCULATION
LSA-500	verb	no	0.67	The ancient car fell apart	A ROAD IS LOUDNESS
LSA-500	verb	no	0.87	The boats moved along shore	KNOWING IS SEEING
LSA-500	verb	no	0.80	The boy grabbed his bike and went home	GOOD IS CLEAN
LSA-500	verb	no	0.69	The building shook from the earthquake	ARGUMENTS ARE AMMUNITION
LSA-500	verb	no	0.36	The clouds gathered on the horizon	HARDNESS IS TRAVELLING
LSA-500	verb	no	0.91	The runners ran through the streets	IMPORTANT IS BIG
LSA-500	verb	no	0.85	The bike moved along the trail	KNOWING IS SEEING
LSA-500	verb	no	0.57	The bread rose to perfection	DOUGH IS A THEATRE
LSA-500	verb	no	0.71	The house decayed over time	DIFFICULTY IS HARDNESS
LSA-500	verb	no	0.79	She opened the gate	WORK IS AGRICULTURE
LSA-500	verb	no	0.26	She cleaned up the spill	GOODNESS IS IRRELEVANCE
LSA-500	verb	no	0.48	She delivered a message	CHANGEABLE IS UNIMPORTANCE
LSA-500	verb	no	0.56	The woman rejected the proposal	ARGUING IS PUNCHING
LSA-500	verb	no	0.93	The man stole their solution	A SOLUTION IS A WAY
LSA-500	verb	no	0.45	They released the prisoner	RELIABILITY IS REPRESSION
LSA-500	verb	yes	0.56	He piloted his dance partner	UNCLEAR IS COHESION
LSA-500	verb	yes	0.88	She devoured the new material	KNOWLEDGE IS FOOD
LSA-500	verb	yes	0.71	He unlocked his old wound	CRITICISING IS CUTTING
LSA-500	verb	yes	0.63	The woman killed the proposal	ARGUING IS WOUNDING
LSA-500	verb	yes	0.93	The man kidnapped their solution	A SOLUTION IS A WAY
LSA-500	verb	yes	0.56	The boats danced on the shore	SUCCESS IS COHESION
LSA-500	verb	yes	0.44	The boy grabbed his bike and flew home	AN ORGANISATION IS A SHIP
LSA-500	verb	yes	0.81	The runners streamed through the streets	A LANDSCAPE IS A BODY
LSA-500	verb	yes	0.59	The plants obeyed the constraints	REPUTE IS DIRT
LSA-500	verb	yes	0.94	The house wilted over time	KNOWING IS SEEING
LSA-500	verb	yes	0.71	He unlocked her old wound	CRITICISING IS CUTTING

Table C.4: Full set of materials used in experiment 2. Fig? denotes whether the sentence is figurative or not and the score is MetID's top-scoring candidate metaphor's score, which was presented in grammatical form as a paraphrase. Note that participants received a random 54 questions from this list, with 6 questions being planted to verify participants' understanding of the study.

Appendix D

LSA Example

All the DSMs are based on the *distributional hypothesis*, but as we have seen, a number of implementations exist. Most DSMs represent meaning as a multi-dimensional feature-vector built from co-occurrences information. These models can be illustrated using an example for LSA [137]. Note that DSMs, are designed to work on large collections of texts, not on the small scale of this example.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Table D.1: Word-document co-occurrence frequencies, $\{X\}$, in the example set documents $d_{1..9}$. Each row lists the number of times the given word occurred in each document. This raw occurrence matrix is the starting point for a number of DSMs, including LSA. Adapted from [137].

LSA starts with a word-document occurrence matrix, $\{X\}$, shown in table D.1. After $\{X\}$ is created from a set of documents, each cell is incremented and its log is taken. Next, the entropy of each row, $-\sum p \log p$ over all entries, is used to normalise the values. The third and defining step in LSA, is to decompose $\{X\}$ using singular value decomposition, which is a kind of principle components analysis. SVD factors the matrix into three matrices such that $\{X\} = \{W\}\{\delta_n\}\{S\}$[1] where each value in this $\{W\}$ and $\{S\}$ is the linear combination of values in the number of desired dimensions, n. By multiplying each value of the decomposed representation, an n-dimensional matrix, $\{\hat{X}\}$, can be re-constructed that represents the condensed semantic space, based on the frequency data from table D.1. This resulting matrix is a set of word "features" which can be analysed by comparing words' vectors using their cosine values. Table D.2 shows the result of a 2-dimensional reconstruction of $\{X\}$ into $\{\hat{X}\}$. Notice that with this method, changing any value in $\{X\}$ will change the entire space of $\{\hat{X}\}$. This is what the developers consider the "latent" effect words have on a semantic space. Observe the resulting correlation between *human* and *user* in $\{\hat{X}\}$ ($r = .94$), which was non-existent in the original co-occurrence matrix.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
human	0.16	0.47	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.09	-0.15	-0.21	0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.27	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Table D.2: The 2-dimensional representation of the semantic space, $\{\hat{X}\}$, resulting from the decomposition of $\{X\}$ (table D.1). Each row is the vector used to define the word, here in two dimensions. Adapted from [137].)

After $\{\hat{X}\}$ is constructed, the result is what is referred to as the semantic space for the corpus C. In this example, the resulting space is two-dimensional, making it easy to present (and visualise), but more often it is a hyper-space of 100, 300, 500 or even 1,000 dimensions. The strength of LSA and related models is that the initial representation is constructed entirely from frequency observations across a collection of texts. Additionally, the statistical manipulations applied to the data result in a representation that generalises definitions of words based on their use.

[1] δ denotes the identity matrix.