A Computational Model of Logical Metonymy

EKATERINA SHUTOVA, University of Cambridge
JAKUB KAPLAN, University of Cambridge
SIMONE TEUFEL, University of Cambridge
ANNA KORHONEN, University of Cambridge

The use of figurative language is ubiquitous in natural language texts and it is a serious bottleneck in automatic text understanding. A system capable of interpreting figurative expressions would be an invaluable addition to the real-world natural language processing (NLP) applications that need to access semantics, such as machine translation, opinion mining, question answering and many others. In this paper we focus on one type of figurative language, logical metonymy, and present a computational model of its interpretation bringing together statistical techniques and the insights from linguistic theory. Compared to previous approaches this model is both more informative and more accurate. The system produces sense-level interpretations of metonymic phrases and then automatically organises them into conceptual classes, or roles, discussed in the majority of linguistic literature on the phenomenon.

Categories and Subject Descriptors: H.3.1 [**Information Systems**]: Information Storage and Retrieval, Content Analysis

General Terms: Figurative language, computational modeling, multiword expressions

Additional Key Words and Phrases: Logical metonymy, semantic interpretation, word sense disambiguation

1. INTRODUCTION

Metonymy is defined as the use of a word or a phrase to stand for a related concept, which is not explicitly mentioned. It is based on contiguity between the concepts and implies a contact or a (rather physical) connection between the entities. Below are some examples of metonymic phrases:

(1) The *pen* is mightier than the *sword*. [Bulwer-Lytton 1839]

(2) He played *Bach*.

(3) He drank *his glass*. [Fass 1991]

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

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

(6) After *three martinis* John was feeling well. [Godard and Jayez 1993]

The metonymic adage in (1) is a classical example. Here the *pen* stands for the press and the *sword* for military power. In (2) *Bach* is used to refer to the composer's music and in (3) the *glass* stands for its *content*, i.e. the actual *drink* (beverage). These are

Authors' address: Computer Laboratory, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK. E-mail: {Ekaterina.Shutova, Simone.Teufel, Anna.Korhonen}@cl.cam.ac.uk, Jakub.Kaplan@gmail.com

DOI:http://dx.doi.org/10.1145/0000000.0000000

examples of *general metonymy*. The sentences (4–6) represent a variation of this phenomenon called *logical metonymy*. Here both *your book*, *his cigarette* and *three martinis* have eventive interpretations, i.e. the noun phrases stand for the events of *reading the book*, *smoking the cigarette* and *drinking three martinis* respectively.

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 (7) and (8) *unconventional*.

(7) The *ham sandwich* is waiting for his check. [Nunberg 1978]

(8) Ask *seat 19* whether he wants to swap. [Markert and Nissim 2006]

These examples illustrate that metonymy is both regular and productive.

As well as general metonymy, logical metonymy is an elliptical construction, i.e. it lacks an element that is recoverable or inferable from the context. However, if general metonymies usually follow a certain conceptual pattern (such as one of the above), logical metonymy arises due to a predicate taking syntactic and semantic arguments of different types. For example, the verb *enjoy* in (4) requires an event 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 (4) 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 our experiments.

The fact that logical metonymy is both highly frequent and productive makes its computational processing an important problem within NLP. In this paper, we focus on the problem of interpretation of logical metonymy and adopt a statistical data-driven approach to it. Our system first derives a set of possible metonymic interpretations from a large corpus. 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]. We then experimentally verify whether this representation is intuitive to humans, by asking human subjects to classify metonymic interpretations into groups of similar concepts.

2. THEORETICAL BACKGROUND

Regular polysemy in general and logical metonymy in particular have long been of considerable interest for lexical semantics. The term logical metonymy captures a range 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)).

(9) a. Mark enjoyed this book.
b. Mark enjoyed *reading* this book.

(10) a. Mark always enjoys his beer.
b. Mark always enjoys *drinking* his beer.

(11) a. Mark enjoyed his cigarette.
b. Mark enjoyed *smoking* his cigarette.

(12) a. Mark enjoyed the cake.
b. Mark enjoyed *eating* the cake.

(13) a. Mark enjoyed the concert.
b. Mark enjoyed *listening to* the concert.

(14) a. a good meal
b. a meal that *tastes* good

(15) a. a good cook
b. a cook that *cooks* well

(16) a. After the movie Mark went straight to bed.
b. After *watching* the movie Mark went straight to bed.

(17) a. After three martinis Mark was feeling well.
b. After *drinking* three martinis Mark was feeling well.

(18) 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 or preposition semantically selects for an argument of type *event*, but however, is combined with a noun phrase syntactically. 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, e.g. *pages, cover* for *book*)
— FORMAL Role (that which distinguishes the object within a larger domain, e.g. *physical object* for *book*)
— TELIC Role (purpose and function of the object, e.g. *read* for *book*)

— AGENTIVE Role (how the object came into being, e.g. *write* for *book*)

Qualia structure plays an important role in the interpretation of different kinds of multiword expressions, most prominent of them being compound nouns (e.g. *cheese knife* is a knife for *cutting* cheese) and logical metonymy. 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.

(19) 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 (19). 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.

(20) *John enjoyed the dictionary.

(21) *John enjoyed the door.

(22) *John began the tree.

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]. However, in reality the interpretation of logical metonymy (like any other linguistic phenomenon) is also a matter of pragmatics, i.e. one can often imagine a possible situation in which a particular phrase can be interpreted (e.g. an artist (John) began drawing a tree interpretation of (22)) and therefore, the validity of such examples is never clear-cut. However, the goal of this paper is to propose a computational method capable of interpreting more common logical metonymies, whose possible meanings can be derived even in isolation from wider context and pragmatics.

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 noun phrase (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.

3. PREVIOUS COMPUTATIONAL APPROACHES

Along with theoretical work, there have been a number of computational accounts of general [Utiyama et al. 2000; Markert and Nissim 2002; Nissim and Markert 2003; Peirsman 2006; Agirre et al. 2007] and logical [Lapata and Lascarides 2003] metonymy. All of these approaches are data-driven. The majority of the approaches to general metonymy (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, followed by [Lapata and Lascarides 2003], used text corpora to automatically derive paraphrases for metonymic expressions. 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 the 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; Pustejovsky 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:

$$P(e, v, o) = \frac{f(v, e) \cdot f(o, e)}{N \cdot f(e)}, \tag{1}$$

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 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 I.

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

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

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

Table I. Interpretations of Lapata and Lascarides for "finish video"

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

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.

4. 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 correspond to paraphrases that a human would give for the metonymic phrase (e.g. *read* for *enjoy a book*). 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 (e.g. *build* for *enjoy book*), 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 we 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 word sense disambiguation (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 start-

ing point in building the sense-based representation. Discovering metonymic interpretations and disambiguating them with respect to word sense is the focus of our first experiment.

The second issue we address in this paper 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 I). Therefore, we 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.

Godard and Jayez 1993 view a metonymic interpretation as 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 I), 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, we 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 our 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 polysemous verbs themselves.

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 our experiments, we 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. Clustering is performed using the information about lexico-syntactic environments, in which metonymic interpretations appear, as features.

We start by reimplementing the method of Lapata and Lascarides 2003, and then extend it by disambiguating the interpretations with respect to WordNet synsets for verb-object metonymic phrases. For this purpose, we 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.

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.

5. 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, our reimplementation of the method differs from the system of Lapata and Lascarides in that we use a more robust parser (RASP [Briscoe et al. 2006]), process a wider range of syntactic structures (coordination, passive), and extract our data from a later version of the BNC. As a result, we expect our system to extract the data more accurately.

5.1. Parameter estimation

The model of Lapata and Lascarides 2003 presented in section 3 is used to create and rank the initial list of ambiguous interpretations. The parameters of the model were estimated from the RASP-parsed BNC, using the grammatical relations (GR) output of RASP for BNC created by Andersen et al. 2008. In particular, we extracted 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 the system 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 $f(v, e)$.

Table II. Possible interpretations of metonymies ranked by our system

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

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 II.

5.2. Comparison with the results of Lapata and Lascarides

We compared the output of our 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 we extracted the data from the BNC parsed by RASP, as opposed to the Cass chunk parser [Abney 1996] utilised by Lapata and Lascarides. Our 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). We then 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).

We evaluated the performance of the system against the judgements elicited from humans in the framework of the experiment of Lapata and Lascarides 2003[3]. The Pearson correlation coefficient between the ranking of our 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 our 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, we conducted a one-tailed t-test to determine if the obtained ranking scores 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.

5.3. 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

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

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]. We 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. Our 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. Our 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 we found evidence supporting this claim. According to our own judgements, 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 our experiments build on the approach of Lapata and Lascarides 2003, we 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 would potentially allow us to rule out some incorrect interpretations, e.g. a resultative *find* for *enjoy book*, as *enjoy* requires an event of the type *activity*. Although we are not testing this hypothesis in the current work, automating this would be an interesting route for extension of our 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 *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).

6. 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. We 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. We propose to calculate synset likelihoods based on the initial likelihood of the ambiguous verbs,

relying on the hypothesis of Zipfian sense frequency distribution and information from WordNet glosses.

6.1. 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.

6.2. 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, we rank the synsets according to their likelihood of being a metonymic interpretation. The sense ranking is based on the probabilities of the verb strings derived by the model of Lapata and Lascarides 2003.

6.2.1. 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 the 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:

$$P_{v,k} = P_v \cdot \frac{1}{k} \tag{2}$$

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.

$$P_v = \sum_{s=1}^{N_v} P_{v,s} \tag{3}$$

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

The problem that arises with (2) 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

Table III. Metonymy interpretations as synsets (for "finish video")

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 IV. Different senses of *direct* (for "finish video")

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

by the Nth generalised harmonic number. Assuming the same notation

$$P_{v,k} = P_v \cdot \frac{1/k}{\sum_{n=1}^{N_v} 1/n} \tag{4}$$

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

$$P_s = \sum_{i=1}^{I_s} P_{v_i,k} \tag{5}$$

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 III.

6.2.2. 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 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 IV, 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

Table V. Metonymic phrases for groups 1 and 2

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

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, as determined experimentally on the development dataset.

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.

6.3. Evaluation

The ranking of the sense-based interpretations was evaluated against a gold standard created with the aid of human annotators. We used the dataset of Lapata and Lascarides 2003 in this experiment. The whole dataset consists of 58 metonymic phrases, 5 of which were used for development purposes and the remaining 53 constituted the test set.

6.3.1. 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. We divided them into 2 groups of 4. Participants in each group annotated three metonymic phrases as shown in Table V.

Materials and task The annotators received written guidelines describing the task (2 pages), which were the only source of information on the experiment. 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.

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

(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)

Fig. 1. Disambiguation gold standard for the phrase "finish video" in the form of WordNet synsets (before clustering)

annotations for the remaining 52 phrases in the dataset were carried out by one of the authors. The gold standard containing correct disambiguated interpretations for the metonymic phrase "finish video" is presented in Figure 1.

6.3.2. Evaluation measure. We 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 (MAP) at top 30 synsets. MAP is defined as follows:

$$MAP = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ji}, \tag{6}$$

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 (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.

6.3.3. Results. We compared the ranking obtained by applying Zipfian sense frequency distribution against that obtained by distributing probability mass over senses uniformly (baseline). We also considered the rankings before and after gloss processing. The results are shown in Table VI. 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 0.79, which suggests that the system is able to reliably disambiguate and re-rank metonymic interpretations. We additionally compared the rankings produced by the system and the baseline using Spearman's rank correlation co-

Table VI. Evaluation of the model ranking

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	0.52
Test set	Zipfian	No	0.74
Test set	Zipfian	Yes	0.79

efficient. The average rank correlation across the test set is 0.82, which suggests that the rankings of the two systems are not independent, however different.

7. 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 interpretations 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, we design feature sets that describe verb senses and test their informativeness using a range of clustering algorithms.

7.1. 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. We 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 our 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 features 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:

$$\mathbb{F}_3 = \{\sum_{i=1}^{K} c_{i1} + \beta f_1, \sum_{i=1}^{K} c_{i2} + \beta f_2, ..., \sum_{i=1}^{K} c_{iN} + \beta f_N\}$$

$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \prod_{i=1}^{K} c_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

We 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:

$$\mathbb{F}_5 = \{\sum_{i=1}^{K} c_{i1} + \beta f_1, \sum_{i=1}^{K} c_{i2} + \beta f_2, ..., \sum_{i=1}^{K} c_{iN} + \beta f_N\}$$

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

7.2. Clustering methods

To cluster the synsets we 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. We 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.

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.

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 we 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 features 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.

Class-based dimensionality reduction: Another issue that needs to be taken into account while designing clustering experiments is data sparseness. The fact that our synset feature vectors are constructed by means of overlap of the feature vectors of the individual verbs amplifies the problem. A possible solution is to apply class-based smoothing to the feature vectors. In other words, we 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, we 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. To do this, we automatically acquire selectional preference classes by means of noun clustering. We use agglomerative clustering algorithm and subject, direct and indirect object relations in which the nouns participate, along with the corresponding verb lemmas, as features. The choice of features is inspired by the results of previous works on noun clustering and selectional preference acquisition [Sun and Korhonen 2009].

Clustering toolkit: The clustering experiments for both nouns and verb synsets 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. Evaluation measures

We 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. We report the values of VI normalised by $\log N$, which brings them into the range $[0, 1]$.

Table VII. Metonymic phrases in development and test sets

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

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.

7.4. Clustering dataset and gold standard

System clustering was evaluated using a subset of the dataset of Lapata and Lascarides 2003. Five most frequent metonymic verbs were chosen to form the experimental data: *begin, enjoy, finish, try, start*. We 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, same as in the disambiguation experiment) and the test set (5 phrases) as shown in Table VII.

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 2. The glosses and the cluster with incorrect interpretations are omitted in this example for the sake of brevity.

We 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 we 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 we calculated the average across the metonymic phrases and the pairs of annotators. The obtained 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.[4] 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

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

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)

Fig. 2. Clustering gold standard for the phrase "finish video"

demonstrate that the task of clustering word senses in the context of logical metonymy is intuitive to humans, but nonetheless, challenging.

7.5. Experiments and results

7.5.1. Development set. To select the best parameter setting we ran the experiments on the development set varying the parameters described in section 7.2 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 VIII. The best result was obtained for the phrase "enjoy concert" as shown in Table IX. The clustering solution produced by the system for the phrase "enjoy concert" is demonstrated in Figure 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, we test the system only using K-means and repeated bisections. Judging from the improvements gained on the majority of the evaluation measures used, the results suggest that feature set 4 is the most informative, although for agglomerative clustering feature set 1 yields a surprisingly good result.

We then also clustered the nouns in the feature sets, as described in section 7.2, to perform class-based dimensionality reduction. We varied the number of noun clusters between 600 and 1400 and the best results obtained (as evaluated through synset clustering performance) are shown in Table X. These results confirm that feature set 4 consistently yields a better performance. We will use feature set 4 for evaluation on the test set, as it proves to be useful for all three clustering algorithms. However, the dimensionality reduction itself, while demonstrating gains in purity, decreases the system performance according to the other measures. This may be due to the fact that in the design of feature set 4 the problem of data sparsity has been taken into account

Table VIII. Average clustering results (development set)

Algorithm	F. S.	Purity	RI	F-measure	VI
K-means	F1	0.60	0.52	0.54	0.45
No scaling	F2	0.61	0.58	0.61	0.45
Cosine	F3	0.57	0.50	0.54	0.47
	F4	**0.65**	**0.57**	**0.69**	**0.35**
	F5	0.60	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	0.50	0.69	0.31
	F5	0.60	0.46	0.64	0.35

Table IX. Best clustering results (for *enjoy concert*, development set)

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

Table X. Top five results for synset clustering when the number of noun clusters is varied. *Algorithm* here refers to the algorithm used in synset clustering. The algorithm used to cluster the nouns was the agglomerative algorithm.

No. clusters	Algorithm	Feature set	Purity	F-measure	Rand index	VI
1100	K-means	**F4**	**0.69**	**0.58**	**0.59**	**0.51**
1300	K-means	F4	0.69	0.58	0.59	0.52
1200	RB	**F4**	**0.70**	**0.57**	**0.58**	**0.51**
1100	RB	F5	0.68	0.57	0.58	0.52
1100	K-means	F5	0.68	0.57	0.58	0.54

already, which makes the result of the class-based smoothing less evident. The error from the imperfect acquisition of the noun classes, may in turn propagate to synset clustering. However, due to gains in purity, we also consider the best noun clustering configurations (highlighted in Table X) interesting for the final experiment on the test set.

7.5.2. Test set. We present the results for the best system configurations on the test data in Table XI. 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 both with and without dimensionality reduction. Its performance was measured at 0.65 (Purity), 0.52 (Rand index), 0.64 (F-measure) and 0.33 (VI) without dimensionality reduction; the application of dimensionality reduction resulted in the performance of 0.70 (Purity), 0.59 (Rand index), 0.59 (F-measure) and 0.50 (VI). An example of k-means clusterings with and without dimensionality reduction obtained for the metonymic phrase “start letter” using the best performing features (F4) are shown in Figure 4.

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)

Fig. 3. Clustering solution for "enjoy concert". Red, blue and black colors represent gold standard classes

Table XI. Clustering results on the test set

Algorithm	F. S.	Purity	RI	F-measure	VI
Baseline		0.48	0.40	0.31	0.51
K-means no Dim. Red.	**F4**	**0.65**	**0.52**	**0.64**	**0.33**
RB no Dim. Red.	F4	0.63	0.48	0.60	0.37
K-means with Dim. Red.	**F4**	**0.70**	**0.59**	**0.59**	**0.50**
RB with Dim. Red.	F4	0.69	0.58	0.59	0.52
Agreement		0.75	0.67	0.76	0.37

7.5.3. 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.

Cluster 1: (receive have) (get let have) (take read) (get acquire) (send direct) (become go get) (receive get find obtain incur) (mail post send) (learn study read take) (experience receive have get)	**Cluster 1: (receive have) (get let have) (take read) (get acquire) (read say) (write save) (send direct) (become go get) (receive get find obtain incur) (mail post send) (learn study read take) (experience receive have get)**
Cluster 2: (view consider look_at) (print publish) (draft outline) (consider take deal look_at) (put set place pose position lay)	**Cluster 2: (view consider look_at) (print publish) (draft outline) (consider take deal look_at) (put set place pose position lay)**
Cluster 3: (write compose pen indite) (publish write) (read) (compose write) (read say) (write save)	**Cluster 3: (write compose pen indite) (publish write) (compose write) (write) (write_in)**
Cluster 4: (use utilize utilise apply employ) (write_in) (write) (read scan)	**Cluster 4: (use utilize utilise apply employ) (read scan) (read)**

Fig. 4. Clusters produced for *start letter* by the configuration k-means/F4 with (right) and without (left) dimensionality reduction.

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.

The class-based dimensionality reduction yields gains in purity and Rand Index, but however, decreases F-measure and VI. This may indicate that the application of dimensionality reduction results in the production of "cleaner" smaller clusters containing correct interpretations, however, leaving a greater number of correct interpretations behind in a large cluster with the incorrect ones. The qualitative analysis has shown that the main effect of dimensionality reduction is the improved clustering of the synsets consisting of only one or two verbs. This can be seen from the example in Figure 4 by comparing both versions of cluster 3, as well as both versions of cluster 4. This suggests that the synsets with fewer constituent verbs suffer from data sparsity more than synsets with more verbs, as can be expected. Our data analysis has also shown that large synsets (containing four or more verbs) are rarely affected.

Since our test set is relatively small (5 metonymic phrases yielding 150 synsets to cluster), we additionally carried out an analysis of performance variability across the dataset in order to demonstrate the consistency of our results. Table XII gives a comparison of the standard deviations (σ) and the coefficients of variation (C_v) for the best system configuration with and without dimensionality reduction. The coefficients of variation are defined as standard deviations divided by the means of the respective samples, thus providing normalized measurements of variation. The F-measure scores show the lowest variability, however, all the measures demonstrate that the clustering is performed with comparable quality across the dataset.

Table XII. Standard deviations (σ) and coefficients of variation (C_v) for all metrics across individual phrases in the test set, as measured for the best system configuration with and without dimensionality reduction

	K-means without Dim. Red.		K-means with Dim. red.	
Metric	σ	C_v	σ	C_v
Purity	0.099	0.146	0.107	0.153
F-measure	0.059	0.108	0.072	0.122
Rand index	0.084	0.147	0.079	0.133
VI	0.078	0.142	0.063	0.126

8. CONCLUSION AND FUTURE DIRECTIONS

Our 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. We proposed a sense-based representation of the interpretation of logical metonymy and developed a new word sense disambiguation method for the task. We also derived a ranking scheme for verb senses using an unannotated corpus, WordNet sense numbering and glosses. Our system identifies and ranks the disambiguated metonymic interpretations with a mean average precision of 0.79. Although the method was only evaluated on the task of disambiguating metonymic interpretations, it may have a wider applicability in the world of WSD.

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 paper offers experimental evidence to verify these claims. We 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 and intuitive to them. The intuitiveness is supported by the fact that they received minimal instructions.

Having verified this idea empirically, we then moved on to build a computational simulation of the class discovery process. We experimented with a number of clustering algorithms, and proposed feature sets for modelling a particular sense of a verb using a non-disambiguated corpus. We also investigated the issue of linguistically motivated, class-based dimensionality reduction. 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 $= 0.76$.

Overall the results show the effectiveness of our approach to resolution of logical metonymy and both the model ranking and system clustering correlate reliably with human judgements. One 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].

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 con-

texts and vice versa. This largely depends on the semantics of the noun. Consider the following examples:

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

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

The logical metonymy in (23) is perfectly well formed, but the one in (24) seems strange unless it is perceived as sarcastic. This is duse 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 (25).

(25) 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. 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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

M. Carpuat and D. Wu. 2007. 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*. 61–72.

Y. S. Chan, H. T. Ng, and D. Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, 33–40.

S. Clark and D. Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics* 28, 2 (2002), 187–206.

A. Copestake. 2001. The Semi-Generative Lexicon: Limits on Lexical Productivity. In *In Proceedings of the First International Workshop on Generative Approaches to the Lexicon*. 41–49.

A. Copestake and T. Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics* 12 (1995), 15–67.

B. J. Dorr. 1998. Large-Scale Dictionary Construction for ForeignLanguage Tutoring and Interlingual Machine Translation. *Machine Translation* 12, 4 (1998), 271–322.

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

D. Fass. 1997. *Processing Metaphor and Metonymy*. Ablex, Stanford, CA.

B. C. M Fung, K. Wang, and M. Ester. 2003. Large hierarchical document clustering using frequent itemsets. In *In Proceedings of SIAM International Conference on Data Mining 2003 (SDM 2003)*.

D. Godard and J. Jayez. 1993. Towards a proper treatment of coercion phenomena.. In *Sixth Conference of the European Chapter of the ACL*. Utrecht, 168–177.

E. Joanis, S. Stevenson, and D. James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering* 14, 3 (2008), 337–367.

G. Karypis. 2002. *Cluto: A clustering toolkit*. Technical Report. University of Minnesota.

C. Kelly, B. Devereux, and A. Korhonen. 2010. Acquiring Human-like Feature-Based Conceptual Representations from Corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*. Los Angeles, USA.

J. Klavans and M. Kan. 1998. 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)*. Montreal, Quebec, Canada, 680–686.

A. Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. Dissertation. UK.

A. Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of ACL 2003*. Sapporo,Japan.

G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

M. Lapata and C. Brew. 2004. Verb Class Disambiguation Using Informative Priors. *Computational Linguistics* 30, 2 (2004), 45–73.

M. Lapata and A. Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics* 29(2) (2003), 261–315.

A. Lascarides and A. Copestake. 1995. The Pragmatics of Word Meaning. In *Journal of Linguistics*. 387–414.

K. Markert and M. Nissim. 2002. Metonymy resolution as a classification task. In *Proceedings of the conference on Empirical methods in natural language processing (EMNLP '02)*. Association for Computational Linguistics, Morristown, NJ, USA, 204–213.

K. Markert and M. Nissim. 2006. Metonymic proper names: A corpus-based account. In *Corpus-Based Approaches to Metaphor and Metonymy*, A. Stefanowitsch and S. T. Gries (Eds.). Mouton de Gruyter, Berlin.

M. Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98, 5 (2007), 873–895.

M. Nissim and K. Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03)*. Sapporo, Japan, 56–63.

G. Nunberg. 1978. *The pragmatics of reference*. Technical Report. PhD thesis, Indiana University.

D. Ó Séaghdha. 2010. Latent Variable Models of Selectional Preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

M. Pasca and S. Harabagiu. 2001. The informative role of WordNet in open-domain question answering. In *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA, 138–143.

Y. Peirsman. 2006. 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)*. Association for Computational Linguistics, Morristown, NJ, USA, 71–78.

J. Preiss. 2006. *Probabilistic word sense disambiguation analysis and techniques for combining knowledge sources*. Technical Report. Computer Laboratory, University of Cambridge.

J. Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics* 17(4) (1991).

J. Pustejovsky. 1995. *The Generative Lexicon.* MIT Press, Cambridge, MA.

J. Pustejovsky and P. Bouillon. 1995. Logical polysemy and aspectual coercion. *Journal of Semantics* 12 (1995), 133–162.

W. M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *J. Amer. Statist. Assoc.* 66, 336 (1971), 846–850.

S. Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*. 747–753.

S. Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32(2) (2006), 159–194.

H. Schutze and J. O. Pedersen. 1995. Information Retrieval Based on Word Senses. In *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. 161–175.

G. Stern. 1931. *Meaning and Change of Meaning*. Wettergren & Kerbers Förlag, Göteburg.

C. Stokoe, M. P. Oakes, and J. Tait. 2003. 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)*. Toronto, Canada, 159–166.

L. Sun and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proceedings of EMNLP 2009*. Singapore, 638–647.

M. Utiyama, M. Masaki, and I. Hitoshi. 2000. A Statistical Approach to the Processing of Metonymy.. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrucken, Germany.

C. J. van Rijsbergen. 1979. *Information Retrieval, 2nd edition*. Butterworths, London.

Z. Vendler. 1968. *Adjectives and Nominalizations*. Mouton, The Hague.

C. M. Verspoor. 1997. Conventionality-governed logical metonymy.. In *Proceedings of the Second International Workshop on Computational Semantics*. Tilburg, 300–312.

A. Vlachos, A. Korhonen, and Z. Ghahramani. 2009. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *EACL workshop on GEometrical Models of Natural Language Semantics*. Athens.

E. M. Voorhees. 1998. Using WordNet for text retrieval. In *WordNet: An Electronic Lexical Database* (first ed.), C. Fellbaum (Ed.). MIT Press, 285–303.

Y. Zhao and G. Karypis. 2001. *Criterion functions for document clustering: Experiments and analysis*. Technical Report. University of Minnesota.