Technical Report

UCAM-CL-TR-803
ISSN 1476-2986

Number 803

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Computational approaches to figurative language

Ekaterina V. Shutova

August 2011

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted March 2011 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Pembroke College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Computational approaches to figurative language

Ekaterina V. Shutova

Summary

The use of figurative language is ubiquitous in natural language text and it is a serious bottleneck in automatic text understanding. A system capable of interpreting figurative language would be extremely beneficial to a wide range of practical NLP applications. The main focus of this thesis is on the phenomenon of metaphor. I adopt a statistical data-driven approach to its modelling, and create the first open-domain system for metaphor identification and interpretation in unrestricted text. In order to verify that similar methods can be applied to modelling other types of figurative language, I then extend this work to the task of interpretation of logical metonymy.

The metaphor interpretation system is capable of discovering literal meanings of metaphorical expressions in text. For the metaphors in the examples "All of this *stirred* an unfathomable excitement in her" or "a carelessly *leaked* report" the system produces interpretations "All of this provoked an unfathomable excitement in her" and "a carelessly disclosed report" respectively. It runs on unrestricted text and to my knowledge is the only existing robust metaphor paraphrasing system. It does not employ any hand-coded knowledge, but instead derives metaphorical interpretations from a large text corpus using statistical pattern-processing. The system was evaluated with the aid of human judges and it operates with the accuracy of 81%.

The metaphor identification system automatically traces the analogies involved in the production of a particular metaphorical expression in a minimally supervised way. The system generalises over the analogies by means of verb and noun clustering, i.e. identification of groups of similar concepts. This generalisation makes it capable to recognise previously unseen metaphorical expressions in text, e.g. having once seen a metaphor *stir excitement* the system concludes that *swallow anger* is also used metaphorically. The system identifies metaphorical expressions with a high precision of 79%.

The logical metonymy processing system produces a list of metonymic interpretations disambiguated with respect to their word sense. It then automatically organises them into a novel class-based model of logical metonymy inspired by both empirical evidence and linguistic theory. This model provides more accurate and generalised information about possible interpretations of metonymic phrases than previous approaches.

Acknowledgments

The dream has come true – the metaphor system works! Besides inspiration and hard work, this result is also indebted to the help of a number of people. Of course the most crucial help came from my supervisors, Simone Teufel and Anna Korhonen. Both excellent researchers and mentors, they contributed to this work in different but very important ways. Simone provided invaluable methodological guidance, from drafting up experiments to carrying out human evaluations. She gave very constructive and involved comments on this thesis, that changed it a lot and for the better. Anna has been an extraordinary source of support both scientifically and personally. She shared her expertise on computational lexical semantics, and many of the decisions I had to make were shaped by her advice. The amount of skills and knowledge Simone and Anna passed on to me in the last three years is immense, and will definitely have a great impact not only on this thesis, but on all of my future work.

My research also benefited greatly from discussions with Ted Briscoe, Stuart Moore, Diarmuid Ó Séaghdha, Andreas Vlachos, Laura Rimell, Øistein Andersen, Aline Villavicencio and Stephen Clark. Special thanks to Lin Sun, who helped with verb and noun clustering for the metaphor identification experiment, and to Ann Copestake, who suggested I looked at logical metonymy. And of course the results of this work would not have been there without the participation of the volunteer annotators, I will always be very thankful for their help!

I have been extremely fortunate to have Stuart Moore and Johanna Geiss as my officemates. Their moral support, good-humoured attitude and intellectual involvement were incredibly motivating all along. And I would also like to thank all the administrators at the Computer Lab for their assistance and in particular Lise Gough for her endless help and for taking a very human stance on student administration.

Special thanks go to Pembroke College, for creating the best possible environment for me to study, live in and grow as a researcher. The members of the college, and in

particular Arwen Deuss and Becky Coombs, provided excellent academic and personal support whenever needed! I will also be always grateful to Cambridge Overseas Trust (Kapitza Scholarship), UK government (ORS), BP (BP Research Scholarship) and Google (Anita Borg Scholarship) for making this research possible by funding my studies.

Many thanks to my PhD examiners, Ted Briscoe and Katja Markert, who gave very insightful and helpful feedback on this work, for their genuine interest and positive, constructive criticism.

And finally, millions of thanks go to the best of Cambridge friends – Cristina Sisu, David O'Regan, Matt Seigel, Matt Smith (and Taylor), Tom Zawisza and Elizabeth Dearnley. Their love and encouragement gave me great confidence to face whatever tomorrow brings. Their genuine curiosity, broad knowledge and enthusiasm were a never failing source of inspiration. Their wisdom was the securest barrier from making mistakes. And... some of them even voluntarily annotated metaphors! Tom, Dave, Elizabeth and Stuart also proofread the final version of the thesis, on the last night and at a very short notice - I appreciate this immensely (especially after having read the whole thing myself before the viva)! But of course many more people made my time in Cambridge bright and memorable – for that I will always be grateful to David Gordon, Joana Borlido, Simon Calder, Christian Schläpfer, Antonio Garcia Castañeda, Lavr Burin, Milica Gasic, Rogier van Dalen, Colin Kelly, Ed Cannon, Alen Sabyrov, Kelsey Edwardsen, Richard Lines, Roseanne Zhao, Francesco Anesi, Annie Hill and all the other wonderful inhabitants of 6 Grange Road back in the day. Special thanks go to Peter Evan for sharing the final stretch with me.

This thesis is dedicated to my best friend and wonderful mother, for giving me this world and teaching me to be happy in it; to my grand-father, for being an excellent example of creative engineering thought; and to Dave and Bart, who sadly will never read it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our production and comprehension of language is a multi-layered computational process. Humans carry out high-level semantic tasks effortlessly by subconsciously employing a vast inventory of complex linguistic devices, while simultaneously integrating their background knowledge, to reason about reality. An ideal computational model of language understanding would also be capable of performing such high-level semantic tasks.

However, a great deal of natural language processing (NLP) research to date focuses on processing lower-level linguistic information, such as part-of-speech tagging, discovering syntactic structure of a sentence (parsing) or coreference resolution. Another cohort of researchers aim at improving application-based statistical inference (e.g. for machine translation or automatic summarisation). In contrast, there have been far fewer attempts to bring the state-of-the-art NLP technologies together to model the way humans use language to frame high-level reasoning processes, such as for example, creative thought.

Creative thought is often reflected in human communication in the form of *figurative language*, or *tropes*. As opposed to literal language, whose interpretation does not deviate from the words' defined and most frequent senses, the meaning of a trope is not simply composed of the common meanings of its components: its surface form and its underlying semantics do not directly correspond to each other. *Metonymic* phrases, for example, involve the use of a concept to stand for another related one that is not explicitly mentioned (e.g. in the sentence "He played *Bach*" the author is referring to the composer's music). *Logical metonymy* is a subtype of metonymy, whereby an entity may have an eventive interpretation (e.g. "enjoy a book" stands for "enjoy reading a book"). In the case of *metaphor*, one concept is viewed in terms of the properties of another (e.g. in the computer science metaphor "How can I *kill* a process?" the computational process is viewed as a living being). Other tropes include *litote* (understatement), *hyperbole* (overstatement), *synechdoche* (part stands for whole, essentially a subtype of metonymy), *simile* (comparison) and *irony*.

Characteristic to all areas of human activity (from poetic to ordinary to scientific) and

(1) ORIGINAL SENTENCE
All of this *stirred* an uncontrollable excitement in her.

TRANSLATION BY *GOOGLE TRANSLATE*
Все это **перемешало* неконтролируемое возбуждение в ней.

CORRECT TRANSLATION
Все это *вызвало* неконтролируемое возбуждение в ней. (All of this *provoked* an uncontrollable excitement in her.)

(2) ORIGINAL SENTENCE
I *spilled* everything I knew to Bobby.

TRANSLATION BY *GOOGLE TRANSLATE*
Я **пролил* все, что знал, Бобби.

CORRECT TRANSLATION
Я *рассказал* все, что знал, Бобби. (I *told* everything I knew to Bobby)

Figure 1.1: Examples of metaphor translation

thus to all types of discourse, these phenomena become an important problem for natural language processing. As I will show in an empirical study, the use of figurative language is ubiquitous in natural language text, thus making it a serious bottleneck in automatic text understanding. For example, an NLP application which is unaware that a "*leaked* report" is a "disclosed report" and not e.g. a "wet report", would fail further semantic processing of the piece of discourse this phrase appears in. A system capable of recognising and interpreting figurative language in unrestricted text would become an invaluable component of any real-world NLP application that needs to access semantics (e.g. information retrieval (IR), machine translation (MT), question answering (QA), information extraction (IE) and opinion mining). So far, these applications have not employed any metaphor processing techniques and thus often failed to interpret metaphorical data correctly. An example of metaphor translation from English into Russian by a state-of-the-art statistical MT system (Google Translate[1]) is presented in Figure 1.1. For both sentences the MT system produces literal translations of metaphorical terms in English, rather than their literal interpretations. This results in otherwise grammatical sentences being semantically infelicitous, poorly-formed and barely understandable to a native speaker of Russian. The meaning of *stir* in (1) and *spill* in (2) would normally be realised in Russian only via their literal interpretation in the given context (*provoke* and *tell*), as shown under CORRECT TRANSLATION in Figure 1.1. A metaphor processing component could help to avoid such errors.

Examples where metaphor understanding is crucial can also be found in opinion mining, i.e. detection of speaker's attitude to what is said and to the topic. Consider the following sentences.

[1] http://translate.google.com/

(1) a. Government *loosened strangle-hold* on business. (Narayanan, 1999)

b. Government deregulated business. (Narayanan, 1999)

Both sentences describe the same fact. However, the use of metaphor *loosened strangle-hold* in (1a) suggests that the speaker opposes government control of economy, whereas (1b) does not imply this. One can infer the speaker's negative attitude via the presence of a negative word *strangle-hold*. A metaphor processing system would establish the correct meaning of (1a) and thus discover the actual fact to which the speaker has negative attitude.

Despite the importance of figurative language for NLP tasks dealing with semantic interpretation, its automatic processing has received little attention in contemporary NLP, and is far from being a solved problem. The majority of computational approaches to figurative language still exploit ideas articulated two or three decades ago (Wilks, 1978; Lakoff and Johnson, 1980; Pustejovsky, 1991). They often rely on task-specific hand-coded knowledge (Fass, 1991; Martin, 1990; Narayanan, 1997, 1999; Feldman and Narayanan, 2004; Barnden and Lee, 2002; Agerri et al., 2007) and reduce the task to reasoning about a limited domain or a subset of phenomena (Markert and Nissim, 2002; Nissim and Markert, 2003; Peirsman, 2006; Gedigian et al., 2006; Krishnakumaran and Zhu, 2007). So far there has been no robust statistical system operating on unrestricted text. However, state-of-the-art accurate parsing (Briscoe et al., 2006; Clark and Curran, 2007; Klein and Manning, 2003), as well as recent work on computational lexical semantics (Schulte im Walde, 2006; Sun and Korhonen, 2009; Erk and McCarthy, 2009; Davidov et al., 2009; Ó Séaghdha, 2010; Abend and Rappoport, 2010) open many avenues for creation of such a system. This is the research niche the work described in this thesis is intending to fill. The main focus of the thesis is on the computational modelling of metaphor, one of the most frequent and puzzling types of figurative language. However, I additionally verify whether similar methods can be applied to other types of figurative language, exemplified by logical metonymy.

One of the main challenges in automatic processing of figurative expressions is that their production and comprehension require vast amounts of world knowledge. For example, to recognise and interpret the metaphor in the phrase "*leaked* report", one needs to be aware of the fact that reports are not liquid and cannot be physically *leaked*, but instead can normally be *written, read, presented* or *disclosed*. I therefore adopt a corpus-based approach to figurative language. The assumption behind my approach is that such knowledge is contained in textual data found in linguistic corpora, such as British National Corpus (BNC) (Burnard, 2007), American National Corpus (ANC) (Ide and Suderman, 2004), and can be automatically extracted from them. It is represented in the data in the form of distributions of predicate–argument combinations that can be identified by a parser. To extract and process this information, I employ state-of-the-art parsing and

lexical acquisition technologies, as well as design my own task-specific statistical models, using insights from linguistic theory to guide the process.

1.1 Metaphor

It is widely acknowledged in linguistics, philosophy and cognitive science that metaphor is based on *analogy* (Gentner, 1983; Lakoff and Johnson, 1980; Grady, 1997; Narayanan, 1997; Fauconnier and Turner, 2002). Metaphors arise when one concept is viewed in terms of the properties of another. Humans often use metaphor to describe abstract concepts through reference to more concrete or physical experiences. Below are some examples of metaphor.

(2) How can I *kill* a process? (Martin, 1988)

(3) Hillary *brushed aside* the accusations.

(4) I *invested* myself fully in this research.

(5) And then my heart with pleasure *fills*,
And *dances* with the daffodils.
("I wandered lonely as a cloud", William Wordsworth, 1804)

Metaphorical expressions may take a great variety of forms, ranging from conventional metaphors, which we produce and comprehend every day, e.g. those in (2) and (4), to poetic and novel ones, such as (5). In metaphorical expressions, seemingly unrelated features of one concept are attributed to another concept. In the example (2), a *computational process* is viewed as something *alive* and, therefore, its forced termination is associated with the act of killing. In (3) Hillary is not literally "cleaning the space by sweeping accusations". Instead, the accusations lose their validity in that situation, in other words Hillary *rejects* them. The verbs *brush aside* and *reject* both entail the resulting disappearance of their object, which is the shared salient property that makes it possible for this analogy to be lexically expressed as a metaphor.

1.1.1 Conceptual metaphor

Metaphor has traditionally been viewed as an artistic device that lends vividness and distinction to its author's style. This view was first challenged by Lakoff and Johnson (1980), who claimed that it is a productive phenomenon that operates at the level of mental processes. According to Lakoff and Johnson, metaphor is thus not merely a

property of language, i.e. a linguistic phenomenon, but rather a property of thought, i.e. a cognitive phenomenon. This view was subsequently acquired and extended by a multitude of approaches (Grady, 1997; Narayanan, 1997; Fauconnier and Turner, 2002; Feldman, 2006; Pinker, 2007) and the term *conceptual metaphor* was coined to describe it.

The view postulates that metaphor is not limited to similarity-based meaning extensions of individual words, but rather involves reconceptualisation of a whole area of experience in terms of another. Thus metaphor always involves two concepts or conceptual domains: the *target* (also called *topic* or *tenor* in linguistics literature) and the *source* (also called *vehicle*). Consider the examples in (6) and (7).

(6) He *shot down* all of my arguments. (Lakoff and Johnson, 1980)

(7) He *attacked* every weak point in my argument. (Lakoff and Johnson, 1980)

According to Lakoff and Johnson, a mapping of the concept of *argument* to that of *war* is employed in both (6) and (7). The *argument*, which is the target concept, is viewed in terms of a *battle* (or a *war*), the source concept. The existence of such a link allows us to talk about *arguments* using *war* terminology, thus giving rise to a number of metaphors. Conceptual metaphor, or *source-target domain mapping*, is thus a generalisation over a set of individual metaphorical expressions that covers multiple cases in which ways of reasoning about the source domain systematically correspond to ways of reasoning about the target.

1.1.2 Linguistic metaphor

Conceptual metaphor manifests itself in natural language in the form of *linguistic metaphor* (or metaphorical expressions) in a variety of ways. The most common types of linguistic metaphor are *lexical* metaphor, i.e. metaphor at the level of a single word sense (as in the examples (2)–(5)), *multi-word* metaphorical expressions (e.g. "we *go on pilgrimage* with Raleigh or *put out to sea* with Tennyson") or *extended* metaphor, that spans over longer discourse fragments.

Lexical metaphor is by far the most frequent type. In the presence of a certain conceptual metaphor individual words can be used in entirely novel contexts, which results in the formation of new meanings. Consider the following example.

(8) How can we build a 'Knowledge economy' if research is *handcuffed*? (Barque and Chaumartin, 2008)

In this sentence the physical verb *handcuff* is used with an abstract object *research* and its meaning adapts accordingly. Metaphor is a productive phenomenon, i.e. its novel examples continue to emerge in language. However, a large number of metaphorical expressions become conventionalised (e.g. "I cannot *grasp* his way of thinking"). Although metaphorical in nature, their meanings are deeply entrenched in everyday use, and are thus cognitively treated as literal terms. Both novel and conventional metaphors are important for text processing, hence this thesis is concerned with both types. However, fixed non-compositional idiomatic expressions (e.g. *kick the bucket*, *rock the boat*, *put a damper on*) are left aside, since the mechanisms of their formation are no longer productive in modern language and, as such, they are of little interest for the design of a generalisable computational model of metaphor.

Extended metaphor refers to the use of metaphor at the discourse level. A famous example of extended metaphor can be found in William Shakespeare's play "As You Like It", where he first compares the world to a stage and then in the following discourse describes its inhabitants as players. Extended metaphor often appears in literature in the form of an *allegory* or a *parable*, whereby a whole story from one domain is metaphorically transferred onto another in order to highlight certain attributes of the subject or teach a moral lesson.

1.1.3 Computational modelling of metaphor

The focus of this thesis is on lexical metaphor and the computational modelling thereof. From an NLP viewpoint, not all metaphorical expressions are equally important. A metaphorical expression is interesting for computational modelling if its metaphorical sense is significantly distinct from its original literal sense and cannot be interpreted directly (e.g. by existing word sense disambiguation techniques using a predefined sense inventory). The identification of highly conventionalised metaphors (e.g. the verb *impress*, whose meaning originally stems from printing) are not of interest for NLP applications, since their metaphorical senses have long been dominant in language and their original literal senses may no longer be used. A number of conventionalised metaphors, however, require explicit interpretation in order to be understood by computer (e.g. "*throw* an idea", "*polish* the thesis", "*catch* contagion"), as do all novel metaphors. Thus the thesis is concerned with both novel and conventional metaphors, but only considers the cases whereby the literal and metaphorical senses of the word are in clear opposition in common use in contemporary language.

Automatic processing of metaphor can be divided into two subtasks: *metaphor identification*, or *recognition* (distinguishing between literal and metaphorical language in text); and *metaphor interpretation* (identifying the intended literal meaning of a metaphorical expression). An ideal metaphor processing system should address both of these tasks and provide useful information to support semantic interpretation in real-world NLP appli-

cations. In order to be directly applicable to other NLP systems it needs to satisfy the following criteria:

- **provide a representation of metaphor interpretation that can be easily integrated with other NLP systems**: This criterion places constraints on how the metaphor processing task should be defined. The most universally applicable metaphor interpretation would be in the text-to-text form. This means that a metaphor processing system would take raw text as input and provide a simpler text as output, in which metaphors are interpreted.
- **operate on unrestricted running text**: In order to be useful for real-world NLP the system needs to be capable to process real-world data. Rather than only dealing with individual carefully selected clear-cut examples, the system should be fully implemented and tested on free naturally occurring text.
- **be open-domain**: The system needs to cover all domains, genres and topics. Thus it should not rely on any domain-specific information or focus on individual types of instances (e.g. a hand-chosen limited set of source-target domain mappings).
- **be unsupervised or minimally supervised**: To be easily adaptable to new domains, the system needs to be unsupervised or minimally supervised. This means it should not use any task-specific (i.e. metaphor-specific) hand-coded knowledge. The only acceptable exception might be a multi-purpose general-domain lexicon that is already in existence and does not need to be created in a costly manner, although it would be an advantage if no such resource is required.
- **cover all syntactic constructions**: To be robust, the system needs to be able to deal with metaphors represented by all word classes and syntactic constructions.

In this thesis, I address both the metaphor identification and interpretation tasks, resulting in the first integrated domain-independent corpus-based computational model of metaphor. The method is designed with the above criteria in mind. It takes unrestricted text as input and produces textual output. All components of the method are in principle applicable to all part-of-speech classes and syntactic constructions. However, in the framework of this thesis I test the system only on single-word metaphors expressed by a verb. Restricting the scope to verbs is a methodological step aimed at testing the main principles of the proposed approach in a well-defined setting. Metaphor identification and interpretation are first evaluated independently, and then together as a joint system.

My first experiment is concerned with the identification of metaphorical expressions in unrestricted text. Starting from a small set of metaphorical expressions, the system learns the analogies involved in their production using unsupervised methods. It generalises over the exemplified analogies by means of verb and noun clustering, i.e. the identification of

groups of similar concepts. This generalisation allows it to recognise previously unseen metaphorical expressions in text. Consider the following examples:

(9) All of this *stirred* an uncontrollable excitement in her.

(10) Time and time again he would stare at the ground, hand on hip, and then *swallow* his anger and play tennis.[2]

Having once seen the metaphor "*stir* excitement" in (9) my metaphor identification system successfully concludes that "*swallow* anger" in (10) is also used metaphorically.

The next experiment deals with metaphor interpretation. For this purpose I developed an algorithm that discovers literal meanings of metaphorical expressions in text and produces their literal paraphrases, i.e. literal ways of saying the same thing. For example, for metaphors in (11a) and (12a) the system produces paraphrases in (11b) and (12b) respectively.

(11) a. All of this *stirred* an uncontrollable excitement in her.

b. All of this provoked an uncontrollable excitement in her.

(12) a. a carelessly *leaked* report

b. a carelessly disclosed report

My approach to metaphor interpretation is built around the assumption that the meaning of a word in context emerges through interaction with the meaning of the words surrounding it. This assumption is widely accepted in lexical semantics theory (Pustejovsky, 1995; Hanks and Pustejovsky, 2005) and has been exploited for lexical acquisition (Lapata, 2001; Schulte im Walde, 2006; Sun and Korhonen, 2009). It also means that the context itself imposes certain semantic restrictions on the words which can occur within it. With this in mind, I design a context-based probabilistic model for paraphrase selection and acquire paraphrases for metaphorical expressions from a large corpus.

Aside from the computational modelling work, I derive an annotation scheme for both linguistic and conceptual metaphor and create a small metaphor corpus. This corpus serves as a testbed for my experiments. The annotation effort also allows for an empirical verification of some of the most influential theoretical claims about metaphor, which in turn guide system design.

[2]These are a real-world examples taken from the British National Corpus and used by the system.

1.2 Metonymy

Metonymy is defined as the use of a word or a phrase to stand for a related concept, which is not explicitly mentioned. If metaphor is based on *similarity* between the concepts, metonymy builds on *contiguity*. Contiguity and similarity are two kinds of association. Metonymy implies a contact or a (rather physical) connection between the entities, whereas metaphor implies the presence of characteristics in common. Here are some examples of metonymic phrases:

(13) The *pen* is mightier than the *sword*. (Bulwer-Lytton, 1839)

(14) He played *Bach*.

(15) He drank *his glass*. (Fass, 1991)

The metonymic adage in (13) is a classical example. Here the *pen* stands for the press and the *sword* for military power. In (14) *Bach* is used to refer to the composer's music and in (15) the *glass* stands for its *content*, i.e. the actual *drink* (beverage). These are examples of *general metonymy*.

General metonymy is traditionally explained via conventionalised *metonymic patterns* that operate over semantic classes (Stern, 1931; Lakoff and Johnson, 1980; Fass, 1997). Below are some examples of common metonymic patterns.

- PART-FOR-WHOLE (also known as synechdoche), e.g. "I could do with an extra *pair of hands*" (referring to a helper or a worker).
- CONTAINER-FOR-CONTENTS, e.g. "He drank his *glass*".
- PRODUCER-FOR-PRODUCT, e.g. "I bought a *Picasso*".
- PLACE-FOR-EVENT, e.g. "at the time of *Vietnam*, increased spending led to inflation and trade deficit" (Markert and Nissim, 2006).
- PLACE-FOR-PRODUCT, e.g. "He drinks *Bordeaux* with his dinner".
- PLACE-FOR-INHABITANTS , e.g. "*France* is on strike again".
- ORGANISATION-FOR-MEMBERS, e.g. "Last February *NASA* announced [...]" (Markert and Nissim, 2006).
- OBJECT USED-FOR-USER, e.g. "The *sax* has a flu today".

Such pattern-based shifts in meaning happen systematically and are known as *regular polysemy* (Apresjan, 1973), or *sense extension* (Copestake and Briscoe, 1995). However, some metonymic examples emerge only in specific contexts and are less conventionalised than others. Markert and Nissim (2006) call metonymies such as those in (16) and (17) *unconventional*.

(16) The *ham sandwich* is waiting for his check. (Nunberg, 1978)

(17) Ask *seat 19* whether he wants to swap. (Markert and Nissim, 2006)

These examples illustrate that metonymy, as well as metaphor, is both regular and productive.

Along with theoretical work, there have been a number of computational accounts of general metonymy (Utiyama et al., 2000; Markert and Nissim, 2002; Nissim and Markert, 2003; Peirsman, 2006; Agirre et al., 2007). All of these approaches are data-driven and the majority of them (with the exception of Utiyama et al. (2000)) deal only with metonymic proper names, use machine learning and treat metonymy resolution as classification according to common metonymic patterns. In contrast, Utiyama et al. (2000) statistically derive paraphrases of metonymic expressions from a large corpus (e.g. "read the books of Shakespeare" for "read *Shakespeare*").

1.2.1 Logical metomymy

The following examples represent a variation of this phenomenon called *logical metonymy*.

(18) Thank you for the present! I really *enjoyed your book*.

(19) John is *enjoying his cigarette* outside.

(20) After *three martinis* John was feeling well. (Godard and Jayez, 1993)

In these sentences the noun phrases *your book*, *his cigarette* and *three martinis* have eventive interpretations, i.e. they stand for the events of "reading a book", "smoking a cigarette" and "drinking three martinis" respectively.

Logical metonymy is an elliptical construction, i.e. it lacks an element that is recoverable or inferable from the context. It arises due to a predicate taking syntactic and semantic arguments of different types. The verb *enjoy* requires an eventuality as its semantic argument (it is a process that one enjoys), but also allows for an object expressed by a noun phrase syntactically. Thus the noun phrase *your book* in (18) is interpreted as "reading your book". But how would one know that *enjoy a book* means *enjoy reading a book* and

enjoy a cigarette means *enjoy smoking a cigarette*, and not e.g. *enjoy buying a book*, or *enjoy smoking a book*, or *enjoy eating a cigarette*? Humans are capable of interpreting these phrases using their world knowledge and contextual information. Modelling this process is the focus of my experiments on logical metonymy.

As well as metaphor and general metonymy, logical metonymy is both highly frequent and productive[3], which makes its computational processing an important problem within NLP. In this thesis, I focus on the problem of interpretation of logical metonymy and, as in the case of metaphor, adopt a statistical data-driven approach to it. My system first derives a set of possible metonymic interpretations from a large corpus, following Lapata and Lascarides (2003). It then disambiguates them with respect to their word sense using an existing sense inventory, and automatically organises them into a new class-based conceptual model of logical metonymy that is inspired by linguistic theory (Vendler, 1968; Pustejovsky, 1991; Godard and Jayez, 1993). I then experimentally study whether this representation is intuitive to humans, by asking human subjects to classify metonymic interpretations into groups of similar concepts.

1.3 Structure of the thesis

The remainder of this thesis is thematically divided into two parts. The next four chapters present the work on metaphor, beginning with a review of previous work (Chapter 2), before presenting my own work (Chapters 3, 4 and 5). Chapter 2 describes linguistic theories of metaphor that are relevant to computational research, provides a review of metaphor annotation and existing computational models of metaphor and summarises the main claims of the thesis. Chapter 3 is devoted to the annotation scheme and its experimental validation in a setting with multiple annotators. It concludes with a data analysis aimed at an empirical verification of the theoretical claims that formed the basis for my computational models. In Chapters 4 and 5, I introduce algorithms for automatic metaphor identification and interpretation respectively, provide details of their evaluation and discuss the final results.

The second part of the thesis is devoted to logical metonymy. In Chapter 6, I discuss the most prominent theoretical and computational approaches to logical metonymy, along with their shortcomings, and motivate my own solution. Chapter 7 describes the design of my logical metonymy interpretation system and the associated experiments.

Finally, the conclusions of the thesis are presented in Chapter 8, along with suggestions for future research directions in the area of computational modelling of figurative language.

[3]By "productive" here I mean that novel examples of the phenomenon still emerge in language. Thus, this does not contradict the previous accounts (Verspoor, 1997) that claim that logical metonymy is to a large extent conventionalised.

1.4 Note on collaboration

I used the verb clustering method and system of Sun and Korhonen (2009) within my metaphor identification experiments. The extended clustering approach presented in Section 4.4.1 is, therefore, a result of joint work with Lin Sun and Anna Korhonen. All the other theoretical, experimental and composition work involved in the production of the thesis was carried out by the author alone.

1.5 Note on publications

Most of the work presented in this thesis was published at conferences in the field of NLP and cognitive science. The list of publications is included below. Four papers are devoted to metaphor and two to logical metonymy.

- E. Shutova, L. Sun and A. Korhonen. 2010. Metaphor Identification Using Verb and Noun Clustering. In *Proceedings of COLING 2010*, Beijing, China.
- E. Shutova. 2010. Models of Metaphor in NLP. In *Proceedings of ACL 2010*, Uppsala, Sweden.
- E. Shutova and S. Teufel. 2010. Metaphor Corpus Annotated for Source - Target Domain Mappings. In *Proceedings of LREC 2010*, Valetta, Malta.
- E. Shutova. 2010. Automatic Metaphor Interpretation as a Paraphrasing Task. In *Proceedings of NAACL 2010*, Los Angeles, USA.
- E. Shutova and S. Teufel. 2009. Logical Metonymy: Discovering Classes of Meanings. In *Proceedings of the CogSci 2009 Workshop on Semantic Space Models*. Amsterdam, Holland.
- E. Shutova. 2009. Sense-based Interpretation of Logical Metonymy Using a Statistical Method. In *Proceedings of ACL 2009 Student Research Workshop*, Singapore.

Chapter 2

Metaphor background and contributions

In this chapter I introduce the most prominent linguistic theories of metaphor that are relevant to its computational modelling, address the issue of metaphor annotation in corpora and review the previous work on automatic metaphor recognition and interpretation. The chapter concludes with a critique of the existing methods and the motivation of my own approach.

2.1 Metaphor and polysemy

Theorists of metaphor distinguish between two kinds of metaphorical language: *novel* (or *poetic*) metaphors, i.e. those that are imaginative, and *conventionalised* metaphors, i.e. those that are used as a part of an ordinary discourse.

> Metaphors begin their lives as novel poetic creations with marked rhetorical effects, whose comprehension requires a special imaginative leap. As time goes by, they become a part of general usage, their comprehension becomes more automatic, and their rhetorical effect is dulled (Nunberg, 1987, p 198).

Following Orwell (1946), Nunberg calls such metaphors "dead" and claims that they are not psychologically distinct from literally-used terms. The scheme described by Nunberg demonstrates how metaphorical associations capture patterns governing *polysemy*, i.e. the capacity of a word to have multiple meanings. Over time some of the aspects of the target domain are added to the meaning of a term in the source domain, resulting in a (metaphorical) sense extension of this term. Copestake and Briscoe (1995) discuss sense extension mainly based on metonymic examples and model the phenomenon using lexical rules encoding metonymic patterns. They also suggest that similar mechanisms can be

used to account for metaphorical processes. According to Copestake and Briscoe, the conceptual mappings encoded in the sense extension rules would define the limits to the possible shifts in meaning.

However, it is often unclear if a metaphorical instance is a case of broadening of the sense in context due to general vagueness in language, as opposed to formation of an entirely distinct metaphorical sense. Consider the following examples.

(21) a. My tea is *cold*.

b. He is such a *cold* person.

(22) a. As soon as I *entered* the room I noticed the difference.

b. How can I *enter* Emacs?

The sentence (21a) exemplifies the basic sense of *cold* – "at a low temperature, especially when compared to the temperature of the human body, and not hot or warm"[1], whereas *cold* in (21b) should be interpreted metaphorically as "not showing kindness, love or emotion and not friendly". These two senses are linked via the metaphorical mapping between EMOTIONAL STATES and TEMPERATURES.

Enter in (22a) is defined as "to come or go into a particular place". In (22b) this sense stretches to describe dealing with *software*, whereby COMPUTER PROGRAMS are viewed as PHYSICAL SPACES. However, this extended sense of *enter* is not yet sufficiently distinct or conventional to be included in current dictionaries, although this could happen over time.

General-domain lexical resources often include information about metaphorical word senses, however, not systematically and without any accompanying semantic annotation. One such example would be WordNet[2] (Fellbaum, 1998). WordNet is a broad-coverage lexical database, where lexical entries are organised into sets of synonyms, or *synsets*, that are connected into a network. Each synset represents a particular sense of words included in it. For instance, the verb synset "(interpret, construe, see)" encodes the meaning "make sense of; assign a meaning to" realised in e.g. "How do you interpret his behavior?" Each synset is also assigned a lexicographic definition, or a *gloss*. The synsets are linked by other semantic relations, such as hyponymy (a subsumption relationship between more specific and more general words), meronymy (part-to-whole relationship), antonymy (indicating words with opposite meanings). An extract from the resulting semantic network is shown in Figure 2.1. The Figure shows the synset hierarchy describing the concepts of

[1] All subsequent sense definitions in this thesis are taken from the Cambridge Advanced Learner's Dictionary (URL `http://dictionary.cambridge.org/`).

[2] `http://wordnet.princeton.edu/`.

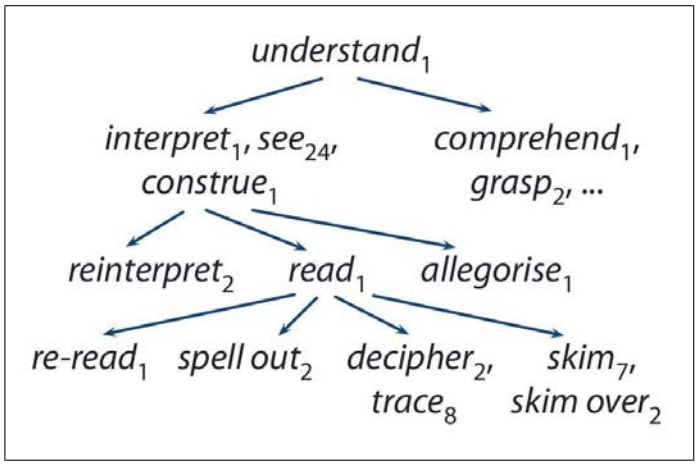

Figure 2.1: An example of synset hierarchy in WordNet

reading and *understanding*. It includes a few metaphorical senses, e.g. the *comprehension* sense of *grasp*, defined in WordNet as "get the meaning of something", or the *reading* sense of *skim*, defined in WordNet as "read superficially".

However, a great deal of metaphorical senses are absent from the current version of WordNet. A number of researchers advocated the necessity of systematic inclusion and mark-up of metaphorical senses in such general-domain lexical resources (Lönneker and Eilts, 2004; Alonge and Castelli, 2003) and claimed that this would be beneficial for the computational modelling of metaphor. Metaphor processing systems could then either use this knowledge or be evaluated against it. Lönneker (2004) mapped the senses from EuroWordNet[3] to the Hamburg Metaphor Database (Lönneker, 2004; Reining and Lönneker-Rodman, 2007) containing examples of metaphorical expressions in German and French. However, currently no explicit information about metaphor is integrated into WordNet for English.

Whereas consistent inclusion in WordNet is in principle possible for conventional metaphorical senses, it is not viable for novel contextual sense alternations. Since metaphor is a productive phenomenon, all possible cases of contextual meaning alternations it results in cannot be described via simple sense enumeration (Pustejovsky, 1995). Computational metaphor processing therefore cannot be approached using the standard word sense disambiguation (WSD) paradigm, whereby the contextual use of a word is classified according to an existing sense inventory. The metaphor interpretation task is inherently more complex and requires generation of new and often uncommon meanings of the metaphorical term based on the context.

[3]EuroWordNet is a multilingual database containing WordNets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The WordNets are structured in the same way as the Princeton WordNet for English. URL: `http://www.illc.uva.nl/EuroWordNet/`.

2.2 Theoretical views on metaphor

Scientific inquiry on the subject of metaphor dates back to Aristotle, who defined the phenomenon in his work *Poetics* as

> the application of a strange term either transferred from the genus and applied to the species or from the species and applied to the genus, or from one species to another, or else by analogy (*Poetics*, section 1457b, translated by Fyfe (1927))

Aristotle, however, did not make a distinction between metaphor and metonymy in their modern sense. While still resonating with some of the principles formulated by Aristotle, the theory of metaphor has since evolved significantly under the influence of linguistic and psychological findings (Black, 1962; Wilks, 1975; Lakoff and Johnson, 1980), and the establishment of the fields of artificial intelligence (Barnden and Lee, 2002; Narayanan, 1997), cognitive science (Haskell, 2002) and neuroscience (Feldman, 2006). The following views on metaphor are prominent in linguistics and philosophy: the comparison view (e.g. the Structure–Mapping Theory of Gentner (1983)), the interaction view (Black, 1962; Hesse, 1966), the selectional restrictions violation view (Wilks, 1975, 1978) and the conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980). All of these approaches share the idea of an interconceptual mapping that underlies the production of metaphorical expressions. Gentner's Structure–Mapping Theory postulates that the ground for metaphor lies in similar properties and relations shared by the two concepts (the target and the source). Tourangeau and Sternberg (1982), however, criticise this view by noting that "everything has some feature or category that it shares with everything else, but we cannot combine just any two things in metaphor" (Tourangeau and Sternberg, 1982, p. 226). The interaction view focuses on the surprise and novelty that metaphor introduces. Its proponents claim that the source concept (or domain) represents a template for seeing the target concept in an entirely new way. The conceptual metaphor theory of Lakoff and Johnson (1980) takes this idea much further by stating that metaphor operates at the level of thought rather than at the level of language, and that it is based on a set of cognitive mappings between source and target domains. Thus Lakoff and Johnson put the emphasis on the structural aspect of metaphor, rather than its decorative function in language that dominated the preceding theories. The selectional restrictions violation view of Wilks (1978) concerns manifestation of metaphor in language. Wilks suggests that metaphor represents a violation of combinatory norms in the linguistic context and that metaphorical expressions can be detected via such violation.

In the remainder of this section I will discuss CMT and the selectional restrictions violation view in more detail, since these two approaches inspired the majority of the existing computational models of metaphor.

2.2.1 Conceptual metaphor theory

The examples in (6) and (7) provided a good illustration of CMT. Lakoff and Johnson explain them via the conceptual metaphor ARGUMENT IS WAR, which is systematically reflected in language in a variety of expressions.

(23) Your claims are *indefensible*. (Lakoff and Johnson, 1980)

(24) I *demolished* his argument. (Lakoff and Johnson, 1980)

(25) I've never *won* an argument with him. (Lakoff and Johnson, 1980)

(26) You disagree? Okay, *shoot*! (Lakoff and Johnson, 1980)

According to CMT, we conceptualise and structure arguments in terms of battle, which systematically influences the way we talk about arguments within our culture. In other words, the conceptual structure behind battle, i.e. that one can shoot, demolish, devise a strategy, win and so on, is metaphorically transferred onto the domain of argument. Lakoff and Johnson claim that such metaphorical associations do not only reveal themselves in language, but to a large extent govern our behavior. For example, one might not only talk about "*demolishing* arguments", but at the same time behave aggressively in the process.

Manifestations of conceptual metaphor are ubiquitous in language and communication. Below are a few other examples of common metaphorical mappings.

- TIME IS MONEY (e.g. "That flat tire *cost* me an hour")
- IDEAS ARE PHYSICAL OBJECTS (e.g. "I can not *grasp* his way of thinking")
- LINGUISTIC EXPRESSIONS ARE CONTAINERS (e.g. "I would not be able to *put* all my feelings *into* words")
- EMOTIONS ARE VEHICLES (e.g. "[...] she was *transported* with pleasure")
- FEELINGS ARE LIQUIDS (e.g. "[...] all of this *stirred* an unfathomable excitement in her")
- LIFE IS A JOURNEY (e.g. "He *arrived* at the end of his life with very little emotional *baggage*")

One of the most widespread types of conceptual metaphor across different cultures is Event Structure Metaphor (ESM) (Lakoff, 1994), which structures the way we reason about events. ESM consists of a number of *primary metaphors* (Grady, 1997), that represent mappings from the concrete domain of physical forces and spatial motion (the

States are Locations (bounded regions in space)
Changes are Movements (into or out of bounded regions)
Causes are Forces
Actions are Self-propelled movements
Purposes are Destinations
Means are Paths (to destinations)
Difficulties are Impediments to motion

Figure 2.2: Metaphorical mappings exemplifying ESM

source domain) to the abstract domain of causes, actions and events (the target domain). Metaphorical mappings that exemplify ESM are presented in Figure 2.2.

Lakoff and Johnson also distinguish between *structural* metaphors (cases where one concept is metaphorically structured in terms of another, such as in the examples above) and *orientational* metaphors, e.g. up-down spatialisation metaphors, such as the following.

- HAPPY IS UP; SAD IS DOWN (e.g. "That *boosted* my spirits", "My spirits *rose*")
- HEALTH IS UP; SICKNESS (DEATH) IS DOWN (e.g. "He is *at the peak* of health")
- VIRTUE IS UP; DEPRAVITY IS DOWN (e.g. "She is an *up-standing* citizen", "That was a *low* trick")

Lakoff and colleagues organised their ideas in a resource called Master Metaphor List (MML) (Lakoff et al., 1991). The list is a collection of source–target domain mappings (mainly those related to mind, feelings and emotions) with corresponding examples of language use. An example entry for the mapping STATES ARE LOCATIONS is shown in Figure 2.3. The mappings in the list are organised in an ontology, e.g. the metaphor PURPOSES ARE DESTINATIONS is a special case of a more general metaphor STATES ARE LOCATIONS. To date MML is the most comprehensive metaphor resource in the linguistic literature.

The list, however, has been criticised for the lack of clear structuring principles of the mapping ontology (Lönneker-Rodman, 2008). The same concept often appears at different taxonomic levels, members of taxonomy are not mutually exclusive and the same classes are referred to by different class labels. This fact and the chosen data representation in the Master Metaphor List make it not directly suitable for computational use. However, CMT is a theoretical account that aims to explain the mechanisms of metaphorical reasoning, and it was not devised with computational modelling in mind.

CMT produced a significant resonance in the fields of philosophy, linguistics, cognitive science and artificial intelligence, including NLP. It inspired novel research (Martin, 1990,

EVENT STRUCTURE States

STATES ARE LOCATIONS

Source Domain: locations

Target Domain: states

He is in love.
What kind of a state was he in when you saw him?
She can stay/remain silent for days.
He is at rest/at play.
He remained standing.
He is at a certain stage in his studies.
What state is the project in?

1. Purposes are Destinations

Alternate names: Desired States are Desired Locations

Note:
This is a sub-metaphor of STATES ARE LOCATIONS

It took him hours to reach a state of perfect concentation.

2. Comparison of States is Comparison of Distance

Note:
See File on this domain: Comparison

Special case 1: Harm is Being in Harmful Location

Note:
See File on this domain: HarmLoc

Special case 2: Existence is a Location (Here)

Note:
See File on this domain: Existence

Special case 3: Opportunities are Open Paths

Figure 2.3: Example entry in the Master Metaphor List

1994; Narayanan, 1997, 1999; Barnden and Lee, 2002; Feldman and Narayanan, 2004; Mason, 2004; Martin, 2006; Agerri et al., 2007), but was also criticised for the lack of consistency and empirical verification (Murphy, 1996; Shalizi, 2003; Pinker, 2007). The sole evidence that Lakoff and Johnson (1980) supported their theory with was a set of carefully selected examples, such as those in the Master Metaphor List. Such examples, albeit clearly illustrating the main tenets of the theory, are not representative. They cannot possibly capture the whole spectrum of metaphorical expressions, and thus do

not provide evidence that the theory can adequately explain the majority of metaphors in real-world texts. A corpus-based study of conceptual metaphor is still needed for the latter purpose.

2.2.2 Selectional restrictions violation view

Lakoff and Johnson do not discuss how metaphors can be recognised in linguistic data. To date, the most influential account of this issue is that of Wilks (1975, 1978). According to Wilks, metaphors represent a violation of *selectional restrictions* (or *preferences*) in a given context. Selectional restrictions are the semantic constraints that a predicate places onto its arguments. Consider the following example.

(27) a. My aunt always drinks her tea on the terrace.

b. My car *drinks* gasoline. (Wilks, 1978)

The verb *drink* normally requires a grammatical subject of type ANIMATE and a grammatical object of type LIQUID, as in example (27a). Therefore, *drink* taking a *car* as a subject in (27b) is an anomaly, which, according to Wilks, indicates a metaphorical use of *drink*.

Although Wilks' idea inspired a number of computational experiments on metaphor recognition (Fass and Wilks, 1983; Fass, 1991; Krishnakumaran and Zhu, 2007), it is important to note that in practice this approach has a number of limitations. Firstly, there are other kinds of non-literalness or anomaly in language that cause a violation of semantic norm, such as metonymies. Thus the method would overgenerate. Secondly, there are kinds of metaphor that do not represent a violation of selectional restrictions, i.e. the approach may also undergenerate. This would happen, for example, when highly conventionalised metaphorical word senses are more frequent than the original literal senses. Due to their frequency, selectional preference distributions of such words in real-world data would be skewed towards the metaphorical senses, e.g. *capture* may select for *ideas* rather than *captives* according to the data. As a result, no selectional preferences violation can be detected in the use of such verbs. Another case where the method does not apply is copula constructions, such as "All the world's a *stage*". And finally, the method does not take into account the fact that interpretation (of metaphor as well as other linguistic phenomena) is always context dependent. For example, the phrase "All men are *animals*" uttered by a biology professor or a feminist would have entirely different interpretations, the latter clearly metaphorical, but without any violation of selectional restrictions.

2.3 Metaphor annotation in corpora

The task of metaphor annotation in corpora can be split into two stages, to reflect two distinct aspects of the phenomenon, i.e. the presence of both conceptual and linguistic metaphor. These stages include the identification of metaphorical senses in text, which requires distinguishing between literal and non-literal meanings, and the assignment of the underlying source-target domain mappings.

Although humans are perfectly capable of producing and comprehending metaphorical expressions, the task of annotating metaphor in text is challenging. This might be due to the variation in its use and external form, as well as the conventionality of many metaphorical senses. Gibbs (1984) suggests that literal and figurative meanings are situated at the ends of a single continuum, along which metaphoricity and idiomaticity are spread. This makes demarcation of metaphorical and literal language fuzzy.

Traditional approaches to metaphor annotation include the manual search for lexical items used metaphorically (Pragglejaz Group, 2007), for source and target domain vocabulary (Deignan, 2006; Koivisto-Alanko and Tissari, 2006; Martin, 2006) or for linguistic markers of metaphor (Goatly, 1997).

The Pragglejaz Group (2007) proposed a metaphor identification procedure (MIP) for human annotators. The procedure involves metaphor annotation at the word level as opposed to identifying metaphorical relations (between words) or source–target domain mappings (between concepts or domains). In order to discriminate between words used metaphorically and literally, the annotators are asked to follow the guidelines presented in Figure 2.4. In the framework of this procedure, the sense of every word in the text is considered as a potential metaphor, and every word is then tagged as literal or metaphorical. Thus such annotation can be viewed as a form of word sense disambiguation with an emphasis on metaphoricity. MIP laid the basis for the creation of the VU Amsterdam Metaphor Corpus[4] (Steen et al., 2010). This corpus is a subset of BNC Baby[5] annotated for linguistic metaphor. Its size is 200,000 words and it comprises four genres: news text, academic text, fiction and conversations. Although this is undoubtedly an important resource, its annotations do not entirely match the definition of metaphor assumed by this thesis. Steen and colleagues are interested in many more aspects of metaphor than those relevant to its computational processing. For example, a large proportion of metaphors annotated in the corpus are borderline cases, whose meaning is highly frequent and long established in language. Consider the following examples.

(28) They want to *show* you that they trust you.

[4] `http://www.ota.ox.ac.uk/headers/2541.xml`

[5] BNC Baby is a four-million-word corpus comprising four different genres: academic, fiction, newspaper and conversation. For more information see `http://www.natcorp.ox.ac.uk/corpus/babyinfo.html`

1. Read the entire text-discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text-discourse.
3.
   - For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   - For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be
     - More concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];
     - Related to bodily action;
     - More precise (as opposed to vague);
     - Historically older;

     Basic meanings are not necessarily the most frequent meanings of the lexical unit.
   - If the lexical unit has a more basic current contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical.

Figure 2.4: Metaphor identification procedure of Pragglejaz Group (2007)

(29) You'll *have* money problems as long as you gamble, Adam.

(30) The Party has not *assisted in* their work.

(31) I do not *regard* property profits as earnings.

The verbs in these sentences are tagged as metaphorical in the corpus, see Figure 2.5. Whether such meanings should be considered metaphorical or not is a debatable issue. Being interested in historical aspects of metaphor, Steen and colleagues tag these cases as metaphorical. In contrast, a model of figurative language within NLP is in the first instance concerned with the interpretation of alternations of traditional meanings in a given context that could not be interpreted using standard word sense disambiguation methods. Therefore, cases of conventional polysemy, such as *show* in (28), are irrelevant to computational modelling, and their annotation in a corpus would be misleading for computational experiments.

```
<p>
<s n=‘‘8’’>
<c type=‘‘PUQ’’>’</c>
<w lemma=‘‘they’’ type=‘‘PNP’’>They </w>
<w lemma=‘‘want’’ type=‘‘VVB’’>want </w>
<w lemma=‘‘to’’ type=‘‘TO0’’>to </w>
<w lemma=‘‘show’’ type=‘‘VVI’’>
<seg function=‘‘mrw’’ type=‘‘met’’ vici:morph=‘‘n’’>show</seg>
</w>
<w lemma=‘‘you’’ type=‘‘PNP’’>you </w>
<w lemma=‘‘that’’ type=‘‘CJT’’>that </w>
<w lemma=‘‘they’’ type=‘‘PNP’’>they </w>
<w lemma=‘‘trust’’ type=‘‘VVB’’>trust </w>
<w lemma=‘‘you’’ type=‘‘PNP’’>you</w>
<c type=‘‘PUN’’>.</c>
<c type=‘‘PUQ’’>’</c>
</s>
</p>
```

Figure 2.5: An example from the VU Amsterdam Metaphor Corpus

The corpus also contains examples of light verb constructions (e.g. "*take* place"), as well as literal meanings (e.g. "his eyebrows *raised* suddenly", "*after* 9 am"), which were tagged as metaphorical. The presence of erroneously annotated examples and the fact that the vast majority of metaphors in the corpus are highly conventional make this resource not directly suitable for the purposes of computational modelling described in this thesis.

Martin (2006) conducted a corpus study in order to confirm that metaphorical expressions occur in text in contexts containing lexical items from source and target domains. The difficulty associated with this approach is that it requires exhaustive lists of source and target domain vocabulary. The analysis was performed on the data from the Wall Street Journal (WSJ) corpus (Charniak et al., 2000) and focused on four conceptual metaphors that occur with considerable regularity in the corpus. These included NUMERICAL VALUE AS LOCATION, COMMERCIAL ACTIVITY AS CONTAINER, COMMERCIAL ACTIVITY AS PATH FOLLOWING and COMMERCIAL ACTIVITY AS WAR. Martin manually compiled the lists of terms characteristic for source and target domains by examining sampled metaphors of these types and then extended them through the use of a thesaurus. He then searched the corpus for sentences containing vocabulary from these lists and checked whether they contain metaphors of the above types. The goal was to evaluate the predictive ability of contexts containing vocabulary from the source domain and the target domain. In addition, Martin estimated the likelihood of a metaphorical expression following another metaphorical expression described by the same map-

ping. The most positive results were obtained for metaphors of the type NUMERICAL VALUE AS LOCATION ($P(Metaphor|Source) = 0.069$, $P(Metaphor|Target) = 0.677$, $P(Metaphor|Metaphor) = 0.703$). The low predictive ability of the source domain vocabulary may be due to the fact that source domains normally refer to our physical experiences. Consequently, the associated vocabulary would tend to occur independently, as opposed to abstract (target) concepts that are more likely to appear in metaphorical constructions.

Wallington et al. (2003) experimented with metaphor annotation in unrestricted text. They employed two teams of annotators and compared externally prescribed definitions of metaphor with intuitive internal ones. Team A was asked to annotate "interesting stretches", whereby a phrase was considered interesting if (1) its significance in the document was non-physical, (2) it could have a physical significance in another context with a similar syntactic frame, (3) this physical significance was related to the abstract one. Team B had to annotate phrases according to their own intuitive definition of metaphor. Apart from metaphorical expressions the respective source-target domain mappings were also to be annotated. For this latter task, the annotators were given a set of mappings from the Master Metaphor List and were asked to assign the most suitable ones. However, the authors do not report the level of interannotator agreement, i.e. the proportion of instances that were tagged similarly by all annotators, nor the coverage of the mappings in the Master Metaphor List on their data.

2.4 Computational approaches to metaphor

A computational approach to metaphor needs to address two tasks: metaphor recognition (or identification) and metaphor interpretation. This section is devoted to previous work in both these areas.

2.4.1 Automatic metaphor recognition

One of the first attempts to automatically identify and interpret metaphorical expressions in text is the approach of Fass (1991). It originates in the idea of Wilks (1978) and utilises hand-coded knowledge. Fass developed a system called `met*`, which is capable of discriminating between literalness, metonymy, metaphor and anomaly. It does this in three stages. First, literalness is distinguished from non-literalness using selectional preference violation as an indicator. In the case that non-literalness is detected, the respective phrase is tested for being metonymic using hand-coded patterns (such as CONTAINER-FOR-CONTENT). If the system fails to recognise metonymy, it proceeds to search the knowledge base for a relevant analogy in order to discriminate metaphorical relations from

Cue	BNC frequency	Sample size	Metaphors	Precision
"metaphorically speaking"	7	7	5	0.71
"literally"	1936	50	13	0.26
"figurative"	125	50	9	0.18
"utterly"	1251	50	16	0.32
"completely"	8339	50	13	0.26
"so to speak"	353	49	35	0.71

Table 2.1: Corpus statistics for linguistic cues

anomalous ones. For example, the sentence in (27b) would be represented in this framework as (*car,drink,gasoline*), which does not satisfy the preference (*animal,drink,liquid*), as *car* is not a hyponym of *animal*. met* then searches its knowledge base for a triple containing a hypernym of both the actual argument and the desired argument and finds (*thing,use,energy_source*), which represents the metaphorical interpretation.

Goatly (1997) identifies a set of linguistic cues, i.e. lexical patterns indicating the presence of a metaphorical expression in running text, such as *metaphorically speaking, utterly, completely, so to speak* and *literally*. However, this approach is likely to find only a small proportion of metaphorical expressions, as the vast majority of them appear without any signaling context. I conducted a corpus study in order to investigate the effectiveness of linguistic cues as metaphor indicators. For each cue suggested by Goatly (1997), I randomly sampled 50 sentences from the BNC containing it and manually annotated them for metaphoricity. The results are presented in Table 2.1. The average precision (i.e. the proportion of identified expressions that were metaphorical) of the linguistic cue method according to these data is 0.40, which suggests that the set of metaphors that this method generates contains a great deal of noise. Thus the cues are unlikely to be sufficient for metaphor extraction on their own, but together with some additional filters, they could contribute to a more complex system.

The work of Peters and Peters (2000) concentrates on detecting figurative language in lexical resources. They mine WordNet (Fellbaum, 1998) for examples of systematic polysemy, which allows them to capture metonymic and metaphorical relations. Their system searches for nodes that are relatively high in the WordNet hierarchy, i.e. are relatively general, and that share a set of common word forms among their descendants. Peters and Peters found that such nodes often happen to be in a metonymic (e.g. *publisher – publication*) or a metaphorical (e.g. *theory – supporting structure*) relation.

The CorMet system (Mason, 2004) is the first attempt at discovering source-target domain mappings automatically. It does this by finding systematic variations in domain-specific selectional preferences, which are inferred from texts on the web. For example, Mason collects texts from the LAB domain and the FINANCE domain, in both of which *pour* would be a characteristic verb. In the LAB domain *pour* has a strong selectional preference for objects of type *liquid*, whereas in the FINANCE domain it selects for *money*. From this

pour
nonliteral cluster
wsj04:7878 N As manufacturers get bigger, they are likely to pour more money into the battle for shelf space, raising the ante for new players.
wsj25:3283 N Salsa and rap music pour out of the windows.
wsj06:300 U Investors hungering for safety and high yields are pouring record sums into single-premium, interest-earning annuities.
literal cluster
wsj59:3286 L Custom demands that cognac be poured from a freshly opened bottle.

Figure 2.6: An example of the data of Birke and Sarkar (2006)

Mason's system infers the domain mapping FINANCE – LAB and the concept mapping MONEY IS LIQUID. He compares the output of his system against the Master Metaphor List and reports a performance of 77% in terms of accuracy, i.e. proportion of correctly induced mappings.

Birke and Sarkar (2006) present a sentence clustering approach for non-literal language recognition, implemented in the TroFi system (Trope Finder). The idea behind their system originates from a similarity-based word sense disambiguation method developed by Karov and Edelman (1998). The latter employs a set of seed sentences annotated with respect to word sense. The system computes similarity between the sentence containing the word to be disambiguated and all of the seed sentences and selects the sense corresponding to the annotation in the most similar seed sentences. Birke and Sarkar adapt this algorithm to perform a two-way classification (literal vs. non-literal), not aiming to distinguish between specific kinds of tropes. An example for the verb *pour* in their database is shown in Figure 2.6. They attain a performance of 0.54 in terms of *F-measure* (van Rijsbergen, 1979). F-measure is a weighted harmonic mean of precision and recall. An F-measure balanced with respect to precision and recall is defined as follows:

$$F = \frac{2PR}{P+R} \tag{2.1}$$

where P is precision that measures the proportion of examples classified as positive that are correctly classified, and R stands for recall that measures the proportion of examples which really are positive that are classified correctly.

The method of Gedigian et al. (2006) discriminates between literal and metaphorical use. The authors trained a maximum entropy classifier for this purpose. They collected their data using FrameNet (Fillmore et al., 2003) and PropBank (Kingsbury and Palmer, 2002) annotations. FrameNet is a lexical resource for English containing information on words' semantic and syntactic combinatory possibilities, or valencies, in each of their senses. PropBank is a corpus annotated with verbal propositions and their arguments. Gedigian et al. (2006) extracted the lexical items whose frames are related to MOTION

and CURE from FrameNet, then searched the PropBank Wall Street Journal corpus (Kingsbury and Palmer, 2002) for sentences containing such lexical items and annotated them with respect to metaphoricity. For example, the verb *run* in the sentence "Texas Air has *run* into difficulty" was annotated as metaphorical, and in "I was doing the laundry and nearly broke my neck *running* upstairs to see..." as literal. Gedigian et al. used PropBank annotation (arguments and their semantic types) as features to train the classifier, and report an accuracy of 95.12%. This result is, however, only 2.22% higher than the performance of the naive baseline assigning majority class to all instances (92.90%). Such high performance of their system can be explained by the fact that 92.90% of the verbs of MOTION and CURE in their data are used metaphorically, thus making the dataset unbalanced with respect to target categories and the task easier.

Both Birke and Sarkar (2006) and Gedigian et al. (2006) focus only on metaphors expressed by a verb. The approach of Krishnakumaran and Zhu (2007) additionally covers metaphors expressed by nouns and adjectives. Krishnakumaran and Zhu use hyponymy relation in WordNet and word bigram counts to predict metaphors at a sentence level. Given a metaphor in copula constructions, or an IS-A metaphor (e.g. the famous quote by William Shakespeare "All the world's a *stage*") they verify if the two nouns involved are in hyponymy relation in WordNet, otherwise this sentence is tagged as containing a metaphor. They also treat expressions containing a verb or an adjective used metaphorically (e.g. "He *planted* good ideas in their minds" or "He has a *fertile* imagination"). For those cases, they calculate bigram probabilities of verb-noun and adjective-noun pairs (including the hyponyms/hypernyms of the noun in question). If the combination is not observed in the data with sufficient frequency, the system tags the sentence as metaphorical. This idea is a modification of the selectional preference view of Wilks, however applied at the bigram level. Alternatively, one could extract verb-object relations from parsed text. Compared to the latter, Krishnakumaran and Zhu (2007) lose a great deal of information. The authors evaluated their system on a set of example sentences compiled from the Master Metaphor List, whereby highly conventionalised metaphors are taken to be negative examples. Thus they do not deal with literal examples as such. Essentially, the distinction Krishnakumaran and Zhu are making is between the senses included in WordNet, even if they are conventional metaphors (e.g. "*capture* an idea"), and those not included in WordNet (e.g. "*planted* good ideas").

2.4.2 Automatic metaphor interpretation

One of the first computational accounts of metaphor interpretation is that of Martin (1990). In his metaphor interpretation, denotation and acquisition system (MIDAS), Martin models the hierarchical organisation of conventional metaphors. The main assumption underlying this approach is that more specific conventional metaphors descend from more general ones. Given an example of a metaphorical expression, MIDAS searches

its database for a corresponding conceptual metaphor that would explain the anomaly. If it does not find any, it abstracts from the example to more general concepts and repeats the search. If a suitable general metaphor is found, it creates a new mapping for its descendant, a more specific metaphor, based on this example. This is also how novel conceptual metaphors are acquired by the system. The metaphors are then organised into a resource called MetaBank (Martin, 1994). The knowledge is represented in MetaBank in the form of *metaphor maps* (Martin, 1988) containing detailed information about source-target concept mappings and empirically derived examples. MIDAS has been integrated with Unix Consultant, a system that answers users' questions about Unix. The system first tries to find a literal answer to the question. If it is not able to, it calls MIDAS, which detects metaphorical expressions via selectional preference violation and searches its database for a metaphor explaining the anomaly in the question.

Another cohort of approaches aims to perform inference about entities and events in the source and target domains for the purpose of metaphor interpretation. These include the KARMA system (Narayanan, 1997, 1999; Feldman and Narayanan, 2004) and the ATT-Meta project (Barnden and Lee, 2002; Agerri et al., 2007). Within both systems the authors developed a metaphor-based reasoning framework in accordance with CMT. The reasoning process relies on manually coded knowledge about the world and operates mainly in the source domain. The results are then projected onto the target domain using the conceptual mapping representation. The ATT-Meta project concerns metaphorical and metonymic description of mental states; reasoning about mental states is performed using first order logic. Their system, however, does not take natural language sentences as input, but hand-coded logical expressions that are representations of small discourse fragments. KARMA in turn deals with a broad range of abstract actions and events and takes parsed text as input.

Veale and Hao (2008) derive a "fluid knowledge representation for metaphor interpretation and generation" called Talking Points. Talking Points is a set of characteristics of concepts belonging to source and target domains and related facts about the world which are acquired automatically from WordNet and from the web. Talking Points are then organised in *Slipnet*, a framework that allows for a number of insertions, deletions and substitutions in definitions of such characteristics in order to establish a connection between the target and the source concepts. This work builds on the idea of *slippage* in knowledge representation for understanding analogies in abstract domains (Hofstadter and Mitchell, 1994; Hofstadter, 1995). Below is an example demonstrating how slippage

operates to explain the metaphor *Make-up is a Western burqa*.

Make-up =>
≡ typically worn by women
≈ expected to be worn by women
≈ must be worn by women
≈ must be worn by Muslim women
Burqa <=

By doing insertions and substitutions, the system arrives from the definition "typically worn by women" at that of "must be worn by Muslim women". Thus it establishes a link between the concepts of *make-up* and *burqa*. Veale and Hao, however, did not evaluate to what extent their system is able to interpret metaphorical expressions in real-world text.

2.5 Main claims and contributions of the thesis

Metaphor understanding is a knowledge-intensive process. Hence its automation requires either an extensive manually-created knowledge-base or a robust knowledge acquisition system. The latter being a hard task, a great deal of metaphor research has resorted to the first option. Hand-coded knowledge has proved useful for both metaphor identification and interpretation (Fass, 1991; Martin, 1990; Narayanan, 1997; Barnden and Lee, 2002; Agerri et al., 2007). However, the systems utilising it can only ever have limited coverage, since it is impossible to capture information about all spheres of life in a manually created database. A number of researchers have thus used statistical modelling to address metaphor understanding (Mason, 2004; Birke and Sarkar, 2006; Gedigian et al., 2006; Krishnakumaran and Zhu, 2007). Although they have not experimented with metaphor in unrestricted text, these approaches were, nonetheless, a significant step on the route to a robust system. They demonstrated that statistical methods and broad-coverage lexical resources can be successfully employed to model at least some aspects of metaphor.

A considerable problem for computational processing of metaphor to date is the lack of a common task definition and a shared dataset. This is true for both metaphor identification and interpretation. Identification experiments sometimes aim to discover interconceptual mappings (Mason, 2004), rather than metaphorical expressions in text, or define the task as discriminating between conventional and novel metaphors (Krishnakumaran and Zhu, 2007), rather than between literal and metaphorical meanings in general. Interpretation experiments range from finding a path between two concepts involved in a metaphorical expression (Veale and Hao, 2008) to finding a common hypernym of the metaphorical term and its intended literal meaning (Fass, 1991). However, a unified task definition and a shared dataset could enable NLP researchers working on metaphor to directly compare their results.

Ideally, a computational metaphor processing task should be aimed at producing a representation of metaphor understanding that can be directly embedded into other NLP applications that could benefit from metaphor resolution. I define metaphor interpretation as a paraphrasing task and aim to build a system that identifies metaphorical expressions in text and produces their literal paraphrases. This would make the system directly applicable to external NLP tasks. For example, it has been already shown that statistical MT systems can benefit from an additional paraphrasing component (Callison-Burch et al., 2006)). By paraphrasing metaphorical expressions, my system would help to accurately translate such examples as "all of this *stirred* an uncontrollable excitement in her" (see Chapter 1), which is likely to improve the overall performance of the MT application.

As opposed to previous approaches that modeled metaphorical reasoning starting from the hand-crafted description and applying it to explain the data, I aim to design a statistical model that captures regular patterns of metaphoricity in a large corpus and thus generalises to unseen examples. Compared to labour-intensive manual efforts, this approach is more robust and, being nearly unsupervised, cost-effective. In contrast to previous statistical approaches, which addressed metaphors of a specific topic or did not consider linguistic metaphor at all (e.g. Mason, 2004), the proposed method covers all metaphors in principle, can be applied to unrestricted text and can be adapted to different domains and genres.

The main contributions of the thesis lie within three areas: human annotation of metaphor, automatic metaphor identification and automatic metaphor interpretation. In particular, this thesis provides

- new task definitions for metaphor identification and interpretation that make them compatible with other NLP systems;
- novel computational models for metaphor identification and interpretation in unrestricted text which do not rely on any manually created knowledge specific to metaphor, but rather employ statistical modelling of linguistic data;
- a new annotation scheme for the identification of metaphorical expressions and the associated source–target domain mappings in unrestricted text, and the first publicly available metaphor corpus annotated for conceptual metaphor.

The focus of the thesis is on single-word metaphors expressed by a verb. Verbs are frequent in language and central to conceptual metaphor. Cameron (2003) conducted a corpus study of the use of metaphor in educational discourse for all parts of speech. She found that verbs account for around 50% of the data, the rest shared by nouns, adjectives, adverbs, copula constructions and multi-word metaphors. This suggests that verb metaphors provide a reliable testbed for both linguistic and computational experiments.

I therefore test the identification and interpretation systems on verb metaphors. I would however expect the presented methods to scale to other parts of speech and to a wide range of syntactic constructions, since they rely on techniques from computational lexical semantics that have been shown effective in modelling not only verb meanings, but also those of nouns and adjectives.

Below, I provide a detailed overview of the main contributions of the thesis.

2.5.1 Metaphor annotation and corpus

The principles of CMT have guided researchers working on metaphor ever since its formulation. Despite this, there still has been no corpus-based study of conceptual metaphor nor a proposal for a comprehensive procedure for annotation of cross-domain mappings. However, corpus-based annotation of conceptual mappings would allow us to see whether the theory can adequately explain real-world data. If successful, such a corpus could provide a new starting point for linguistic and computational experiments on metaphor.

The annotation scheme presented in the thesis is a step towards filling this gap. It is a joint scheme for identification of metaphorical expressions and source-target domain mappings. The procedure does not rely on predefined metaphorical mappings, but instead makes use of independent sets of common source and target domain categories. This results in a more flexible model of metaphorical associations than, for instance, that of Wallington et al. (2003), which is limited to a set of mappings exemplified in the Master Metaphor List.

The annotation was carried out on real-world texts taken from the BNC, representing various genres. I tested the scheme in an experimental setting involving multiple annotators and measured their agreement on the task. The focus of the study is on single-word metaphors expressed by a verb. The annotators were asked to (1) classify the verbs in the text into two categories: metaphorical or literal and (2) identify the interconceptual mapping for each verb they tagged as metaphorical. For the second task, the annotators were given precompiled lists of suggested source and target domain labels, from which they selected the categories that – in their judgement – described the source and target concepts best or introduced their own category if the relevant list did not contain the desired one. I expect the assignment of domain labels to be the most challenging part of the annotation process. The main goal of the study is thus to verify whether such labels can be assigned consistently.

Only a part of the corpus was annotated by multiple annotators, to measure reliability. The rest of the dataset was annotated by myself. Besides verbal metaphors, this annotation also captured metaphors expressed by nouns, adjectives and adverbs, in order to estimate metaphor statistics across part-of-speech classes and syntactic constructions. Verbal metaphorical expressions annotated in the corpus provide a testbed for system

development. The design and evaluation of the annotation scheme, as well as the resulting corpus are described in Chapter 3. A summary of all human experiments, including annotation, is presented in Appendix A.

2.5.2 Metaphor identification system

The first task for metaphor processing within NLP is its identification in text. Previous approaches to this problem either utilise hand-coded knowledge (Fass, 1991; Krishnakumaran and Zhu, 2007) or reduce the task to searching for metaphors of a specific domain defined a priori (e.g. MOTION metaphors) in a specific type of discourse, e.g. the Wall Street Journal (Gedigian et al., 2006). In contrast, the search space in my experiments is the entire British National Corpus and the domain of the expressions identified is unrestricted. In addition, the developed technique does not rely on any hand-crafted lexical or world knowledge, but rather captures metaphoricity by means of verb and noun clustering in a data-driven manner.

The motivation behind the use of clustering methods for the metaphor identification task lies in CMT. The patterns of conceptual metaphor (e.g. FEELINGS ARE LIQUIDS) always operate on semantic classes, i.e. groups of related concepts, defined by Lakoff and Johnson as conceptual domains (e.g. FEELINGS include *love, anger, hatred* etc.; LIQUIDS include *water, tea, petrol, beer* etc.) Thus modelling metaphorical mechanisms in accordance with CMT would involve capturing such semantic classes automatically.

Previous research on corpus-based lexical semantics has shown that it is possible to automatically induce semantic word classes from corpus data via clustering of contextual cues (Pereira et al., 1993; Lin, 1998; Schulte im Walde, 2006). The consensus is that the lexical items showing similar behavior in a large body of text most likely have related meanings. Clustering words according to their distribution in particular syntactic contexts in a corpus is known as *distributional clustering*.

The method behind the metaphor identification system presented in this thesis relies on distributional clustering. Noun clustering, specifically, is central to the approach. It is traditionally assumed that noun clusters produced using distributional clustering contain concepts which are similar to each other. This is, however, true only in part. There exist two types of concepts: *concrete*, i.e. those denoting physical entities or physical experiences (e.g. *chair, apple, house, rain*) and *abstract*, that do not physically exist at any particular time or place, but rather exist as a type of thing or as an idea (e.g. *justice, love, democracy*). It is the abstract concepts that tend to be described metaphorically, rather than concrete concepts. Humans use metaphor attempting to gain a better understanding of an abstract concept by comparing it to their physical experiences.

As a result, abstract concepts expose different distributional behavior in a corpus. This in turn affects the application of clustering techniques and the obtained clusters for concrete

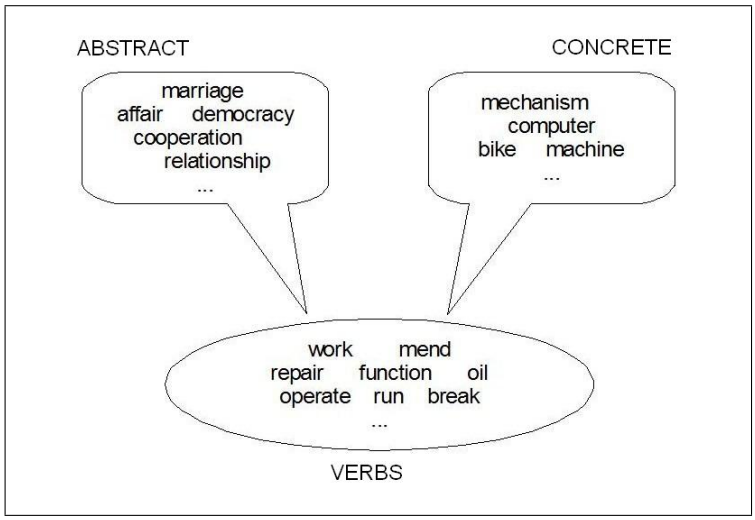

Figure 2.7: Cluster of target concepts associated with MECHANISM

and abstract concepts would be structured differently. Consider the example in Figure 2.7. The figure shows a cluster containing concrete concepts (on the right) that are various kinds of mechanisms; a cluster containing verbs co-occurring with mechanisms in the corpus (at the bottom); and a cluster containing abstract concepts (on the left) that tend to co-occur with these verbs. Such abstract concepts albeit having quite distinct meanings (e.g. *marriage* and *democracy*), are observed in similar lexico-syntactic environments. This is due to the fact that they are systematically used metaphorically with the verbs from the domain of MECHANISM. Hence, they are automatically assigned to the same cluster. The following examples illustrate this phenomenon in textual data.

(32) Our relationship is not really *working*.

(33) Diana and Charles did not succeed in *mending* their marriage.

(34) The *wheels* of Stalin's regime were *well oiled* and already *turning*.

Such a structure of the abstract clusters can be explained by the fact that *relationships*, *marriages*, *collaborations* and *political systems* are all cognitively mapped to the same source domain of MECHANISM. In contrast to concrete concepts, such as *tea, water, coffee, beer, drink, liquid*, that are clustered together due to meaning similarity, abstract concepts tend to be clustered together if they are associated with the same source domain. I define this phenomenon as *clustering by association* and it becomes central to the system design. The expectation is that clustering by association would allow the harvesting of new target domains that are associated with the same source domain, and thus identify new metaphors. To confirm this hypothesis I carry out a qualitative and quantitative analysis of the output of a state-of-the-art noun clustering system. The details of this analysis are presented in Chapter 4.

The system starts from a small set of *seed* metaphorical expressions, i.e. annotated metaphors (such as those in (32) or (33)), which serve as training data. Note that seed annotation only concerns linguistic metaphors; metaphorical mappings are not annotated. The system then (1) creates source domains describing these examples by means of verb clustering (such as the verb cluster in Figure 2.7); (2) identifies new target domains associated with the same source domain by means of noun clustering and (3) establishes a link between the source and the target clusters based on the examples.

Thus the system captures metaphorical associations implicitly. It generalises over the associated domains by means of verb and noun clustering. The obtained clusters then represent source and target concepts between which metaphorical associations hold. The knowledge of such associations is then used to identify new metaphorical expressions in a large corpus.

In addition to this, I build a selectional preference-based metaphor filter. This idea stems from the view of Wilks (1978), namely that metaphor represents a violation of selectional preferences in a given context. My method, however, is a modification of Wilks' view. The filter assumes that the verbs exhibiting weak selectional preferences, i.e. verbs co-occurring with any argument class in linguistic data, e.g. *remember, influence* etc., generally have no or only weak potential of being a metaphor. It has been previously shown that it is possible to quantify verb selectional preferences on the basis of corpus data (e.g. Resnik, 1993). Thus, once the candidate metaphors are identified in the corpus using clustering methods, the ones displaying weak selectional preferences can be filtered out.

In summary, the system (1) starts from a seed set of metaphorical expressions exemplifying a range of source–target domain mappings; (2) performs noun clustering in order to harvest various target concepts associated with the same source domain; (3) creates a source domain verb lexicon by means of verb clustering; (4) searches the corpus for metaphorical expressions describing the target domain concepts using the verbs from the source domain lexicon; (5) filters out the candidates exposing weak selectional preference strength as non-metaphorical.

I test the system starting with a collection of metaphorical expressions representing verb-subject and verb-direct object constructions, where the verb is used metaphorically. The quality of metaphor identification is evaluated with the aid of human judges. The system is then also compared to a baseline built upon WordNet, which demonstrates that the method reaches beyond synonymy and captures novel metaphors not directly related to any of those seen in the seed set (e.g. having seen a metaphor "*stir* excitement" the system concludes that the verb in "*swallow* anger" is also used metaphorically).

2.5.3 Metaphor interpretation system

As is the case in metaphor identification, the majority of approaches to metaphor interpretation also rely on task-specific hand-coded knowledge (Fass, 1991; Martin, 1990; Narayanan, 1997, 1999; Feldman and Narayanan, 2004; Barnden and Lee, 2002; Agerri et al., 2007) and produce interpretations in a non-textual format (Veale and Hao, 2008). However, the ultimate objective of automatic metaphor processing is a type of interpretation that can be directly embedded into other systems to enhance their performance. I thus define metaphor interpretation as a paraphrasing task and build a system that automatically derives literal paraphrases for metaphorical expressions in unrestricted text. My method is also distinguished from previous work in that it does not rely on any hand-crafted knowledge about metaphor, but in contrast is corpus-based and employs automatically induced selectional preferences.

The metaphor paraphrasing task can be divided into two subtasks: (1) identifying *paraphrases*, i.e. other ways of expressing the same meaning in a given context, and (2) discriminating between literal and metaphorical paraphrases. Consequently, the proposed approach is theoretically grounded in two ideas:

- The meaning of a word in context emerges through interaction with the meaning of the words surrounding it. This assumption is widely accepted in lexical semantics theory (Pustejovsky, 1995; Hanks and Pustejovsky, 2005) and has been exploited for lexical acquisition (Schulte im Walde, 2006; Sun and Korhonen, 2009). It suggests that the context itself imposes a certain expectation on which words can occur within it. Given a large amount of linguistic data, it is possible to model this expectation in probabilistic terms (Lapata, 2001). This can be done by deriving a ranking scheme for possible paraphrases that fit or do not fit in a specific context based on word co-occurrence evidence. This is how initial paraphrases are generated within the metaphor interpretation system.

- Literalness can be detected via strong selectional preference. This idea is a mirror-image of the selectional preference violation view of Wilks (1978), who suggested that a violation of selectional preferences indicates a metaphor. The key information that selectional preferences provide is whether there is an association between the predicate and its potential argument and how strong it is. A literal paraphrase would normally come from the target domain (e.g. "understand the explanation") and be strongly associated with the target concept, whereas a metaphorical paraphrase would belong to the source domain (e.g. "*grasp* the explanation") and be associated with the concepts from this source domain more strongly than with the target concept. Thus I use a selectional preference model to measure the semantic fit of the generated paraphrases into the given context as opposed to all other contexts. The highest semantic fit then indicates the most literal paraphrase.

Thus the context-based probabilistic model is used for paraphrase generation and the selectional preference model for literalness detection. The key difference between the two models is that the former favours the paraphrases that co-occur with the words in the context more frequently than other paraphrases do, and the latter favours the paraphrases that co-occur with the words from the context more frequently than with any other lexical items in the corpus. This is the main intuition behind the approach.

The system thus incorporates the following components:

- **a context-based probabilistic model** that acquires paraphrases for metaphorical expressions from a large corpus;
- **a WordNet similarity component** that filters out the irrelevant paraphrases, e.g. the antonymous ones that can occur within the same context;
- **a selectional preference model** that discriminates literal paraphrases from the metaphorical ones.

The paraphrasing system was first evaluated on its own on a set of manually annotated metaphorical expressions against human judgements. Subsequently, it was combined with the metaphor identification method into an integrated metaphor processing system and the two systems were tested operating together. The details of the system design, implementation and evaluation are presented in Chapter 5.

Chapter 3

Annotation of linguistic and conceptual metaphor

Besides creating a dataset for system evaluation, metaphor annotation can shed light on how conceptual metaphor manifests itself in linguistic data. Most examples of metaphorical expressions and associated mappings in linguistic literature are carefully selected to clearly demonstrate the interconceptual correspondences. However, such examples do not adequately illustrate the behavior of the phenomena in real-world text, for which a corpus-based account is needed. I expect the annotation study presented in this chapter to reveal (1) how intuitive the conceptual metaphor explanation of linguistic metaphors is for human annotators and whether it is possible to consistently annotate interconceptual mappings; (2) what are the main difficulties that the annotators experience during the annotation process; (3) whether one conceptual metaphor is sufficient to explain a linguistic metaphor or a chain of conceptual metaphors is needed; and (4) what proportion of metaphorical expressions can be explained using the proposed lists of most general source and target categories suggested in the MML.

This chapter starts by describing the dataset and the annotation scheme used to identify both linguistic and conceptual metaphor in text, and then presents the annotation reliability study conducted in a setting with multiple annotators and the analysis of the resulting corpus.

3.1 Data

The annotation study was conducted on a set of texts taken from the British National Corpus (BNC) (Burnard, 2007). BNC is a 100 million word corpus containing samples of written (90%) and spoken (10%) British English from the second half of the 20th century. The data for it was gathered from a wide range of sources and the corpus is balanced with respect to genre, style and topic. As such, it provides a suitable platform for the

development of a metaphor corpus, aimed at the study of metaphor in real-world texts in contemporary English.

To collect the data for the metaphor corpus I sampled texts from the BNC representing various genres: fiction, newspaper and journal articles, essays on politics, international relations and sociology, and radio broadcast (transcribed speech). This allowed for a study of metaphor in diverse discourse. The total size of the corpus annotated is 13,642 words.

3.2 Annotation scheme

The task is to identify both linguistic metaphors and the corresponding conceptual metaphors. The annotation process will, therefore, operate in two stages. First, lexical items are classified as either metaphorically or literally used. Then, for all cases of metaphorical use the appropriate source-target domain mappings need to be assigned. The annotation scheme thus addresses two problems: the distinction between literal and metaphorical language in text and the formalisation of human conceptualisation of metaphorical mappings.

3.2.1 Main principles and challenges

The main challenges in developing such a metaphor annotation procedure are the choice of the definition of metaphor and a suitable inventory of source and target domain categories used to assign the mappings.

- **Definition of metaphor** As already mentioned in section 2.1, the distinction between metaphorical and literal meanings is not always clear-cut. A large number of metaphorical expressions are conventionalised to the extent that they are perceived as literal by most native speakers (e.g. "He *found out* the truth"). Some approaches consider only novel expressions to be truly metaphorical (Krishnakumaran and Zhu, 2007), whereas others consider any linguistic expression to be metaphorical where an underlying analogy can be identified (Steen et al., 2010).

  This thesis assumes a definition of metaphor informed by the needs of NLP. This means that both novel and conventional metaphors are interesting for annotation; however, only including the conventional cases where both literal and metaphorical senses are commonly used and stand in clear opposition in contemporary language.

- **Inventory of categories** The primary question one faces when trying to derive an annotation scheme for metaphorical associations is defining a set of source and target domain categories. As opposed to the previous approach of Wallington et al.

(2003), who used a predefined set of fixed mappings from the MML (e.g. LIFE IS A JOURNEY), in my scheme both source (e.g. JOURNEY) and target (e.g. LIFE) domains can be chosen independently. I expect that this will allow for higher flexibility of annotation and thus provide a better reflection of human intuitive conceptualisation of metaphor, as well as the identification of novel mappings.

The main properties of categories to consider while designing and evaluating such an annotation scheme are their coverage and specificity. The inventory of categories should cover a wide range of topics and genres. The categories themselves should be at the right level of generality, i.e. not too general (to ensure they are sufficiently informative for the task), but at the same time not too specific (to ensure they provide high coverage of the data).

The remainder of this section describes how the annotation scheme was developed and tested with these principles in mind.

3.2.2 Source and target domain categories

To date the most comprehensive resource of metaphorical mappings is the Master Metaphor List (Lakoff et al., 1991). Its source and target domain categories were repeatedly adopted for linguistics and NLP research (Barnden and Lee, 2002; Lönneker, 2004). Following these approaches, I relied on a subset of categories from the Master Metaphor List to construct the inventory of categories for annotation.

I selected a number of general categories from the MML, e.g. LOCATION, CONTAINER, JOURNEY, LIFE, TIME, RELATIONSHIP, and arranged them into source and target concept lists. These lists were then given as suggested categories to annotators. Suggested source and target concepts are shown in Tables 3.1 and 3.2 respectively. The expectation is that the categories in these lists would account for a large proportion of metaphorical data, i.e. provide high, albeit not exhaustive, coverage. In order to test their coverage, I conducted a pilot study on a small text sample (2,750 words) from the BNC. I annotated metaphorical expressions and the corresponding interconceptual mappings in these texts using the categories from the suggested source and target concept lists. The study revealed that the target concept list accounted for 76% of metaphorical expressions in these texts, whereas the source concept list had a 100% coverage. Such discrepancy can be explained by the fact that target categories, which tend to describe abstract concepts, are significantly less restricted than source categories that stand for our physical experiences. In other words, we can use metaphor to talk about an unlimited number of abstract things, whereas the entities, events and processes to which we compare them are limited to the actual physical experience we all share. Thus the set of potential target concepts is likely to be significantly larger and harder to predict. To account for this, the annotators, although strongly encouraged to use categories from the provided lists, were

Source concepts
PHYSICAL OBJECT
LIVING BEING
ADVERSARY/ENEMY
LOCATION
DISTANCE
CONTAINER
PATH
PHYSICAL OBSTACLE (e.g. barrier)
DIRECTIONALITY: e.g. UP/DOWN
BASIS/PLATFORM
DEPTH
GROWTH/RISE
SIZE
MOTION
JOURNEY
VEHICLE
MACHINE/MECHANISM
STORY
LIQUID
POSSESSIONS
INFECTION
VISION

Table 3.1: Suggested source concepts

allowed to introduce novel categories in cases where they felt no category from the lists could adequately explain the instance. This step is also crucial for the identification of novel unconventional mappings.

3.2.3 Annotation procedure

Metaphor annotation is carried out at the word level. The proposed annotation scheme is based on some of the principles of the metaphor identification procedure developed by Pragglejaz Group (2007). I adopt their definition of a basic sense of a word and their approach to distinguishing basic senses from metaphorical ones. I modify and extend the procedure to identify source-target domain mappings by comparing the contexts in which a word appears in its basic and metaphorical senses. Besides assigning labels to metaphorical associations, this stage of the procedure then feeds back into the metaphor identification process and acts as an additional constraint on metaphoricity.

Since the experiments presented in the thesis focus on metaphors expressed by a verb, the annotation procedure, although in principle suitable for the analysis of all parts of speech, was tailored to verb metaphors. The procedure used as part of annotation guidelines is presented below.

Target concepts
LIFE
DEATH
TIME/MOMENT IN TIME
FUTURE
PAST
CHANGE
PROGRESS/EVOLUTION/DEVELOPMENT
SUCCESS/ACCOMPLISHMENT
CAREER
FEELINGS/EMOTIONS
ATTITUDES/VIEWS
MIND
IDEAS
KNOWLEDGE
PROBLEM
TASK/DUTY/RESPONSIBILITY
VALUE
WELL-BEING
SOCIAL/ECONOMIC/POLITICAL SYSTEM
RELATIONSHIP

Table 3.2: Suggested target concepts

1. For each verb establish its meaning in context and try to imagine a more basic meaning of this verb in other contexts. As defined in the framework of MIP (Pragglejaz Group, 2007) basic meanings are normally:
   - more concrete;
   - related to bodily action;
   - more precise (as opposed to vague);
   - historically older.
2. If you can establish a basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically. Try to identify a mapping between the source domain (where the basic meaning comes from) and the target domain (the concepts forming the context of the verb in front of you) using the provided lists of source and target categories. Record the mapping. If you fail to identify a mapping, reconsider whether the sense is really metaphorical in this context.

The following example illustrates how the procedure operates in practice.

(35) If he asked her to post a letter or buy some razor blades from the chemist, she was transported with pleasure.

In this sentence one needs to annotate the four verbs that are underlined.

- The first 3 verbs are used in their basic sense, i.e. literally (*ask* in the context of "a person asking another person a question or a favour"; *post* in the context of "a person posting/sending a letter by post"; *buy* in the sense of "making a purchase"). Thus they are tagged as literal.
- The verb *transport*, however, in its basic sense is used in the context of "goods being transported/carried somewhere by a vehicle". The context in this sentence involves "a person being transported by a feeling", which contrasts with the basic sense in that the agent of *transporting* is an EMOTION (the target concept) as opposed to a VEHICLE (the source concept). Thus one can infer that the use of *transport* in this sentence is metaphorical and the associated interconceptual mapping is EMOTIONS – VEHICLES.

3.3 Annotation reliability

After an annotation scheme has been developed its reliability needs to be verified. Reliability of a scheme can be assessed by comparing annotations carried out by multiple annotators independently (Krippendorff, 1980). This section describes an experiment where the same small portion of the metaphor corpus was annotated by several participants.

3.3.1 Data

A text sample from the BNC (text ID: ACA) was selected for the reliability study. Since the focus of the study is on single-word metaphors expressed by a verb, the first part of the annotation task can be viewed as verb classification according to whether the verbs are used metaphorically or literally. However, some verbs inherently have a weak potential, or no potential at all, to be used metaphorically, and as such the study is not concerned with them. The following verb classes were excluded: (1) auxiliary verbs; (2) modal verbs; (3) aspectual verbs (e.g. *begin, start, finish*); and (4) light verbs (e.g. *take, give, put, get, make*).

3.3.2 Annotation experiment

Subjects Three independent volunteer annotators participated in the experiment. They were native speakers of English and held a graduate degree in linguistics or computer science.

Material and Task The subjects were given the same text from the BNC which was an essay on sociology. The text contained 142 verbs to annotate, that were underlined. They were asked to (1) classify verbs as metaphorical or literal, and (2) identify the source-target domain mappings for the verbs they marked as metaphorical. They received two lists of suggested categories describing source and target concepts, and asked to select one from each list, in a way that described the metaphorical mapping best. Along with this they were allowed to introduce new categories if they felt none of the given categories expressed the mapping well enough. The annotation was done electronically using colour highlighting and inserting category labels in Microsoft Word[1].

Guidelines and Training The annotators received written guidelines (2 pages) and were asked to do a small annotation exercise (2 sentences: 1 example sentence and 1 sentence to annotate, containing 8 verbs in total). The goal of the exercise was to ensure they were at ease with the annotation format. Both annotation guidelines and the exercise are reproduced in Appendix B.

3.3.3 Interannotator agreement

Semantic annotations involve interpretation on the part of the participant and are thus inherently subjective. It is therefore essential to report *interannotator agreement*, that quantifies the similarity of the annotations produced by different annotators. I evaluated reliability of the proposed annotation scheme by assessing interannotator agreement in terms of κ statistic (Siegel and Castellan, 1988) on both tasks separately.

Kappa statistic As opposed to simple percentage of identically tagged instances, κ measures agreement by factoring out that expected by chance. It is calculated as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{3.1}$$

where $P(A)$ is the proportion of times that k annotators agree and $P(E)$ the proportion of times one would expect them to agree by chance. If there is perfect agreement among the annotators, then $\kappa = 1$, whereas if there is no agreement besides what is expected by chance, then $\kappa = 0$. κ is negative when annotators agree less than expected by chance. The amount of agreement one would expect to happen by chance depends on the number and distribution of the categories used for annotation.[2] The values of κ are traditionally interpreted using the following scale (Landis and Koch, 1977): 0 - 0.20 indicates slight agreement, 0.21 - 0.40 fair agreement, 0.41 - 0.60 moderate agreement, 0.61 - 0.80 substantial agreement, 0.81 - 1 means the agreement is almost perfect.

[1] All subsequent human experiments reported in the thesis were carried out electronically using Microsoft Word.

[2] For more details on how chance agreement is calculated see Siegel and Castellan (1988).

Annotator	Metaphors	Annotated mappings	Target chosen from list	Source chosen from list
A	53	53	52	52
B	39	39	39	37
C	58	51	42	25

Table 3.3: Differences in annotations

Results The number of metaphors and their conceptual mappings as annotated by the participants are shown in Table 3.3. The average proportion of the cases where a conceptual metaphor could be annotated for a given linguistic metaphor (across the three annotators) was 95%, whereas that using the categories from the provided lists was 82%.

The reliability of the scheme was first measured for the task of metaphor identification and then for the assignment of interconceptual mappings. The identification of metaphorical verbs yielded a reliability of $\kappa = 0.64$ ($n = 2; N = 142; k = 3$), where n stands for the number of categories, N for the number of instances annotated and k for the number of annotators. This level of agreement is considered substantial.

The measurement of the agreement in the second task appeared less straightforward. It was complicated by the fact that each annotator only assigned conceptual mappings to a set of verbs that in their judgement were metaphorical. These sets were not identical for all annotators. Thus, the agreement on the assignment of source and target domain categories was calculated only using the instances that all annotators considered to be metaphorical. This yielded a total of 30 conceptual mappings to compare.

One of the annotators (C) found the provided categories insufficient. Although trying to use them where possible, he nonetheless had to introduce a large number of categories of his own to match his intuitions. In addition, he did not assign any mapping for 7 metaphorical expressions. Both of these issues complicated the comparison of his annotation to those of the other annotators. Thus, his labelling of the mappings was excluded from the calculation of kappa statistic for agreement on conceptual metaphor annotation. However, his data was qualitatively analysed along with the rest.

The resulting overall agreement on the assignment of conceptual metaphor was thus $\kappa = 0.57$ ($n = 26; N = 60; k = 2$), whereby the agreement was stronger on the choice of the target categories ($\kappa = 0.60$ ($n = 14; N = 30; k = 2$)) than the source categories ($\kappa = 0.54$ ($n = 12; N = 30; k = 2$)).

Analysis of annotations Analysing cases of disagreement during metaphor identification suggests that the main source of disagreement was the conventionality of some metaphorical uses. These include expressions whose metaphorical etymology can be clearly traced, but the senses are lexicalised (e.g. "*fall* silent", "the end is *coming*") and thus perceived by some annotators as literal.

According to the annotators' informal feedback on the experiment, they found the task of identifying linguistic metaphor relatively straightforward, whereas the task of assigning

the respective conceptual metaphor appeared more difficult. The analysis of annotations has shown that one of the sources of disagreement in the latter task was the presence of partially overlapping categories in the target concept list. For example, the categories of PROGRESS and SUCCESS, or VIEWS, IDEAS and METHODS were often confused. This level of granularity was chosen following the Master Metaphor List. However, the annotated data suggests that, for the purpose of annotation of conceptual mappings, such categories may be joined into more general categories without significant information loss (e.g. VIEWS, IDEAS and METHODS can be covered by a single category IDEAS). This would increase mutual exclusivity of categories and thus lead to a more consistent annotation. Based on the observations in the data and the annotators' feedback, the source and target lists were refined to ensure no or minimal overlap between the categories, while maximally preserving their informativeness. As a post-hoc experiment, the labels in the annotations were mapped to this new set of categories and the annotations were compared again. The agreement rose to $\kappa = 0.61$ ($n = 23; N = 60; k = 2$), as expected.

Further examples of similarities and differences in the annotations are given in Figure 3.1. As the examples illustrate, the annotators tend to agree on whether a verb is used metaphorically or literally (with the exception of the verb *catch* tagged as literal by Annotator B). Their choices of source and target domain categories, however, vary. The annotators often choose the same target domain, although they refer to it by different (overlapping) labels, e.g. IDEA/THOUGHT/VIEW or TIME/MOMENT IN TIME. Annotator C introduced a more general category PERCEPTION, rather than using the more specific category VISION provided in the list, or DISEASE instead of the suggested category INFECTION. Thus they tend to choose categories that are intuitively related and the variation of the target domain labels is rather due to the granularity of categories used. In contrast, the choice of the source domain labels exhibits more conceptual variation. Annotator A tends to assign a general category PHYSICAL OBJECT to all instances appearing within the context related to physical activity, whereas Annotator B opts for finer-grained categories, as well as conceptualising the context in terms of events and actions rather than objects. These observations suggest that, although the annotators may share some of the intuitions with respect to conceptual metaphor, the explicit labelling of the latter in text is a challenging task.

3.4 Corpus data analysis

In order to create a dataset for experimentation, as well as perform a more comprehensive data analysis, I annotated a larger corpus using the above annotation scheme. The corpus contains 761 sentences and 13,642 words. The text used for the reliability study constituted a part of the corpus. This allowed me to measure my own agreement with the external annotators. The agreement on the identification of linguistic metaphor was

Annotator A

The Impressionist painters caught (IDEA -- PHYSICAL OBJECT) the contagion, and the new race of photographers tried to seize (MOMENT IN TIME -- PHYSICAL OBJECT) the fleeting moment and make it stay (MOMENT IN TIME -- PHYSICAL OBJECT). Cultures and historical periods differ ( ) greatly in their concepts of time and the continuity of life. We live ( ) in a century imprinted (IDEA -- PHYSICAL OBJECT) on the present, which regards (VIEWS -- VISION) the past as little more than the springboard from which we were launched (PROGRESS -- MOTION) on our way.

Annotator B

The Impressionist painters caught ( ) the contagion, and the new race of photographers tried to seize (TIME – ACTION) the fleeting moment and make it stay (TIME – MOTION). Cultures and historical periods differ ( ) greatly in their concepts of time and the continuity of life. We live ( ) in a century imprinted (VIEWS – MECHANISM) on the present, which regards (VIEWS – VISION) the past as little more than the springboard from which we were launched (PROGRESS – MOTION) on our way.

Annotator C

The Impressionist painters caught (IDEA – DISEASE) the contagion, and the new race of photographers tried to seize (MOMENT IN TIME – PHYSICAL OBJECT) the fleeting moment and make it stay (MOMENT IN TIME – PHYSICAL OBJECT). Cultures and historical periods differ ( ) greatly in their concepts of time and the continuity of life. We live ( ) in a century imprinted (TIME – PAGE) on the present, which regards (THOUGHT - PERCEPTION) the past as little more than the springboard from which we were launched (PASSAGE OF TIME – JOURNEY) on our way.

Figure 3.1: Example of similarities and differences in annotation

$\kappa = 62$ ($n = 2; N = 142; k = 4$), whereas that on the choice of source and target domain categories reached $\kappa = 0.58$ ($n = 22; N = 56; k = 3$). An extract from the annotated corpus is shown in Appendix C.

As an additional experiment, I also annotated nouns, adjectives and adverbs in the corpus

Text	ID	Genre	Sent.	Words	Met–rs	Met./Sent.	Verb m.
Hand in Glove, Goddard	G0N	Literature	335	3927	41	0.12	30
After Gorbachev, White	FYT	Politics	45	1384	23	0.51	17
Today newspaper	CEK	News	116	2086	48	0.41	30
Tortoise by Candlelight, Bawden	HH9	Literature	79	1366	12	0.15	10
The Masks of Death, Cecil	ACA	Sociology	60	1566	70	1.17	42
Radio broadcast (current affairs)	HM5	Speech	58	1828	10	0.17	7
Language and Literature journal	J85	Article	68	1485	37	0.54	28
Total			761	13642	241	0.32	164

Table 3.4: Corpus statistics for metaphor

as metaphorical or literal using the same procedure. This was done in order to investigate how metaphor can be expressed by other word classes, to gather metaphor statistics across a wider range of syntactic constructions and to estimate the relative proportion of verbal metaphors across genres (the study by Cameron (2003) only concerned metaphor in educational discourse). In what follows I will describe statistics of the resulting corpus and attempt to identify common traps in the annotation of source-target domain mappings in real-world text.

3.4.1 Metaphor statistics across genres and syntactic constructions

Metaphor frequency was calculated as a number of metaphors relative to the number of sentences in the text. The results presented in Table 3.4 indicate that metaphor is overall an extremely frequent phenomenon - it appears on average in every third sentence. This makes its automatic analysis indispensable for a wide range of NLP applications. An interesting finding is that fiction texts seem to contain fewer metaphors than other genres. However, it should be noted that the frequency metric used is biased towards genres with longer sentences, and fiction texts contain some dialogues consisting of short phrases. In addition, the dialogues themselves tend to contain mainly literal language, as opposed to author's descriptions where metaphors are more frequent. Overall, therefore, fiction contains relatively fewer metaphorical expressions than other genres.

The last column of Table 3.4 shows the proportion of verb metaphors across genres. The distribution of their frequency over genres appears similar to that of other part of speech classes. However, it should be noted that metaphors expressed by a verb are by a large margin the most frequent type and constitute 68% of all metaphorical expressions in the corpus.

Frequency	Source concepts
0.23	MOTION
0.13	VISION/SEEING
0.13	LIVING BEING
0.13	GROWTH/RISE
0.07	SPEED
0.03	DIRECTIONALITY: e.g. UP/DOWN
0.03	BASIS/PLATFORM
0.03	LOCATION
0.03	DISTANCE
0.03	MACHINE/MECHANISM
0.03	PHYSICAL OBJECT
...	

Table 3.5: Distribution of source concepts

Frequency	Target concepts
0.27	ATTITUDES/VIEWS
0.13	CHANGE
0.12	TIME/MOMENT IN TIME
0.12	PROGRESS/EVOLUTION/DEVELOPMENT
0.05	BEHAVIOUR
0.05	SUCCESS/ACCOMPLISHMENT
0.05	FUTURE
0.05	CAREER
0.03	SOCIAL/ECONOMIC/POLITICAL SYSTEM
0.03	IDEAS
0.03	METHODS
0.03	KNOWLEDGE
0.02	DEATH
0.02	PAST

Table 3.6: Distribution of target concepts

3.4.2 Mappings statistics

It is also interesting to look at the distributions of the source and target categories in the text annotated by the three annotators, shown in Tables 3.5 and 3.6 respectively. The topic of the text (in this case sociology) has an evident influence on the kind of mappings that can be observed in this text.

The most frequent source domain of MOTION was mainly mapped onto the target concepts of CHANGE, PROGRESS, CAREER and SUCCESS. TIME was generally associated with DISTANCE, and the MOMENT IN TIME with LOCATION. VIEWS and IDEAS were viewed as either LIVING BEINGS or PHYSICAL OBJECTS. A large proportion of the mappings identified match those exemplified in the Master Metaphor List, but some of the mappings suggested by the annotators are novel, e.g. EMPHASIS IS A PHYSICAL FORCE; SITUATION IS A PICTURE etc.

3.4.3 Metaphor and metonymy

An interesting issue observed in the data is the combination of metaphor and metonymy within a phrase. Consider the following example:

(36) We live in a century *imprinted* on the present, which *regards* the past as little more than the springboard from which we were *launched* on our way. (BNC: ACA)

In this sentence the verbs *imprint*, *regard* and *launch* are used metaphorically according to all annotators. However, the noun *present* can be interpreted as a general metonymy referring to the people who live in the present, rather than the time period. In the latter case, the verb *regard* would receive a different, more conventional interpretation. This in turn is likely to affect the annotation of the corresponding conceptual metaphor and may even result in *regard* being tagged as literally used.

3.4.4 Challenges for mapping annotation

The current study also revealed a number of difficulties in the annotation of source-target domain mappings in real-world text. These are presented below.

Level of abstraction

One of the major steps in the design of the annotation scheme for conceptual metaphor is the construction of the inventory of categories that generalise across many metaphorical expressions. However, given a set of examples, it is often unclear at which level of abstraction the source and target categories should stand. Consider the following sentence.

(37) Sons aspired to *follow* ((CAREER or LIFE) IS A (PATH or JOURNEY)) in their fathers' trades or professions.

Here the verb *follow* is used metaphorically; the best generalisations for both source and target domains are, however, not obvious. This metaphor can be characterised by a more precise mapping of CAREER IS A PATH, as well as the general one of LIFE IS A JOURNEY. These two mappings are related, however, the nature of this relationship is not entirely clear. Martin (1990) discusses hierarchical organisation of conceptual metaphors and models it in terms of subsumption. Lakoff and Johnson (1980) point out cases of entailment relations between mappings, e.g. the metaphor TIME IS MONEY entails TIME IS A VALUABLE COMMODITY or TIME IS A LIMITED RESOURCE. This entailment is based on the fact that the source concepts in the latter mappings are properties of MONEY. However, the more general metaphor LIFE IS A JOURNEY does not

strictly entail or subsume the metaphor CAREER IS A PATH. CAREER is not necessarily a property of LIFE, but is part of one possible life scenario. Fauconnier and Turner (2002) view metaphor in terms of such discrete scenarios within the domains, rather then in terms of continuous domains themselves. Originating in the source domain, the scenarios can then be applied to reason about the target domain. Thus certain scenarios from the domain of JOURNEY can be projected onto the domain of LIFE, e.g. describing the concept of CAREER through that of a PATH.

Chains of mappings

In some cases chains of mappings are necessary to explain a metaphorical expression. Consider the following example:

(38) The Impressionist painters *caught the contagion* [..] (BNC: ACA)

In this sentence the phrase *catch the contagion* is used metaphorically. The interpretation of this metaphor triggers two conceptual mappings, namely IDEAS/VIEWS – INFECTION and INFECTION – PHYSICAL OBJECT. This chain-like association structure intuitively seems natural to a human. At the same time, though, it adds additional complexity to the annotation process, since the number of associations involved may vary. However, it should be noted that the cases where chains of mappings are necessary to explain a metaphorical expression are rare, and only three examples of this phenomenon were found in the corpus.

3.5 Conclusion

The annotation experiment described in this chapter has shown that metaphor is frequent in language. This provides support for my hypothesis that accurate and robust models of metaphor should improve system performance in various NLP tasks. Another important finding is that a large proportion of linguistic metaphors (68%) are represented by verbs, which provides a post-hoc justification for my choice of verbal constructions as the testbed for the metaphor experiments in this thesis.

The second issue that was investigated is how conceptual metaphor manifests itself in language. The annotation experiments described in this chapter are the first empirical study of conceptual metaphor in real-world text. Although the annotators reach some overall agreement on the annotation of interconceptual mappings, they experienced a number of difficulties, e.g. the problem of finding the right level of abstraction for the categories. Awareness of these issues can potentially feed back to CMT or other theoretical accounts of metaphor.

Such problems also need to be taken into account when designing a computational model of metaphor that relies on CMT. The difficulties in category assignment for conceptual metaphor suggest that it is hard to consistently assign explicit labels to source and target domains, even though the interconceptual associations exist in some sense and are intuitive to humans. I therefore believe that these domains and their mappings should be modeled implicitly. This idea motivates the design of the metaphor identification algorithm presented in the following chapter.

Finally, the corpus created here provides a new dataset for linguistic and computational research on conceptual metaphor. In this thesis, however, I focus on computational modelling of linguistic metaphor and will only use the corpus for the intrinsic evaluation of the metaphor identification and interpretation systems.

Chapter 4

Automatic metaphor identification

This chapter is devoted to the description of the metaphor identification method and its evaluation. The analysis of conceptual mappings in unrestricted text, described in the previous chapter, while confirming some aspects of CMT, uncovered a number of fundamental difficulties. One of these is the choice of the level of abstraction and granularity of categories (i.e. labels for source and target domains). This suggests that it is hard to define a comprehensive inventory of labels for source and target domains. Thus a computational model of metaphorical associations should not rely on explicit domain labels. Unsupervised methods allow us to recover patterns in data without assigning any explicit labels to concepts, and thus to model interconceptual mappings implicitly. Since the focus of the experiments presented in the thesis is on verb-object and verb-subject constructions, verb and noun clustering become central to the approach. Target domains are represented as clusters of nouns, and source domains are modelled as characteristic source domain vocabularies, e.g. clusters of verbs pertaining to a domain. Consider the following example:

(39) All of this *stirred* an uncontrollable excitement in her.

The linguistic metaphor "*stir* excitement" can be explained by the conceptual metaphor FEELINGS ARE LIQUIDS. The clusters describing these concepts are shown in Figure 4.1. The noun cluster corresponding to FEELINGS (on the left) is metaphorically linked to the verb cluster containing actions from the domain of LIQUIDS (at the bottom). Besides "*stir* excitement", this link captures other metaphorical expressions representing the same source-target mapping, e.g. "*boiling* with rage". The identification method is based on such linking. The new target domains associated with the given source domain are identified relying on the idea of clustering by association introduced in section 2.5.2. The method consists of the following steps:

- A small number of **seed metaphorical expressions** are the input to the system. They exemplify a range of source–target domain mappings.

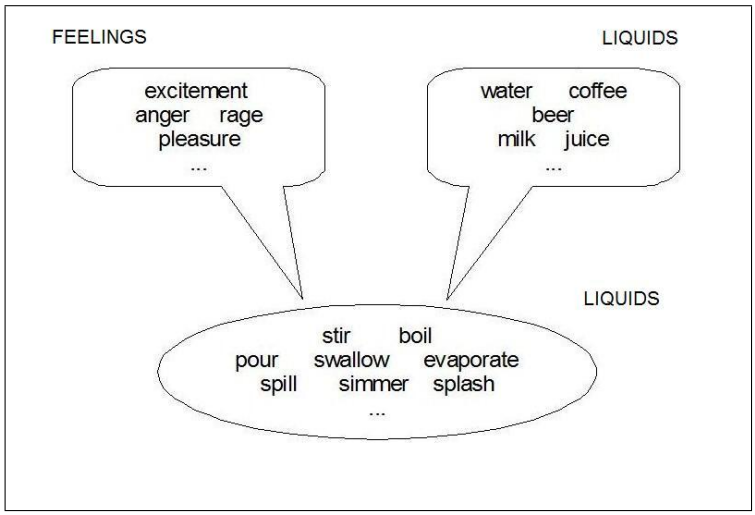

Figure 4.1: Clusters for the conceptual metaphor FEELINGS ARE LIQUIDS

- **Noun clustering** is used for the identification of new target domains associated with the given source domain.

- **Verb clustering** is used to collect the source domain vocabulary. A source domain is represented within the model as a set of actions and events possible in it.

- **Corpus search** is used to identify new metaphorical expressions in unrestricted text on the basis of the information about metaphorical mappings obtained by clustering in the previous two steps.

- A **selectional preference strength filter** discards the candidates which have a low potential of being a metaphor.

The system was tested on a collection of metaphorical expressions representing verb - subject and verb - direct object constructions, where the verb is used metaphorically. It was then evaluated with the aid of human judges and compared to a baseline built upon WordNet. Such evaluation demonstrates that the method reaches beyond synonymy, i.e. captures novel metaphors not directly related to any of those seen in the seed set. For example, starting from the metaphor "*reflect* concern" the system tags the verbs in "*obscure* determination" and "*disguise* intention" as metaphorical. Since the current seed set is relatively small, this experiment is intended as a proof of concept, rather than as a creation of a full-scale high-recall system. The system has been shown to expand significantly on the seed set and identify previously unseen, novel, non-synonymous metaphors.

In this chapter, I first provide some background of the NLP methods used and then present a data analysis conducted in order to verify the hypothesis of clustering by association, which underlies the approach. This is followed by a description of the experimental data, the proposed method and the evaluation of the system.

```
(40) she was transported with pleasure
ncsubj head=transport+ed_VVN_25 dep=she_PPHS1_23
aux head=transport+ed_VVN_25 dep=be+ed_VBDZ_24
iobj head=transport+ed_VVN_25 dep=with_IW_26
conj dep=transport+ed_VVN_25
passive head=transport+ed_VVN_25
dobj head=with_IW_26 dep=pleasure_NN1_27
```

Figure 4.2: Grammatical relations output of RASP

4.1 Methodological background

This section introduces the existing NLP methods and tools that the metaphor identification approach builds on.

4.1.1 Parsing

Since metaphorical meanings of words emerge only in relation to other words in their context, and not independently thereof, it is crucial to be able to detect syntactic relations in text. Thus both metaphor identification and interpretation systems described in the thesis rely on parsed, i.e. syntactically analysed, text. The data is parsed using the Robust Accurate Statistical Parser (RASP) (Briscoe et al., 2006), which is able to deal with unrestricted text. In preparation for parsing, RASP can also perform various low-level tasks, such as sentence boundary detection, tokenisation, part-of-speech tagging and morphological analysis. RASP generates sentence parses that can be output in a variety of formats, including a set of grammatical relations (GRs) associated with a particular analysis. I use the GR output of RASP to identify dependencies between lexical items in my data. Consider the following example from the BNC (BNC text ID is given in brackets):

(40) If he asked her to run an errand, to post a letter or buy some razor blades from the chemist, she was *transported* with pleasure. (BNC: HH9)

The list of GRs that RASP produces for the metaphorical expression "she was *transported* with pleasure" is shown in Figure 4.2. The GR output consists of a named relation (e.g. `ncsubj` for verb-subject relation), a head and a dependent. Both head and dependent are expressed as a concatenation of their lemmas (e.g. *transported*, *she*) and their part-of-speech (PoS) tags from CLAWS tagset (Leech et al., 1994).

For the metaphorical verb *transport* in (40), RASP outputs indirect object and subject relations as follows. The indirect object relation is output in the form of two separate

co-indexed GRs (`iobj` followed by `dobj`), which then need to be combined into a single relation. The relation between the verb *transport* and its semantic object *she* is expressed in passive voice and is tagged by RASP as an `ncsubj` GR (non-clausal subject). Passive voice of this verb is, however, indicated by a separate GR (`passive`), from which one can infer that *she* is an object of *transport*, rather than the subject.

4.1.2 Distributional clustering

Clustering refers to a family of methods that partition data points into disjoint clusters, with points in the same cluster ideally having high similarity and points in different clusters ideally having low similarity. Each data point is represented within a clustering paradigm as a set of characteristic features, also known as a *feature vector*. Clustering algorithms thus operate on feature matrices, i.e. sets of feature vectors for all data points, and are *unsupervised*, i.e. they do not require labelled training data.

The choice of characteristic features for the objects to cluster depends on the task. For word clustering in NLP, the set of linguistic environments in which a word occurs in a corpus has traditionally been used as features. The idea of clustering words based on their syntactic context originates from the work of Levin (1993), who manually grouped verbs exposing similar *diathesis alternations*. Diathesis alternations are regular variations in verb-argument structure (e.g. *break* in "She broke the window" and "The window broke"). The verb-argument structure is encoded in *subcategorisation frames* (SCFs), i.e. the number and types of arguments the verb can take. Levin has shown that verbs exhibiting similar variation of subcategorisation frames tend to form coherent semantic classes. Since Levin published her classification, there has been a number of attempts to automatically classify verbs into semantic classes based on contextual cues using supervised and unsupervised approaches (Lin, 1998; Brew and Schulte im Walde, 2002; Korhonen et al., 2003; Schulte im Walde, 2006; Joanis et al., 2008; Sun and Korhonen, 2009). Similar methods were also applied to acquisition of noun classes from corpus data (Hindle, 1990; Rooth et al., 1999; Pantel and Lin, 2002; Bergsma et al., 2008).

The metaphor identification system described in this chapter relies on the clustering method of Sun and Korhonen (2009). They use a rich set of syntactic and semantic features (GRs, SCFs and verb selectional preferences) and spectral clustering, a method particularly suitable for the resulting high dimensional feature space. This algorithm has proved to be effective in previous verb clustering experiments (Brew and Schulte im Walde, 2002) and in other NLP tasks involving high dimensional data (Chen et al., 2006).

Spectral clustering partitions objects relying on their similarity matrix. Given a set of data points, the similarity matrix records similarities between all pairs of points. The system of Sun and Korhonen (2009) constructs similarity matrices using the *Jensen-Shannon divergence* as a measure. Jensen-Shannon divergence between two feature vectors w_i and

w_j is defined as follows:

$$JSD(w_i, w_j) = \frac{1}{2}D(w_i||m) + \frac{1}{2}D(w_j||m), \tag{4.1}$$

where D is the Kullbach-Leibler distance, and m is the average of the w_i and w_j. Kullbach-Leibler distance between two feature vectors is calculated as follows:

$$D(w_i||w_j) = \sum_{n=1}^{N} w_{i_n} \log \frac{w_{i_n}}{w_{j_n}}, \tag{4.2}$$

where w_i and w_j are the two feature vectors and N is their dimensionality. Kullbach-Leibler distance is not symmetrical, and Jensen-Shannon divergence has been proposed as the symmetrised version of it.

Spectral clustering can be viewed in abstract terms as partitioning of a graph G over a set of words W. The weights on the edges of G are the similarities S_{ij}. The similarity matrix S thus represents the adjacency matrix for G. The clustering problem is then defined as identifying the optimal partition, or *cut*, of the graph into clusters, such that the intra-cluster weights are high and the inter-cluster weights are low. The system of Sun and Korhonen (2009) uses the MNCut algorithm of Meila and Shi (2001) for this purpose.

Sun and Korhonen evaluated their approach on 204 verbs from 17 Levin classes and obtained an F-measure of 80.4, which is the state-of-the-art performance level. The metaphor identification system uses the method of Sun and Korhonen to cluster both verbs and nouns, however, significantly extending its coverage to unrestricted general-domain data.

4.1.3 Selectional preference induction

The idea of selectional preferences (SPs) has long existed in generative (Katz and Fodor, 1964; Chomsky, 1965) and computational (Grishman et al., 1986) linguistics. However, the wide spread of interest to the phenomenon in NLP was triggered by the work of Resnik (1993). Resnik was the first to combine the knowledge of semantic classes (at that stage predefined by an ontology) with the statistical methods from information theory. He viewed selectional preferences as probability distributions over all potential arguments of a predicate, rather than a single argument class (or a limited set of argument classes) assigned to the predicate. This new setting enabled corpus-based statistical learning of selectional preferences, mainly concentrated on the preferences of verbs for their nominal arguments.

Resnik models selectional preferences of a verb in probabilistic terms as the difference between the posterior distribution of noun classes in a particular relation with the verb and their prior distribution in that syntactic position irrespective of the identity of the verb.

He quantifies this difference using the Kullbach-Leibler distance and defines *selectional preference strength* as follows:

$$S_R(v) = D(P(c|v)||P(c)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}, \tag{4.3}$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb and R is the grammatical relation in question. In order to quantify how well a particular argument class fits the verb, Resnik defines another measure called *selectional association*:

$$A_R(v, c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)}, \tag{4.4}$$

which stands for the contribution of a particular argument class to the overall selectional preference strength of a verb.

The key issue in the induction of selectional preferences is the choice of word classes in terms of which SPs should be represented, and the acquisition thereof. Resnik (1993) used WordNet to define the SP classes, as well as to map the words in the corpus to those classes. A number of approaches followed him in this (Li and Abe, 1998; Clark and Weir, 1999; Abney and Light, 1999; Ciaramita and Johnson, 2000). Modelling classes in this way involves identifying the WordNet concepts that most accurately describe the SPs for a given predicate. This is performed by finding the optimal cut, i.e. the right level of generality in the WordNet hierarchy. Another route was the automatic acquisition of selectional preference classes from corpus data, e.g. by means of clustering (Rooth et al., 1999; Bergsma et al., 2008) or similarity-based methods (Erk, 2007; Peirsman and Padó, 2010). Although the majority of studies focus on verbal preferences, some look at SPs exhibited by nouns, adjectives, and prepositions (Brockmann and Lapata, 2003; Zapirain et al., 2009; Schulte im Walde, 2010; Ó Séaghdha, 2010).

Selectional preferences find a range of applications in NLP including word sense disambiguation (Resnik, 1997; McCarthy and Carroll, 2003; Wagner et al., 2009), resolving ambiguous syntactic attachments (Hindle and Rooth, 1993), semantic role labelling (Gildea and Jurafsky, 2002; Zapirain et al., 2009, 2010), natural language inference (Zanzotto et al., 2006; Pantel et al., 2007) and detecting multi-word expressions (Mccarthy et al., 2007).

The metaphor identification system presented further on in this chapter combines the measure proposed by Resnik (1993) with the automatic noun class induction method of Sun and Korhonen (2009) in order to acquire and quantify verb selectional preferences for their nominal subjects and objects.

4.2 Clustering by association hypothesis

Abstract concepts that are associated with the same source domain are often related to each other in the way we conceptualise them, but their meanings are not necessarily synonymous or even semantically close. Compare, for example, the target concepts MARRIAGE and POLITICAL REGIME. Although not directly related in semantic space, they are both metaphorically mapped to the source concept of MECHANISM (e.g. "Our marriage is not really *working*" or "The *wheels* of Stalin's regime were well *oiled* and already *turning*"). The expectation is that the target concepts' conceptual relatedness should manifest itself in the examples of language use, i.e. they will appear in similar lexico-syntactic environments. As a result, target concepts associated with one source domain will appear together in one cluster produced using distributional clustering. Since the identification of new metaphorical associations is based on this hypothesis I first conducted a data study on the output of a state-of-the-art noun clustering method. I analysed the clusters produced by this method in order to confirm that abstract concepts get clustered together if they are associated with the same source domain and to estimate how frequently this happens.

4.2.1 Data

The study was carried out using the output of the noun clustering algorithm of Sun and Korhonen (2009). Sun and Korhonen's system clusters nouns using their subject, direct and indirect object relations with verbs in a corpus as features. A feature set for each noun contains verb lemmas indexed by the type of grammatical relation. Their system allows a choice of the number of nouns to cluster and the number of classes (N). For my study, I selected the 2000 most frequent nouns in the BNC and varied the number of clusters (100 or 200 clusters) in order to determine a suitable cluster granularity for the metaphor identification task.

I then randomly sampled 6 target concepts from the metaphor corpus described in the previous chapter and extracted all conceptual mappings in which they participated. For example, for the concept of LIFE two mappings were found in the corpus: LIFE IS A JOURNEY and LIFE IS A STORY. In total eleven mappings we extracted; they are presented in Table 4.1. The goal of the study was to verify whether other concepts in the cluster containing the target concept are associated with the source domains given in the mappings. To do this I extracted the clusters containing these target concepts from the output of Sun and Korhonen's (2009) method. Some of these clusters are shown in Figure 4.3. Each member of these clusters was analysed for possible association with the respective source domains.

Metaphorical mappings
MARRIAGE IS A MECHANISM
MIND IS A CONTAINER
IDEA IS A PHYSICAL OBJECT
IDEA IS A LIVING BEING
IDEA IS A STRUCTURE
DEATH IS A LIVING BEING
DEATH IS THE END
LIFE IS A JOURNEY
LIFE IS A STORY
FUTURE IS A CONTAINER
FUTURE IS A LOCATION

Table 4.1: The list of sampled conceptual mappings

Conceptual mapping: MARRIAGE IS A MECHANISM
Cluster: consensus relation tradition partnership resistance foundation alliance friendship contact reserve unity link peace bond myth identity hierarchy relationship connection balance marriage democracy defense faith empire distinction coalition regime division
Conceptual mapping: LIFE IS A STORY; JOURNEY
Cluster: politics practice trading reading occupation profession sport pursuit affair career thinking life
Conceptual mapping: FUTURE IS A LOCATION; CONTAINER
Cluster: lifetime quarter period century succession stage generation decade phase interval future
Conceptual mapping: DEATH IS A LIVING BEING; END
Cluster: defeat fall death tragedy loss collapse decline disaster destruction fate

Figure 4.3: Noun clusters of Sun and Korhonen (2009)

4.2.2 Hypothesis evaluation

The degree of association of the members of the clusters with a given source domain was evaluated in terms of precision on the set of hypothesised mappings. Precision measures the proportion of examples correctly classified as positive (in this case as associated with the given source domain) against all examples classified as positive. For each concept in a cluster, I verified that it is associated with the respective source domain by finding (i.e. thinking of) a corresponding metaphorical expression and annotating the concepts accordingly. The precision of the cluster's association with the source concept was calculated as a proportion of the associated concepts in it. Based on these results I computed the Average Precision (AP) for each level of clustering granularity as follows:

$$AP = \frac{1}{M} \sum_{j=1}^{M} \frac{\#\text{associated concepts in cluster } c_j}{|c_j|}, \tag{4.5}$$

Metaphorical mappings	$N = 100$	$N = 200$
MARRIAGE IS A MECHANISM	0.69	0.68
MIND IS A CONTAINER	0.86	0.28
IDEA IS A PHYSICAL OBJECT	0.85	0.71
IDEA IS A LIVING BEING	0.90	1.00
IDEA IS A STRUCTURE	1.00	1.00
DEATH IS A LIVING BEING	0.58	0.80
DEATH IS THE END	0.53	0.90
LIFE IS A STORY	0.54	0.83
LIFE IS A JOURNEY	0.21	0.92
FUTURE IS A LOCATION	0.47	0.82
FUTURE IS A CONTAINER	0.39	0.82
Average Precision	0.64	0.80

Table 4.2: Purity of clusters in precision

where M is the number of mappings and c_j is the cluster of target concepts corresponding to mapping j.

The results for each mapping and level of granularity are shown in Table 4.2. These results suggest that smaller clusters ($N = 200$) are generally more accurate ($AP = 0.8$). However, for the mapping MIND IS A CONTAINER, this setting yields a cluster with body parts that can be physical containers, as opposed to abstract concepts viewed as a CONTAINER. This makes this cluster poorly associated with its source concept. Apart from this error, the results look very promising: dissimilar members of abstract clusters tend to be associated with the same source domains. This confirms the hypothesis of clustering by association.

4.3 Experimental data

This section describes the data used by the metaphor identification system. The system takes a list of seed phrases as input. Seed phrases contain manually annotated linguistic metaphors. The system generalises from these linguistic metaphors to the respective conceptual metaphors by means of clustering. This generalisation is then used to harvest a large number of new metaphorical expressions in unseen text. Thus the data needed for the experiment consists of a seed set, datasets of verbs and nouns that are subsequently clustered, and an evaluation corpus.

4.3.1 Seed phrases

The seed phrases used in the experiment were extracted from the manually annotated metaphor corpus, described in the previous chapter. The resulting seed set consists of 62

phrases that are single-word metaphors representing verb - subject and verb - direct object relations, where a verb is used metaphorically. The seed phrases include, for instance, "*stir* excitement", "*reflect* enthusiasm", "*grasp* theory", "*cast* doubt", "*suppress* memory", "*throw* remark" (verb - direct object constructions), and "campaign *surged*", "factor *shaped* [...]", "tension *mounted*", "ideology *embraces*", "changes *operated*", "example *illustrates*" (subject - verb constructions). The seed phrases were manually annotated for grammatical relations.

4.3.2 Verb and noun datasets

The noun dataset used for clustering consists of the 2000 most frequent nouns in the BNC, as in the data study in section 4.2. The 2000 most frequent nouns cover most common target categories and their linguistic realisations. BNC represents a suitable source for such nouns since the corpus is balanced with respect to genre, style and theme.

The verb dataset is a subset of VerbNet (Kipper et al., 2006). VerbNet is the largest resource for general-domain verbs organised into semantic classes as proposed by Levin (1993). The dataset includes all the verbs in VerbNet with the exception of highly infrequent ones. The frequency of the verbs was estimated from the data collected by Korhonen et al. (2006) for the construction of VALEX lexicon, which to date is one of the largest automatically created verb resources. This data was gathered from five corpora: the British National Corpus (previous edition) (Leech, 1992), the North American News Text Corpus (Graff, 1995), the Guardian corpus, the Reuters corpus (Rose et al., 2002) and the data used for TREC-4 (Harman, 1995) and TREC-5 (Voorhees and Harman, 1996). The verbs from VerbNet that appear less than 150 times in this data were excluded. The resulting dataset consists of 1610 general-domain verbs.

4.3.3 Evaluation corpus

The search space for metaphor identification was the RASP-parsed British National Corpus. I used the GR output of RASP for the BNC created by Andersen et al. (2008). The system searched the corpus for the source and target domain vocabulary within a particular grammatical relation (verb - direct object or verb - subject).

4.4 Method

The main components of the method include (1) distributional clustering of verbs and nouns, (2) search through the parsed corpus, and (3) selectional preference-based filtering. This section provides a description of these components.

4.4.1 Verb and noun clustering

I used the system of Sun and Korhonen (2009)[1] to perform clustering. Sun and Korhonen only evaluated their approach on 204 verbs from 17 Levin classes. Metaphor identification in unrestricted text, however, requires a much broader coverage. Thus a new clustering experiment was performed, applying the method to a considerably larger dataset of 1610 verbs.

Feature extraction

The first task of a word clustering experiment is to select the descriptive features of the instances to be clustered, and to extract them from linguistic data.

Verb clustering For verb clustering, the best performing features from Sun and Korhonen (2009) were adopted. These include automatically acquired verb subcategorisation frames parameterised by their selectional preferences. These features were obtained using the SCF acquisition system of Preiss et al. (2007). The system tags and parses corpus data using the RASP parser (Briscoe et al., 2006) and extracts SCFs from the produced grammatical relations using a rule-based classifier which identifies 168 SCF types for English verbs. It produces a lexical entry for each verb and SCF combination occurring in corpus data. The selectional preference classes were obtained by clustering nominal arguments appearing in the subject and object slots of verbs in the resulting lexicon.

Noun clustering Following previous works on semantic noun classification (Pantel and Lin, 2002; Bergsma et al., 2008), grammatical relations were used as features for noun clustering. More specifically, the frequencies of nouns and verb lemmas appearing in the subject, direct object and indirect object relations in the RASP-parsed BNC were included in the feature vectors.

Clustering

I experimented with different clustering granularities and found that the number of clusters set to 200 is the most suitable setting for both nouns and verbs in my task. This was done by means of qualitative analysis of the clusters as representations of source and target domains. Examples of the resulting clusters are shown in Figures 4.4 (nouns) and 4.5 (verbs) respectively. The noun clusters represent target concepts associated with the same source concept (some suggested source concepts are given in Figure 4.4, although the system only captures those implicitly). The verb clusters contain lists of source domain vocabulary.

[1] I worked on the clustering part of this experiment in collaboration with Lin Sun, another PhD student at the Computer Laboratory, who specialises in clustering methods for NLP and bioinformatics, and one of the authors of the method of Sun and Korhonen (2009). This is the only piece of collaborative work included in this thesis; the rest was done by the author alone.

Source: MECHANISM
Target Cluster: consensus relation tradition partnership resistance foundation alliance friendship contact reserve unity link peace bond myth identity hierarchy relationship connection balance marriage democracy defense faith empire distinction coalition regime division
Source: PHYSICAL OBJECT; LIVING BEING; STRUCTURE
Target Cluster: view conception theory concept ideal belief doctrine logic hypothesis interpretation proposition thesis assumption idea argument ideology conclusion principle notion philosophy
Source: STORY; JOURNEY
Target Cluster: politics practice trading reading occupation profession sport pursuit affair career thinking life
Source: LIQUID
Target Cluster: disappointment rage concern desire hostility excitement anxiety passion doubt panic delight anger fear curiosity shock terror surprise pride happiness pain enthusiasm alarm hope memory love satisfaction sympathy spirit frustration impulse instinct warmth beauty ambition thought guilt emotion sensation horror feeling laughter suspicion pleasure
Source: LIVING BEING; END
Target Cluster: defeat fall death tragedy loss collapse decline disaster destruction fate

Figure 4.4: Clustered target concepts

Source Cluster: sparkle glow widen flash flare gleam darken narrow flicker shine blaze bulge
Source Cluster: gulp drain stir empty pour sip spill swallow drink pollute seep flow drip purify ooze pump bubble splash ripple simmer boil tread
Source Cluster: polish clean scrape scrub soak
Source Cluster: kick hurl push fling throw pull drag haul
Source Cluster: rise fall shrink drop double fluctuate dwindle decline plunge decrease soar tumble surge spiral boom
Source Cluster: initiate inhibit aid halt trace track speed obstruct impede accelerate slow stimulate hinder block
Source Cluster: work escape fight head ride fly arrive travel come run go slip move

Figure 4.5: Clustered verbs (source domains)

4.4.2 Corpus search

Once the clusters have been obtained, the system proceeds to search the corpus for source and target domain terms within verb-object (both direct and indirect) and verb-subject relations. Its task is to classify grammatical relations as metaphorical or non-metaphorical relying on the source and target domain vocabulary in the associated clusters. This search is performed on the BNC parsed by RASP. Consider the following example sentence

```
(41) Change was greatly accelerated - CHANGE IS MOTION
ncsubj head=accelerate+ed_VVN_25 dep=change_NN1_22
aux head=accelerate+ed_VVN_25 dep=be+ed_VBDZ_23
ncmod head=accelerate+ed_VVN_25 dep=greatly_RR_24
conj dep=accelerate+ed_VVN_25
passive head=accelerate+ed_VVN_25
```

Figure 4.6: Grammatical relations output for metaphorical expressions

extracted from the BNC (BNC text ID is given in brackets, followed by the hypothetical conceptual metaphor[2]):

(41) Few would deny that in the nineteenth century change was greatly *accelerated*. (ACA) – CHANGE IS MOTION

The relevant GRs identified by the parser are presented in Figure 4.6. The relation between the verb *accelerate* and its semantic object *change* in (41) is expressed in the passive voice and is, therefore, tagged by RASP as an `ncsubj` GR. Since this GR contains terminology from associated source (MOTION) and target (CHANGE) domains, it is marked as metaphorical and so is the term *accelerate*, which belongs to the source domain of MOTION.

4.4.3 Selectional preference strength filter

In the previous step a set of candidate verb metaphors and the associated grammatical relations were extracted from the BNC. These now need to be filtered based on selectional preference strength. Following Wilks (1978), the phenomenon of metaphor can be seen as a violation of verb selectional preferences. However, not all verbs have an equally strong capacity to constrain their arguments. For instance, *remember, accept, choose* have a weak preference for their direct objects, i.e. they can take arguments of most semantic classes. The expectation is that, for this reason, not all verbs would be equally prone to metaphoricity, but only the ones exhibiting strong selectional preferences. Exploiting this observation should therefore enable the system to filter out a number of candidate expressions which are less likely to be used metaphorically.

To do this, I automatically acquire selectional preference distributions for verb - subject and verb - direct object relations from the RASP-parsed BNC. The noun clusters obtained using Sun and Korhonen's (2009) method as described above form the selectional preference classes. To quantify selectional preferences, I adopt the selectional preference strength (SPS) measure of Resnik (1993) defined in (4.3). The probabilities $P(c|v)$ (the

[2]The labels are assigned by myself for the sake of clarity. The system does not assign any labels.

SPS	Verb
1.3175	undo
1.3160	bud
1.3143	deplore
1.3138	seal
1.3131	slide
1.3126	omit
1.3118	reject
1.3097	augment
1.3094	frustrate
1.3087	restrict
1.3082	employ
1.3081	highlight
1.3081	correspond
1.3056	dab
1.3053	assist
1.3043	neglect
...	

Table 4.3: Verbs with weak direct object SPs

noun class given the verb) and $P(c)$ (the noun class appearing in that syntactic position regardless of the identity of the verb) were estimated from the corpus data as follows:

$$P(c|v) = \frac{f(v,c)}{\sum_k f(v,c_k)}, \tag{4.6}$$

$$P(c) = \frac{f(c)}{\sum_k f(c_k)}, \tag{4.7}$$

where $f(v,c)$ is the number of times the predicate v co-occurs with the argument class c in the relation R, and $f(c)$ is the number of times the argument class occurs in the relation R regardless of the identity of the predicate.

Thus for each verb, its SPS can be calculated for specific grammatical relations. This measure was used to filter out the verbs with weak selectional preferences. Depending on whether the potential metaphor is within the verb-subject or verb-direct object relation, the SPS of the given verb was computed for subject or direct object arguments respectively. The optimal selectional preference strength threshold was set experimentally on a small held-out dataset (via qualitative analysis of the data) and approximates to 1.32. The system excludes expressions containing the verbs with preference strength below this threshold from the set of candidate metaphors. Examples of verbs with weak and strong direct object SPs are shown in Tables 4.3 and 4.4 respectively. Given the SPS threshold of 1.32, the filter discards 31% of candidate expressions initially identified in the corpus.

SPS	Verb	SPS	Verb
...			
3.0810	aggravate	2.9434	coop
3.0692	dispose	2.9326	hobble
3.0536	rim	2.9285	paper
3.0504	deteriorate	2.9212	sip
3.0372	mourn	...	
3.0365	tread	1.7889	schedule
3.0348	cadge	1.7867	cheat
3.0254	intersperse	1.7860	update
3.0225	activate	1.7840	belt
3.0085	predominate	1.7835	roar
3.0033	lope	1.7824	intensify
2.9957	bone	1.7811	read
2.9955	pummel	1.7805	unnerve
2.9868	disapprove	1.7776	arrive
2.9838	hoover	1.7775	publish
2.9824	beam	1.7775	reason
2.9807	amble	1.7774	bond
2.9760	diversify	1.7770	issue
2.9759	mantle	1.7760	verify
2.9730	pulverize	1.7734	vomit
2.9604	skim	1.7728	impose
2.9539	slam	1.7726	phone
2.9523	archive	1.7723	purify
2.9504	grease	...	

Table 4.4: Verbs with strong direct object SPs

4.5 Evaluation and discussion

In order to show that the described metaphor identification method generalises well over the seed set and that it operates beyond synonymy, its output was compared to that of a baseline using WordNet. In the baseline system, WordNet synsets represent source and target domains. The quality of metaphor identification for both the system and the baseline was evaluated in terms of precision with the aid of human judges.

As well as the quality of annotations, the coverage of a system is also crucial for it to be useful for other NLP tasks. Although the current experiment was not intended to provide a high coverage, ready-to-use system, but rather a proof of concept of metaphor identification using clustering by association, it is still important to assess how broadly the method expands on the seed set. The coverage of the method was therefore evaluated in the following two ways:

- by measuring recall of metaphor identification in the corpus. Recall is calculated as a proportion of metaphors correctly identified by the system over the total number

of metaphors in the corpus. The recall-based evaluation of the system output was carried out on a small text sample from the BNC annotated by the author.

- by estimating the proportion of identified metaphors that are not synonymous to any of those seen in the seed set, according to WordNet. This type of evaluation was carried out in order to quantify how well clustering methods are suited in principle to identify new metaphors not directly related to those in the seed set.

4.5.1 Comparison with WordNet baseline

The baseline system was implemented using synonymy information from WordNet to expand on the seed set. Source and target domain vocabularies were thus represented as sets of synonyms of verbs and nouns in seed expressions. The baseline system then searched the corpus for phrases composed of lexical items belonging to those vocabularies. For example, given a seed expression "*stir* excitement", the baseline finds phrases such as "*arouse* fervour, *stimulate* agitation, *stir* turmoil" etc. However, it is not able to generalise over the concepts to broad semantic classes, e.g. it does not find other FEELINGS such as *rage, fear, anger, pleasure*. This, however, is necessary to fully characterise the target domain. Similarly, in the source domain, the system only has access to direct synonyms of *stir*, rather than to other verbs characteristic to the domain of LIQUIDS, e.g. *pour, flow, boil* etc.

To compare the coverage achieved by the system using clustering to that of the baseline in quantitative terms, I estimated the number of WordNet synsets, i.e. different word senses, in the metaphorical expressions captured by the two systems. I found that the baseline system covers only 13% of the data identified using clustering. This is due to the fact that it does not reach beyond the concepts present in the seed set. In contrast, most metaphors tagged by the clustering method (87%) are non-synonymous to those in the seed set and some of them are novel. Together, these metaphors represent a considerably wider range of meanings. Given the seed metaphors "*stir* excitement, *throw* remark, *cast* doubt", the system identifies previously unseen expressions "*swallow* anger, *hurl* comment, *spark* enthusiasm" etc. as metaphorical. Tables 4.5 and 4.6 show examples of how the system and the baseline expand on the seed set respectively. Examples of full sentences containing metaphors annotated by the system are shown in Figure 4.7. 21% of the expessions identified by the system do not have their corresponing metaphorical senses included in WordNet, such as "*spark* enthusiasm"; the remaining 79% are, however, more common conventional metaphors.

Seed phrase	Harvested metaphors	BNC frequency
reflect concern (V-O):	reflect concern	78
	reflect interest	74
	reflect commitment	26
	reflect preference	22
	reflect wish	17
	reflect determination	12
	reflect intention	8
	reflect willingness	4
	reflect sympathy	3
	reflect loyalty	2
	disclose interest	10
	disclose intention	3
	disclose concern	2
	disclose sympathy	1
	disclose commitment	1
	disguise interest	6
	disguise intention	3
	disguise determination	2
	obscure interest	1
	obscure determination	1
cast doubt (V-O):	cast doubt	197
	cast fear	3
	cast suspicion	2
	catch feeling	3
	catch suspicion	2
	catch enthusiasm	1
	catch emotion	1
	spark fear	10
	spark enthusiasm	3
	spark passion	1
	spark feeling	1
campaign surged (S-V):	campaign surged	1
	charity boomed	1
	effort decreased	1
	expedition doubled	1
	effort doubled	1
	campaign shrank	1
	campaign soared	1
	drive spiraled	1

Table 4.5: Examples of seed set expansion by the system

4.5.2 Evaluation against human judgements

In order to assess the quality of metaphor identification by both systems their output was assessed by human judgements. For this purpose, I randomly sampled sentences containing metaphorical expressions as annotated by the system and by the baseline and

Seed phrase	Harvested metaphors	BNC frequency
reflect concern (V-O):	reflect concern	78
	ponder business	1
	ponder headache	1
	reflect business	4
	reflect care	2
	reflect fear	19
	reflect worry	3
cast doubt (V-O):	cast doubt	197
	cast question	11
	couch question	1
	drop question	2
	frame question	21
	purge doubt	2
	put doubt	12
	put question	151
	range question	1
	roll question	1
	shed doubt	2
	stray question	1
	throw doubt	35
	throw question	17
	throw uncertainty	1
campaign surged (S-V):	campaign surged	1
	campaign soared	1

Table 4.6: Examples of seed set expansion by the baseline

CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **swallow his anger** and play tennis.
AD9 3205 He tried to **disguise the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.
AMA 349 We will **halt the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.
ADK 634 **Catch their interest** and **spark their enthusiasm** so that they begin to see the product's potential.
K2W 1771 The committee heard today that gangs regularly **hurled** abusive **comments** at local people, making an unacceptable level of noise and leaving litter behind them.

Figure 4.7: Sentences tagged by the system (metaphors in bold)

asked human annotators to decide whether these were metaphorical or not.

Obtaining the judgements

Participants Five volunteer subjects participated in the experiment. They were all native speakers of English and had no formal training in linguistics.

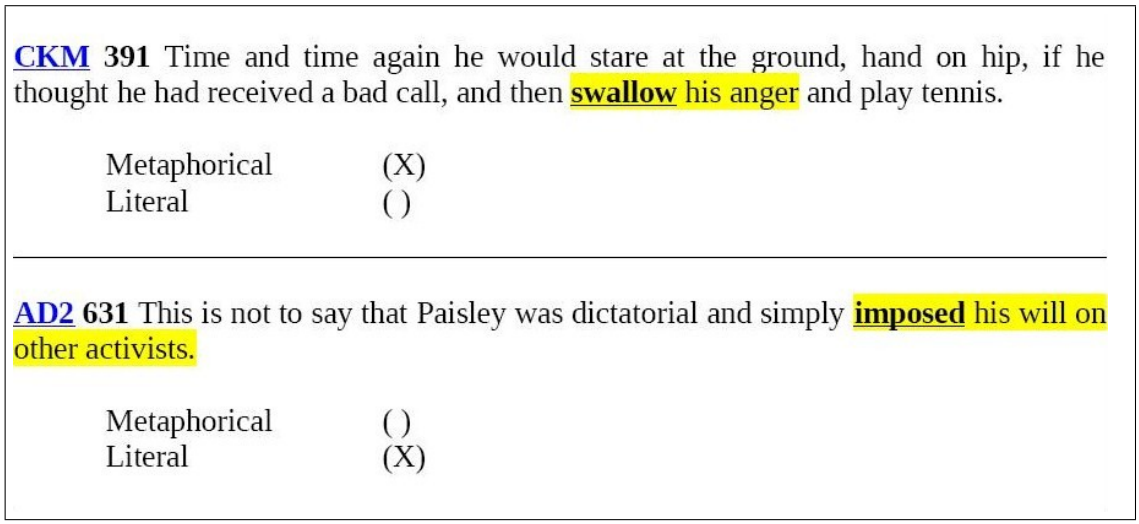

Figure 4.8: Evaluation of metaphor identification

Materials The subjects were presented with a set of 78 randomly sampled sentences annotated by the two systems. 50% of the dataset were the sentences annotated by the identification system and the remaining 50% by the baseline; and the sentences were randomised. The annotation was done electronically in Microsoft Word. An example of annotated sentences is given in Figure 4.8. A sample answer is shown in Appendix D.

Task The subjects were asked to mark which of the expressions were metaphorical in their judgement.

Guidelines The participants were encouraged to rely on their own intuition of what a metaphor is in the annotation process. However, additional guidance in the form of the following definition of metaphor was also provided (cf. Chapter 3):

1. For each verb establish its meaning in context and try to imagine a more basic meaning of this verb in other contexts. Basic meanings normally are: (1) more concrete; (2) related to bodily action; (3) more precise (as opposed to vague); (4) historically older.

2. If you can establish a basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically.

Interannotator agreement Reliability was measured at $\kappa = 0.63$ ($n = 2, N = 78$, $k = 5$). The data suggests that the main source of disagreement between the annotators was the presence of conventional metaphors, e.g. verbs such as *adopt, convey, decline*.

Results

The system performance was then evaluated against the elicited judgements in terms of precision. The system output was compared to the gold standard constructed by

Judge	Precision
A	0.69
B	0.74
C	0.87
D	0.59
E	0.79
Average	0.74

Table 4.7: System precision computed pairwise

merging the judgements, whereby the expressions tagged as metaphorical by at least three annotators were considered to be correct. This resulted in $P = 0.79$, with the baseline attaining $P = 0.44$. In addition, the system tagging was compared to that of each annotator pairwise, yielding an average $P = 0.74$, as shown in Table 4.7.

In order to compare system performance to the human ceiling, pairwise agreement was additionally calculated in terms of precision between the majority gold standard and each judge. This corresponds to an average of $P = 0.94$.

To show that the system performance is significantly different from that of the baseline, I annotated additional 150 instances identified by both systems for correctness and conducted a one-tailed t-test for independent samples. The difference is statistically significant with $t = 4.11$ ($df = 148, p < 0.0005$).

4.5.3 Recall-based evaluation

In order to evaluate the coverage of the system, a recall study is necessary. However, since there is no large metaphor-annotated corpus available, it is hard to evaluate the recall of the system fully and accurately. The creation of such a corpus is expensive, thus I could only estimate recall of the method approximately in the following two ways:

- by annotating metaphorical expressions in a text sample from the BNC and estimating the recall of metaphor identification in those texts. Texts from 4 different sources (BNC: ACA, FYT, HY5, HY8) were annotated for this purpose. There is an overlap with texts used for the creation of the metaphor corpus, from which the seed set was extracted. However, the text extracts themselves were different. The total number of words in the text sample was 15,669; of these, 222 were verbal metaphors. The system identified 5 metaphorical expressions in these texts, which yields a recall of 2.25%.

- by calculating the expected number of metaphors in the BNC based on the statistics from the metaphor annotation described in the previous chapter. The metaphor corpus is a subset of the BNC, consisting of 13,642 words in total and containing 164 verbal metaphors. Thus the relative number of metaphors to the number of

words can be calculated, which can then be generalised to the whole of the BNC (100,000,000 words). Based on this data the expected number of verbal metaphors in the whole BNC is 1,202,170. The system identifies a total of 4,456 metaphorical expressions with a precision of 79% estimated on a random sample. The expected number of correctly tagged metaphors is therefore 3,520, suggesting an overall recall of 0,29%.

The above scores, although estimated only approximately, suggest that the overall recall of the system is low. One explanation for this is the lack of training data. Starting with a seed set of only 62 examples, the system expands significantly on the seed set and identifies the total of 4,456 metaphorical expressions in the BNC. This suggests that the method has the potential to attain a broad coverage of the corpus given a large and representative seed set.

4.5.4 Discussion

The observed discrepancy in precision between the clustering approach and the baseline can be explained by the fact that a large number of metaphorical senses are included in WordNet. This means that in WordNet synsets source domain verbs appear together with more abstract terms. For instance, the metaphorical sense of *shape* in the phrase "*shape* opinion" is part of the synset "(determine, shape, mold, influence, regulate)". This results in the low precision of the baseline system, since it tags literal expressions (e.g. "influence opinion") as metaphorical, assuming that all verbs from the synset belong to the source domain.

Precision errors in the output of the clustering method were also concentrated around the problem of conventionality of some metaphorical verbs, such as those in "*hold* views, *adopt* traditions, *tackle* a problem". This conventionality is reflected in the data in that such verbs are frequently used in their "metaphorical" contexts. As a result, they are clustered together with literally used terms. For instance, the verb *tackle* is found in a cluster with *solve, resolve, handle, confront, face* etc. This results in the system tagging "resolve a problem" as metaphorical if it has previously seen "*tackle* a problem". Such errors are, however, rare.

A number of system errors affecting its precision are also due to cases of general polysemy and homonymy of both verbs and nouns. For example, the noun *passage* can mean both "the act of passing from one state or place to the next" and "a section of text; particularly a section of medium length", as defined in WordNet. Sun and Korhonen's (2009) method performs hard clustering, i.e. it does not distinguish between different word senses. Hence the noun *passage* occurred in only one cluster, containing concepts like *thought, word, sentence, expression, reference, address, description* etc. This cluster models the "textual" meaning of *passage*. As a result of sense ambiguity within the cluster,

given the seed phrase "she *blocked* the thought", the system tags such expressions as "block passage", "impede passage", "obstruct passage", "speed passage" as metaphorical.

The errors that cause low recall of the system are of a different nature. While noun clustering considerably expands the seed set by identifying new associated target concepts (e.g. given the seed metaphor "*sell* soul" it identifies "*sell* skin" and "*launch* pulse" as metaphorical), the verb clusters sometimes miss a certain proportion of source domain vocabulary. For instance, given the seed metaphor "example *illustrates*", the system identifies the following expressions: "history *illustrates*", "episode *illustrates*", "tale *illustrates*", "combination *illustrates*", "event *illustrates*" etc. However, it does not capture obvious verb-based expansions, such as "tale *pictures*", "episode *portrays*", present in the BNC. This is one of the problems that leads to a lower recall of the system.

Nevertheless, in many cases the system benefits not only from dissimilar concepts within the noun clusters used to detect new target domains, but also from dissimilar concepts in the verb clusters. Verb clusters produced automatically relying on contextual features may contain lexical items with distinct, or even opposite meanings (e.g. *throw* and *catch*, *take off* and *land* etc.) However, they tend to belong to the same semantic domain (e.g. verbs of dealing with LIQUIDS, verbs describing a FIGHT etc.) It is the diversity of verb meanings within the domain cluster that allows the generalisation from a limited number of seed expressions to a broader spectrum of previously unseen and novel metaphors, non-synonymous to those in the seed set.

4.6 Conclusion

In this chapter I presented a novel approach to metaphor identification in unrestricted text. Starting from a limited set of metaphorical seeds, the system captures the regularities behind their production and annotates a large number and wide range of previously unseen metaphors in the corpus.

This is the first system that identifies metaphorical expressions in unrestricted text. It does not rely on any hand-coded knowledge (besides the seed set) and operates with a high precision of 0.79. By comparing its coverage to that of a WordNet baseline, I showed that the method reaches beyond synonymy and generalises well over the source and target domains. Although the system has been tested only on verb-subject and verb-object metaphors at this stage, the described identification method should be similarly applicable to a wider range of syntactic constructions. This expectation rests on the fact that both distributional clustering and selectional preference induction techniques have been shown to model the meanings of a range of word classes (Hatzivassiloglou and McKeown, 1993; Boleda Torrent and Alonso i Alemany, 2003; Brockmann and Lapata, 2003; Zapirain et al., 2009). Extending the system to deal with metaphors represented by other word classes and constructions is part of future work.

The fact that the approach is seed-dependent is one of its possible limitations, affecting the coverage of the system. The small size of the seed set, with which the system has been tested so far, is one of the reasons why the current recall is low. A larger and more representative seed set is likely to increase the coverage considerably. In addition, since the precision of the system was measured on the dataset produced by expanding individual seed expressions, I would expect the expansion of other, new seed expressions to yield a comparable quality of annotations. Incorporating new seed expressions is thus likely to result in increasing recall without a significant loss in precision.

The system at this stage was not intended as a high-coverage ready-made component that could be plugged into other NLP applications, but rather as a proof of concept of the proposed method. I have shown that the method leads to a considerable expansion on the seed set, operates with high precision, i.e. produces high quality annotations, as well as identifying fully novel metaphorical expressions relying only on the knowledge of source-target domain mappings that it learns automatically.

Finally, despite the low overall recall in the BNC, the system harvests a large and relatively clean set of metaphorical expressions from the corpus. These annotations provide a new platform for the development and testing of other metaphor systems. The metaphor paraphrasing system described in the next chapter will be tested on this automatically created dataset, along with the manually annotated metaphor corpus.

Chapter 5

Automatic metaphor interpretation

Metaphor interpretation is defined in this thesis as a paraphrasing task. As opposed to explicit identification of source-target domain mappings, this form of interpretation is more natural to a human, as well as easily applicable within NLP. In this chapter, I describe a metaphor paraphrasing system, whose task is to automatically acquire literal and more common paraphrases for metaphorical expressions from a large corpus. The system does this in the following stages:

- it produces a list of all possible paraphrases for a metaphorical expression (induced automatically from a large corpus);
- it ranks the paraphrases according to their likelihood derived from the corpus;
- it filters out paraphrases based on their similarity to the metaphorical term; similarity is defined as sharing a common hypernym within three levels in the WordNet hierarchy;
- it discriminates between literal and metaphorical paraphrases using a selectional preference model and re-ranks the paraphrases, de-emphasising the metaphorical ones and emphasising the literal ones;
- it disambiguates the sense of the paraphrases using the WordNet inventory of senses.

The system thus incorporates two probabilistic models – the context-based probabilistic model, used for paraphrase generation, and the selectional preference model, used to detect literal paraphrases. The key difference between the two models is that the former favours paraphrases that co-occur with words in the context more frequently than other paraphrases do, and the latter favours paraphrases that co-occur with words from the context more frequently than with any other lexical items in the corpus.

The context-based model together with the WordNet filter constitute a metaphor paraphrasing baseline. By comparing the final system to this baseline, I demonstrate that

simple context-based substitution, even supplied by extensive knowledge contained in lexical resources, is not sufficient for metaphor interpretation and that a selectional preference model is needed to establish the literalness of the paraphrases.

The system was first tested on a collection of manually annotated metaphorical expressions in verb - subject and verb - direct object constructions where the verb is used metaphorically. I evaluated the quality of the paraphrases with the aid of human judges in two different experimental settings. The first setting involved direct judgements of system output by humans, whereas in the second setting paraphrases were elicited from humans independently of system output. These paraphrases were then merged into a gold standard to which the system output was compared. Such a twofold evaluation covers both the precision of the system at its top-ranked paraphrases, judged directly by humans, as well as its recall, evaluated by gold standard comparison.

I subsequently applied paraphrasing to the output of the metaphor identification system, described in Chapter 4, and evaluated the performance of the resulting integrated system. It was also evaluated with the help of human judges, who were asked to compare the original sentences to the paraphrased ones and to assess both the correctness and literalness of the paraphrases. First, a small experiment was conducted in a setting with multiple judges in order to measure their agreement on the task, and then the system was evaluated on a larger dataset against the judgements of one person only.

This chapter first provides an overview of paraphrasing and lexical substitution and relates these tasks to the the problem of metaphor interpretation. It then describes the experimental data used to develop and test the paraphrasing system and the method itself, and finally, concludes with the system evaluation and the presentation of results.

5.1 Paraphrasing and lexical substitution

Paraphrasing can be viewed as a text-to-text generation problem, whereby a new piece of text is produced conveying the same meaning as the original text. Paraphrasing can be carried out at multiple levels, i.e. sentence-, phrase- and word-levels, and may involve both syntactic and lexical transformations. Paraphrasing by replacing individual words in a sentence is known as *lexical substitution* (McCarthy, 2002). Since, in this thesis, I address the phenomenon of metaphor at a single-word level, my task is close in nature to lexical substitution. The task of lexical substitution originates from word sense disambiguation. The key difference between the two is that while WSD makes use of a predefined sense-inventory to characterise the meaning of a word in context, lexical substitution is aimed at automatic induction of meanings. Thus the goal of lexical substitution is to generate the set of semantically valid substitutes for the word. Consider the following sentences from (Preiss et al., 2009).

(42) His parents felt that he was a bright boy.

(43) Our sun is a bright star.

Bright in (42) can be replaced by the word *intelligent*. However, the same replacement in the context of (43) will not produce a felicitous sentence. A lexical substitution system needs to (1) find a set of candidate synonyms for the word and (2) select the candidate that matches the context of the word best.

Both sentence- or phrase-level paraphrasing and lexical substitution find a wide range of applications in NLP. These include summarisation (Knight and Marcu, 2000; Zhou et al., 2006), information extraction (Shinyama and Sekine, 2003), machine translation (Kurohashi, 2001; Callison-Burch et al., 2006), text simplification (Carroll et al., 1999), question answering (McKeown, 1979; Lin and Pantel, 2001) and textual entailment (Sekine et al., 2007). Consequently, there has been a plethora of NLP approaches to paraphrasing (McKeown, 1979; Meteer and Shaked, 1988; Dras, 1999; Barzilay and McKeown, 2001; Lin and Pantel, 2001; Barzilay and Lee, 2003; Quirk et al., 2004; Bolshakov and Gelbukh, 2004; Kauchak and Barzilay, 2006; Zhao et al., 2009; Kok and Brockett, 2010) and lexical substitution (McCarthy and Navigli, 2007, 2009; Erk and Padó, 2009; Preiss et al., 2009; Toral, 2009; McCarthy et al., 2010).

Among paraphrasing methods one can distinguish (1) rule-based approaches, that rely on a set of hand-crafted (McKeown, 1979; Zong et al., 2001) or automatically learned (Lin and Pantel, 2001; Barzilay and Lee, 2003; Zhao et al., 2008) paraphrasing patterns; (2) thesaurus-based approaches, that generate paraphrases by substituting words in the sentence by their synonyms (Bolshakov and Gelbukh, 2004; Kauchak and Barzilay, 2006); (3) natural language generation-based approaches (Kozlowski et al., 2003; Power and Scott, 2005), that transform a sentence into its semantic representation and generate a new sentence from it; and (4) SMT-based methods (Quirk et al., 2004), operating as monolingual MT. A number of approaches to lexical substitution rely on manually constructed thesauri to find sets of candidate synonyms (McCarthy and Navigli, 2007), while others address the task in a fully unsupervised fashion. In order to derive and rank candidate substitutes, the latter systems made use of distributional similarity measures (Pucci et al., 2009; McCarthy et al., 2010), vector space models of word meaning (Erk and Padó, 2009; Cao and Basili, 2009) or statistical learning techniques, such as Hidden Markov Models (HMMs) and n-grams (Preiss et al., 2009).

The metaphor interpretation task is different from the word sense disambiguation task, since it is impossible to predefine a set of senses of metaphorical words, in particular for novel metaphors. Instead, the correct substitute for the metaphorical term needs to be generated in a data-driven manner, as for lexical substitution. The metaphor paraphrasing task differs from lexical substitution in two main ways. Firstly, a suitable substitute needs to be used literally in the target context, or at least more conventionally than the original

word. Secondly, by definition, the substitution is not required to be a synonym of the metaphorical word. Moreover, for my task this is not even desired, since there is the danger that synonymous paraphrasing may result in another metaphorical expression, rather than the literal interpretation of the original one. Metaphor paraphrasing therefore presents an additional challenge in comparison to lexical substitution, namely that of discriminating between literal and metaphorical substitutes. This second, harder and not previously addressed task is the main focus of the work presented in this chapter. The remainder of the chapter is devoted to the description of the metaphor paraphrasing experiment.

5.2 Experimental data

The paraphrasing system is tested on verb - subject and verb - direct object metaphorical expressions. These were extracted from the manually annotated metaphor corpus described in Chapter 3. In order to avoid extra noise, I enforced some additional selection criteria. All phrases were included unless they fell in one of the following categories:

- Phrases where the subject or object referent is unknown (e.g. containing pronouns such as in "in which they [changes] *operated*") or represented by a named entity (e.g. "Then Hillary *leapt* into the conversation"). These cases were excluded from the dataset since their processing would involve the use of additional modules for coreference resolution and named entity recognition, which in turn may introduce additional noise into the system.

- Phrases whose metaphorical meaning is realised solely in passive constructions (e.g. "sociologists have been *inclined* to [..]"). These cases were excluded since for many such examples it was hard for humans to produce a literal paraphrase realised in the form of the same syntactic construction. Thus their paraphrasing was deemed to be an unfairly hard task for the system.

- Multiword metaphors (e.g. "*go on pilgrimage* with Raleigh or *put out to sea* with Tennyson"). The current system is designed to paraphrase single-word, lexical metaphors. In the future the system needs to be modified to process multiword metaphorical expressions, this is however outside the scope of the current experiments.

The resulting dataset contains 62 metaphorical expressions. Here are some examples of extracted phrases: "memories were *slipping away*"; "*hold* the truth *back*"; "*stirred* an unfathomable excitement"; "factors *shape* results"; "*mending* their marriage"; "*brushed aside* the accusations".

In addition, metaphor paraphrasing was applied to a dataset that was automatically created using the metaphor identification system described in Chapter 4. Since the identification system operates with low recall, is was impossible to evaluate metaphor paraphrasing

on continuous text. Instead, the evaluation was carried out on individual sentences annotated by the system, extracted from the corpus. The whole dataset comprises 4,456 such sentences.

5.3 Method

The system takes phrases containing annotated single-word metaphors as input, where a verb is used metaphorically, its context is used literally. It generates a list of possible paraphrases of the verb that can occur in the same context and ranks them according to their likelihood, as derived from the corpus. It then identifies shared features of the paraphrases and the metaphorical verb using the WordNet hierarchy and removes unrelated concepts. It then identifies the literal paraphrases among the remaining candidates based on the verb's automatically induced selectional preferences and the properties of the context.

5.3.1 Context-based paraphrase ranking model

Terms replacing the metaphorical verb v will be called its interpretations i. I model the likelihood L of a particular paraphrase as a joint probability of the following events: the interpretation i co-occurring with the other lexical items from its context $w_1, ..., w_N$ in syntactic relations $r_1, ..., r_N$ respectively.

$$L_i = P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)), \tag{5.1}$$

where $w_1, ..., w_N$ and $r_1, ..., r_N$ represent the fixed context of the term used metaphorically in the sentence. In the system output, the context $w_1, ..., w_N$ will be preserved, and the verb v will be replaced by the interpretation i.

I assume statistical independence between the relations of the terms in a phrase. For instance, for a verb that stands in a relation with both a subject and an object, the verb - subject and verb - direct object relations are considered to be independent events within the model. This yields the following approximation:

$$P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) = P(i) \cdot P((w_1, r_1)|i) \cdot ... \cdot P((w_N, r_N)|i). \tag{5.2}$$

The probabilities can be calculated using maximum likelihood estimation

$$P(i) = \frac{f(i)}{\sum_k f(i_k)}, \tag{5.3}$$

$$P(w_n, r_n|i) = \frac{f(w_n, r_n, i)}{f(i)}, \tag{5.4}$$

where $f(i)$ is the frequency of the interpretation irrespective of its arguments, $\sum_k f(i_k)$ is the number of times its part of speech class is attested in the corpus and $f(w_n, r_n, i)$ is the number of times the interpretation co-occurs with context word w_n in relation r_n. By performing appropriate substitutions into (5.2) one obtains

$$P(i,(w_1,r_1),(w_2,r_2),...,(w_N,r_N)) = \frac{f(i)}{\sum_k f(i_k)} \cdot \frac{f(w_1,r_1,i)}{f(i)} \cdot ... \cdot \frac{f(w_N,r_N,i)}{f(i)} = \frac{\prod_{n=1}^{N} f(w_n,r_n,i)}{(f(i))^{N-1} \cdot \sum_k f(i_k)}. \quad (5.5)$$

This model is then used to rank the possible replacements of the term used metaphorically in the fixed context according to the data. The parameters of the model were estimated from the RASP-parsed BNC using the grammatical relations output created by Andersen et al. (2008).

5.3.2 WordNet filter

The context-based model described in 5.3.1 overgenerates and hence there is a need to further narrow down the results. It is acknowledged in the linguistics community that metaphor is, to a great extent, based on similarity between the concepts involved (Gentner et al., 2001). I exploit this fact to refine paraphrasing. After obtaining the initial list of possible substitutes for the metaphorical term, the system filters out the terms whose meanings do not share any common properties with that of the metaphorical term. Consider the computer science metaphor "*kill* a process", which stands for "terminate a process". The basic sense of *kill* implies an *end* or *termination* of life. Thus *termination* is the shared element of the metaphorical verb and its literal interpretation.

Such an overlap of properties can be identified using the hyponymy relations in the WordNet taxonomy. Within the initial list of paraphrases, the system selects the terms that are hypernyms of the metaphorical term, or share a common hypernym with it. To maximise the accuracy, I restrict the hypernym search to a depth of three levels in the taxomomy. Table 5.1 shows the filtered lists of paraphrases for some of the test phrases, together with their log-likelihood. Selecting the highest ranked paraphrase from this list as a literal interpretation will serve as a baseline.

5.3.3 Re-ranking based on selectional preferences

The lists which were generated contain some irrelevant paraphrases (e.g. "contain the truth" for "*hold back* the truth") and some paraphrases where the substitute itself metaphorically used (e.g. "*suppress* the truth"). However, as the task is to identify the literal interpretation, the system should remove these.

Log-likelihood	Replacement
Verb-DirectObject	
hold back truth:	
-13.09	contain
-14.15	conceal
-14.62	suppress
-15.13	hold
-16.23	keep
-16.24	defend
stir excitement:	
-14.28	create
-14.84	provoke
-15.53	make
-15.53	elicit
-15.53	arouse
-16.23	stimulate
-16.23	raise
-16.23	excite
-16.23	conjure
leak report:	
-11.78	reveal
-12.59	issue
-13.18	disclose
-13.28	emerge
-14.84	expose
-16.23	discover
Subject-Verb	
campaign surge:	
-13.01	run
-15.53	improve
-16.23	soar
-16.23	lift

Table 5.1: The list of paraphrases with the initial ranking

One way of dealing with both problems simultaneously is to use selectional preferences of the verbs. Verbs used metaphorically are likely to demonstrate semantic preference for the source domain, e.g. *suppress* would select for MOVEMENTS (political) rather than IDEAS, or TRUTH, (the target domain), whereas the ones used literally for the target domain, e.g. *conceal* would select for TRUTH. Selecting the verbs whose preferences the noun in the metaphorical expression matches best should allow to filter out non-literalness, as well as unrelated terms.

I automatically acquired selectional preference distributions of the verbs in the paraphrase lists (for verb - subject and verb - direct object relations) from the RASP-parsed BNC. As in the identification experiment, I derived selectional preference classes by clustering 2000 most frequent nouns in the BNC into 200 clusters using Sun and Korhonen's (2009)

Association	Replacement
Verb-DirectObject	
hold back truth:	
0.1161	conceal
0.0214	keep
0.0070	suppress
0.0022	contain
0.0018	defend
0.0006	hold
stir excitement:	
0.0696	provoke
0.0245	elicit
0.0194	arouse
0.0061	conjure
0.0028	create
0.0001	stimulate
≈ 0	raise
≈ 0	make
≈ 0	excite
leak report:	
0.1492	disclose
0.1463	discover
0.0674	reveal
0.0597	issue
≈ 0	emerge
≈ 0	expose
Subject-Verb	
campaign surge:	
0.0086	improve
0.0009	run
≈ 0	soar
≈ 0	lift

Table 5.2: The list of paraphrases re-ranked using selectional preferences

algorithm. In order to quantify how well a particular argument class fits the verb, I adopted the selectional association measure proposed by Resnik (1993). To remind the reader, selectional association is defined as follows (repeated from 4.4):

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)}, \tag{5.6}$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb and S_R is the overall selectional preference strength of the verb in the grammatical relation R.

I use selectional association as a measure of semantic fitness, i.e. literalness, of the paraphrases. The paraphrases are re-ranked based on their selectional association with the noun in the context. Those paraphrases that are not well suited or used metaphorically

are dispreferred within this ranking. The new ranking is shown in Table 5.2. The expectation is that the paraphrase in the first rank (i.e. the verb with which the noun in the context has the highest association) represents a literal interpretation.

5.3.4 Sense disambiguation

Having identified literal interpretations, the system can also perform their word sense disambiguation with respect to WordNet sense inventory. This is an additional property resulting from the use of the WordNet filter, rather than the central goal of the system. However, disambiguated metaphorical interpretations would be a useful source of information for NLP applications relying on word senses, e.g. for the tasks of information retrieval (Voorhees, 1998; Schutze and Pedersen, 1995; Stokoe et al., 2003), question answering (Pasca and Harabagiu, 2001) or machine translation (Chan et al., 2007; Carpuat and Wu, 2007).

WSD of the interpretations is performed by selecting WordNet nodes containing them that share a common hypernym with the metaphorical verb. The list of disambiguated interpretations for a random selection of phrases from my dataset is shown in Table 5.3.

5.4 Evaluation and discussion

The metaphor paraphrasing system was applied to both the manually annotated and the automatically created datasets. The evaluation on manually annotated metaphorical expressions is intended to assess the quality of paraphrasing only. In contrast, applying the paraphrasing system to a set of automatically identified metaphors provides an indication of how well the two systems perform when run together in a pipeline.

5.4.1 Evaluation on manually annotated dataset

I evaluated the quality of paraphrasing with the help of human judges in two different experimental settings. The first setting involved direct judgments of system output by humans. In the second setting, the subjects did not have access to system output and had to provide their own literal paraphrases for the metaphorical expressions in the dataset. The system was then evaluated against human judgements in Setting 1 and a paraphrasing gold standard created by merging annotations in Setting 2.

Setting 1: Direct judgement of system output

Materials and task The subjects were presented with a set of sentences containing metaphorical expressions and the top-ranked paraphrases produced by the system and by

Met. Expression	Top Int.	Its WordNet Sense
Verb-DirectObject		
stir excitement	provoke	**(arouse-1 elicit-1 enkindle-2 kindle-3 evoke-1 fire-7 raise-10 provoke-1)** - call forth (emotions, feelings, and responses): "arouse pity"; "raise a smile"; "evoke sympathy"
reflect concern	manifest	**(attest-1 certify-1 manifest-1 demonstrate-3 evidence-1)** - provide evidence for; stand as proof of; show by one's behavior, attitude, or external attributes: "The buildings in Rome manifest a high level of architectural sophistication"; "This decision demonstrates his sense of fairness"
brush aside accusation	reject	**(reject-1)** - refuse to accept or acknowledge: "we reject the idea of starting a war"; "The journal rejected the student's paper"
leaked report	disclose	**(unwrap-2 disclose-1 let_on-1 bring_out-9 reveal-2 discover-6 expose-2 divulge-1 break-15 give_away-2 let_out-2)** - make known to the public information that was previously known only to a few people or that was meant to be kept a secret: "The auction house would not disclose the price at which the van Gogh had sold"; "The actress won't reveal how old she is"
spell out reason	specify	**(specify-4 particularize-1 particularise-1 specialize-2 specialise-2)** - be specific about "Could you please specify your criticism of my paper?"
Verb-Subject		
campaign *surged*	improve	**(better-3 improve-2 ameliorate-2 meliorate-2)** - to make better: "The editor improved the manuscript with his changes"
tension *mounted*	lift	**(rise-1 lift-4 arise-5 move_up-2 go_up-1 come_up-6 uprise-6)** - move upward: "The fog lifted"; "The smoke arose from the forest fire"; "The mist uprose from the meadows"

Table 5.3: Disambiguated paraphrases produced by the system

the baseline, randomised. They were asked to mark as correct the paraphrases that have the same meaning as the term used metaphorically if they are used literally in the given context.

Subjects Seven volunteers participated in the experiment. They were all native speakers of English (one bilingual) and had little or no linguistics expertise.

Interannotator agreement The reliability was measured at $\kappa = 0.62$ ($n = 2, N = 95, k = 7$).

System evaluation against judgements I then evaluated the system performance against their judgments in terms of Precision at Rank 1, $P(1)$. Precision at Rank (1) measures the proportion of correct literal interpretations among the paraphrases in rank 1. The results are shown in Table 5.4. The system identifies literal paraphrases with a $P(1) = 0.81$ and the baseline with a $P(1) = 0.55$. I then conducted a one-tailed Sign test (Siegel and Castellan, 1988) that showed that this difference in performance is statistically

Relation	System	Baseline
Verb-DirectObject	0.79	0.52
Verb-Subject	0.83	0.57
Average	0.81	0.55

Table 5.4: System and baseline precision at rank (1)

significant ($N = 15, x = 1, p < 0.001$).

Setting 2: Creation of a paraphrasing gold standard

Materials and task The subjects were presented with a set of sentences containing metaphorical expressions and asked to write down all suitable literal paraphrases for the highlighted metaphorical verbs that they could think of.

Subjects Five volunteer subjects participated in this experiment, who were different from the ones employed in the previous setting. They were all native speakers of English and some of them had a linguistics background (postgraduate-level degree in English).

Gold Standard The elicited paraphrases combined together can be interpreted as a gold standard. For instance, the gold standard for the phrase "*brushed aside* the accusations" consists of the verbs *rejected, ignored, disregarded, dismissed, overlooked, discarded*.

System evaluation by gold standard comparison The system output was compared against the gold standard using *mean average precision* (MAP) as a measure. MAP is defined as follows:

$$MAP = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ji}, \tag{5.7}$$

where M is the number of metaphorical expressions, N_j is the number of correct paraphrases for the metaphorical expression, P_{ji} is the precision at each correct paraphrase (the number of correct paraphrases among the top i ranks). First, average precision is estimated for individual metaphorical expressions, and then the mean is computed across the dataset. This measure allows us to assess ranking quality beyond rank 1, as well as the recall of the system. As compared to the gold standard, MAP of the paraphrasing system is 0.62 and that of the baseline is 0.56.

Discussion

Given that the metaphor paraphrasing task is open-ended, any gold standard elicited on the basis of it cannot be exhaustive. Some of the correct paraphrases may not occur to subjects during the experiment. As an example, for the phrase "*stir* excitement" most subjects suggested only one paraphrase "create excitement", which is found in rank 3 suggesting an average precision of 0.33 for this phrase. However, the top ranks of the

system output are occupied by *provoke* and *stimulate*, which are intuitively correct, more precise paraphrases, despite none of the subjects having thought of them. Such examples contribute to the fact that the system's MAP is significantly lower than its precision at rank 1, since a number of correct paraphrases proposed by the system are not included in the gold standard.

The selectional preference-based re-ranking yields a considerable improvement in precision at rank 1 (26%) over the baseline. However, this component is also responsible for some errors of the system. One of the potential limitations of selectional preference-based approaches to metaphor paraphrasing is the presence of verbs exhibiting weak selectional preferences. This means that these verbs are not strongly associated with any of their argument classes. As noted in Chapter 4, such verbs tend to be used literally, and are therefore suitable paraphrases. However, our selectional preference model de-emphasises them and, as a result, they are not selected as literal paraphrases despite matching the context. This type of error is exemplified by the phrase "*mend* marriage". For this phrase, the system ranking overruns the correct top suggestion of the baseline, "improve marriage", and outputs "*repair* marriage" as the most likely literal interpretation, although it is in fact a metaphorical use. This is likely to be due to the fact that *improve* exposes a moderate selectional preference strength.

A second type of error is triggered by the conventionality of certain metaphorical verbs. Since they frequently co-occur with the target noun class in the corpus, they receive high association score with that noun class. This results in a high ranking of conventional metaphorical paraphrases. Examples of top-ranked metaphorical paraphrases include "*confront* a question" for "*tackle* a question", "*repair* marriage" for "*mend* marriage", "example *pictures*" for "example *illustrates*".

The above errors concern non-literalness of the produced paraphrases. A less frequently occurring error was paraphrasing with a verb that has a different meaning. One such example was the metaphorical expression "tension *mounted*", for which the system produced a paraphrase "tension *lifted*", which has the opposite meaning. This error is likely to have been triggered by the WordNet filter, whereby one of the senses of *lift* would have a common hypernym with the metaphorical verb *mount*. This results in *lift* not being discarded by the filter, and subsequently ranked top due to the conventionality of the expression "tension *lifted*".

Another important issue that the paraphrase analysis brought to the foreground is the influence of wider context on metaphorical interpretation. The current system processes only the information contained within the GR of interest, discarding the rest of the context. However, for some cases this is not sufficient and the analysis of a wider context is necessary. For instance, given the phrase "scientists *focus*" the system produces a paraphrase "scientists think", rather than the more likely paraphrase "scientists study". Such ambiguity of *focus* could potentially be resolved by taking its wider context into

account. The context-based paraphrase ranking model described in section 5.3.1 allows to incorporate multiple relations of the metaphorical verb in the sentence.

Evaluating WSD of the paraphrases

As discussed, the system outputs paraphrases along with the corresponding WordNet synsets, i.e. it performs WSD. I also evaluated the quality of this aspect of the output. Due to the lack of human annotators for this task, I estimated the accuracy of WSD using my own judgements. The system-derived sense (synset) of the top-ranked paraphrase was considered correct if it represented a valid literal interpretation of the metaphorical verb in the given context.

The precision of WSD was estimated as the proportion of correctly identified senses in rank 1 and it is 0.81. All errors in synset assignment corresponded to those in the assignment of the paraphrase itself and there were no cases of incorrect disambiguation of paraphrases. Examples where paraphrases and the corresponding synsets were assigned incorrectly include "tension *lifted* (rise-1 lift-4 arise-5 move_up-2 go_up-1 come_up-6 uprise-6)" for the metaphorical expression "tension *mounted*", or "relate (refer-2 pertain-1 relate-2 concern-1 come_to-2 bear_on-1 touch-4 touch_on-2 have_to_do_with-1) past" for "*regard* past". For the latter expression, the correct paraphrase "consider past" was found in rank 5 correctly disambiguated by the synset (think-1 believe-2 consider-6 conceive-2), which is defined in WordNet as "judge or regard; look upon; judge".

5.4.2 Evaluation of integrated system

Up to now, the identification and the paraphrasing systems were evaluated individually as modules. The paraphrasing system was then applied to a large dataset of metaphorical expressions identified automatically by the identification system. This allowed for a more accurate estimate of paraphrasing quality, as well as the evaluation of the performance of the two systems when they are integrated together. Some of the expressions identified and paraphrased by the integrated system are shown in Figure 5.1. The system output was compared against human judgements in two phases. In phase 1, a small sample of sentences containing metaphors identified and paraphrased by the system was judged by multiple judges. In phase 2, a larger sample of phrases was judged by myself. Agreement of my own judgements with the other judges was measured on the data from phase 1.

Phase 1: small sample, multiple judges

Subjects Three volunteer subjects participated in the experiment. They were all native speakers of English and had no formal training in linguistics.

CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **swallow his anger** and play tennis.
CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **suppress his anger** and play tennis.

AD9 3205 He tried to **disguise the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.
AD9 3205 He tried to **hide the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.

AMA 349 We will **halt the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.
AMA 349 We will **prevent the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.

J7F 77 An economist would **frame this question** in terms of a cost-benefit analysis: the maximisation of returns for the minimum amount of effort injected.
J7F 77 An economist would **phrase this question** in terms of a cost-benefit analysis: the maximisation of returns for the minimum amount of effort injected.

EEC 1362 In it, Younger stressed the need for additional alternatives to custodial sentences, which had been implicit in the decision to ask the Council to **undertake the enquiry**.
EEC 1362 In it, Younger stressed the need for additional alternatives to custodial sentences, which had been implicit in the decision to ask the Council to **initiate the enquiry**.

A1F 24 Moreover, Mr Kinnock **brushed aside the suggestion** that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.
A1F 24 Moreover, Mr Kinnock **dismissed the suggestion** that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.

Figure 5.1: Metaphors identified (in red) and paraphrased (in blue) by the system

Materials and task Subjects were presented with a set of sentences containing metaphorical expressions identified by the system and their paraphrases, as shown in Figure 5.2. There were 35 such sentences in the sample. They were asked to do the following:

1. Compare the sentences, decide whether the highlighted expressions have the same meaning and record this in the box provided;

2. Decide whether the verbs in both sentences are used metaphorically or literally and tick the respective boxes.

For the second task, the same definition of metaphor as in the identification evaluation (cf. section 4.5.2) was provided for guidance. Annotation was carried out electronically in Microsoft Word. A sample answer is shown in Appendix F.

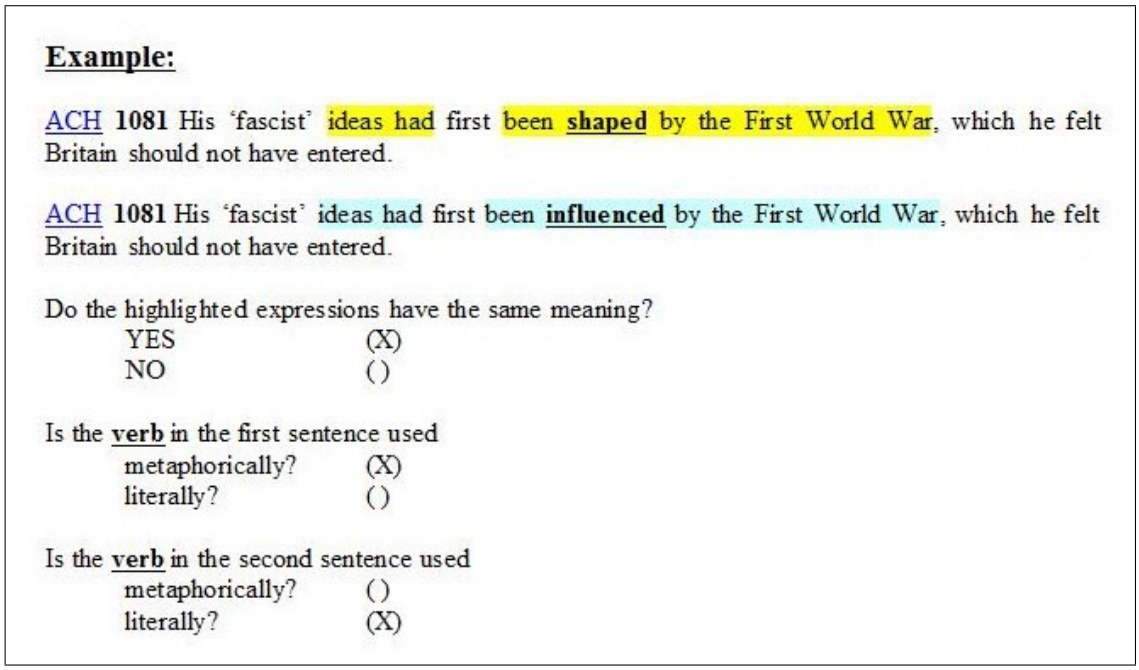

Figure 5.2: Evaluation of metaphor identification and paraphrasing

Interannotator agreement The reliability of annotations was evaluated independently for judgements on similarity of paraphrases and their literalness. The interjudge agreement on the task of distinguishing metaphoricity from literalness was measured at $\kappa = 0.53$ ($n = 2, N = 70, k = 3$). On the paraphrase (i.e. meaning retention) task, reliability was measured at $\kappa = 0.63$ ($n = 2, N = 35, k = 3$).

System performance I then evaluated the integrated system performance against their judgments in terms of accuracy. Accuracy in this task measures the proportion of metaphors both identified and paraphrased correctly in the given set of sentences. Human judgements were merged into a majority gold standard, which consists of those instances that were considered correct (i.e. identified metaphor correctly paraphrased by the system) by at least two judges. Compared to this majority gold standard, the integrated system operates with an accuracy of 0.66. The overall proportion of paraphrases that retained the meaning and resulted in a literal paraphrase, i.e. including literal paraphrasing of literal expressions in original sentences, is 0.71. The average human agreement with the majority gold standard in terms of accuracy is 0.80 on the literalness judgements and 0.89 on the meaning retention judgements.

Phase 2: larger sample, one judge

The system was also evaluated on a larger sample of automatically annotated metaphorical expressions (200 sentences) using my own judgements produced following the procedure from phase 1. I measured in how far my judgements agree with the judges employed in phase 1. The agreement on the meaning retention was measured at $k = 0.59$ ($n = 2, N =$

Tagging case	Acceptability	Percentage
Correct paraphrase: metaphorical –> literal	✔	56%
Correct paraphrase: literal –> literal	✔	18%
Correct paraphrase: literal –> metaphorical	✗	1%
Correct paraphrase: metaphorical –> metaphorical	✗	8%
Incorrect paraphrase	✗	17%

Table 5.5: Integrated system performance

$35, k = 4$) and that on the literalness of paraphrases at $k = 0.54$ ($n = 2, N = 70, k = 4$).

On this larger dataset, the system achieved an accuracy of 0.56 on metaphor to literal paraphrasing and 0.74 on correct paraphrasing resulting in a literal expression (including metaphor-to-literal and literal-to-literal paraphrasing). The proportions of different tagging cases are shown in Table 5.5. The table also shows acceptability of tagging cases. Acceptability indicates whether this type of system paraphrasing would cause an error when hypothetically integrated with an external NLP application or not. Cases where the system produces correct literal paraphrases for metaphorical expressions identified in the text would benefit another NLP application, whereas cases where literal expressions are correctly paraphrased by other literal expressions are considered neutral. Both such cases are deemed acceptable, since they increase or preserve literalness of the text. All other tagging cases introduce errors, thus they are marked as unacceptable.

The accuracy of metaphor to literal paraphrasing (0.56) indicates the level of informative contribution of the system, whereas the overall accuracy of correct paraphrasing resulting in a literal expression (0.74) the level of its acceptability within NLP.

Discussion

The results of integrated system evaluation suggest that, although the system is capable of providing useful information about metaphor for an external text processing application, it also introduces errors in the text by incorrect paraphrasing, as well as producing metaphorical paraphrases. If the latter errors are rare (1%), the errors of the former type are sufficiently frequent (17%) to make the metaphor system less desirable for use in NLP. It is therefore important to address such errors.

The reasons for incorrect paraphrasing are manifold and concern both the metaphor identification and paraphrasing components. One of the central problems stems from the initial tagging of literal expressions as metaphorical by the identification system. The paraphrasing system is not designed with literal-to-literal paraphrasing in mind. When it receives literal expressions which have been incorrectly identified as input, it searches for a more literal paraphrase for them. However, not all literally used words have suitable substitutes in the given context. For instance, the literal expression “approve conclusion” is incorrectly paraphrased as “evaluate conclusion”. Such paraphrasing is due to the fact

that there are no synonym options for *approve* in the context of *conclusion*.

Similar errors occur when metaphorical expressions do not have any single-word literal paraphrases, e.g. "country *functions* according to [..]". This is, however, a more fundamental problem for metaphor paraphrasing as a task. In such cases, the system, nonetheless, attempts to produce a substitute with approximately the same meaning, which often leads to either metaphorical or incorrect paraphrasing. For instance, "country *functions*" is paraphrased by "country runs", with suggestions in lower ranks being "country *works*" and "country *operates*".

Some errors that occur at the paraphrasing level are also due to general word sense ambiguity of certain verbs or nouns. Consider the following paraphrasing example, where (44a) shows an automatically identified metaphor and (44b) its system-derived paraphrase:

(44) a. B71 852 Craig Packer and Anne Pusey of the University of Chicago have continued to *follow* the life and loves of these Tanzanian lions.

b. B71 852 Craig Packer and Anne Pusey of the University of Chicago have continued to succeed the life and loves of these Tanzanian lions.

This error results from the fact that the verb *succeed* has a high selectional preference for *life* in one of its senses ("attain success or reach a desired goal") and is similar to *follow* in WordNet in another of its senses ("be the successor (of)"). The system merges these two senses in one, resulting in an incorrect paraphrase.

One automatically identified example exhibited interaction of metaphor with metonymy at the interpretation level, which was already mentioned in Chapter 3. In the phrase "*break* word", the verb *break* is used metaphorically (although conventionally) and the noun *word* is a metonymy standing for *promise*. This affected paraphrasing in that the system searched for verbs denoting actions that could be done with words, rather than promises, and suggested the paraphrase "interrupt word(s)". This paraphrase is interpretable in the context of a person giving a speech, but not in the context of a person giving a promise. However, this was the only case of metonymy in the analysed data.

Another issue that the evaluation on a larger dataset revealed is the limitations of the WordNet filter used in the paraphrasing system. Despite being a wide-coverage general-domain database, WordNet does not include information about all possible relations that exist between particular word senses. This means that some of the correct paraphrases suggested by the context-based model get discarded by the WordNet filter due to missing information in WordNet. For instance, the system produces no paraphrase for the metaphors "*hurl* comment", "*spark* enthusiasm" and "*magnify* thought", that it correctly identified. This problem motivates the exploration of possible WordNet-free solutions for similarity detection in the metaphor paraphrasing task. The system could either rely entirely on such a solution, or back off to it in cases when the WordNet-based system fails.

5.5 Conclusion

In this chapter, I presented an approach to metaphor interpretation and a system that produces literal paraphrases for metaphorical expressions. My method is distinguished from previous work in that it operates on unrestricted open-domain text and produces interpretations in textual format. It also does not rely on any metaphor-specific hand-crafted knowledge (only the general lexical knowledge from WordNet), but in contrast employs automatically induced selectional preferences.

The described system is the first of its kind and it is capable of paraphrasing metaphorical expressions with high precision (0.81). The current test set consists of verb - subject and verb - direct object metaphors only, but there is no reason why it should not be possible to extend the system to other parts of speech and a wider range of syntactic constructions. The context-based model is suited to all part-of-speech classes and types of relations. Selectional preferences have been previously successfully acquired not only for verbs, but also for nouns, adjectives and even prepositions (Brockmann and Lapata, 2003; Zapirain et al., 2009; Ó Séaghdha, 2010).

The proposed representation of metaphor interpretation is directly transferable to other NLP applications that could benefit from the inclusion of a metaphor processing component. In section 5.4.2, I evaluated the paraphrasing system run in conjunction with the identification system, described in Chapter 4, and judged the level of applicability of the integrated system as 0.74. This means that, for this percentage of automatically identified instances, the system produces correct literal paraphrases.

A data analysis revealed a number of errors the system makes. A large proportion of these errors can be explained and addressed in the next version of the system. Overall, the results suggest that the system can in principle provide useful and accurate information about metaphor to other NLP applications relying on lexical semantics.

Chapter 6

Logical metonymy background and contributions

This and the following chapters address a different type of figurative language, logical metonymy. As in case of metaphor, I first introduce the phenomenon and describe the related theoretical and computational work, and then present my own approach.

6.1 Theoretical background

Regular polysemy has long been of considerable interest for lexical semantics, and so has been one of its frequent types, logical metonymy. The term logical metonymy captures a class of phenomena where a noun phrase is used to stand for an event associated with this noun phrase. Below are a few examples of logical metonymic phrases (under (a)) and their usual interpretations (under (b)).

(45) a. Mark enjoyed this book.

b. Mark enjoyed *reading* this book.

(46) a. Mark always enjoys his beer.

b. Mark always enjoys *drinking* his beer.

(47) a. Mark enjoyed his cigarette.

b. Mark enjoyed *smoking* his cigarette.

(48) a. Mark enjoyed the cake.

b. Mark enjoyed *eating* the cake.

(49) a. Mark enjoyed the concert.

b. Mark enjoyed *listening to* the concert.

(50) a. a good meal

b. a meal that *tastes* good

(51) a. a good cook

b. a cook that *cooks* well

(52) a. After the movie Mark went straight to bed.

b. After *watching* the movie Mark went straight to bed.

(53) a. After three martinis Mark was feeling well.

b. After *drinking* three martinis Mark was feeling well.

(54) a. After the lecture Mark looked tired.

b. After *listening to* the lecture Mark looked tired.

In all of these phrases a shift of meaning happens in a systematic way. The metonymic verb, adjective or preposition semantically selects for an argument of type *event*, but however, is combined with a noun phrase syntactically. For instance, in (45a) the verb *enjoy* requires an eventuality as its argument, and therefore, its noun phrase complement is interpreted as an event of "*reading* the book". This is *metonymy* in the sense that one phrase is used to stand for another related one, and it is *logical* because it is triggered by semantic type constraints that the verb, adjective or preposition places onto its arguments. This is known in linguistics as a phenomenon of *type coercion*. Many existing approaches to logical metonymy explain systematic syntactic ambiguity of metonymic verbs (such as *enjoy*) or prepositions (such as *after*) by means of type coercion (Pustejovsky, 1991, 1995; Briscoe et al., 1990; Verspoor, 1997; Godard and Jayez, 1993). The actual interpretations (events), according to these approaches, are suggested by *lexical defaults* associated with the noun in the complement. Within his Generative Lexicon theory, Pustejovsky (1991) models these lexical defaults in the form of the *qualia structure* of the noun. As set out by Pustejovsky the qualia structure of a noun specifies the following aspects of its meaning:

- CONSTITUTIVE Role (the relation between an object and its constituents)
- FORMAL Role (that which distinguishes the object within a larger domain)

- TELIC Role (purpose and function of the object)
- AGENTIVE Role (how the object came into being)

For the problem of logical metonymy telic and agentive roles are of particular interest. For example, the noun *book* would have *read* specified as its telic role and *write* as its agentive role in its qualia structure. Lexical defaults are inherited within the semantic class hierarchy and are activated in the absence of contradictory pragmatic information (Briscoe et al., 1990). For example, all the nouns belonging to the class LITERATURE (e.g. *book*, *story*, *novel* etc.) will have *read* specified as their telic role. In some cases lexical defaults can, however, be overridden by context. Consider the following example taken from Lascarides and Copestake (1995).

(55) My goat eats anything. He really enjoyed your book.

Here it is clear that "the goat enjoyed *eating* the book" and not "*reading* the book", which is enforced by the context. Such cases, however, are rare.

This shows that logical metonymy is both conventionalised (e.g. conventional telic interpretations such as "enjoy *reading* the book"), as well as productive, i.e. new metonymic interpretations emerge outside of ordinary context, as in (55). A number of approaches discuss *semi-productivity* of the phenomenon (Copestake and Briscoe, 1995; Copestake, 2001). Not all nouns that have evident telic and agentive roles can be equally combined with aspectual verbs. Consider the following examples.

(56) *John enjoyed the dictionary.

(57) *John enjoyed the door.

(58) *John began the bridge.

These examples suggest that there are certain conventional constraints on realisation and interpretation of logical metonymy. Such constraints were discussed in a number of studies (Pustejovsky, 1991; Godard and Jayez, 1993; Pustejovsky and Bouillon, 1995; Copestake and Briscoe, 1995; Verspoor, 1997; Copestake, 2001). While Pustejovsky's treatment of logical metonymy within the Generative Lexicon theory evolves around the rich semantics of the head noun in the metonymic phrase, other approaches perceive linguistic constraints on interpretations as inherent to the semantics of metonymic verbs (Copestake and Briscoe, 1995; Pustejovsky and Bouillon, 1995). Godard and Jayez (1993) claim that possible interpretations represent a kind of a modification to the object referred to by the NP, more specifically, that the object usually "comes into being", "is consumed", or "undergoes a change of state". All of these approaches view metonymic interpretation

at the level of individual words, as opposed to Vendler (1968), who points out that in some cases a group of verbs is needed to fully interpret metonymic phrases. He gives examples of adjective-noun metonymic constructions, e.g. "fast scientist" can be interpreted as both "a scientist who does experiments quickly" and "publishes fast (and a lot)" at the same time.

Verspoor (1997) conducted an empirical study of logical metonymy in real-world text. She explored the data regarding logical metonymy from the Lancaster Oslo/Bergen (LOB) Corpus[1] and the British National Corpus for aspectual verbs *begin* and *finish*. She investigated how frequent the use of logical metonymy is for these verbs, as well as how often the resulting constructions can be interpreted based on the head noun's qualia structure. Verspoor came to a conclusion that for these two aspectual verbs the interpretation of logical metonymy is indeed restricted to either agentive events or conventionalised telic events associated with the noun complement and that the vast majority of uses are conventional.

6.2 Computational models of logical metonymy

Utiyama et al. (2000) and then Lapata and Lascarides (2003) used text corpora to automatically derive interpretations of metonymic phrases. Utiyama et al. (2000) used a statistical model for the interpretation of general metonymies for Japanese. Given a verb-object metonymic phrase, such as *read Shakespeare*, they searched for entities the object could stand for, such as *plays of Shakespeare*. They considered all the nouns co-occuring with the object noun and the Japanese equivalent of the preposition *of*. Utiyama and his colleagues tested their approach on 75 metonymic phrases taken from the literature and report the resulting precision of 70.6%, whereby an interpretation was considered correct if it made sense in some imaginary context.

Lapata and Lascarides (2003) extend this approach to interpretation of logical metonymies containing aspectual verbs (e.g. "begin the book") and polysemous adjectives (e.g. "good meal" vs. "good cook"). The intuition behind their approach is similar to that of Pustejovsky (1991, 1995), namely that there is an event not explicitly mentioned, but implied by the metonymic phrase ("begin to *read* the book", or "the meal that *tastes* good" vs. "the cook that *cooks* well"). They used the BNC parsed by the Cass parser (Abney, 1996) to extract events (verbs) co-occuring with both the metonymic verb (or adjective) and the noun independently and ranked them in terms of their likelihood according to the data. The likelihood of a particular interpretation was calculated as follows:

$$P(e, v, o) = \frac{f(v, e) \cdot f(o, e)}{N \cdot f(e)}, \tag{6.1}$$

where e stands for the eventive interpretation of the metonymic phrase, v for the metonymic verb and o for its noun complement. $f(e)$, $f(v, e)$ and $f(o, e)$ are the respective corpus

[1] http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM

Metonymic Phrase	Interpretations	Log-probability
finish video	film	-19.65
	edit	-20.37
	shoot	-20.40
	view	-21.19
	play	-21.29
	stack	-21.75
	make	-21.95
	programme	-22.08
	pack	-22.12
	use	-22.23
	watch	-22.36
	produce	-22.37

Table 6.1: Interpretations of Lapata and Lascarides (2003) for "finish video"

frequencies. $N = \sum_i f(e_i)$ is the total number of verbs in the corpus. The list of interpretations Lapata and Lascarides (2003) report for the phrase "finish video" is shown in Table 6.1.

Lapata and Lascarides produced ranked lists of interpretations for 58 metonymic phrases. This dataset was compiled by selecting 12 verbs that allow logical metonymy[2] from the lexical semantics literature and combining each of them with 5 nouns. This yields 60 phrases, which were then manually filtered, excluding 2 phrases as non-metonymic.

They compared their results to paraphrase judgements elicited from humans. The subjects were presented with three interpretations for each metonymic phrase (from high, medium and low probability ranges) and were asked to associate a number with each of them reflecting how good they found the interpretation. They report a correlation of 0.64, whereby the inter-subject agreement was 0.74. It should be noted, however, that such an evaluation scheme is not very informative as Lapata and Lascarides calculate correlation only on 3 data points for each phrase out of many more yielded by the model. It fails to take into account the quality of the list of top-ranked interpretations, although the latter is deemed to provide the right answer. In comparison, the fact that Lapata and Lascarides initially select the interpretations from high, medium or low probability ranges makes achieving a high correlation between the model rankings and human judgements significantly easier.

6.3 Alternative interpretation of logical metonymy

The approach of Lapata and Lascarides (2003) produces a list of non-disambiguated verbs representing possible interpretations of a metonymic phrase. Some of them indeed

[2] *attempt, begin, enjoy, finish, expect, postpone, prefer, resist, start, survive, try, want*

correspond to paraphrases that a human would give for the metonymic phrase. However, to provide useful information to NLP applications dealing with semantics this work can be improved on in two main ways.

- Firstly, the lists of possible interpretations produced by the system of Lapata and Lascarides need to be filtered. They contain a certain proportion of incorrect interpretations (i.e. noise), as well as synonymous ones.
- Secondly, in order to obtain the actual meaning of the metonymic phrase, the interpretations need to be disambiguated with respect to their word sense. Using sense-based interpretations of logical metonymy as opposed to ambiguous verbs could benefit other NLP applications that rely on disambiguated text (e.g. for the tasks of information retrieval (Voorhees, 1998; Schutze and Pedersen, 1995; Stokoe et al., 2003), question answering (Pasca and Harabagiu, 2001) or machine translation (Chan et al., 2007; Carpuat and Wu, 2007)).

Thus I propose an alternative representation of interpretation of logical metonymy consisting of a list of verb senses that map to WordNet synsets and develop a word sense disambiguation method for this task. Besides performing WSD, this method also allows to filter out irrelevant interpretations yielded by the model of Lapata and Lascarides. However, the list of non-disambiguated interpretations similar to the one Lapata and Lascarides produce is a necessary starting point in building the sense-based representation. Discovering metonymic interpretations and disambiguating them with respect to word sense is the focus of my first experiment on logical metonymy.

The second issue that the thesis addresses is the design of a class-based model of logical metonymy and its verification against human judgements. The class-based model of logical metonymy is both application-driven and theoretically grounded. NLP applications would benefit from the class-based representation since it provides more accurate and generalised information about possible interpretations of metonymic phrases that can be adapted to particular contexts the phrases appear in. Class-based models of semantics are frequently created and used in NLP (Clark and Weir, 2002; Korhonen et al., 2003; Lapata and Brew, 2004; Schulte im Walde, 2006; Ó Séaghdha, 2010). Verb classifications specifically have been used to support a number of NLP tasks, e.g. machine translation (Dorr, 1998), document classification (Klavans and Kan, 1998) and subcategorisation acquisition (Korhonen, 2002). Besides providing meaningful generalisations over concepts, class-based models also improve the accuracy of statistical generalisations over corpus data (Brown et al., 1992). They address the issue of data sparseness, which is a bottleneck for statistical learning from limited amounts of data.

The class-based model also takes into account the constraints on logical metonymy pointed out in linguistics literature (Vendler, 1968; Pustejovsky, 1991, 1995; Godard and Jayez,

1993; Verspoor, 1997). To remind, Pustejovsky (1991) explains the interpretation of logical metonymy by means of lexical defaults associated with the noun complement in the metonymic phrase. He models these lexical defaults in the form of the qualia structure of the noun, whereby telic and agentive roles are of particular relevance for logical metonymy. For example, the noun *book* would have *read* specified as its telic role and *write* as its agentive role in its qualia structure. Nevertheless, multiple telic and agentive roles can exist and be valid interpretations, as suggested by Verspoor (1997) and confirmed by the data of Lapata and Lascarides (see Table 6.1). Therefore, I propose that these lexical defaults should be represented in the form of classes of interpretations (e.g. {*read, browse, look through*} vs. {*write, compose, pen*}) rather than single word interpretations (e.g. *read* and *write*) as suggested by Pustejovsky (1991).

According to Godard and Jayez (1993), a metonymic interpretation represents a modification to the object referred to by the NP, i.e. the object "comes into being", "is consumed", or "undergoes a change of state". This conveys an intuition that a sensible metonymic interpretation should fall under one of those three classes.

Comparing the interpretations Lapata and Lascarides obtained for the phrase "finish video" (Table 6.1), one can clearly distinguish between the meanings pertaining to the creation of the video, e.g. *film, shoot, take*, and those denoting using the video, e.g. *watch, view, see*. However, the classes based on Pustejovsky's telic and agentive roles do not explain the interpretation of logical metonymy for all cases. Neither does the class division proposed by Godard and Jayez (1993). For example, the most intuitive interpretation for the metonymic phrase "he *attempted the peak*" is *reach*, which does not fall under any of these classes. It is hard to exhaustively characterise all possible classes of interpretations. Therefore, I treat this as an unsupervised clustering problem rather than a classification task and choose a theory-neutral data-driven approach to it. The objective of my second experiment is to model the class division structure of metonymic interpretations and experimentally verify whether the obtained data conforms to it.

In order to discover such classes, the interpretations are automatically clustered to identify groups of related meanings. The automatic class discovery is carried out using disambiguated interpretations produced in the previous step. This is motivated by the fact that it is verb senses that form classes rather than ambiguous verbs. It is possible to model verb senses starting from non-disambiguated verbs using *soft clustering*, i.e. a clustering algorithm that allows for one object to be part of different clusters, as opposed to *hard clustering*, whereby each object can belong to one cluster only. However, previous approaches to soft clustering of verbs have proved that this is a challenging task (Schulte im Walde, 2000), whereas much success has ben achieved in hard clustering (Korhonen et al., 2003; Schulte im Walde, 2006; Joanis et al., 2008; Sun and Korhonen, 2009). Thus in my experiments, I create verb clusters by performing hard clustering of verb senses instead of soft clustering of ambiguous verbs, and expect this method to yield a better model of verb meaning. As it was the case in metaphor experiments, clustering is performed using

the information about lexico-syntactic environments, in which metonymic interpretations appear, as features.

Both the disambiguation method and the class-based model are evaluated individually against human judgements. Humans are presented with a set of verb senses the system produces as metonymic interpretations and asked to (1) remove the irrelevant interpretations and (2) cluster the remaining ones. Their annotations are then used for the creation of a gold standard for the task. The performance of the system is subsequently evaluated against this gold standard. The details of the system design and evaluation are presented in the next chapter.

Chapter 7

Logical metonymy experiments

This chapter describes the experiments on logical metonymy. I first extend the method of Lapata and Lascarides (2003) by disambiguating the interpretations with respect to WordNet synsets for verb-object metonymic phrases. For this purpose, I develop a ranking scheme for the synsets using a non-disambiguated corpus, address the issue of sense frequency distribution and utilise information from WordNet glosses to refine the ranking.

In the second experiment, the produced sense-based interpretations are automatically clustered based on their semantic similarity. In addition, I verify whether the class-based representation of the interpretation of logical metonymy is intuitive to humans. This chapter first describes my reimplementation of the method of Lapata and Lascarides and then the disambiguation and clustering experiments.

7.1 Extracting ambiguous interpretations

The method of Lapata and Lascarides (2003) is reimplemented to obtain a set of candidate interpretations (ambiguous verbs) from a non-annotated corpus. However, my reimplementation of the method differs from the system of Lapata and Lascarides in that I use a more robust parser (RASP), process a wider range of syntactic structures (coordination, passive), and extract my data from a later version of the BNC. As a result, I expect my system to extract the data more accurately.

7.1.1 Parameter estimation

The model of Lapata and Lascarides (2003) presented in section 6.2 is used to create and rank the initial list of ambiguous interpretations. As in metaphor experiments, the parameters of the model were estimated from the BNC, using the grammatical relations output of RASP for BNC created by Andersen et al. (2008). In particular, I extracted

finish video		**enjoy book**	
Interpretations	Log-prob	Interpretations	Log-prob
view	-19.68	read	-15.68
watch	-19.84	write	-17.47
shoot	-20.58	work on	-18.58
edit	-20.60	look at	-19.09
film on	-20.69	read in	-19.10
film	-20.87	write in	-19.73
view on	-20.93	browse	-19.74
make	-21.26	get	-19.90
edit of	-21.29	re-read	-19.97
play	-21.31	talk about	-20.02
direct	-21.72	see	-20.03
sort	-21.73	publish	-20.06
look at	-22.23	read through	-20.10
record on	-22.38	recount in	-20.13

Table 7.1: Possible interpretations of metonymies ranked by my system

all direct and indirect object relations for the nouns from the metonymic phrases, i.e. all the verbs that take the head noun in the complement as an object (direct or indirect), in order to obtain the counts for $f(o, e)$ from Lapata and Lascarides' model. Relations expressed in the passive voice and with the use of coordination were also extracted. The verb-object pairs attested in the corpus only once were discarded, as well as the verb *be*, since it does not add any semantic information to the metonymic interpretation. In the case of indirect object relations, the verb was considered to constitute an interpretation together with the preposition, e.g. for the metonymic phrase "enjoy the city" the correct interpretation is *live in* as opposed to *live*.

As the next step I identified all possible verb phrase (VP) complements of the metonymic verb (both progressive and infinitive), which represent $f(v, e)$. This was done by searching for `xcomp` relations in the GRs output of RASP, in which the metonymic verb participates in any of its inflected forms. Infinitival and progressive complement counts were summed up to obtain the final frequency $f(v, e)$.

After the frequencies $f(v, e)$ and $f(o, e)$ were obtained, possible interpretations were ranked according to the model of Lapata and Lascarides (2003). The top interpretations for the metonymic phrases "enjoy book" and "finish video" together with their log-probabilities are shown in Table 7.1.

7.1.2 Comparison with the results of Lapata and Lascarides

I compared the output of my reimplementation of the model on Lapata and Lascarides' dataset with their own results obtained from the authors. The major difference between the two systems is that I extracted the data from the BNC parsed by RASP, as opposed

to the Cass chunk parser (Abney, 1996) utilised by Lapata and Lascarides. My system finds approximately twice as many interpretations as theirs and covers 80% of their lists (the system fails to find only some of the low-probability range verbs of Lapata and Lascarides). Then I compared the rankings of the two implementations using the Pearson correlation coefficient and obtained the average correlation of 0.83 (over all metonymic phrases from the dataset of Lapata and Lascarides).

I evaluated the performance of the system against the judgements elicited from humans in the framework of the experiment of Lapata and Lascarides (2003)[1]. The Pearson correlation coefficient between the ranking of my system and the human ranking equals to 0.62 (the inter-subject agreement on this task is 0.74). This is slightly lower than the number achieved by Lapata and Lascarides (0.64). Such a difference is likely to be caused by the fact that my system does not find some of the low-probability range verbs that Lapata and Lascarides included in their test set, and thus those interpretations get assigned a probability of 0. In addition, I conducted a one-tailed t-test to determine if the obtained counts were significantly different from those of Lapata and Lascarides. The difference is statistically insignificant ($t=3.6$; $df=180$; $p<.0005$), and the output of the system is deemed acceptable to be used for further experiments.

7.1.3 Data analysis

There has been a debate in linguistics literature as whether it is the noun or the verb in the metonymic phrase that determines the interpretation (Pustejovsky, 1991; Copestake and Briscoe, 1995). Pustejovsky's theory of noun qualia explains the contribution of the noun to the semantics of the whole phrase. However, it has been also pointed out that different metonymic verbs also place their own requirements on the interpretation of logical metonymy (Godard and Jayez, 1993; Pustejovsky and Bouillon, 1995; Copestake and Briscoe, 1995). I analysed the sets of interpretations for metonymic phrases extracted from the corpus using the method of Lapata and Lascarides (2003) with respect to such requirements. My data suggests the following classification criteria for metonymic verbs:

- ***Control* vs. *raising***. Consider the phrase "require poetry". *Require* is a typical object raising verb and, therefore, the most obvious interpretation of this phrase would be "require someone to *learn/recite* poetry", rather than "require to *hear* poetry" or "require to *learn* poetry", as suggested by the model of Lapata and Lascarides. Their model does not take into account raising syntactic frame and as such its interpretation of raising metonymic phrases will be based on the wrong kind of corpus evidence and lead to ungrammaticality. My expectation, however, is that control verbs tend to form logical metonymies more frequently. By analyzing the lists of control and raising verbs compiled by Boguraev and Briscoe (1987) I found

[1]For a detailed description of the human evaluation setup see Lapata and Lascarides (2003), pp 12-18.

evidence supporting this claim. Only 20% of raising verbs can form metonymic constructions (e.g. *expect, allow, request, require* etc.), while others cannot (e.g. *appear, seem, consider* etc.) This finding complies with the view previously articulated by Pustejovsky and Bouillon (1995). Due to both this finding and the fact that my experiments build on the approach of Lapata and Lascarides (2003), I gave preference to control verbs when compiling a dataset to develop and test the system.

- ***Activity* vs. *result***. Some metonymic verbs require the reconstructed event to be an *activity* (e.g. *begin writing the book*), while others require a *result* (e.g. *attempt to reach the peak*). This distinction potentially allows to rule out some incorrect interpretations, e.g. a resultative *find* for *enjoy book*, as *enjoy* requires an event of the type *activity*. Although I am not testing this hypothesis in the current work, automating this would be an interesting route for extension of my experiments in the future.

- ***Telic* vs. *agentive* vs. *other*** events. Another interesting observation captures the constraints that the metonymic verb imposes on the reconstructed event in terms of its function. While some metonymic verbs require rather *telic* events (e.g., *enjoy, want, try*), others have strong preference for *agentive* (e.g., *start*). However, for some categories of verbs it is hard to define a particular type of the event they require (e.g., *attempt the peak* should be interpreted as *attempt to reach the peak*, which is neither telic nor agentive).

7.2 Disambiguation experiments

The reimplementation of the method of Lapata and Lascarides produces interpretations in the form of ambiguous strings representing collectively all senses of the verb. The aim is, however, to construct the list of verb senses that are correct interpretations for the metonymic phrase. I assume the WordNet synset representation of a sense and map the ambiguous interpretations to WordNet synsets. This is done by searching the obtained lists for verbs, whose senses are in hyponymy and synonymy relations with each other according to WordNet and recording the respective senses.

After word sense disambiguation of the interpretations is completed, one needs to derive a new likelihood ranking for the resulting senses. Since there is no word sense disambiguated corpus available which would be large enough to reliably extract statistics for metonymic interpretations, the new ranking scheme is needed to estimate the likelihood of a WordNet synset as a unit from a non-disambiguated corpus. The calculation of synset likelihoods is based on the initial likelihood of the ambiguous verbs, relying on the hypothesis of Zipfian sense frequency distribution and information from WordNet glosses.

7.2.1 Generation of candidate senses

It has been recognised (Pustejovsky, 1991, 1995; Godard and Jayez, 1993) that correct interpretations tend to form semantic classes, and therefore, they should be related to each other by semantic relations, such as synonymy or hyponymy. The right senses of the verbs in the context of the metonymic phrase were obtained by searching the WordNet database for the senses of the verbs in the list that are in synonymy, hypernymy and hyponymy relations and storing the corresponding synsets in a new list of interpretations. If one synset was a hypernym (or hyponym) of the other, then both synsets were stored.

For example, for the metonymic phrase "finish video" the interpretations *watch, view* and *see* are synonymous, therefore the synset containing `(watch(3) view(3) see(7))` was stored. This means that sense 3 of *watch*, sense 3 of *view* and sense 7 of *see* would be correct interpretations of the metonymic expression.

The obtained number of synsets ranges from 14 ("try shampoo") to 1216 ("want money") for the whole dataset of Lapata and Lascarides (2003).

7.2.2 Ranking the senses

A problem arises with the obtained lists of synsets in that they contain different senses of the same verb. However, few verbs have such a range of meanings that their two different senses could represent two distinct metonymic interpretations (e.g., in case of *take* interpretation of "finish video", *shoot* sense and *look at, consider* sense are both acceptable interpretations, the second obviously being dispreferred). In the majority of cases the occurrence of the same verb in different synsets means that the list still needs filtering.

In order to do this I rank the synsets according to their likelihood of being a metonymic interpretation. The sense ranking is largely based on the probabilities of the verb strings derived by the model of Lapata and Lascarides (2003).

Zipfian sense frequency distribution

The probability of each ambiguous verb from the initial list represents the sum of probabilities of all senses of this verb. Hence this probability mass needs to be distributed over senses first. The sense frequency distribution for most words has been argued to be closer to Zipfian, rather than uniform or any other distribution (Preiss, 2006). This means that the first senses will be favored over the others, and the frequency of each sense will be inversely proportional to its rank in the list of senses (i.e. sense number, since word senses are ordered in WordNet by frequency). Thus the sense probability can be expressed as follows:

$$P_{v,k} = P_v \cdot \frac{1}{k} \tag{7.1}$$

Synset and its Gloss	Log-prob
(watch-v-1) - look attentively; "watch a basketball game"	-4.56
(view-v-2 consider-v-8 look-at-v-2) - look at carefully; study mentally; "view a problem"	-4.66
(watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6) - see or watch; "view a show on television"; "This program will be seen all over the world"; "view an exhibition"; "Catch a show on Broadway"; "see a movie"	-4.68
(film-v-1 shoot-v-4 take-v-16) - make a film or photograph of something; "take a scene"; "shoot a movie"	-4.91
(edit-v-1 redact-v-2) - prepare for publication or presentation by correcting, revising, or adapting; "Edit a book on lexical semantics"; "she edited the letters of the politician so as to omit the most personal passages"	-5.11
(film-v-2) - record in film; "The coronation was filmed"	-5.74
(screen-v-3 screen-out-v-1 sieve-v-1 sort-v-1) - examine in order to test suitability; "screen these samples"; "screen the job applicants"	-5.91
(edit-v-3 cut-v-10 edit-out-v-1) - cut and assemble the components of; "edit film"; "cut recording tape"	-6.20

Table 7.2: Metonymy interpretations as synsets (for "finish video")

where k is the sense number and P_v is the likelihood of the verb string being an interpretation according to the corpus data, i.e.

$$P_v = \sum_{k=1}^{N_v} P_{v,k} \tag{7.2}$$

where N_v is the total number of senses for the verb in question.

The problem that arises with (7.1) is that the inverse sense numbers $(1/k)$ do not add up to 1. In order to circumvent this, the Zipfian distribution is commonly normalised by the Nth generalised harmonic number. Assuming the same notation

$$P_{v,k} = P_v \cdot \frac{1/k}{\sum_{n=1}^{N_v} 1/n} \tag{7.3}$$

Once we have obtained the sense probabilities $P_{v,k}$, we can calculate the likelihood of the whole synset

$$P_s = \sum_{i=1}^{I_s} P_{v_i,k} \tag{7.4}$$

where v_i is a verb in the synset s and I_s is the total number of verbs in the synset s. The verbs suggested by WordNet, but not attested in the corpus in the required environment, are assigned the probability of 0. Some output synsets for the metonymic phrase "finish video" and their log-probabilities are demonstrated in Table 7.2.

Gloss processing

The model in the previous section penalises synsets that are incorrect interpretations. However, it can not discriminate well between the ones consisting of a single verb. By

Synset and its Gloss	Log-prob
(direct-v-1) - command with authority; "He directed the children to do their homework"	-6.65
(target-v-1 aim-v-5 place-v-7 direct-v-2 point-v-11) - intend (something) to move towards a certain goal; "He aimed his fists towards his opponent's face"; "criticism directed at her superior"; "direct your anger towards others, not towards yourself"	-7.35
(direct-v-3) - guide the actors in (plays and films)	-7.75
(direct-v-4) - be in charge of	-8.04

Table 7.3: Different senses of *direct* (for "finish video")

default it favours the sense with a smaller sense number in WordNet. This poses a problem for the examples such as *direct* for the phrase "finish video": our list contains several senses of it as shown in Table 7.3, and their ranking is not satisfactory. The only correct interpretation in this case, sense 3, is assigned a lower likelihood than the senses 1 and 2.

The most relevant synset can be found by using the information from WordNet glosses (the verbal descriptions of concepts, often with examples). The system searched for the glosses containing terms related to the noun in the metonymic phrase, here *video*. Such related terms would be its direct synonyms, hyponyms, hypernyms, meronyms or holonyms according to WordNet. The system assigned more weight to the synsets whose gloss contained related terms. In our example the synset `(direct-v-3)`, which is the correct metonymic interpretation, contained the term *film* in its gloss and was therefore selected. Its likelihood was multiplied by the factor of 10.

However, the glosses do not always contain the related terms; the expectation is that they will be useful in the majority of cases, not in all of them.

7.2.3 Evaluation

The ranking of the sense-based interpretations was evaluated against a gold standard created with the aid of human annotators.

Dataset

Five most frequent metonymic verbs were chosen to form the experimental data: *begin, enjoy, finish, try, start*. I randomly selected 10 metonymic phrases containing these verbs from the dataset of Lapata and Lascarides (2003) and split them into the development set (5 phrases) and the test set (5 phrases) as shown in Table 7.4.

Development Set	Test Set
enjoy book	enjoy story
finish video	finish project
start experiment	try vegetable
finish novel	begin theory
enjoy concert	start letter

Table 7.4: Metonymic phrases in development and test sets

Group 1	Group 2
finish video	finish project
start experiment	begin theory
enjoy concert	start letter

Table 7.5: Metonymic phrases for groups 1 and 2

Gold standard

The gold standards were created for the top 30 synsets obtained for each metonymic phrase after ranking. This threshold was set experimentally: the recall of correct interpretations among the top 30 synsets is 0.75 (average over metonymic phrases from the development set). This threshold allows to filter out a large number of incorrect interpretations. The gold standards for the evaluation of both synset ranking and the class-based model presented further on were created simultaneously in one annotation experiment.

Annotators Eight volunteer annotators participated in the experiment. All of them were native speakers of English and non-linguists. I divided them into 2 groups of 4. Participants in each group annotated three metonymic phrases as shown in Table 7.5.

Materials and task The annotators received written guidelines describing the task (2 pages), which were the only source of information on the experiment. The full version of the guidelines is shown in Appendix G. For each metonymic phrase the annotators were presented with a list of top 30 synsets produced by the system and asked to do the following.

- For each synset in the list, decide whether it was a plausible interpretation of the metonymic phrase in an imaginary context and remove the synsets that are not plausible interpretations.
- cluster the remaining ones according to their semantic similarity

Interannotator agreement The interannotator agreement was assessed in terms of f-measure (computed pairwise and then averaged across the annotators) and κ. The agreement in group 1 was F-measure = 0.76 and $\kappa = 0.56$ ($n = 2, N = 90, k = 4$); in group 2 – F-measure = 0.68 and $\kappa = 0.51$ ($n = 2, N = 90, k = 4$). This yielded the average agreement of F-measure = 0.72 and $\kappa = 0.53$. The interannotator agreement for the clustering part of the experiment will be reported in the next section.

(film-v-1 shoot-v-4 take-v-16)
(film-v-2)
(produce-v-2 make-v-6 create-v-6)
(direct-v-3)
(work-at-v-1 work-on-v-1)
(work-v-5 work-on-v-2 process-v-6)
(make-v-3 create-v-1)
(produce-v-1 bring-forth-v-3)
(watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6)
(watch-v-1)
(view-v-2 consider-v-8 look-at-v-2)
(analyze-v-1 analyse-v-1 study-v-1 examine-v-1 canvass-v-3 canvas-v-4)
(use-v-1 utilize-v-1 utilise-v-1 apply-v-1 employ-v-1)
(play-v-18 run-v-10)
(edit-v-1 redact-v-2)
(edit-v-3 cut-v-10 edit-out-v-1)
(screen-v-3 screen-out-v-1 sieve-v-1 sort-v-1)
(work-through-v-1 run-through-v-1 go-through-v-2)

Figure 7.1: Disambiguation gold standard for the phrase "finish video" (before clustering)

Subsequently, their annotations were merged into a gold standard, whereby an interpretation was considered correct if at least three annotators tagged it as such. The annotations for the remaining four phrases in the dataset were carried out by the author. The gold standard containing correct disambiguated interpretations for the metonymic phrase "finish video" is presented in Figure 7.1.

Evaluation measure

I evaluated the performance of the system against the gold standard. The objective was to find out if the synsets were distributed in such a way that the plausible interpretations appear at the top of the list and the incorrect ones at the bottom. The evaluation was performed in terms of mean average precision at top 30 synsets. To remind (cf. Chapter 5), MAP is defined as follows:

$$MAP = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ji}, \tag{7.5}$$

where M is the number of metonymic phrases, N_j is the number of correct interpretations for the metonymic phrase, P_{ji} is the precision at each correct interpretation (the number of correct interpretations among the top i ranks). First, the average precision was computed for each metonymic phrase independently. Then the mean values were calculated for the development and the test sets.

The motivation behind computing MAP instead of precision at a fixed number of synsets

Dataset	Verb Probability Mass Distribution	Gloss Processing	MAP
Development set	Uniform	No	0.51
Development set	Zipfian	No	0.65
Development set	Zipfian	Yes	0.73
Test set	Uniform	No	0.56
Test set	Zipfian	No	0.77
Test set	Zipfian	Yes	0.83

Table 7.6: Evaluation of the model ranking

(e.g. top 30) is that the number of correct interpretations varies dramatically for different metonymic phrases. MAP essentially evaluates how many good interpretations appear at the top of the list, which takes this variation into account.

Results

I compared the ranking obtained by applying Zipfian sense frequency distribution against that obtained by distributing probability mass over senses uniformly (baseline). I also considered the rankings before and after gloss processing. The results are shown in Table 7.6. These results demonstrate the positive contribution of both Zipfian distribution and gloss processing to the ranking. MAP of the system on the test set is 0.83, which suggests that the system is able to reliably disambiguate and re-rank metonymic interpretations.

I additionally compared the rankings produced by the system and the baseline using Spearman's rank correlation coefficient. The average rank correlation across the test set is 0.86, which suggests that the rankings of the two systems are not independent, however different. I then also compared the rankings using the Wilcoxon Signed Ranks test (Siegel and Castellan, 1988). For two paired samples (e.g. the rankings of the system and the baseline), Wilcoxon Signed Ranks test considers differences in scores with respect to direction, or sign (positive, negative, no change), similarly to the Sign test, as well as magnitude, which makes it appropriate to compare rankings. The test was carried out on all of the phrases from the test set, resulting in $z = 0.47$ ($T^+ = 3223, N = 116, p = 0.31$). This suggests that the overall difference is not statistically significant ($\alpha = 0.05, p > \alpha$), however for one of the metonymic phrases ("enjoy story") the difference between the system and the baseline rankings was shown to be significant ($z = 1.91, T^+ = 119, N = 28, p = 0.02$).

7.3 Clustering experiments

The obtained lists of synsets constitute the basis for creating a class-based representation of the interpretation of logical metonymy. Besides, identifying meaningful clusters of in-

terpretations this would allow us to filter out irrelevant senses. For example, the synset "(target-v-1 aim-v-5 place-v-7 direct-v-2 point-v-11) - intend (something) to move towards a certain goal" for "finish *directing* a video" is not likely to be semantically similar to any other synset in the list. Clustering relying on the distances in semantic feature space may be able to reveal such cases.

The challenge of our clustering task is that one needs to cluster verb senses as opposed to non-disambiguated verbs and, therefore, needs to model the distributional information representing a single sense given a non-disambiguated corpus. In this experiment I design feature sets that describe verb senses and test their informativeness using a range of clustering algorithms.

7.3.1 Feature extraction

The goal is to cluster synsets with similar distributional semantics together. The features were extracted from the BNC parsed by RASP. The feature sets comprise the nouns co-occurring with the verbs in the synset in subject and object relations. The object relations were represented by the nouns co-occurring with the verb in the same syntactic frame as the noun in the metonymic phrase (e.g. `indirect object` with the preposition *in* for *live in the city*, `direct object` for *visit the city*). These nouns together with the co-occurrence frequencies were used as features for clustering. The subject and object relations were marked respectively. The feature vectors for synsets were constructed from the feature vectors of the individual verbs included in the synset. I will use the following notation to describe the feature sets:

$$\begin{aligned}
&\mathbb{V}_1 = \{c_{11}, c_{12}, ..., c_{1N}\} \\
&\mathbb{V}_2 = \{c_{21}, c_{22}, ..., c_{2N}\} \\
&\vdots \\
&\mathbb{V}_K = \{c_{K1}, c_{K2}, ..., c_{KN}\}
\end{aligned}$$

where K is the number of the verbs in the synset, $\mathbb{V}_1, ..., \mathbb{V}_K$ are the feature sets of each verb, N is the total number of features (ranges from 18517 to 20661 in my experiments) and c_{ij} are the corpus counts. The following feature sets were taken to represent the whole synset.

Feature set 1 - the union of the features of all the verbs of the synset.

$$\mathbb{F}_1 = \mathbb{V}_1 \cup \mathbb{V}_2 \cup \ldots \cup \mathbb{V}_K$$

The counts are computed as follows:

$$\mathbb{F}_1 = \{\sum_{i=1}^{K} c_{i1}, \sum_{i=1}^{K} c_{i2}, ..., \sum_{i=1}^{K} c_{iN}\}$$

This feature set is the most naive representation of a synset, the problem with it being that it contains features describing irrelevant senses of the verbs. Such irrelevant features can be filtered out by taking an intersection of the nouns of all the verbs in the synset. This yields the following feature set:

Feature set 2 - the intersection of the feature sets of the verbs in the synset.

$$\mathbb{F}_2 = \mathbb{V}_1 \cap \mathbb{V}_2 \cap \ldots \cap \mathbb{V}_K$$

The counts are computed as follows:

$$\mathbb{F}_2 = \{f_1, f_2, ..., f_N\}$$
$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \prod_{i=1}^{K} c_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

This would theoretically be a comprehensive representation. However, in practice the system is likely to run into the problem of data sparseness and some synsets end up with very limited feature vectors, or no feature vectors at all. The next feature set is an attempt to accommodate this problem.

Feature set 3 - union of features as in feature set 1, reweighted in favor of overlapping features.

$$\mathbb{F}_3 = \mathbb{V}_1 \cup \mathbb{V}_2 \cup \ldots \cup \mathbb{V}_K \cup \beta * (\mathbb{V}_1 \cap \mathbb{V}_2 \cap \ldots \cap \mathbb{V}_K) = \mathbb{F}_1 \cup \beta * \mathbb{F}_2$$

where β is the weighting coefficient for the overlapping features. The counts are computed as follows:

$$\mathbb{F}_3 = \{\sum_{i=1}^{K} c_{i1} + \beta f_1, \sum_{i=1}^{K} c_{i2} + \beta f_2, ..., \sum_{i=1}^{K} c_{iN} + \beta f_N\}$$
$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \prod_{i=1}^{K} c_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

I experimented with different values of β from the range [1..10] and found $\beta = 5$ to be the optimal setting for this parameter.

The feature sets 4 and 5 are also motivated by the problem of sparse data. But the intersection of features is calculated pairwise, instead of an overall intersection.

Feature set 4 - pairwise intersections of the feature sets of the verbs in the synset.

$$\mathbb{F}_4 = (\mathbb{V}_1 \cap \mathbb{V}_2) \cup \ldots \cup (\mathbb{V}_1 \cap \mathbb{V}_K)$$
$$\cup(\mathbb{V}_2 \cap \mathbb{V}_3) \cup \ldots \cup (\mathbb{V}_2 \cap \mathbb{V}_K) \cup \ldots$$
$$\cup(\mathbb{V}_{K-2} \cap \mathbb{V}_{K-1}) \cup (\mathbb{V}_{K-1} \cap \mathbb{V}_K)$$

The counts are computed as follows:

$$\mathbb{F}_4 = \{f_1, f_2, ..., f_N\}$$
$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \exists x, y | c_{xj} \cdot c_{yj} \neq 0, x, y \in [1..K], x \neq y; \\ 0 & \text{otherwise.} \end{cases}$$

Feature set 5 - the union of features as in feature set 1, reweighted in favor of overlapping features (pairwise overlap).

$$\mathbb{F}_5 = \mathbb{F}_1 \cup \beta * \mathbb{F}_4$$

where β is the weighting coefficient for the overlapping features. The counts are computed as follows:

$$\mathbb{F}_5 = \{\sum_{i=1}^{K} c_{i1} + \beta f_1, \sum_{i=1}^{K} c_{i2} + \beta f_2, ..., \sum_{i=1}^{K} c_{iN} + \beta f_N\}$$
$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \exists x, y | c_{xj} \cdot c_{yj} \neq 0, x, y \in [1..K], x \neq y; \\ 0 & \text{otherwise.} \end{cases}$$

7.3.2 Clustering methods

To cluster the synsets I experimented with the following clustering algorithms and configurations:

Clustering algorithms: Synsets were clustered using both partitional (K-means, repeated bisections) and agglomerative clustering. *K-means* first randomly selects a number of cluster centroids. Then it assigns each data point to a cluster with the nearest centroid and recomputes the centroids. This process is repeated until the clustering solution stops changing. *Repeated bisections* algorithm partitions the data points by performing a sequence of binary divisions in a way that optimises the chosen criterion function. *Agglomerative* clustering, in contrast, is performed by joining the nearest pairs of objects (or clusters of objects) in a hierarchical fashion. The similarity of clusters in agglomerative clustering is judged using *single link* (the minimum distance between elements in each cluster), *complete link* (the maximum distance between elements in each cluster) or *group average* (the mean distance between elements in each cluster) methods.

Similarity measures: Cosine similarity function and Pearson Correlation coefficient were used to determine similarity of the feature vectors. They are computed as follows:

$$\text{Cosine}(v,u) = \frac{\sum_{i=1}^{n} v_i u_i}{\sqrt{\sum_{i=1}^{n} (v_i)^2}\sqrt{\sum_{i=1}^{n} (u_i)^2}}$$

$$\text{Corr}(v,u) = \frac{n\sum_{i=1}^{n} v_i u_i - \sum_{i=1}^{n} v_i \sum_{i=1}^{n} u_i}{\sqrt{n\sum_{i=1}^{n} (v_i)^2 - (\sum_{i=1}^{n} v_i)^2}\sqrt{n\sum_{i=1}^{n} (u_i)^2 - (\sum_{i=1}^{n} u_i)^2}}$$

where v and u are the two feature vectors and n is the number of features.

Criterion function: The goal is to maximise intra-cluster similarity and to minimise inter-cluster similarity. I use the function ε_2 (Zhao and Karypis, 2001) defined in the following way

$$\epsilon_2 = \min \sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} sim(v,u)}{\sqrt{\sum_{v,u \in S_i} sim(v,u)}}$$

where S is the set of objects to cluster, S_i is the set of objects in cluster i, n_i is the number of objects in cluster i, k is the number of clusters and sim stands for the chosen similarity measure. As such, the numerator represents inter-cluster similarity and the denominator intra-cluster similarity.

Feature matrix scaling: Feature space in NLP clustering tasks usually has a large number of dimensions, that are not equally informative. Hence, clustering may benefit from the prior identification and emphasis of the most discriminative features. This process is known as *feature matrix scaling*, which I perform in the following ways:

- IDF paradigm: the counts of each column are scaled by the $\log_2$ of the total number of rows divided by the number of rows the feature appears in (this scaling scheme only uses the frequency information inside the matrix). The effect is to de-emphasise columns that appear in many rows and are, therefore, not very discriminative features.

- Preprocess the matrix by dividing initial counts for each noun by the total number of occurrences of this noun in the whole BNC. The objective is again to decrease the influence of generally frequent nouns that are also likely to be ambiguous features.

The number of clusters: The number of clusters (k) for each metonymic phrase was set manually according to the number observed in the gold standard.

The clustering experiments were performed using the Cluto toolkit (Karypis, 2002). Cluto has been widely applied in NLP, mainly for document classification tasks, but also for a number of experiments on lexical semantics (Baroni et al., 2008).

7.3.3 Evaluation measures

I will call the gold standard partitions *classes* and the clustering solution suggested by the model a set of *clusters*. The following measures were used to evaluate clustering:

Purity (Zhao and Karypis, 2001) is calculated as follows

$$\text{Purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

where $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, ..., c_j\}$ is the set of classes, N is the number of objects to cluster. Purity evaluates only the homogeneity of the clusters, i.e. the average proportion of similar objects within the clusters. High purity is easy to achieve when the number of clusters is large. As such, it does not provide a measure for the trade off between the quality of clustering and the number of classes.

F-Measure was introduced by van Rijsbergen (1979) and adapted to the clustering task by Fung et al. (2003). It matches each class with the cluster that has the highest precision and recall. Using the same notation as above

$$F(\mathbb{C}, \Omega) = \sum_j \frac{|c_j|}{N} \max_k \{F(c_j, \omega_k)\}$$

$$F(c_j, \omega_k) = \frac{2 \cdot P(c_j, \omega_k) \cdot R(c_j, \omega_k)}{P(c_j, \omega_k) + R(c_j, \omega_k)}$$

$$R(c_j, \omega_k) = \frac{|\omega_k \cap c_j|}{|c_j|}$$

$$P(c_j, \omega_k) = \frac{|\omega_k \cap c_j|}{|\omega_k|}$$

Recall represents a portion of objects of class c_j assigned to cluster ω_k and precision the portion of objects in cluster ω_k belonging to the class c_j.

Rand Index (Rand, 1971). An alternative way of looking at clustering is to consider it as a series of decisions for each pair of objects, whether these two objects belong to the same cluster or not. For N objects there will be $N(N-1)/2$ pairs. One needs to calculate the number of true positives (TP) (similar objects in the same cluster), true negatives (TN) (dissimilar objects in different clusters), false positives (FP) (dissimilar objects in the same cluster) and false negatives (FN) (similar objects in different clusters). Rand Index corresponds to accuracy: it measures the percentage of decisions that are correct considered pairwise.

$$RI = \frac{TP + TN}{TP + FP + TN + FN}$$

Variation of Information (Meilă, 2007) is an entropy-based measure defined as follows:

$$VI(\Omega, \mathbb{C}) = H(\Omega|\mathbb{C}) + H(\mathbb{C}|\Omega)$$

where $H(\mathbb{C}|\Omega)$ is the conditional entropy of the class distribution given the proposed clustering, $H(\Omega|\mathbb{C})$ is the opposite.

$$H(\Omega|\mathbb{C}) = -\sum_j \sum_k \frac{|\omega_k \cap c_j|}{N} \log \frac{|\omega_k \cap c_j|}{|\omega_k|}$$

$$H(\mathbb{C}|\Omega) = -\sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{|\omega_k \cap c_j|}{|c_j|}$$

where $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, ..., c_j\}$ is the set of classes, N is the number of objects to cluster. I report the values of VI normalised by $\log N$, which brings them into the range $[0, 1]$.

It is easy to see that VI is symmetrical. This means that it accounts for both homogeneity (only similar objects within the cluster) and completeness (all similar objects are covered by the cluster). In the perfectly homogeneous case the value of $H(\mathbb{C}|\Omega)$ is 0, in the perfectly complete case the value of $H(\Omega|\mathbb{C})$ is 0. The values are maximal (and equal to $H(\mathbb{C})$ and $H(\Omega)$ respectively) when the clustering gives no new information and the class distribution within each cluster is the same as the overall class distribution. This measure provides an adequate evaluation of the clustering solutions where the number of clusters is different from that in the gold standard.

7.3.4 Clustering gold standard

System clustering was evaluated using the same dataset as in the disambiguation experiment (see Table 7.4, page 124). The clustering gold standard was created in conjunction with the disambiguation gold standard for the top 30 synsets from the lists of interpretations. It consists of a number of clusters containing correct interpretations in the form of synsets and a cluster containing incorrect interpretations. The cluster containing incorrect interpretations is considerably larger than the others for the majority of metonymic phrases. The gold standard exemplified for the metonymic phrase "finish video" is presented in Figure 7.2. The glosses and the cluster with incorrect interpretations are omitted in this example for the sake of brevity.

I estimated the inter-annotator agreement by comparing the annotations pairwise (each annotator with each other annotator) and assessed it using the same clustering evaluation measures as the ones used to assess the system performance. In order to compare the groupings elicited from humans I added the cluster with the interpretations they excluded as incorrect to their clustering solutions. This was necessary, as the metrics used require that all annotators' clusterings contain the same objects (all 30 interpretations). Within each group the clustering partition of the annotator exhibiting the highest agreement with the remaining annotators as computed pairwise was selected for the gold standard.

After having evaluated the agreement pairwise for each metonymic phrase I calculated the average across the metonymic phrases and the pairs of annotators. The obtained

Cluster 1:
(film-v-1 shoot-v-4 take-v-16)
(film-v-2)
(produce-v-2 make-v-6 create-v-6)
(direct-v-3)
(work-at-v-1 work-on-v-1)
(work-v-5 work-on-v-2 process-v-6)
(make-v-3 create-v-1)
(produce-v-1 bring-forth-v-3)

Cluster 2:
(watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6)
(watch-v-1)
(view-v-2 consider-v-8 look-at-v-2)
(analyze-v-1 analyse-v-1 study-v-1 examine-v-1 canvass-v-3 canvas-v-4)
(use-v-1 utilize-v-1 utilise-v-1 apply-v-1 employ-v-1)
(play-v-18 run-v-10)

Cluster 3:
(edit-v-1 redact-v-2)
(edit-v-3 cut-v-10 edit-out-v-1)
(screen-v-3 screen-out-v-1 sieve-v-1 sort-v-1)
(work-through-v-1 run-through-v-1 go-through-v-2)

Figure 7.2: Clustering gold standard for the phrase "finish video"

agreement equals 0.75 in terms of purity, 0.67 in terms of Rand index, 0.76 in terms of F-measure and 0.37 in terms of VI[2]. It should be noted, however, that the number of clusters produced varies from annotator to annotator and the chosen measures (except for VI) penalise this. The obtained results for inter-annotator agreement demonstrate that the task of clustering word senses in the context of logical metonymy is intuitive to humans, but nonetheless, challenging.

7.3.5 Experiments and results

Development set

To select the best parameter setting I ran the experiments on the development set varying the parameters described in section 7.3.2 for feature sets 1 to 5. The system clustering solutions were evaluated for each metonymic phrase separately; the average values for the best clustering configurations for each algorithm and each feature set on the development set are given in Table 7.7. The best result was obtained for the phrase "enjoy concert" as

[2]Please note normalised VI values are in the range [0,1] and the lower values indicate better clustering quality

Algorithm	F. S.	Purity	RI	F-measure	VI
K-means	F1	0.6	0.52	0.54	0.45
No scaling	F2	0.61	0.58	0.61	0.45
Cosine	F3	0.57	0.5	0.54	0.47
	F4	**0.65**	**0.57**	**0.69**	**0.35**
	F5	0.6	0.54	0.57	0.44
RB	F1	0.61	0.51	0.58	0.43
No scaling	F2	0.62	0.57	0.63	0.44
Cosine	F3	0.63	0.52	0.61	0.40
	F4	**0.64**	**0.56**	**0.70**	**0.34**
	F5	0.61	0.52	0.59	0.42
Agglomerative	F1	0.61	0.47	0.76	0.33
No scaling	F2	0.61	0.57	0.70	0.44
Cosine	F3	0.61	0.47	0.64	0.35
Group average	F4	0.63	0.5	0.69	0.31
	F5	0.6	0.46	0.64	0.35

Table 7.7: Average clustering results (development set)

Algorithm	F. S.	Purity	RI	F-measure	VI
K-means	F1	0.7	0.54	0.58	0.35
No scaling	F2	0.67	0.48	0.57	0.36
Cosine	F3	0.7	0.54	0.58	0.35
	F4	**0.73**	**0.70**	**0.88**	**0.19**
	F5	0.7	0.54	0.58	0.35

Table 7.8: Best clustering results (for *enjoy concert*, development set)

shown in Table 7.8. The clustering solution produced by the system for the phrase "enjoy concert" is demonstrated in Figure 7.3.

The performance of the system is similar across the algorithms. However, the agglomerative algorithm tends to produce single object clusters and one large cluster containing the rest, which is strongly dispreferred. For this reason, I test the system only using K-means and repeated bisections. The obtained results suggest that feature set 4 is the most informative, although for agglomerative clustering feature set 1 yields a surprisingly good result. I will use feature set 4 for evaluation on the test set, as it proves to be useful for all three clustering algorithms.

Cluster 1:
(provide-v-2 supply-v-3 ply-v-1 cater-v-1)
(know-v-5 experience-v-2 live-v-6)
(attend-v-2 take-care-v-3 look-v-6 see-v-14)
(watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6)
(give-v-8 gift-v-2 present-v-7)
(give-v-32)
(hold-v-3 throw-v-11 have-v-8 make-v-26 give-v-6)
(bet-v-2 wager-v-1 play-v-30)
(watch-v-2 observe-v-7 follow-v-13 watch-over-v-1 keep-an-eye-on-v-1)
(leave-v-6 allow-for-v-1 allow-v-5 provide-v-5)
(present-v-4 submit-v-4)
(give-v-3)
(refer-v-2 pertain-v-1 relate-v-2 concern-v-1 come-to-v-2 bear-on-v-1 touch-v-4 touch-on-v-2 have-to-doe-with-v-1)
(include-v-2)
(yield-v-1 give-v-2 afford-v-2)
(supply-v-1 provide-v-1 render-v-2 furnish-v-1)
(perform-v-3)
(see-v-5 consider-v-1 reckon-v-3 view-v-1 regard-v-1)
(deem-v-1 hold-v-5 view-as-v-1 take-for-v-1)
(determine-v-8 check-v-21 find-out-v-3 see-v-9 ascertain-v-3 watch-v-7 learn-v-6)
(learn-v-2 hear-v-2 get-word-v-1 get-wind-v-1 pick-up-v-5 find-out-v-2 get-a-line-v-1 discover-v-2 see-v-6)
(feed-v-2 give-v-24)
(include-v-1)

Cluster 2:
(play-v-3)
(play-v-18 run-v-10)
(act-v-10 play-v-25 roleplay-v-1 playact-v-1)
(play-v-14)

Cluster 3:
(attend-v-1 go-to-v-1)
(watch-v-1)
(entertain-v-2 think-of-v-2 toy-with-v-1 flirt-with-v-1 think-about-v-2)

Figure 7.3: Clustering solution for “enjoy concert”. Red, blue and black colors represent gold standard classes

Algorithm	F. S.	Purity	RI	F-measure	VI
Baseline		0.48	0.40	0.31	0.51
K-means	**F4**	**0.65**	**0.52**	**0.64**	**0.33**
RB	F4	0.63	0.48	0.60	0.37
Agreement		0.75	0.67	0.76	0.37

Table 7.9: Clustering results on the test set

Test set

I present the results for the best system configuration on the test data in Table 7.9. The system clustering was compared to that of a baseline built using a simple heuristic and an upper bound set by the inter-annotator agreement. The baseline assigns synsets that contain the same verb string to the same cluster. The system outperforms the naive baseline, but does not reach the upper bound. K-means algorithm yields the best result of 0.65 (Purity), 0.52 (Rand index), 0.64 (F-measure) and 0.33 (VI).

Discussion

A particularity of our clustering task is that our goal is to eliminate incorrect interpretations as well as assign the correct ones to their classes based on semantic similarity. The cluster containing incorrect interpretations is often significantly larger than the other clusters. The overall trend is that the system selects correct interpretations and assigns them to smaller clusters, leaving the incorrect ones in one large cluster, as desired.

A common error of the system is that the synsets that contain different senses of the same verb often get clustered together. This is due to the fact that the features are extracted from a non-disambiguated corpus, which results in the following problems: (1) the verbs are ambiguous, therefore, the features, as extracted from the corpus, represent all the senses of the verb in one feature set. The task of dividing this feature set into subsets describing particular senses of the verb is very hard; (2) the features themselves (the nouns) are ambiguous (different senses of a noun can co-occur with different senses of a verb), which makes it very hard to distribute the counts realistically over verb senses.

However, it is not always the case that synsets with overlapping verbs get clustered together (in 38% of all cases the same verb string is assigned to different clusters). This demonstrates the contribution of the presented feature sets. More importantly, synsets containing different verbs are often assigned to the same cluster, when the sense is related (mainly for feature sets 2 and 4), which was the goal of clustering.

7.4 Conclusion

This chapter presented two experiments on resolution of logical metonymy. First, I described a system producing disambiguated interpretations of logical metonymy with respect to word sense. In the framework of the experiment I developed a novel scheme for estimating the likelihood of a WordNet synset as a unit from a non-disambiguated corpus and demonstrated its effectiveness for the task of interpretation of logical metonymy.

Using this sense-based representation I then created a class-based model of metonymic interpretation. This model is intuitive to humans and complies with the results of theoretical research on logical metonymy in linguistics. Within the clustering experiment, I addressed the issue of modelling distributional semantics of single senses represented in the form of WordNet synsets using a non-disambiguated corpus. The obtained results proved the effectiveness of the proposed method.

Chapter 8

Conclusions

In this chapter, I briefly summarise the main contributions of the presented work and suggest future research directions in the field of computational processing of figurative language.

8.1 Contributions of this thesis

The experiments described in this thesis are the first attempt at open-domain automatic processing of metaphor in free text. In contrast to previous approaches, the presented systems do not rely on any hand-crafted knowledge specific to metaphor, and instead employ statistical modelling of linguistic data. I demonstrated that the proposed methods provide accurate non-trivial information about metaphor, and additionally verified whether similar techniques can be used to address another type of figurative language as well, specifically logical metonymy.

I explored figurative language from three different perspectives: that of linguistics, i.e. how creative thought is expressed in natural languages (by metaphor annotation and data analysis); that of computer science, i.e. how computational mechanisms of creative thought can be simulated (by implementing metaphor and logical metonymy processing systems); and that of cognitive science, i.e. to what extent this simulation replicates the conceptual structures intuitive to humans (by studying human annotation of metaphorical associations and clustering metonymic interpretations). Thus, the novelties introduced in the thesis concern aspects of all these areas. The main contributions are listed below.

8.1.1 Metaphor

A new annotation scheme for metaphorical mappings: In Chapter 2, I presented and critiqued a number of previous accounts of metaphor annotation. I also noted that

to date there has been no proposal for a flexible scheme that would allow annotation of metaphorical associations in arbitrary text. Wallington et al. (2003) aimed at assigning a preconstructed mapping from the Master Metaphor List to a metaphorical expression, a task that can be viewed as a classification of metaphorical expressions with respect to their topic. In addition, Wallington et al. (2003) did not report any inter-annotator agreement on this task. The annotation scheme I presented in Chapter 3 is designed with open-domain metaphor annotation in mind. Metaphorical mappings are annotated by explicit context comparison, and source and target domain labels are assigned to the contexts independently, rather than in the form of a preconstructed mapping. I tested the reliability of the scheme in a setting with multiple annotators. The agreement for metaphor identification was $\kappa = 0.64$, and for annotation of source–target domain mappings $\kappa = 0.57$.

A new metaphor corpus: I annotated metaphorical expressions and the corresponding associations using the above annotation scheme in a 13,642 word subset of the BNC, which covers various genres: fiction, newspaper and journal articles, essays on politics, international relations and sociology, radio broadcast (transcribed speech). To my knowledge, this is the first available corpus annotated for source–target domain mappings. Part of the corpus (1,566 words) was annotated by three independent annotators to allow a measurement of their agreement on the task; the rest of it was annotated by the author. I provided an analysis of the data and discussed the observed patterns of disagreement in Chapter 3. Although the corpus is not sufficiently large to serve as training data for a supervised machine learning algorithm, it nevertheless provides a suitable dataset for the evaluation of the future metaphor processing systems.

The first corpus-based study of conceptual metaphor: The annotation experiment described in Chapter 3 is the first empirical study of conceptual metaphor in unrestricted text. Annotators reached some agreement on the assignment of source-target domain mappings, which suggests that metaphorical associations may exist. However, data analysis revealed a number of problems that conceptual metaphor theory does not explain. For example, the annotators experienced great difficulty choosing the right level of abstraction of source and target domain categories involved in the mapping. This suggests that it is hard to assign explicit labels to source and target domains. This finding motivated the implicit modelling of source and target domain categories within my computational approach to metaphor.

The notion of clustering by association: An important theoretical contribution of the thesis and the core of my implementation is the idea of clustering by association. Previous approaches to noun clustering based on contextual cues aimed to discover classes of nouns with similar meanings. By analysing corpus data I discovered that while concrete concepts do cluster by meaning similarity, the principle by which abstract concepts are clustered is association with the same source domain. For instance, the non-synonymous concepts of *marriage* and *democracy* may be found in one cluster since they are both viewed as

mechanisms. It is exactly this association that makes it possible for them to appear in similar contexts, such as with verbs *function* and *work*. This finding is interesting for both linguistics and lexical acquisition within NLP. My computational approach to metaphor identification is entirely built around this idea.

The first fully automated model for metaphor identification in unrestricted text: One of the most significant contributions of the present work is the design of a novel minimally supervised metaphor identification algorithm (Chapter 4). Starting from a small seed set of manually annotated metaphorical expressions, the system harvests a large number of metaphors of similar syntactic structure from a corpus. The method is distinguished from previous work in that it does not employ any hand-crafted knowledge, other than the initial seed set, but, in contrast, captures metaphoricity by means of verb and noun clustering. While the recall in the current experiment was low, the system has been shown to operate with a high precision of 0.79 and to discover novel metaphors, non-synonymous to any of the seed phrases. I expect that increasing the size of the seed set would make it possible to improve the recall of the system without significant loss in precision.

The first computational model for metaphor paraphrasing: Another significant contribution is my approach to metaphor interpretation (Chapter 5). As opposed to previous accounts, I defined metaphor interpretation as a paraphrasing task. My algorithm produces literal paraphrases for metaphorical expressions with an accuracy of 0.81. Other NLP applications that can directly use the output from such a metaphor processing system, e.g. MT, IE or opinion mining as discussed in Chapter 1. My method also differs from previous work in that it does not rely on any hand-crafted knowledge about metaphor, but in contrast employs a probabilistic context-based model for paraphrase selection and automatically induced selectional preferences.

8.1.2 Logical metonymy

A new sense disambiguation method for logical metonymy interpretation: My approach to logical metonymy is an extension of that of Lapata and Lascarides (2003), which generates a list of interpretations with their likelihood derived from a corpus. These interpretations are string-based, i.e. they are not disambiguated with respect to word sense. I proposed a sense-based representation of the interpretation of logical metonymy and developed a new word sense disambiguation method for the task. I also derived a ranking scheme for verb senses using an unannotated corpus, WordNet sense numbering and glosses. My system identifies and ranks the disambiguated metonymic interpretations with a mean average precision of 0.83.

Evidence for existing linguistic claims about class-based structure behind metonymic interpretations: It has previously been suggested in linguistic literature

that default metonymic interpretations tend to form semantic classes (Vendler, 1968; Pustejovsky, 1991; Godard and Jayez, 1993). This thesis offers experimental evidence to verify these claims. I conducted a human experiment, in which human subjects were asked to cluster possible interpretations into classes. Their agreement was measured at F-measure = 0.76. This indicates that the class-based structure behind metonymic interpretation might be learnable by humans. The intuitiveness is supported by the fact that they received minimal instructions.

Computational model of class-based logical metonymy: Having verified this idea empirically, I then moved on to build a computational simulation of the class discovery process. I experimented with a number of clustering algorithms, and proposed feature sets for modelling a particular sense of a verb using a non-disambiguated corpus. The system clusters the senses with an F-measure of 0.64 as compared to the gold standard, given that human agreement is F-measure = 0.76.

8.2 Future directions

The eighties and the nineties provided us with a wealth of ideas on the structure and mechanisms of metaphor and metonymy. The computational approaches formulated back then are still highly influential, although their use of task-specific hand-coded knowledge is becoming increasingly less common. The last decade witnessed a significant technological leap in natural language computation, whereby manually crafted rules gradually gave way to more robust corpus-based statistical methods. This is also the case for metaphor and metonymy research. In this thesis, I presented the first fully automated corpus-based approach to metaphor identification and interpretation. However, some important problems remain unsolved. This section is devoted to the limitations and extensions of the current work, as well as suggestions for new experiments.

8.2.1 Metaphor

The experiments presented in Chapters 4 and 5 yielded encouraging results and revealed many avenues for future research. These include both direct extensions to the described systems, as well as new experiments on metaphor processing and its applications.

Limitations and extensions of the current work

While the current work met with success, it was so far small in scale only dealing with metaphors expressed by a verb in verb-subject and verb-direct object constructions. Restricting the scope to verbs was a methodological step aimed at testing the main principles of the proposed approach in a well-defined setting and it was done without loss

of generality. However, I expect the presented solutions to scale well to other syntactic constructions. This is due to the fact that both identification and interpretation algorithms rely on the concept of distributional semantics. Distributional information of word co-occurrences in corpora has been shown to be indicative of word meanings for all parts of speech classes, including the tasks of distributional clustering (Hatzivassiloglou and McKeown, 1993; Boleda Torrent and Alonso i Alemany, 2003) and selectional preference acquisition (Brockmann and Lapata, 2003; Zapirain et al., 2009; Ó Séaghdha, 2010).

Such an extension of the identification system would require the creation of a seed set exemplifying more syntactic constructions and the corpus search over further grammatical relations (e.g. verb - indirect object relations "she was *transported* with pleasure, Hillary *leapt* in the conversation", adjectival modifier - noun relations "*slippery* mind, *deep* unease, *heavy* loss", noun - PP complement relations "a *fraction* of self-control, a *foot* of a mountain", verb - VP complement relations "*aching* to begin the day", copula constructions "Death is the sorry end of the human story, not a mysterious prelude to a new one" etc.) Besides noun and verb clustering, it would also be necessary to perform clustering of adjectives and adverbs. Clusters of verbs, adjectives, adverbs and concrete nouns would then represent source domains within the model. The data study described in Chapter 3 suggested that it is sometimes difficult to choose the optimal level of abstraction of domain categories that would generalise well over the data. Although the system does not explicitly assign any domain labels, its domain representation is still restricted by the fixed level of generality of source concepts, defined by the chosen cluster granularity. To relax this constraint, one could attempt to automatically optimise cluster granularity to fit the data more accurately and to ensure that the generated clusters explain the metaphorical expressions in the data more comprehensively. A hierarchical clustering algorithm (e.g. Yu et al., 2006) could be employed for this purpose. Besides this, it would be desirable to be able to generalise metaphorical associations learned from one type of syntactic constructions across all syntactic constructions, without providing explicit seed examples for the latter. For instance, given the seed phrase "*stir* excitement", representing the conceptual mapping FEELINGS ARE LIQUIDS, the system should be able to discover not only that phrases such as "*swallow* anger" are metaphorical, but that phrases such as "*ocean* of happiness" are as well.

The extension of the paraphrasing system to other syntactic constructions would involve the extraction of further grammatical relations from the corpus, such as those listed above, and their incorporation into the context-based paraphrase selection model. Extending both the identification system and the paraphrasing system would require the application of the selectional preference model to other word classes. Although Resnik's selectional association measure has been used to model selectional preferences (SPs) of verbs for their nominal arguments, it is in principle a generalisable measure of word association. Information-theoretic word association measures, e.g. mutual information (Church and Hanks, 1990), have been successfully applied to a range of syntactic constructions in a

number of NLP tasks (Hoang et al., 2009; Baldwin and Kim, 2010). This suggests that applying a distributional association measure, such as the one proposed by Resnik, to other part-of-speech classes should still result in a realistic model of semantic fitness, which in our terms corresponds to a measure of "literalness" of the paraphrases.

In addition, the selectional preference model can be improved by using an SP acquisition algorithm that can handle word sense ambiguity, e.g. (Rooth et al., 1999; Ó Séaghdha, 2010; Reisinger and Mooney, 2010). The current approach relies on SP classes produced by hard clustering and fails to accurately model word senses of generally polysemous words. This resulted in a number of errors in metaphor paraphrasing and it therefore needs to be addressed in the future.

The current version of the metaphor paraphrasing system still relies on some hand-coded knowledge in the form of WordNet. WordNet has been criticised for a lack of consistency, high granularity of senses and negligence with respect to some important semantic relations (Lenat et al., 1995). In addition, WordNet is a general-domain resource, that is less suitable if one wanted to apply the system to specific-domain data. Such an application could however be desirable, for instance, to assist ontology extraction for that domain. For all of the above reasons it would be preferable to develop a WordNet-free fully automated approach to metaphor resolution. Vector space models of word meaning (Erk, 2009; Rudolph and Giesbrecht, 2010) might provide a solution, as they have proved efficient in general paraphrasing and lexical substitution settings (Erk and Padó, 2009). The feature similarity component of the paraphrasing system that is currently based on WordNet could be replaced by such a model.

Another crucial problem that needs to be addressed is the coverage of the identification system. To enable high usability of the system it is necessary to perform high-recall processing. One way to improve the coverage is the creation of a larger, more diverse seed set. While it is hardly possible to describe the whole variety of metaphorical language, it is possible to compile a set representative of (1) all most common source–target domain mappings and (2) all types of syntactic constructions that exhibit metaphoricity. The existing metaphor resources, primarily Master Metaphor List, and examples from the linguistic literature about metaphor, could be a sensible starting point on a route to such a dataset. Having a diverse seed set should enable the identification system to attain a far broader coverage of the corpus than that reported in the current experiment.

Extrinsic task-based evaluation

Once the system is more robust, an extrinsic evaluation could be performed in order to verify its usefulness for NLP, e.g. by evaluating its impact on MT performance. MT constitutes a good platform for such an evaluation since it has been previously shown to benefit from an additional paraphrasing component (Callison-Burch et al., 2006). This

would imply integrating metaphor processing with a state-of-the-art statistical MT system. In cases where the MT system is not sufficiently confident about translation, it can use the metaphor system to verify whether metaphor is present and, when needed, paraphrase it with a literal paraphrase. The expectation is that the paraphrase will be easier to translate correctly (see the example in Chapter 1), which will improve the overall performance of MT.

Another interesting testbed for the metaphor system is educational applications. For example, the system could be deployed to automatically detect and assess creativity in students' essays. The quality of its output could be tested by comparing the scores it assigns to essays to human ratings of the essays. Along with assessment, the application of metaphor system to this data would also allow to investigate how students' learning of creative devices correlates with the acquisition of other linguistic competencies. In fact, educational and e-learning companies, such as Educational Testing Service (ETS) and Education First (EF), have already expressed an interest in such an application if the metaphor system to their data and we are currently exploring possibilities of collaboration.

Metaphor processing in NLP

Metaphor accounts for a whole range of processes in natural language: from how the language evolves by means of metaphorical sense extension (Copestake and Briscoe, 1995), to our use of metaphor as a persuasion device (Beigman Klebanov and Beigman, 2010). However, the popularisation of its study in the field of NLP encounters a number of barriers. The most significant of them is the lack of a unified task definition and a large publicly available metaphor corpus suited to the needs of NLP. The existence of the same task and a common dataset would enable researchers to directly compare their systems. In this thesis, I formulated proposals addressing both of these issues: I defined metaphor interpretation as a paraphrasing task and described a reproducible metaphor annotation scheme. However, the corpus I annotated is relatively small due to the limited annotation resources at my disposal and it is necessary to create a larger corpus. It should be balanced with respect to genre and exemplify all linguistic aspects of metaphor important for robust text processing.

Given a large metaphor corpus, I see the computational work on metaphor proceeding along the following lines:

- explicit or implicit acquisition of an extensive set of valid metaphorical associations from linguistic data via statistical pattern matching;
- metaphor recognition in the unseen unrestricted text using the knowledge of these associations;
- interpretation of the identified metaphorical expressions by deriving the closest literal paraphrase.

Ideas for new experiments

An interesting linguistic study would be that of how metaphor is expressed across languages and cultures. Patterns of metaphor translation found in parallel corpora could shed new light on human conceptualisation of metaphor, stripping off its culture-specific linguistic properties and revealing its associative mechanisms shared across cultures. For example, a study of such patterns might give an indication of the level at which source and target domains should be categorised (which I found to be a hard problem in monolingual analysis of conceptual metaphor, as discussed in Chapter 3). Parallel corpora provide an ideal environment for a study of both the cognitive nature of metaphor and its linguistic properties developed within individual cultures.

The focus of this thesis has been lexical metaphor. Lexical metaphor is by far the most frequent kind in natural language text, and thus the most relevant for NLP. However, humans often use extended metaphor to achieve a certain communicative goal. In the case of extended metaphor, one metaphorical theme is spread through the entire discourse. This type of metaphor is particularly characteristic of the genre of political debate, and consequently news articles, as well as literature. There have been attempts at modelling extended metaphor in political discourse using game-theoretic assumptions (Beigman Klebanov and Beigman, 2010). For instance, Beigman Klebanov and Beigman show that besides adding vividness to the shared imagery, metaphor can also guide the conversation. A metaphor suggested by one speaker is often taken up by his opponent. This highlights the importance of modelling metaphor at the discourse level, i.e. extended metaphor. Modelling extended metaphor would involve discovering complete and coherent scenarios, i.e. sequences of events, in linguistic data from one domain (source) that are metaphorically mapped to another domain (target). A comparison to a story taken from the source domain gives new significance to the events in the target domain. The conceptual structure of metaphorical mappings is likely to operate in a similar manner as with lexical metaphor. Scenarios could be represented and automatically extracted in the form of e.g. event chains (Chambers and Jurafsky, 2008) and narrative schemas (Chambers and Jurafsky, 2009). The presence of the same event chain in distinct domains may indicate the use of extended metaphor. This information may in turn ease the interpretation of the discourse fragment. In addition, modelling metaphor at a story level opens new routes to the comparison of metaphor in language to other analogy-based creative processes, such as poetry or visual art.

8.2.2 Logical metonymy

In this section, I suggest a number of extensions to the experiments on logical metonymy interpretation described here, as well as present some of my thoughts on its generation.

Limitations and extensions of the current interpretation experiments

The clustering experiment presented in Chapter 7 suffers from the problem of data sparseness. The fact that the synset feature vectors are constructed by means of overlap of the feature vectors of the individual verbs amplifies the problem. A possible solution would be to apply class-based smoothing to the feature vectors. In other words, one could back-off to the broad classes of nouns and represent the features of a verb in the form of its selectional preferences. To build a feature vector of a synset, one would then need to find common preferences of its verbs. Representing features in the form of semantic classes can also be viewed as a linguistically motivated way of dimensionality reduction for feature matrices, as the dimensions belonging to the same class are merged. While this is potentially a promising experiment, there is always a risk of introducing additional errors into the system due to incorrectly acquired selectional preference classes. I am currently exploring this extension with an undergraduate student at the Computer Laboratory working under my supervision.

Another limitation of the current approach is that the clustering algorithms presented here require the prior specification of the number of clusters. Since the number of classes of interpretations identified by humans varies across metonymic phrases, the next step would be to apply a clustering algorithm that determines the number of clusters automatically. This can be achieved by using Bayesian non-parametric models, e.g. Dirichlet Process Mixture Models, that have proved effective for verb clustering (Vlachos et al., 2009).

In addition to this, I intend to perform a more comprehensive evaluation. I will intrinsically test the system on a larger data set using the clustering evaluation techniques described in Chapter 7, as well as perform an extrinsic evaluation through integration with another NLP system that operates on word sense disambiguated text.

Generation

An interesting task for logical metonymy processing is its generation. Due to both the frequency with which the phenomenon occurs and the naturalness it gives to our speech, logical metonymy becomes an important problem for text generation. At first glance, this seems relatively straightforward: there is a limited set of verbs that have the property to coerce their nominal arguments to an eventive interpretation. However, semantically not all aspectual verbs can be combined with all nouns in all contexts and vice versa. This largely depends on the semantics of the noun. Consider the following examples:

(59) I *finished the dictionary* and will get paid soon!

(60) Thank you for the present! *I really *enjoyed this dictionary.*

The logical metonymy in (59) is perfectly well formed, but the one in (60) seems strange unless it is perceived as sarcastic. This is due to the fact that some processes are not characteristic to some concepts, e.g. dictionaries do not tend to be enjoyed, they are rather used for practical purposes. Such properties of concepts, and thus their restrictions on logical metonymy could, however, be induced from corpora using word co-occurrence information (Kelly et al., 2010). But this would not solve the problem entirely. Consider the sentence in (61).

(61) My goat eats anything. He really *enjoyed your dictionary.*

Here *the goat enjoys the dictionary* in a perfectly grammatical manner, which shows that the restrictions on the use of metonymic verbs can easily be overridden by context. The reader is able to derive the metonymic interpretation *eat*, and he or she knows that *eating* is normally *enjoyable*, thus for him or her this is a semantically valid sentence. This process is indicative of how language interpretation operates in general. Therefore, a study and computational account of the above issues would be an invaluable contribution to the way we conceive natural language semantics and model it within NLP.

8.2.3 The role of the thesis in modelling human creativity

> I think that metaphor really is a key to explaining thought and language. [..] Our powers of analogy allow us to apply ancient neural structures to newfound subject matter, to discover hidden laws and systems in nature, and not least, to amplify the expressive power of language itself. (Pinker, 2007)

In this quote Pinker (2007) suggests that the same cognitive mechanisms, namely those of metaphor, being generally characteristic to human thought, underlie both the production of figurative language, artistic creativity and scientific discovery. Veale and Hao (2008) were the first to empirically substantiate this idea. They successfully applied the analogy-based methods, which Hofstadter (1995) introduced to describe mathematical reasoning, for the explanation of metaphorical examples in language.

Whilst agreeing that metaphorical reasoning may operate in a similar manner across disciplines, I, however, approach the problem of its modelling from a different perspective. I aim to model metaphor processing mechanisms by primarily exploring their reflection in natural language, and then from the latter draw conclusions about the former. Studying linguistic properties of metaphorical processes is beneficial for a number of reasons. First of all, computational mechanisms of natural language semantics are well studied, as opposed to the semantics of symbolism in visual art, or abstract analogies in science. Secondly, there is a plethora of NLP tools available, which allow for statistical generalisations over linguistic data, that in turn lay the basis for computational modelling. Finally

and importantly, there are vast quantities of linguistic data that is balanced with respect to genre, style, subject area and function; under a number of assumptions such data is fully representative of human use of language and world knowledge.

I believe that, to a certain degree, the models presented in this thesis capture regular patterns of how creative thought operates using linguistic media. These include the model of metaphorical associations in the form of clusters of source and target concepts within the metaphor identification method and the models of context fitness and concept similarity employed within the metaphor interpretation method. Being entirely data-driven, these methods could be adapted to deal with other subject matters, e.g. to model interactions of symbols in visual art in a similar manner to the interactions of word meanings in linguistic context. In his book *Early Writings on Visual Language*, Cohn (2003) attempted to model the syntax of visual images in terms of the generative grammar of Chomsky (1957) that in its time transformed the face of contemporary linguistics. According to Cohn,

> the visual language exhibits markedly different *representations* of conceptual phenomena, while still governed under the same *properties* as symbolic linguistic structures. (Cohn, 2003)

If Cohn is concerned with the rules according to which a complete image can be formally built out of its constituents, my methods would deal with how the meaning of a piece of art emerges due to analogy-based creative processes, not the least important of which is metaphoricity. In order to experimentally verify this hypothesis, I plan to integrate the statistical metaphor identification method described in the thesis with the state-of-the-art image processing techniques, aiming to automatically reveal the non-literal use of symbols in visual art. Consider, for example, the painting "The Son of Man" by René Magritte, shown in Figure 8.1. Here, non-literalness is introduced by the presence of an apple hovering in front of a man's face. Thus, instead of expected imagery (a human face) in a given context (the body of the man), we see an apple, which is a violation of visual norms. Common combinations of objects and their visual contexts can be statistically learned from large quantities of visual data in an unsupervised way (e.g. using object co-occurrence and image clustering techniques). This would allow to automatically detect such violations and thus predict non-literal use of objects in an image, as well as the involved analogies.

Besides making our thoughts more vivid and filling our communication with richer imagery, metaphors also play an important structural role in our reasoning. Thus, my long term research goal is to build a complete and consistent computational intelligence model accounting for the way metaphors organise our conceptual system, in terms of which we think and act. Maybe one day such algorithms would enable computers to effectively carry out human-like communication, make important scientific discoveries, as well as understand, create and possibly even appreciate art.

Figure 8.1: René Magritte - The Son of Man (1964)

Bibliography

O. Abend and A. Rappoport. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010.

S. Abney. Partial parsing via finite-state cascades. In J. Carroll, editor, *Workshop on Robust Parsing*, pages 8–15, Prague, 1996.

S. Abney and M. Light. Hiding a Semantic Hierarchy in a Markov Model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, ACL*, pages 1–8, 1999.

R. Agerri, J.A. Barnden, M.G. Lee, and A.M. Wallington. Metaphor, inference and domain-independent mappings. In *Proceedings of RANLP-2007*, pages 17–23, Borovets, Bulgaria, 2007.

E. Agirre, L. Marquez, and R. Wicentowski, editors. *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, 2007. Association for Computational Linguistics.

A. Alonge and M. Castelli. Encoding information on metaphoric expressions in WordNet-like resources. In *Proceedings of the ACL 2003 Workshop on Lexicon and Figurative Language*, pages 10–17, 2003.

O. E. Andersen, J. Nioche, E. Briscoe, and J. Carroll. The BNC parsed with RASP4UIMA. In *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

J. D. Apresjan. Regular polysemy. *Linguistics*, 142:5–32, 1973.

T. Baldwin and S. N. Kim. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

J.A. Barnden and M.G. Lee. An artificial intelligence approach to metaphor understanding. *Theoria et Historia Scientiarum*, 6(1):399–412, 2002.

M. Baroni, S. Evert, and A. Lenci, editors. *Proceedings of ESSLLI Workshop on Distributional Lexical Semantics*, Hamburg, Germany, 2008.

L. Barque and F. Chaumartin. Regular Polysemy in WordNet. In *LDV Forum Band 21 (1)*, 2008.

R. Barzilay and L. Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 16–23, Edmonton, Canada, 2003.

R. Barzilay and K. R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Toulouse, France, 2001.

B. Beigman Klebanov and E. Beigman. A game-theoretic model of metaphorical bargaining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010. Association for Computational Linguistics.

S. Bergsma, D. Lin, and R. Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 59–68, Honolulu, Hawaii, 2008.

J. Birke and A. Sarkar. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *In Proceedings of EACL-06*, pages 329–336, 2006.

M. Black. *Models and Metaphors*. Cornell University Press, 1962.

B. Boguraev and E. Briscoe. Large lexicons for natural language processing: utilising the grammar coding system of the *Longman Dictionary of Contemporary English*. *Computational Linguistics*, 13(4):219–240, 1987.

G. Boleda Torrent and L. Alonso i Alemany. Clustering adjectives for class acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 9–16, Budapest, Hungary, 2003.

I. A. Bolshakov and A. Gelbukh. Synonymous paraphrasing using wordnet and internet. In *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004*, pages 312–323. Springer, 2004.

C. Brew and S. Schulte im Walde. Spectral clustering for German verbs. In *Proceedings of EMNLP*, 2002.

E. Briscoe, A. Copestake, and B. Boguraev. Enjoy the paper: lexical semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 42–47, Helsinki, 1990.

E. Briscoe, J. Carroll, and R. Watson. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80, 2006.

C. Brockmann and M. Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 27–34, Budapest, Hungary, 2003.

P. F. Brown, P. V. de Souza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December 1992.

E. Bulwer-Lytton. *Richelieu; Or the Conspiracy: A Play in Five Acts*. Saunders and Otley, London, 1839.

L. Burnard. *Reference Guide for the British National Corpus (XML Edition)*. 2007. URL `http://www.natcorp.ox.ac.uk/XMLedition/URG/`.

C. Callison-Burch, P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of NAACL*, 2006.

L. Cameron. *Metaphor in Educational Discourse*. Continuum, London, 2003.

D. De Cao and R. Basili. Combining distributional and paradigmatic information in a lexical substitution task. In *Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence*, 2009.

M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007*, pages 61–72, 2007.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL*, pages 269–270, 1999.

N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, 2008.

N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, 2009.

Y. S. Chan, H. T. Ng, and D. Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June 2007.

E. Charniak, D. Blaheta, N. Ge, K. Hall, J. Hale, and M. Johnson. *BLLIP 1987-89 WSJ Corpus Release 1.* Linguistic Data Consortium, Philadelphia, 2000.

J. Chen, D. Ji, C. Lim Tan, and Z. Niu. Unsupervised relation disambiguation using spectral clustering. In *Proceedings of COLING/ACL*, 2006.

N. Chomsky. *Syntactic Structures.* Mouton, The Hague, 1957.

N. Chomsky. *Aspects of the theory of syntax.* MIT Press, Cambridge, MA, 1965.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M. Ciaramita and M. Johnson. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of COLING 2000*, pages 187–193, Saarbrücken, Germany, 2000.

S Clark and J. R. Curran. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

S. Clark and D. Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.

S. Clark and D. Weir. An iterative approach to estimating frequencies over a semantic hierarchy. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 258–265, 1999.

N. Cohn. *Early Writings on Visual Language.* Emaki Productions, USA, 2003.

A. Copestake. The semi-generative lexicon: Limits on lexical productivity. In *In Proceedings of the First International Workshop on Generative Approaches to the Lexicon*, pages 41–49, 2001.

A. Copestake and T. Briscoe. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67, 1995.

D. Davidov, R. Reichart, and A. Rappoport. Superior and efficient fully unsupervised pattern-based concept acquisition using an unsupervised parser. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 2009.

A. Deignan. The grammar of linguistic metaphors. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin, 2006. Mouton de Gruyter.

B. J. Dorr. Large-scale dictionary construction for foreignlanguage tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322, 1998. ISSN 0922-6567.

M. Dras. Tree adjoining grammar and the reluctant paraphrasing of text. Technical report, PhD thesis, Macquarie University, Australia, 1999.

K. Erk. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Conference of the Association for Computational Linguistics (ACL 2007)*, 2007.

K. Erk. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 57–65. Association for Computational Linguistics, 2009.

K. Erk and D. McCarthy. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, 2009.

K. Erk and S. Padó. Paraphrase assessment in structured vector space: exploring parameters and datasets. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 57–65. Association for Computational Linguistics, 2009.

D. Fass. *Processing Metaphor and Metonymy*. Ablex, Stanford, CA, 1997.

D. Fass. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90, 1991.

D. Fass and Y. Wilks. Preference semantics, ill-formedness, and metaphor. *Computational Linguistics*, 9(3-4):178–187, 1983.

G. Fauconnier and M. Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2002.

J. Feldman and S. Narayanan. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392, 2004.

J. A. Feldman. *From Molecule to Metaphor: A Neural Theory of Language*. The MIT Press, 2006.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition, 1998.

C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003.

B. C. M Fung, K. Wang, and M. Ester. Large hierarchical document clustering using frequent itemsets. In *In Proceedings of SIAM International Conference on Data Mining 2003 (SDM 2003)*, 2003.

W. H. Fyfe. *Aristotle: Poetics*. Loeb Classical Library, Harvard University Press, Cambridge, MA, 1927.

M. Gedigian, J. Bryant, S. Narayanan, and B. Ciric. Catching metaphors. In *In Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York, 2006.

D. Gentner. Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7:155–170, 1983.

D. Gentner, B. F. Bowdle, P. Wolff, and C. Boronat. Metaphor is like analogy. In D. Gentner, K.J. Holyoak, and B.N. Kokinov, editors, *The analogical mind: Perspectives from cognitive science*, pages 199–253. MIT Press, 2001.

R. Gibbs. Literal meaning and psychological theory. *Cognitive Science*, 8:275–304, 1984.

D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288, September 2002.

A. Goatly. *The Language of Metaphors*. Routledge, London, 1997.

D. Godard and J. Jayez. Towards a proper treatment of coercion phenomena. In *Sixth Conference of the European Chapter of the ACL*, pages 168–177, Utrecht, 1993.

J. Grady. Foundations of meaning: primary metaphors and primary scenes. Technical report, PhD thesis, University of California at Berkeley, 1997.

D. Graff. North american news text corpus. *Linguistic Data Consortium*, 1995.

R. Grishman, L. Hirschman, and N. T. Nhan. Discovery procedures for sublanguage selectional patterns: initial experiments. *Computational Linguistics*, 12(3):205–215, July 1986.

P. Hanks and J. Pustejovsky. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82, 2005.

D. Harman. Overview of the fourth text retrieval conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference*, 1995.

R. E. Haskell. Cognitive science and the origin of lexical metaphor. *Theoria et Historia Scientiarum*, 6(1):291–331, 2002.

V. Hatzivassiloglou and K. R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *In Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182, 1993.

M. Hesse. *Models and Analogies in Science*. Notre Dame University Press, 1966.

D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, ACL '90, Pittsburgh, Pennsylvania, 1990.

D. Hindle and M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120, March 1993.

H. H. Hoang, S. N. Kim, and M. Kan. A re-examination of lexical association measures. In *Proceedings of the Workshop on Multiword Expressions*, pages 31–39. Association for Computational Linguistics, 2009.

D. Hofstadter. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. HarperCollins Publishers, 1995.

D. Hofstadter and M. Mitchell. The Copycat Project: A model of mental fluidity and analogy-making. In K.J. Holyoak and J. A. Barnden, editors, *Advances in Connectionist and Neural Computation Theory*, Ablex, New Jersey, 1994.

N. Ide and K. Suderman. The american national corpus first release. In *Proceedings of the Fourth Language Resources and Evaluation Conference*, pages 1681–1684, 2004.

E. Joanis, S. Stevenson, and D. James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367, 2008.

Y. Karov and S. Edelman. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59, 1998.

G. Karypis. Cluto: A clustering toolkit. Technical report, University of Minnesota, 2002.

J. J. Katz and J. A. Fodor. The structure of semantic theory. In J. J. Katz and J. A. Fodor, editors, *The structure of language*, pages 479–518, Prentice Hall, Englewood Cliffs, NJ, 1964.

D. Kauchak and R. Barzilay. Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, New York, New York, 2006.

C. Kelly, B. Devereux, and A. Korhonen. Acquiring human-like feature-based conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, Los Angeles, USA, 2010.

P. Kingsbury and M. Palmer. From TreeBank to PropBank. In *Proceedings of LREC-2002*, Gran Canaria, Canary Islands, Spain, 2002.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extensive classifications of English verbs. In *Proceedings of the 12th EURALEX International Congress*, 2006.

J. Klavans and M. Kan. Role of verbs in document analysis. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 680–686, Montreal, Quebec, Canada, 1998.

D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, 2003.

K. Knight and D. Marcu. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710, 2000.

P. Koivisto-Alanko and H. Tissari. Sense and sensibility: Rational thought versus emotion in metaphorical language. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin, 2006. Mouton de Gruyter.

S. Kok and C. Brockett. Hitting the right paraphrases in good time. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 145–153, Los Angeles, California, 2010.

A. Korhonen. *Subcategorization Acquisition*. PhD thesis, UK, 2002.

A. Korhonen, Y. Krymolowski, and Z. Marx. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL 2003*, Sapporo,Japan, 2003.

A. Korhonen, Y. Krymolowski, and T. Briscoe. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC 2006*, 2006.

R. Kozlowski, K. F. McCoy, and K. Vijay-Shanker. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 1–8, Sapporo, Japan, 2003.

K. Krippendorff. *Content Analysis*. SAGE Publications, Beverly Hills, CA, 1980.

S. Krishnakumaran and X. Zhu. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY, 2007.

S. Kurohashi. SENSEVAL-2 Japanese translation task. In *Proceedings of the SENSEVAL-2 workshop*, pages 37–44, 2001.

G. Lakoff. What is metaphor. In J. A. Barnden and K. J. Holoyak, editors, *Advances in Connectionist and Neural Computation Theory: Analogy, Metaphor and Reminding*, Norwood, NJ, 1994. Ablex.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago, 1980.

G. Lakoff, J. Espenson, and A. Schwartz. The master metaphor list. Technical report, University of California at Berkeley, 1991.

J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

M. Lapata. The acquisition and modeling of lexical knowledge: A corpus-based investigation of systematic polysemy. Technical report, PhD thesis, University of Edinburgh, 2001.

M. Lapata and C. Brew. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(2):45–73, 2004.

M. Lapata and A. Lascarides. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315, 2003.

A. Lascarides and A. Copestake. The pragmatics of word meaning. In *Journal of Linguistics*, pages 387–414, 1995.

G. Leech. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13, 1992.

G. Leech, R. Garside, and M. Bryant. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, COLING '94, pages 622–628, 1994.

D. Lenat, G. Miller, and T. Yokoi. CYC, WordNet, and EDR: critiques and responses. *Commun. ACM*, 38(11):45–48, 1995.

B. Levin. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, 1993.

H. Li and N. Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics*, 24(2):217–244, 1998.

D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, 1998.

D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360, 2001.

B. Lönneker. Lexical databases as resources for linguistic creativity: Focus on metaphor. In *Proceedings of the LREC 2004 Workshop on Language Resources for Linguistic Creativity*, pages 9–16, Lisbon, Portugal, 2004.

B. Lönneker and C. Eilts. A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information. In *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 157–162, Brno, Czech Republic, 2004.

B. Lönneker-Rodman. The hamburg metaphor database project: issues in resource creation. *Language Resources and Evaluation*, 42(3):293–318, 2008.

K. Markert and M. Nissim. Metonymy resolution as a classification task. In *Proceedings of the conference on Empirical methods in natural language processing*, EMNLP '02, pages 204–213, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

K. Markert and M. Nissim. Metonymic proper names: A corpus-based account. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin, 2006. Mouton de Gruyter.

J. H. Martin. Representing regularities in the metaphoric lexicon. In *Proceedings of the 12th conference on Computational linguistics*, pages 396–401, 1988.

J. H. Martin. Metabank: A knowledge-base of metaphoric language conventions. *Computational Intelligence*, 10:134–149, 1994.

J. H. Martin. *A Computational Model of Metaphor Interpretation.* Academic Press Professional, Inc., San Diego, CA, USA, 1990.

J. H. Martin. A corpus-based analysis of context effects on metaphor comprehension. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin, 2006. Mouton de Gruyter.

Z. J. Mason. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44, 2004.

D. McCarthy. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8*, WSD '02, pages 109–115, 2002.

D. McCarthy and J. Carroll. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654, 2003.

D. McCarthy and R. Navigli. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159, 2009.

D. McCarthy and R. Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, 2007.

D. Mccarthy, S. Venkatapathy, and A. K. Joshi. Detecting compositionality of verb-object combinations using selectional preferences. In *In Proceedings of the 200 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379, 2007.

D. McCarthy, B. Keller, and R. Navigli. Getting synonym candidates from raw data in the english lexical substitution task. In *Proceedings of the 14th EURALEX International Congress*, Leeuwarden, The Netherlands, 2010.

K. R. McKeown. Paraphrasing using given and new information in a question-answer system. In *Proceedings of the 17th annual meeting on Association for Computational Linguistics*, ACL '79, pages 67–72, La Jolla, California, 1979.

M. Meila and J. Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001.

M. Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

M. Meteer and V. Shaked. Strategies for effective paraphrasing. In *Proceedings of the 12th conference on Computational linguistics - Volume 2*, COLING '88, pages 431–436, Budapest, Hungry, 1988.

G. L. Murphy. On metaphoric representation. *Cognition*, 60:173–204, 1996.

S. Narayanan. Knowledge-based Action Representations for Metaphor and Aspect (KARMA). Technical report, PhD thesis, University of California at Berkeley, 1997.

S. Narayanan. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of AAAI 99)*, pages 121–128, Orlando, Florida, 1999.

M. Nissim and K. Markert. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 56–63, Sapporo, Japan, 2003.

G. Nunberg. The pragmatics of reference. Technical report, PhD thesis, Indiana University, 1978.

G. Nunberg. Poetic and prosaic metaphors. In *Proceedings of the 1987 workshop on Theoretical issues in natural language processing*, pages 198–201, 1987.

D. Ó Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010.

G. Orwell. Politics and the English Language. *Horizon*, 1946.

P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.

P. Pantel, R. Bhagat, T. Chklovski, and E. Hovy. Isp: Learning inferential selectional preferences. In *In Proceedings of NAACL 2007*, 2007.

M. Pasca and S. Harabagiu. The informative role of WordNet in open-domain question answering. In *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*, pages 138–143, Pittsburgh, PA, 2001.

Y. Peirsman. Example-based metonymy recognition for proper nouns. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06, pages 71–78, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Y. Peirsman and S. Padó. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California, 2010.

F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of ACL-93*, pages 183–190, Morristown, NJ, USA, 1993.

W. Peters and I. Peters. Lexicalised systematic polysemy in wordnet. In *Proceedings of LREC 2000*, Athens, 2000.

S. Pinker. *The Stuff of Thought: Language as a Window into Human Nature*. Viking Adult, USA, September 2007.

R. Power and D. Scott. Automatic generation of large-scale paraphrases. In *In Proceedings of IWP*, pages 73–79, 2005.

Pragglejaz Group. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39, 2007.

J. Preiss. Probabilistic word sense disambiguation analysis and techniques for combining knowledge sources. Technical report, Computer Laboratory, University of Cambridge, 2006.

J. Preiss, T. Briscoe, and A. Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL-2007*, volume 45, page 912, 2007.

J. Preiss, A. Coonce, and B. Baker. HMMs, GRs, and n-grams as lexical substitution techniques: are they portable to other languages? In *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, MCTLLL '09, pages 21–27, Borovets, Bulgaria, 2009.

D. Pucci, M. Baroni, F. Cutugno, and A. Lenci. Unsupervised lexical substitution with a word space model. In *Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence*, 2009.

J. Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, MA, 1995.

J. Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4), 1991.

J. Pustejovsky and P. Bouillon. Logical polysemy and aspectual coercion. *Journal of Semantics*, 12:133–162, 1995.

C. Quirk, C. Brockett, and W. Dolan. Monolingual machine translation for paraphrase generation. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, 2004.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

A. Reining and B. Lönneker-Rodman. Corpus-driven metaphor harvesting. In *Proceedings of the HLT/NAACL-07 Workshop on Computational Approaches to Figurative Language*, pages 5–12, Rochester, New York, 2007.

J. Reisinger and R. Mooney. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1173–1182, Cambridge, Massachusetts, 2010.

P. Resnik. Selectional preference and sense disambiguation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington, D.C., 1997.

P. Resnik. *Selection and Information: A Class-based Approach to Lexical Relationships.* PhD thesis, Philadelphia, PA, USA, 1993.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL 99*, pages 104–111, 1999.

T.G. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus Volume 1-from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 29–31, 2002.

S. Rudolph and E. Giesbrecht. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916, Uppsala, Sweden, 2010. Association for Computational Linguistics.

S. Schulte im Walde. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 747–753, 2000.

S. Schulte im Walde. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194, 2006.

S. Schulte im Walde. Comparing computational models of selectional preferences - second-order co-occurrence vs. latent semantic clusters. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

H. Schutze and J. O. Pedersen. Information retrieval based on word senses. In *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

S. Sekine, K. Inui, I. Dagan, B. Dolan, D. Giampiccolo, and B. Magnini, editors. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague, June 2007.

C. R. Shalizi. Analogy and Metaphor, 2003. URL `http://www.cscs.umich.edu/~crshalizi/notabene/analogy.html`.

Y. Shinyama and S. Sekine. Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 65–71, Sapporo, Japan, 2003.

S. Siegel and N. J. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company, New York, USA, 1988.

G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, and T. Pasma. *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins, Amsterdam/Philadelphia, 2010.

G. Stern. *Meaning and Change of Meaning*. Wettergren & Kerbers Förlag, Göteburg, 1931.

C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 159–166, Toronto, Canada, 2003.

L. Sun and A. Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP 2009*, pages 638–647, Singapore, August 2009.

A. Toral. The lexical substitution task at evalita 2009. In *Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence*, 2009.

R. Tourangeau and R. Sternberg. Understanding and appreciating metaphors. *Cognition*, 11:203–244, 1982.

M. Utiyama, M. Masaki, and I. Hitoshi. A statistical approach to the processing of metonymy. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrucken, Germany, 2000.

C. J. van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, 1979.

T. Veale and Y. Hao. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of COLING 2008*, pages 945–952, Manchester, UK, 2008.

Z. Vendler. *Adjectives and Nominalizations*. Mouton, The Hague, 1968.

C. M. Verspoor. Conventionality-governed logical metonymy. In *Proceedings of the Second International Workshop on Computational Semantics*, pages 300–312, Tilburg, 1997.

A. Vlachos, A. Korhonen, and Z. Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *EACL workshop on GEometrical Models of Natural Language Semantics*, Athens, 2009.

E. M. Voorhees. Using WordNet for text retrieval. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 285–303. MIT Press, 1998.

E.M. Voorhees and D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *Proceedings of the Fifth Text Retrieval Conference*, 1996.

W. Wagner, H. Schmid, and S. Schulte Im Walde. Verb sense disambiguation using a predicate-argument clustering model. In *Proceedings of CogSci Workshop on Semantic Space Models (DISCO)*, Amsterdam, Holland, 2009.

A. M. Wallington, J. A. Barnden, P. Buchlovsky, L. Fellows, and S. R. Glasbey. Metaphor Annotation: A Systematic Study. Technical report, School of Computer Science, The University of Birmingham, 2003.

Y. Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74, 1975.

Y. Wilks. Making preferences more active. *Artificial Intelligence*, 11(3):197–223, 1978.

K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. *Advances in Neural Information Processing Systems*, 18, 2006.

F. M. Zanzotto, M. Pennacchiotti, and M. T. Pazienza. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 849–856, Sydney, Australia, 2006.

B. Zapirain, E. Agirre, and L. Màrquez. Generalizing over lexical features: selectional preferences for semantic role classification. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 73–76, 2009.

B. Zapirain, E. Agirre, L. Màrquez, and M. Surdeanu. Improving semantic role classification with selectional preferences. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 373–376, Los Angeles, California, 2010.

S. Zhao, H. Wang, T. Liu, and S. Li. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *In Proceedings of ACL-08:HLT*, pages 780–788, 2008.

S. Zhao, X. Lan, T. Liu, and S. Li. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 834–842, Suntec, Singapore, 2009.

Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, University of Minnesota, 2001.

L. Zhou, C. Lin, D. S. Munteanu, and E. Hovy. PARAEVAL: Using paraphrases to evaluate summaries automatically. In *Proceedings of HLT-NAACL*, pages 447–454, 2006.

C. Zong, Y. Zhang, and K. Yamamoto. Approach to spoken chinese paraphrasing based on feature extraction. In *Proceedings of NLPRS*, pages 551–556, 2001.

Appendix A

Summary of human experiments

Chapter	Judge	# Metaphors	Type	Aim	Agreement (κ)	Conclusion
3	3 subjects	39–53	Annotation	Linguistic metaphor	0.64	Can be reliably annotated using the procedure
3	3 subjects	39–53	Annotation	Conceptual metaphor	0.57	Challenging task
3	ES	241	Annotation	Linguistic metaphor	0.62	Creation of the corpus
3	ES	241	Annotation	Conceptual metaphor	0.58	Creation of the corpus
4	5 subjects	78	Judgement	Evaluation of identification system	0.63	Identification system operates with high precision
4	ES	222	Annotation	Recall study	–	Identification system currently operates with low recall
5	7 subjects	49	Judgement	Evaluation of paraphrasing output	0.62	Literal paraphrasing with high precision at rank (1)
5	5 subjects	49	Elicitation	Creation of a paraphrasing gold standard	–	High recall of paraphrases
5	ES	49	Judgement	WSD of paraphrases	–	100% accurate WSD
5	3 subjects	35	Judgement	Evaluation of integrated system	0.63; 0.53	Integrated system works well, although can be improved (see section 5.4.2)
5	ES	200	Judgement	Evaluation of integrated system	0.59; 0.54	Integrated system works well, although can be improved (see section 5.4.2)
7	8 subjects	6	Judgement	WSD of metonymic interpretations	0.53	System disambiguates and re-ranks senses well
7	8 subjects	6	Clustering	Clustering of metonymic interpretations	–	Humans agree on the task; system clusters metonymic interpretations well
7	ES	10	Judgement	WSD of metonymic interpretations	0.53	System disambiguates and re-ranks senses well
7	ES	10	Clustering	Clustering of metonymic interpretations	–	System clusters metonymic interpretations with an F-measure of 0.64

Table A.1: Summary of human experiments

Appendix B

Metaphor annotation guidelines

Task The focus of our study is on single-word metaphors expressed by a verb. You will be given 2 texts. Please annotate all the verbs which are underlined in the texts.

- **Step 1**: classify the verbs in the text into two categories: **metaphorical** or **literal**.

  Consider an example "How can I *kill* a process?", where the verb *kill* is used metaphorically. Metaphors arise when one concept is viewed in terms of the properties of another. In our example the *computational process* is viewed as something *alive* and, therefore, its forced termination is associated with the act of killing. Therefore, *kill* should be tagged as metaphorical.

- **Step 2**: identify the **interconceptual mapping** for each expression you tag as metaphorical.

  The association between the concepts of COMPUTATIONAL PROCESS and LIVING BEING is called an **interconceptual mapping**, whereby the concepts are the *target* and the *source* concepts respectively. We compiled lists of categories that are generally frequent source and target concepts. The source categories are given in red and the target categories in blue. Select the categories from the lists that you think describe the source and target concepts best or suggest your own category if the list does not include your judgement.

Procedure To discriminate between the verbs used **metaphorically** and **literally** use the following procedure:

1. For each verb that is underlined establish its meaning in context and try to imagine a more **basic** meaning of this verb on other contexts. Basic meanings are normally:

   - more concrete;
   - related to bodily action;

- more precise (as opposed to vague);
- historically older;

2. If you can establish the basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically. Try to identify a mapping between the source domain (where the basic meaning comes from) and the target domain (the concepts forming the context of the verb in front of you). Record the mapping. If you fail to identify a mapping, reconsider whether the sense is metaphorical in this context.

Example This is how the annotation procedure works on the following example sentence:

(62) If he asked her to post a letter or buy some razor blades from the chemist, she was transported with pleasure.

- The first 3 verbs are used in their basic sense, i.e. literally (*ask* in the context of "a person asking another person a question or a favour"; *post* in the context of "a person posting/sending a letter by post"; *buy* in the sense of "making a purchase"). So they are tagged as literal.

- The verb *transport*, however, in its basic sense is used in the context of "the goods being transported/carried somewhere by a vehicle". The context in this sentense involves "a person being transported by a feeling", which contrasts the basic sense in that the agent of *transporting* is an EMOTION (the target concept) as opposed to a VEHICLE (the source concept). Thus, we can infer that the use of *transport* in this sentence is **metaphorical** and the associated interconceptual mapping is EMOTIONS – VEHICLES.

Below is the exercise that the subjects were asked to do prior to annotation.

Metaphor annotation example. Text ID: HH9.

Please annotate the underlined verbs that are used metaphorically by highlighting them in colour. Then select the source and target categories for them from the lists and type them in brackets next to the verb.

ANNOTATED SENTENCE:

If he asked ( ) her to post ( ) a letter or buy ( ) some razor blades from the chemist, she was transported (FEELINGS, EMOTIONS — VEHICLES) with pleasure.

PLEASE ANNOTATE:

Just to see ( ) his silk dressing-gown hanging ( ) on the back of the door or his watch lying ( ) on the edge of the basin, stirred ( ) an unfathomable excitement in her.

Appendix C

An extract from the metaphor corpus

Metaphor annotation. Text ID: ACA

‘ Thou know'st 't is common, all that lives must die, Passing through nature to eternity. ’ Hamlet

When we are going ( ) on a journey to a strange country, our state of mind and the nature of our preparations are determined ( ) by what we think ( ) we shall find ( ) there and, in particular, by whether we have friends and relations living ( ) there. It can not be otherwise with the approach of death ; whether we go ( ) on pilgrimage with Raleigh or put out ( ) to sea with Tennyson, the metaphor of travel is one with which the poets have made us familiar. This book, therefore, is not only about how men and women in the nineteenth century faced (MOMENT IN TIME – LIVING BEING) the hour of death ; it is also about what they expected ( ) to encounter ( ) on the further shore. It is a subject that historians and sociologists, stressing ( ) the activities of men's lives, have been inclined ( ) to overlook (IDEA – PHYSICAL OBJECT). Scant attention has been given to the way in which man's attitude to death feeds back (FEELINGS, ATTITUDES – SUBSTANCES; LIFE – CONTAINER) into his life and so exerts ( ) an influence upon society. For this is a link that can not be uncoupled ( ); ‘in my end is my beginning’. How we live ( ) is inescapably linked (IDEA – PHYSICAL OBJECT) to what we think ( ) about our origin and our destiny, even if we only discuss ( ) these mysteries late at night in the company of a few friends. Those hardy souls in the present century who ignore ( ) the mysteries and regard ( ) themselves as random atoms, moving ( ) purposelessly in a world of blind chance, must necessarily behave ( ) differently from those who, like so many in the nineteenth century, believed ( ) that they inhabited ( ) an ordered world in which they had moral duties to perform ( ), even if these were obscurely glimpsed (DUTY – PHYSICAL OBJECT) and seldom accomplished ( ). The contrasting attitudes become ( ) more vivid if we try to envisage ( ) men and women of the two epochs as they lie ( ) (as all must) on their death-beds. There is probably no such thing as a typical death-bed in any century; but it is not difficult to take two representative scenarios, separated ( ) from one another by about one hundred

years. In the first, the friends and relatives of the dying ( ) man, frequently accompanied ( ) by young children, are standing ( ) within earshot of his bed, listening ( ) either to his last words or, if he has already fallen (CALM -- DOWN) silent, to the prayers of the pastor, who is celebrating ( ) the triumph of Christian immortality over death and the devil. The beneficiary of their prayers dies ( ) in his own home ; his relations, or servants, lay out ( ) the body and watch ( ) beside it until it is placed ( ) in the coffin and taken for interment in accordance with the rites of the Church. An account of the last hours and words is written ( ) for the edification of those unable to be present. In our day there is little of the earlier resignation in the face of death; we resist (DEATH -- ENEMY) it to the limits of medical technology, and our obituaries speak ( ) of our 'gallant struggle'. Death, for us, is the sorry end of the human story, not the mysterious prelude to a new one. A high proportion of us die ( ) not at home, but in hospitals, clinics and special institutions for the terminally ill. There our friends and relatives sometimes come ( ) to visit ( ) us and speak ( ) resolutely of trivial matters ; children seldom come ( ), because that might seem ( ) ' morbid ', especially if the illness is known ( ) to be terminal. When the end is coming (MOMENT IN TIME – LIVING BEING), screens are placed ( ) round the bed ; in any case there are unlikely to be any 'last words', because most of us die ( ) under some form of sedation. The undertaker then takes over ( ), makes up ( ) our faces and carries ( ) the coffin to the crematorium. There a hybrid service takes place, to which even a confirmed atheist could hardly object ( ). If there is an address, it is often given by a layman, who commends ( ) what we did in life and skims discreetly over (IDEAS -- LIQUID) the question of survival, if any. As the coffin slides ( ) into the furnace, we try to restrict ( ) our emotional involvement — sometimes at considerable psychological cost. The contrast is striking, and can not be dismissed ( ) as irrelevant to the social and other problems that we confront (PROBLEM -- ENEMY) in the last decade of our century. It is not the purpose of this book to study ( ) these problems, but rather the changes that in an earlier century began to alter ( ) our attitude towards death. It is not enough to assume ( ) that all that has happened ( ) is that we no longer believe ( ) in hell, and that mutes, carrying ( ) black ostrich

plumes, are out of favour. The changes have been more fundamental and some of them may have affected ( ) us in ways that we do not immediately recognise ( ). All centuries, of course, have been centuries of change ; but few would deny that in the nineteenth century change was greatly accelerated (CHANGE – MOTION); that much was apparent to the more perceptive of those living ( ) at that time. Some felt ( ) that they were hurrying (TIME -- PATH) into an epoch of unprecedented enlightenment, in which better education and beneficent technology would ensure ( ) wealth and leisure for all. This was Herbert Spencer's view, namely that an upward evolutionary process was inherent in the human condition. To others, including Tennyson and Arnold, it seemed as if ‘ the ringing grooves of change ’ were carrying (CHANGE – VEHICLE (MOTION)) them at break-neck speed into a future full of uncertainty and alarm. Browning was more hopeful, but he, too, was impressed ( ) by the transiency of the world, flashing (WORLD -- PICTURE) past the carriage windows : ‘ *Must the rose sigh ‘ Pluck — I perish ! ’ Must the eve weep ‘ Gaze — I fade !* ’ ’ The Impressionist painters caught (INFECTION – PHYSICAL OBJECT) the contagion, and the new race of photographers tried to seize (MOMENT IN TIME – PHYSICAL OBJECT) the fleeting moment and make it stay (MOMENT IN TIME – LIVING BEING). Cultures and historical periods differ ( ) greatly in their concepts of time and the continuity of life. We live ( ) in a century imprinted (LIFE – STORY) on the present, which regards ( ) the past as little more than the springboard from which we were launched (PAST – PLATFORM) on our way. Ours, for better or for worse, is the century of youth. Earlier centuries, in contrast, had an appreciation of the past that embodied ( ) more than nostalgia or antiquarian interest. Age and experience were valued ( ) in the belief that length of days had provided ( ) some guidance on how to live ( ) and what to expect ( ) in the life to come ( ). For the life on earth of each individual was not a finite entity, complete in itself, but a transition to another mode of existence ; the gateway to that unknown land was death — Mors Janna Vitae , as the memorial tablets had it. This belief was fostered ( ) by the churches, the floors and walls of which were incised ( ) with the records of those who had gone ( ) before, but it was expressed ( ) positively in the family. Families cherished ( )

their forbears, whether these had lived ( ) in humble cottages or in manor houses. Sons aspired ( ) to follow (CAREER, LIFE – PATH, JOURNEY) in their fathers' trades or professions. The landed gentry planted (DEVELOPMENT – GROWTH (BIOLOGICAL)) for their grandchildren avenues of hardwood that they themselves would never see ( ). In the nineteenth century this leisurely view of the pageant of time began to speed up ( ). People became ( ) more mobile, both physically and socially ; men wished ( ) to rise (PROGRESS – GROWTH (BIOLOGICAL), RISE) in the world. Young men with feet on the ladder, provided ( ) by the Industrial Revolution, looked ( ) askance at the old-fashioned ways of their fathers. Growing ( ) more acquisitive in the present, they prepared ( ) to disown (ACCEPTANCE – OWNERSHIP) the past ; the future was to be different, both for themselves and their children, and they had to run (TIME – PATH) to catch up (TIME – PATH) with it. Improving ( ) life expectancy gave them every hope of doing so, especially if they belonged ( ) to the rising (PROGRESS – GROWTH (BIOLOGICAL), RISE) middle-class. For the middle-class was both the agent and product of these changes. The rise of the middle-class was not, on the whole, predicated ( ) on an aspiration to join ( ) the aristocracy, whose way of life, especially during the Regency, met (FEELINGS, ATTITUDES – LIVING BEINGS) with a good deal of disapprobation ; but it was determined ( ) by the resolute intention of the new men to distance (DIFFERENCE – DISTANCE) themselves in every possible way from the working-class, out of which so many of them had raised (PROGRESS – GROWTH (BIOLOGICAL), RISE) themselves. Their rise during the Industrial Revolution was expressed ( ) in capital accumulation ; the status of the aristocracy still derived ( ) from birth and ownership of land. The new men were not aping ( ) the landed gentry; they were basing (ACCOMPLISHMENT – PLATFORM, BASIS) their careers upon the infrastructure provided ( ) by urban Britain. There was no coherent ideology embracing (BELIEFS – COVER) the entire middle-class, but there were two ideologies that subsumed ( ) its more active sectors. The older one was that of the Evangelicals and Dissenters, of whom more will be written ( ) in chapter three. The newer ideology was that of the followers of Jeremy Bentham (1745–1832), the so-called Utilitarians ; it was far from sharing (VIEWS,

IDEAS – PROPERTY) a common world view with the Evangelicals, but there were certain social issues, such as abolition of slavery, on which concerted action was possible. Macaulay was one of those who had a foot in both camps. Before we consider ( ) how these ideologies affected ( ) attitudes to death, we must glance (WORLD – PICTURE) at the social and economic changes that provided ( ) the context within which they operated (SOCIAL, ECONOMIC, POLITICAL SYSTEM – MECHANISM, MACHINE). In 1801 the population of England and Wales was under nine million, of whom the great majority lived ( ) in rural communities outside London. Seventy years later the population had risen (PROGRESS – GROWTH (BIOLOGICAL), RISE) to 22.7 million, of whom 62% lived ( ) in towns and cities.

Appendix D

Evaluation of metaphor identification

Evaluation of Automatic Metaphor Identification

Task:

The following sentences contain expressions that the system tagged as metaphorical (in yellow). In those expressions **verbs** are used metaphorically (or not). Please evaluate the tagging of the system by ticking the box next to "Metaphorical" if you think the expression in yellow is a metaphor, and "Literal" if not.

We suggest that you rely on your own intuition on metaphor in the first place. However, you could also use the following procedure for some guidance:
1.For each underlined verb in yellow establish its meaning in context and try to imagine a more basic meaning of this verb on other contexts. Basic meanings normally are:

(1) more concrete;
(2) related to bodily action;
(3) more precise (as opposed to vague);
(4) historically older.

2.If you can establish the basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically.

****

Examples:

ACH 1081 His 'fascist' ideas had first been **shaped** by the First World War, which he felt Britain should not have entered.

Metaphorical (X)
Literal ( )

CEK 15 Immediately some commentators claimed that she and Prince Charles had succeeded in **mending** their marriage.

Metaphorical (X)
Literal ( )

CGH 1281 Often the oval **shaped** body with its waving flagella can be seen quite clearly darting around among the intestinal debris obtained from your fish.

Metaphorical ( )
Literal (X)

****

Please evaluate the expressions below:

CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **swallow** his anger and play tennis.

Metaphorical (X)
Literal ( )

AD2 631 This is not to say that Paisley was dictatorial and simply **imposed** his will on other activists.

Metaphorical ( )
Literal (X)

AND 322 It's almost as if some teachers **hold** the belief that the best parents are those that are docile and ignorant about the school, leaving the professionals to get on with the job.

Metaphorical (X)
Literal ( )

K54 2685 And it **approved** the recommendation by Darlington Council not to have special exemptions for disabled drivers.

Metaphorical ( )
Literal (X)

G0G 1306 It would be natural to assume that this attempt to create a rift between Offa and the papacy occurred before the visit of the legates in 786 and that the visit was part of a process of reconciliation, but this is not wholly justified for Hadrian's letter could date to the late rather than the mid-780s, and **reflect** hostility to one or more of a number of Offa's actions.

Metaphorical (X)
Literal ( )

CGY 735 All that remained for theory was to explore the details of determination while avoiding the tendency of bourgeois ideology to **obscure** determination by the material base.

Metaphorical (X)
Literal ( )

AD9 **3205** He tried to **disguise** the anxiety he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.

Metaphorical	(X)
Literal	( )

A1F **24** Moreover, Mr Kinnock **brushed aside** the suggestion that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.

Metaphorical	(X)
Literal	( )

FB1 **701** Moreover, radicals with reservations about the socialist credentials of the USSR **confronted** the dilemma that every word of criticism aligned them with the 'reactionary' views dominant in the West

Metaphorical	(X)
Literal	( )

AMA **349** We will **halt** the reduction in NHS services for long-term care and community health services which support elderly and disabled patients at home.

Metaphorical	( )
Literal	(X)

B71 **852** Craig Packer and Anne Pusey of the University of Chicago have continued to **follow** the life and loves of these Tanzanian lions.

Metaphorical	(X)
Literal	( )

CK1 **92** To understand this we have to see that Mill is **basing** his answer to one question on his answer to another.

Metaphorical	( )
Literal	(X)

JXS **1794** She tried to **cast** the thought from her, but it stayed stubbornly with her as he took another small step towards her.

Metaphorical (X)
Literal ( )

ADK **634** **Catch** their interest and **spark** their enthusiasm so that they begin to see the product's potential.

Metaphorical (X)
Literal ( )

K2W **1771** The committee heard today that gangs regularly **hurled** abusive comments at local people, making an unacceptable level of noise and leaving litter behind them.

Metaphorical (X)
Literal ( )

H9Y **2173** [..] my instinct was to **convey** the spirit and feeling of Grenfell, without resorting to mimicry.

Metaphorical (X)
Literal ( )

EE9 **965** [..] not to be 'at liberty to make far-reaching proposals that the Unionist party would **decline** to support'.

Metaphorical ( )
Literal (X)

HGF **2874** The teacher came in like a colossus and the class **shrank** into a shivering line.

Metaphorical (X)
Literal ( )

B1W **265** The above example **illustrates** the impact of the credit creation multiplier.

Metaphorical (X)
Literal ( )

EF1 **1272** He was panicking; and panic invariably **leads** to mistakes.

Metaphorical (X)
Literal ( )

BN4 **126** It has also **adopted** nuclear power as a solution to the greenhouse effect.

Metaphorical (X)
Literal ( )

G1G **1105** This, it is argued, **hinders** rational thought and leads to the avoidance and downgrading of non-military solutions to world problems.

Metaphorical ( )
Literal (X)

A6L **1688** I've also **adopted** the philosophy that I must develop somebody to do my job better than I have done it.

Metaphorical (X)
Literal ( )

J7C **785** Lord Denning suggested an alternative approach: the court should try to **construe** a contract out of all the communications passing between the parties.

Metaphorical ( )
Literal (X)

B1F **1429** It is as if we put on our telescopes of the mind and **magnify** infinitely our habitual thoughts.

Metaphorical (X)
Literal ( )

G04 **3020** Why do we **keep** the truth from them?

Metaphorical (X)
Literal ( )

CLN 500 Print propaganda aimed then to **stimulate** 'agitation' in the hope of shaping or sustaining antislavery public policies 'by a loud, strong and solemn expression of the public opinion'and establish direct 'influence'with individuals who exercised power and authority.

Metaphorical (X)
Literal ( )

A8D 29 Dai Qing's name was singled out as one of the 'tiny handful of people' who had 'colluded with foreign forces, ganged up among themselves at home and made ideological, public opinion and organisational preparations for years to **stir up** turmoil in China, overthrow the leadership of the Communist Party and subvert the socialist people's republic.'

Metaphorical (X)
Literal ( )

ABA 925 In consequence, the project became a Whitehall 'camel'(a vehicle designed by a committee) instead of having a chief designer able to **impose** his personal decisiveness on the final product.

Metaphorical ( )
Literal (X)

CM0 1225 This factor **determines** the answer to our next question, that of the frequency and extent to which companies use search firms.

Metaphorical ( )
Literal (X)

HA2 2654 Theodora **formed** her resolution.

Metaphorical (X)
Literal ( )

AJ6 296 But the committee has denied that it is attempting to **influence** the outcome of the election.

Metaphorical ( )
Literal (X)

B0R **463** Coleridge and Sara **fixed** their wedding for early October, and in the meantime began looking for a home.

Metaphorical (X)
Literal ( )

J1K **167** If O'Keeffe had been able to **ignore** the criticism of her art that Stieglitz's ideas initiated, she might not have felt compelled to limit her experimentation with abstraction in the early 1920s.

Metaphorical ( )
Literal (X)

G17 **710** He paused, and then gestured as if to **dismiss** the subject.

Metaphorical (X)
Literal ( )

H84 **2637** 'I do not believe that you will simply **drop** this case.'

Metaphorical (X)
Literal ( )

EEC **1362** In it, Younger stressed the need for additional alternatives to custodial sentences, which had been implicit in the decision to ask the Council to **undertake** the enquiry.

Metaphorical ( )
Literal (X)

J7F **77** An economist would **frame** this question in terms of a cost-benefit analysis: the maximisation of returns for the minimum amount of effort injected.

Metaphorical (X)
Literal ( )

GVF **455** Starting from the top, a typical kung fu attack would **concentrate** on the eyes, nose, throat, larynx, sternum, heart, solar plexus and groin.

Metaphorical (X)
Literal ( )

Appendix E

Evaluation of metaphor paraphrasing

Evaluation of Automatic Metaphor Paraphrasing – 1

Task:

You are given a list of sentences containing a metaphorical expression (in yellow) where **verbs** (**in bold and underlined**) are used metaphorically. These are followed by a number of paraphrases the system produced for the verbs. Please evaluate the paraphrases by typing YES in the box next to them if the paraphrase means the same thing as the metaphorical expression and is used literally in the given context, and NO if not. Note that in the list of paraphrases for a metaphorical expression there can be more than one (or none) that are correct.

If you can think of a better paraphrase (a verb) than the ones suggested by the system, please add it in the slot "MINE" (if you can not come up with a **better** paraphrase, don't worry, just leave the slot empty!)

****

Example:

Soviet socialism was **held** to have solved the national question [..]

socialism was thought	(YES)
socialism was accepted	(NO)

MINE: socialism was considered

****

Please evaluate:

She couldn't **hold** the truth **back.**

conceal the truth	(YES)
contain the truth	(NO)

MINE:

All of this **stirred** an unfathomable excitement in her.

provoked excitement	(YES)
created excitement	(YES)

MINE:

Gorbachev **inherited** a state, which [..].

acquired a state	(NO)
got a state	(NO)

MINE:

Central decisions could be **imposed** by the government.
decisions could be enforced (YES)
decisions could be communicated (NO)

MINE:

The writer [..] **reflected** some of these concerns.
showed concerns (YES)
manifested concerns (NO)

MINE:

Their views **reflect** enthusiasm among the British people.
show enthusiasm (YES)
indicate enthusiasm (YES)

MINE:

There are other factors which will have **shaped** this result
factors influenced this result (YES)
factors made this result (NO)
factors determined this result (NO)

MINE:

[..] that she and Prince Charles had succeeded in **mending** their marriage.
repairing their marriage (YES)
improving their marriage (NO)

MINE:

Leister has **brought** pleasure to millions with his fine sportsmanship and personal bravery.
got pleasure (NO)
taken pleasure (NO)

MINE: given pleasure

They did not want to **waste** their time and [..]
spend their time (YES)
expend their time (YES)

MINE:

[..] how man and women in the nineteenth century **faced** the hour of death.

turned the hour	(NO)
lied the hour	(NO)
met the hour	(YES)

MINE:

[..] historians and sociologists **stressing** activities of men's lives [..]

showing activities	(NO)
emphasizing activities	(YES)

MINE:

[sociologists] were inclined to **overlook** this subject.

omit this subject	(YES)
drop this subject	(YES)

MINE: ignore this subject

[..] duties were obscurely **glimpsed** and seldom accomplished.

duties were seen	(YES)

MINE:

[he] **skims** discretely **over** the question of survival.

reads the question	(YES)
touches the question	(YES)

MINE:

[..] contrast can not be **dismissed**.

contrast can not be changed	(NO)
contrast can not be dropped	(NO)

MINE:

[..] in the nineteenth century change was greatly **accelerated**

change was quickened	(NO)

MINE:

Impressionist painters **caught** the contagion
took the contagion (NO)

MINE:

The photographers tried to **seize** the fleeting moment
take the moment (NO)
capture the moment (YES)

MINE:

The present **regards** the past as a little more than the springboard from which we were launched on our way.
relates the past (NO)
sees the past (YES)
views the past (YES)

MINE:

[..] aspired to **follow** their fathers' professions
pursue professions (YES)
take professions (NO)
practice professions (YES)

MINE:

---

they were prepared to **disown** the past
repudiate the past (YES)

MINE:

---

they **base** their careers on [..]
establish their careers (YES)
locate their careers (NO)

MINE:

[..] irrelevant to the social and other problems that we **confront**.
encounter problems (YES)
present problems (NO)

MINE:

The report was carefully **leaked**
the report was disclosed (YES)
the report was revealed (YES)

MINE:

Hillary **brushed aside** the accusations and [..]
rejected the accusations (NO)
ignored the accusations (YES)

MINE:

[..] the speech act theory **developed** by Austin [..]
theory formulated (YES)
theory produced (NO)

MINE:

[..] use it to **tackle** different questions.
confront questions (YES)
face questions (YES)

MINE:

The reasons for this superiority are never **spelled out**.
the reasons are never specified (YES)
the reasons are never written (NO)

MINE:

This aspect of Petrey's thinking is also **illustrated** in his discussion [..]
aspect is elaborated (NO)
aspect is exemplified (YES)
aspect is shown (YES)

MINE:

This theory is not easy to **grasp** [..]
understand theory (YES)
hold theory (NO)

MINE:

[..] **fixing** these terms clearly in their minds.
specifying these terms (NO)
defining these terms (YES)

MINE:

These terms are not easy to **grasp**
terms easy to understand (YES)

MINE:

The lack of explicitness will surely **limit** the significance of the book.
determine the significance (YES)
hold the significance (NO)

MINE:

She gripped the steering-wheel tighter and managed to **block out** the thought
block the thought (YES)

MINE:

She had deliberately chosen the outfit likely to **reinforce** the general perception of her as [..]
strengthen the perception (YES)

MINE:

[..] **imposing** a fraction of her normal self-control
enforcing self-control (NO)

MINE:

[..] he is never in doubt about agreements he **reaches** or deals he makes.
agreements he attains (NO)
agreements he makes (YES)

MINE:

The only question she could **frame** was [..]
phrase question (YES)
put question (NO)

MINE:

To be fair I was as **opposed** to the idea as he was.
contradicted the idea (NO)
refuted the idea (NO)

MINE:

Your husband **broke** our agreement, Mrs Abberley..
terminated agreement (YES)
got agreement (NO)

MINE:

[she] marched out of the room **throwing** a partying remark
sending a remark (YES)
making a remark (YES)

MINE:

[..] social and economic changes that **operated** [..]
changes occurred (YES)

MINE:

There are many well-chosen examples which clearly **illustrate** [..]
examples picture (NO)
examples show (YES)

MINE:

---

scientists **focus** on [..]
scientists think (NO)
scientists concentrate (YES)

MINE:

The man's voice **cut in** – 'Do you believe me now , Mr Abberley?'

voice interrupted	(YES)

MINE:

[..] and then the memory **slipped away**

the memory passed	(YES)
the memory left	(NO)

MINE:

The Clinton **campaign surged** again and he easily won the Democratic nomination.

campaign improved	(YES)
campaign ran	(NO)

MINE:

The **tension mounted** around what seemed such a small ring for two massive men.

tension lifted	(NO)
tension increased	(YES)
tension rose	(YES)

MINE:

Thank you SO much for participation!!!

If you would like to give any feedback you're welcome to do it here:

Evaluation of Automatic Metaphor Paraphrasing – 2

Task:

You are given a list of sentences containing a metaphorical expression (in yellow) where **verbs** (**in bold and underlined**) are used metaphorically. Please give literal paraphrases for these verbs that mean the same thing in the context of the metaphorical expression. There can be a number of literal paraphrases for a given expression or no paraphrases.

Please note that it is **only the verbs** that you need to paraphrase, leaving the context fixed!

Please record all of the paraphrases you can think of in the slot "MINE" (if you can not come up with a paraphrase, don't worry, just leave the slot empty!)

****

Examples:

Soviet socialism was **held** to have solved the national question [..]

MINE: socialism was considered, socialism was thought

She couldn't **hold** the truth **back.**

MINE: conceal the truth

****

Please paraphrase:

All of this **stirred** an unfathomable excitement in her.

MINE: aroused an excitement, caused an excitement, precipitated an excitement, generated an excitement

---

Gorbachev **inherited** a state, which [..].

MINE: governed a state, took control of a state, assumed control of a state, won power in a state

---

Central decisions could be **imposed** by the government.

MINE: decisions could be made, laws could be enforced

The writer [..] **reflected** some of these concerns.

MINE: Expressed some of these concerns, Manifested some of these concerns

---

Their views **reflect** enthusiasm among the British people.

MINE: Exhibit enthusiasm, show enthusiasm, demonstrate enthusiasm, manifested enthusiasm

---

There are other factors which will have **shaped** this result

MINE: influenced this result, caused this result, precipitated this result

---

[..] that Diana and Prince Charles had succeeded in **mending** their marriage.

MINE: repairing their marriage, improving their marriage, reinstating their marriage

---

The did not want to **waste** their time and [..]

MINE:

[..] how man and women in the nineteenth century **faced** the hour of death.

MINE: reacted to their imminent death, reflected on

[..] historians and sociologists **stressing** activities of men's lives [..]

MINE: emphasising activites

[sociologists] were inclined to **overlook** this subject.

MINE: ignore this subject, neglect this subject

[..] these duties were obscurely **glimpsed** and seldom accomplished.

MINE: duties were rarely seen, duties were seldom witnessed

[he] **skims** discretely **over** the question of survival.

MINE: superficially circumvents the issue, inadequately addresses the issues

---

[..] this contrast can not be **dismissed**.

MINE: contrast can not be neglected, contrast can not be ignored, contast cannot be discarded

---

[..] in the nineteenth century change was greatly **accelerated**

MINE: *this is not a metaphor by my understanding*

Impressionist painters **caught** the contagion

MINE: adopted the fashion,

The photographers tried to **seize** the fleeting moment

MINE: photograph the instantaneous occurrence, capture the brief moment, record the short moment

---

The present **regards** the past as a little more than the springboard from which we were launched on our way.

MINE: the present considers the past to be, the present's view of the past is

---

[..] aspired to **follow** their fathers' professions

MINE: inherit their father's professions

---

they were prepared to **disown** the past in favour of the bright future

MINE: forget the past, reject the past, disallow the past,

---

they were **basing** their careers on the industrial platform [..]

MINE: building their careers, formatting their careers

[..] irrelevant to the social and other problems that we **confront**.

MINE: that we face, that we meet, that we hit,

The report was carefully **leaked**

MINE: carefully reported, carefully disseminated, carefully published

---

Hillary **brushed aside** the accusations and [..]

MINE: overlooked the accusations, rejected the accusations, ignored the accusations

---

[..] the speech act theory **developed** by Austin [..]

MINE: theory invented, theory build, theory constructed

[..] use it to **tackle** different questions.

MINE: answer, attempt, solve

The reasons for this superiority are never **spelled out**.

MINE: explained, described, made clear,

This aspect of Petrey's thinking is also **illustrated** in his discussion [..]

MINE: described, shown, illuminated

This theory is not easy to **grasp** [..]

MINE: understand, digest

[..] **fixing** these terms clearly in their minds.

MINE: putting these terms, enforcing these terms, establishing these terms

---

The lack of explicitness will surely **limit** the significance of the book.

MINE: inhibit the significance, hinder the significance, withold the significance

She gripped the steering-wheel tighter and managed to **block out** the thought

MINE: ignore the thought

She had deliberately chosen the outfit likely to **reinforce** the general perception of her as [..]

MINE: establish, imprint,

[..] **imposing** a fraction of her normal self-control

MINE: enforcing, exhibiting

[..] he is never in doubt about agreements he **reaches** or deals he makes.

MINE: makes, seals,

The only question she could **frame** was [..]

MINE: ask, pose, articulate

To be fair I was as **opposed** to the idea as he was.

MINE: rejecting of, set against,

Your husband **broke** our agreement, Mrs Abberley..

MINE: neglected, ignored, obviated

[she] marched out of the room **throwing** a partying remark

MINE: making, speaking, offering

[..] the context in which these social and economic changes **operated** [..]

MINE: happened, took place, occurred

There are many well-chosen examples which clearly **illustrate** [..]

MINE: describe, show, clarify, demonstrate

---

scientists **focus** on [..]

MINE: specialise, concentrate

The man's voice **cut in** – 'Do you believe me now , Mr Abberley?'

MINE: interjected, interrupted, spoke

[..] and then the memory **slipped away**

MINE: was forgotten, was lost, dispersed

The Clinton campaign **surged** again and he easily won the Democratic nomination.

MINE: increased, climaxed,

The tension **mounted** around what seemed such a small ring for two massive men.

MINE: hightened, increased, ameliorated

Thank you SO much for participation!!!

If you would like to give any feedback you're welcome to do it here:

Some words, in my opinion at least, may be considered metaphorical from a etymological perspective only.

For example “to reflect” in “reflected some of his concerns” is metaphorical if one considers it in the context of the the generally quite literal use of the French verb reflechir from which it originates. However I consider that at this stage “reflected” genuinely has a meaning congruent to those paraphrases I suggested in English and its literal meaning is not confined to light or fluid waves bouncing off an object.

Appendix F

Evaluation of integrated system performance

Evaluation of Metaphor Identification and Paraphrasing

Task:

The metaphors in the following sentences were identified (in yellow) and paraphrased (in blue) by the system. In the highlighted expressions **verbs** can be used metaphorically or literally.

You need to

1. compare the sentences, decide whether the highlighted expressions have the same meaning and record this in the box provided;
2. decide whether the **verbs** in both sentences are used metaphorically or literally and tick the respective boxes.

For the second decision, we suggest that you mainly rely on your own intuition on metaphor. However, you could also use the following procedure for guidance:

1. For each underlined verb in yellow (or blue) establish its meaning in context and try to imagine a more basic meaning of this verb on other contexts. Basic meanings normally are:
   (1) more concrete;
   (2) related to bodily action;
   (3) more precise (as opposed to vague);
   (4) historically older.
2. If you can establish the basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically.

****

Example:

ACH **1081** His 'fascist' ideas had first been **shaped** by the First World War, which he felt Britain should not have entered.

ACH **1081** His 'fascist' ideas had first been **influenced** by the First World War, which he felt Britain should not have entered.

Do the highlighted expressions have the same meaning?

YES	(X)
NO	( )

Is the **verb** in the first sentence used

metaphorically?	(X)
literally?	( )

Is the **verb** in the second sentence used

metaphorically?	( )
literally?	(X)

Please evaluate the sentences below:

CKM **391** Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **swallow** his anger and play tennis.

CKM **391** Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **suppress** his anger and play tennis.

Do the highlighted expressions have the same meaning?

YES	(X)
NO	( )

Is the **verb** in the first sentence used

metaphorically?	(X)
literally?	( )

Is the **verb** in the second sentence used

metaphorically?	()
literally?	(X)

AD2 **631** This is not to say that Paisley was dictatorial and simply **imposed** his will on other activists.

AD2 **631** This is not to say that Paisley was dictatorial and simply **enforced** his will on other activists.

Do the highlighted expressions have the same meaning?

YES	( )
NO	(X)

Is the **verb** in the first sentence used

metaphorically?	( )
literally?	(X)

Is the **verb** in the second sentence used

metaphorically?	( )
literally?	(X)

AND **322** It's almost as if some teachers **hold** the belief that the best parents are those that are docile and ignorant about the school, leaving the professionals to get on with the job.

AND **322** It's almost as if some teachers **apply** the belief that the best parents are those that are docile and ignorant about the school, leaving the professionals to get on with the job.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

K54 **2685** And it **approved** the recommendation by Darlington Council not to have special exemptions for disabled drivers.

K54 **2685** And it **passed** the recommendation by Darlington Council not to have special exemptions for disabled drivers.

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

G0G **1306** It would be natural to assume that this attempt to create a rift between Offa and the papacy occurred before the visit of the legates in 786 and that the visit was part of a process of reconciliation, but this is not wholly justified for Hadrian's letter could date to the late rather than the mid-780s, and **reflect** hostility to one or more of a number of Offa's actions.

G0G **1306** It would be natural to assume that this attempt to create a rift between Offa and the papacy occurred before the visit of the legates in 786 and that the visit was part of a process of reconciliation, but this is not wholly justified for Hadrian's letter could date to the late rather than the mid-780s, and **show** hostility to one or more of a number of Offa's actions.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

CGY **735** All that remained for theory was to explore the details of determination while avoiding the tendency of bourgeois ideology to **obscure** determination by the material base.

CGY **735** All that remained for theory was to explore the details of determination while avoiding the tendency of bourgeois ideology to **hide** determination by the material base.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ()

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ()

AD9 **3205** He tried to **disguise** the anxiety he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.

AD9 **3205** He tried to **hide** the anxiety he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ()

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

A1F **24** Moreover, Mr Kinnock **brushed aside** the suggestion that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.

A1F **24** Moreover, Mr Kinnock **dismissed** the suggestion that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

FB1 **701** Moreover, radicals with reservations about the socialist credentials of the USSR **confronted** the dilemma that every word of criticism aligned them with the 'reactionary' views dominant in the West

FB1 **701** Moreover, radicals with reservations about the socialist credentials of the USSR **presented** the dilemma that every word of criticism aligned them with the 'reactionary' views dominant in the West

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

AMA **349** We will **halt** the reduction in NHS services for long-term care and community health services which support elderly and disabled patients at home.

AMA **349** We will **prevent** the reduction in NHS services for long-term care and community health services which support elderly and disabled patients at home.

Do the highlighted expressions have the same meaning?
YES ( )

NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

B71 **852** Craig Packer and Anne Pusey of the University of Chicago have continued to **follow** the life and loves of these Tanzanian lions.

B71 **852** Craig Packer and Anne Pusey of the University of Chicago have continued to **succeed** the life and loves of these Tanzanian lions.

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ( )

CK1 **92** To understand this we have to see that Mill is **basing** his answer to one question on his answer to another.

CK1 **92** To understand this we have to see that Mill is **establishing** his answer to one question on his answer to another.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

JXS **1794** She tried to **cast** the thought from her, but it stayed stubbornly with her as he took another small step towards her.

JXS **1794** She tried to **put** the thought from her, but it stayed stubbornly with her as he took another small step towards her.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

ADK **634** **Catch** their interest and spark their enthusiasm so that they begin to see the product's potential.

ADK **634** **Ignite** their interest and spark their enthusiasm so that they begin to see the product's potential.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ( )

H9Y **2173** [..] my instinct was to **convey** the spirit and feeling of Grenfell, without resorting to mimicry.

H9Y **2173** [..] my instinct was to **bring** the spirit and feeling of Grenfell, without resorting to mimicry.

Do the highlighted expressions have the same meaning?

YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ()

EE9 **965** [..] not to be 'at liberty to make far-reaching proposals that the Unionist party would **decline** to support'.

EE9 **965** [..] not to be 'at liberty to make far-reaching proposals that the Unionist party would **reject** to support'.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

HGF **2874** The teacher came in like a colossus and the class **shrank** into a shivering line.

HGF **2874** The teacher came in like a colossus and the class **contracted** into a shivering line.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ( )

B1W **265** The above example **illustrates** the impact of the credit creation multiplier.

B1W **265** The above example **pictures** the impact of the credit creation multiplier.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ()

EF1 **1272** He was panicking; and panic invariably **leads** to mistakes.

EF1 **1272** He was panicking; and panic invariably **stimulates** mistakes.

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ( )

BN4 **126** It has also **adopted** nuclear power as a solution to the greenhouse effect.

BN4 **126** It has also **assumed** nuclear power as a solution to the greenhouse effect.

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )

literally? (X)

G1G **1105** This, it is argued, **hinders** rational thought and leads to the avoidance and downgrading of non-military solutions to world problems.

G1G **1105** This, it is argued, **keeps** rational thought and leads to the avoidance and downgrading of non-military solutions to world problems.

Do the highlighted expressions have the same meaning?
YES ( )
NO (X)

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ()

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ()

A6L **1688** I've also **adopted** the philosophy that I must develop somebody to do my job better than I have done it.

A6L **1688** I've also **espoused** the philosophy that I must develop somebody to do my job better than I have done it.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ()

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

J7C **785** Lord Denning suggested an alternative approach: the court should try to **construe** a contract out of all the communications passing between the parties.

J7C **785** Lord Denning suggested an alternative approach: the court should try to **interpret** a contract out of all the communications passing between the parties.

Do the highlighted expressions have the same meaning?
YES (X)

NO ()

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ()

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

G04 **3020** Why do we **keep** the truth from them?

G04 **3020** Why do we **preserve** the truth from them?

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? (X)
literally? ( )

CLN **500** Print propaganda aimed then to **stimulate** 'agitation' in the hope of shaping or sustaining antislavery public policies 'by a loud, strong and solemn expression of the public opinion'and establish direct 'influence'with individuals who exercised power and authority.

CLN **500** Print propaganda aimed then to **cause** 'agitation' in the hope of shaping or sustaining antislavery public policies 'by a loud, strong and solemn expression of the public opinion'and establish direct 'influence'with individuals who exercised power and authority.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? ( )
literally? (X)

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

Appendix G

Gold standard annotation guidelines for logical metonymy

Data Description Consider a phrase *enjoy a book*. It is obvious to the reader that its meaning extends to *enjoy reading a book* or *enjoy writing a book* depending on the context. This is what we call an interpretation of such phrases.

The evaluators are presented with the top 30 interpretations for each phrase. Some interpretations exemplified for the phrase *enjoy ... book* are demonstrated in Table G.1. The interpretations are represented in the form of *synonym sets*, each synonym set being a set of verbs with the same meaning. Some verbs appear in more than one synonym set. This means that the system suggests different senses of a verb as interpretations. You should identify which senses are correct interpretations (there could be more than one).

For example, the synonym set **(write-v-1 compose-v-3 pen-v-1 indite-v-1)** should be read as *enjoy writing a book* and the synonym set **(work-v-5 work_on-v-2 process-v-6)** as *enjoy working on a book*. You can understand the meaning of each synonym set by looking at the verbs in brackets and the associated description including examples.

Task You are given 3 metonymic phrases and for each a list of 30 possible interpretations produced by computer (the attached text files). For each synonym set in the list you need to decide whether it is a plausible interpretation of the metonymic phrase in an imaginary context.

Some interpretations are similar actions in the context of the metonymic phrase (e.g. *write a book, work on a book, produce a book* etc.) Group those together. In the example in the Table 7.2 the correct interpretations are shown in colors, the incorrect ones are left black. Each color indicates related meanings.

Returning the Results

- Remove the interpretations that are incorrect in your judgement (remove the whole synonym sets, not the verbs inside a set).
- Among the remaining interpretations mark the ones that you think are similar actions respectively (e.g., by highlighting them in the same color or in some other way convenient for you).

ENJOY ... A BOOK

synonym set and its Gloss (definition) - the synonym set is correct if it can fill the gap above
(write-v-1 compose-v-3 pen-v-1 indite-v-1) - produce a literary work; "She composed a poem"; "He wrote four novels"
(read-v-1) - interpret something that is written or printed; "read the advertisement"; "Have you read Salman Rushdie?"
(publish-v-3 write-v-3) - have (one's written work) issued for publication; "How many books did Georges Simenon write?"; "She published 25 books during her long career"
(read-v-2 say-v-4) - have or contain a certain wording or form; "The passage reads as follows"; "What does the law say?"
(read-v-3) - look at, interpret, and say out loud something that is written or printed; "The King will read the proclamation at noon"
(read-v-4 scan-v-7) - obtain data from magnetic tapes; "This dictionary can be read by the computer"
(search-v-1 seek-v-2 look_for-v-1) - try to locate or discover, or try to establish the existence of; "The police are searching for clues"; "They are searching for the missing man in the entire county"
(take-v-6 read-v-6) - interpret something in a certain way; convey a particular meaning or impression; "I read this address as a satire"; "How should I take this message?"; "You can't take credit for this!"
(learn-v-4 study-v-4 read-v-7 take-v-25) - be a student of a certain subject; "She is reading for the bar exam"
(read-v-8 register-v-5 show-v-9 record-v-3) - indicate a certain reading; of gauges and instruments; "The thermometer showed thirteen degrees below zero"; "The gauge read 'empty'"
(work_at-v-1 work_on-v-1) - to exert effort in order to do, make, or perform something; "the child worked at the multiplication table until she had it down cold"
(write-v-2) - communicate or express by writing; "Please write to me every week"
(analyze-v-1 analyse-v-1 study-v-1 examine-v-1 canvass-v-3 canvas-v-4) - consider in detail and subject to an analysis in order to discover essential features or meaning; "analyze a sonnet by Shakespeare"; "analyze the evidence in a criminal trial"; "analyze your real motives"
(use-v-1 utilize-v-1 utilise-v-1 apply-v-1 employ-v-1) - put into service; make work or employ for a particular purpose or for its inherent or natural purpose; "use your head!"; "we only use Spanish at home"; "I can't use this tool"; "Apply a magnetic field here"; "This thinking was applied to many projects"
(choose-v-1 take-v-10 select-v-1 pick_out-v-1) - pick out, select, or choose from a number of alternatives; "Take any one of these cards"; "Choose a good husband for your daughter"; "She selected a pair of shoes from among the dozen the salesgirl had shown her"
(consider-v-3 take-v-13 deal-v-2 look_at-v-1) - take into consideration for exemplifying purposes; "Take the case of China"; "Consider the following case"
(work-v-5 work_on-v-2 process-v-6) - shape, form, or improve a material; "work stone into tools"; "process iron"; "work the metal"
(produce-v-2 make-v-6 create-v-6) - create or manufacture a man-made product; "We produce more cars than we can sell"; "The company has been making toys for two centuries"

Table G.1: Metonymy interpretations as synonym sets (for *enjoy book*)