A Probabilistic Account of Logical Metonymy

Maria Lapata*
University of Edinburgh

Alex Lascarides†
University of Edinburgh

In this paper we investigate logical metonymy, i.e., constructions where the argument of a word in syntax appears to be different from that argument in logical form (e.g., enjoy the book *means enjoy reading the book, and* easy problem *means a problem that is easy to solve). The systematic variation in the interpretation of such constructions suggests a rich and complex theory of composition on the syntax/semantics interface (Pustejovsky, 1995). Linguistic accounts of logical metonymy typically fail to exhaustively describe all the possible interpretations, or they don't rank those interpretations in terms of their likelihood. In view of this, we acquire the meanings of metonymic verbs and adjectives from a large corpus and propose a probabilistic model which provides a ranking on the set of possible interpretations. We identify the interpretations automatically by exploiting the consistent correspondences between surface syntactic cues and meaning. We evaluate our results against paraphrase judgements elicited experimentally from humans, and show that the model's ranking of meanings correlates reliably with human intuitions.*

1 Introduction

Much work in lexical semantics has been concerned with accounting for regular polysemy, i.e., the regular and predictable sense alternations certain classes of words are subject to (Apresjan, 1973). It has been argued that in some cases, the different interpretations must arise from the *interaction* between the semantics of the words during syntactic composition, rather than by exhaustively listing all the possible senses of a word in distinct lexical entries (Pustejovsky, 1991). The class of phenomena which Pustejovsky (1991; 1995) has called logical metonymy is one such example. In the case of logical metonymy additional meaning arises for particular verb/noun and adjective/noun combinations in a systematic way: the verb (or adjective) semantically selects for an event-type argument, which is a different semantic type to that denoted by the noun. Nevertheless, the value of this event is predictable from the semantics of the noun. An example of verbal logical metonymy is given in (1) and (2): (1a) usually means (1b) and (2a) usually means (2b).

(1) a. Mary finished the cigarette.
    b. Mary finished smoking the cigarette.
(2) a. Mary finished her beer.
    b. Mary finished drinking her beer.

Note how the events in these examples correspond to the purpose of the object denoted by the noun: the purpose of a cigarette is to smoke it and the purpose of a beer is to drink it. Similarly,

* Division of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: mlap@cogsci.ed.ac.uk
† Division of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: alex@cogsci.ed.ac.uk

(3a) means a problem that is easy to solve; (3b) means a language that is difficult to learn, speak, or write; (3c) means a cook that cooks well, (3d) means a soup that tastes good, (3e) is someone who programmes fast, and (3f) is a plane that flies quickly.

(3) a. easy problem
  b. difficult language
  c. good cook
  d. good soup
  e. fast programmer
  f. fast plane

The interpretations of logical metonymies can typically be rendered with a paraphrase, as we have indicated for the above examples. Verb-nouns are paraphrased with a progressive or infinitive VP which is the complement of the polysemous verb (e.g., *smoking* in (1b)) and whose object is the NP figuring in the verb-noun combination (e.g., *cigarette* in (1b)). Adjective-noun combinations are usually paraphrased with a verb modified by the adjective in question or its corresponding adverb. For example, an *easy problem* is a problem that is easy to solve or a problem that one can solve easily (see (3a)).

Logical metonymy has been extensively studied in the lexical semantics literature. Previous approaches have focused on descriptive (Vendler, 1968) or theoretical accounts (Pustejovsky, 1991; Pustejovsky, 1995; Briscoe, Copestake, and Boguraev, 1990), on the linguistic constraints on the phenomenon (Godard and Jayez, 1993; Pustejovsky and Bouillon, 1995; Copestake and Briscoe, 1995; Copestake, 2001), and the influence of discourse context on the interpretation of metonymies (Briscoe, Copestake, and Boguraev, 1990; Lascarides and Copestake, 1998; Verspoor, 1997). McElree et al. (2001) investigated the on-line processing of metonymic expressions; their results indicate that humans display longer reading times for sentences like (1a) than for sentences like (1b).

There are at least two challenges in providing an adequate account for logical metonymy. The first concerns *semi-productivity*: there is a wealth of evidence that metonymic constructions are partially conventionalized, and so resolving metonymy entirely via pragmatic reasoning (e.g., by computing the purpose of the object that is denoted by the noun according to real world knowledge) will over-generate the possible interpretations (Hobbs et al., 1993). For example, the logical metonymies in (4) sound odd, even though pragmatics suggests an interpretation (because world knowledge assigns a purpose to the object denoted by the NP):

(4) a. ?John enjoyed the dictionary.
  b. ?John enjoyed the door.
  c. ?John began/enjoyed the highway.
  d. ?John began the bridge.

Sentence (4a) is odd because the purpose of dictionaries is to refer to them, or to consult them. These are (point-like) achievements and cannot easily combine with *enjoy* which has to be true of an event with significant duration. Domain knowledge assigns doors, highways, and bridges a particular purpose, and so the fact that the sentences in (4b)–(4d) are odd indicates that metonymic interpretations are subject to conventional constraints (Godard and Jayez, 1993).

The second challenge concerns the *diversity* of possible interpretations of metonymic constructions. This diversity is attested across and within metonymic constructions. Metonymic verbs and adjectives are able to take on different meanings depending on their local context, viz., the noun or noun class they select as objects (in the case of verbs) or modify (in the case of adjectives). Consider the examples in (1) where the meaning of the verb *finish* varies depending on the object it selects. Similarly, the adjective *good* receives different interpretations when modifying the nouns *cook* and *soup* (see (3c,d)).

While we've observed that some logical metonymies are odd even though pragmatics suggests an interpretation (e.g., (4c)), Vendler (1968) acknowledges that other logical metonymies have more than one plausible interpretation. In order to account for the meaning of adjective-noun combinations Vendler (1968, 92) points out that "in most cases not one verb, but a family of verbs is needed". For example *fast scientist* can mean a scientist who does experiments quickly, publishes quickly, and so on.

Vendler (1968) further observes that the noun figuring in an adjective-noun combination is usually the subject or object of the paraphrasing verb. Although *fast* usually triggers a verb-subject interpretation (see (3e,f)), *easy* and *difficult* trigger verb-object interpretations (see (3a,b)). An *easy problem* is usually a problem that one solves easily (so *problem* is the object of *solve*), and a *difficult language* is a language that one learns, speaks, or writes with difficulty (so *language* is the object of learn, speak, and write). Adjectives like *good* allow either verb-subject or verb-object interpretations: a *good cook* is a cook who cooks well whereas *good soup* is a soup that tastes good.

All of these interpretations of *fast scientist*, *difficult language* or *good soup* seem highly plausible 'out-of-context', though one interpretation may be favored over another in a particular context. In fact, in sufficiently rich contexts, pragmatics can even override the conventional interpretations: Lascarides and Copestake (1998) suggest that (5c) means (5d) and not (5e):

(5) a. All the office personnel took part in the company sports day last week.
b. One of the programmers was a good athlete, but the other was struggling to finish the courses.
c. The fast programmer came first in the 100m.
d. The programmer who runs fast came first in the 100m.
e. The programmer who programs fast came first in the 100m.

The discourse context can also ameliorate highly marked logical metonymies, such as (4c):

(6) a. John uses two highways to get to work every morning.
b. He first takes H-280 and then H-101.
c. He always enjoys H-280,
d. but the traffic jams on H-101 frustrate him.

Arguably the most influential account of logical metonymy is Pustejovsky's (1991; 1995) theory of the Generative Lexicon. He avoids enumerating the various senses for adjectives like *fast* and verbs like *finish* by exploiting a rich lexical semantics for nouns. The lexical entry for an artifact-denoting noun includes a *qualia structure*: this specifies key features of the word's meaning that are in some sense derivable from real world knowledge but which are lexicalized so as to influence conventional processes. The qualia structure includes a telic role (i.e., the purpose of the object denoted by the noun) and an agentive role (i.e., the event which brought the object into existence). Thus the lexical entry for *book* includes a telic role with a value equivalent to *read* and an agentive role with a value equivalent to *write*, whereas for *cigarette* the telic role is equivalent to *smoke* and the agentive role is equivalent to *roll* or *manufacture*.

When *finish* combines with an object-denoting NP, a metonymic interpretation is constructed where the missing information is provided by the qualia structure of the NP. More technically, semantic composition of *finish* with *cigarette* causes the semantic type of the noun to be *coerced* into its telic event (or its agentive event), and the semantic relation corresponding to the metonymic verb (*finish*) predicates over this event. This results in an interpretation of (1a) equivalent to (1b). Verbs like *begin* and *enjoy* behave in a similar way. *Enjoy the book* can mean enjoy reading the book because of *book*'s telic role or enjoy writing the book because of *book*'s agentive role. In fact, the agentive reading is less typical for *book* than the telic one, but for other nouns the opposite is true. For instance, *begin the tunnel* can mean begin building the tunnel, but the interpretation that is equivalent to begin going through the tunnel is highly marked. There is also variation in the relative likelihood of interpretations among different metonymic verbs. We return to this issue shortly. The adjective/noun combinations are treated along similar lines. Thus the logical polysemy of words like *finish* and *fast* is not accounted for by exhaustive listing.[1]

In contrast to the volume of theoretical work on logical metonymy, very little empirical work has tackled the topic. Briscoe et al. (1990) investigate the presence of verbal logical metonymies in naturally occurring text by looking into data extracted from the Lancaster-Oslo/Bergen corpus (LOB, one million words). Verspoor (1997) undertakes a similar study in the British National Corpus (BNC, 100 million words). Both studies investigate how widespread the use of logical metonymy is, and how far the interpretation for metonymic examples can be recovered from the head noun's qualia structure, assuming one knows what the qualia structure for any given noun is. None of these studies are concerned with the automatic generation of interpretations for logical metonymies and the determination of their likelihood.

Although conceptually elegant, Pustejovsky's (1995) theory of the Generative Lexicon does not aim to provide an exhaustive description of the telic roles that a given noun may have. However, these are crucial for interpreting verb-noun and adjective-noun metonymies. In contrast to Vendler (1968), who acknowledges the fact that logical metonymies may trigger more than one interpretation (in other words, there may be more than one possible event associated with the noun in question), Pustejovsky implicitly assumes that nouns or noun classes have one—perhaps default—telic role without however systematically investigating the relative degree of ambiguity of the various cases of logical metonymy (e.g., the 'out of context' possible readings for *fast scientist* suggest that *fast scientist* exhibits a higher degree of semantic ambiguity than *fast plane*). One could conceivably represent this by the generality of the semantic type of the telic role in the various nouns (e.g., assign the telic role of *scientist* a relatively general type of event compared with that for *plane*). But this simply transfers the problem: the degree of generality in lexical representation is highly idiosyncratic and we need to acquire this information from linguistic evidence; furthermore, for nouns with a general telic role, pragmatics would have to do 'more work' to augment the general interpretation with a more specific one, and ideally we need to compute this in a systematic way. Even in theories where more than one interpretation is provided (see Vendler, 1968), no information is given with respect to the relative likelihood of these interpretations.

Pustejovsky's account also doesn't predict the degree of variation of interpretations for a given noun among the different metonymic verbs: for example, the fact that *begin the house* is, intuitively at least, more likely to resolve to an agentive-role interpretation (i.e., begin building the house) than a telic-role interpretation (i.e., begin living in the house), while the reverse is true of *enjoy the house*. Ideally, we would like a model of logical metonymy which reflects this variation in interpretation.

In this paper we aim to complement the theoretical work on the interpretation of logical metonymy by addressing the following questions: (a) Can the meanings of metonymic adjective-noun and verb-noun combinations be acquired automatically from corpora? (b) Can we constrain the number of interpretations by providing a ranking on the set of possible meanings? (c) Can we determine if an adjective has a preference for a verb-subject or verb-object interpretation? We provide a probabilistic model which uses distributional information extracted from a large corpus to automatically interpret logical metonymies without recourse to pre-existing taxonomies or manually annotated data.

1 Other lexical accounts, such as Copestake and Briscoe (1995), differ from Pustejovsky's (1995) in that the 'coercion' is treated as internal to the semantics of the metonymic verb or adjective rather than the noun; motivation for this comes from *co-predication* data; such as the acceptability of *fast and intelligent typist* and *John picked up and finished his beer*.

The differences among the theoretical accounts—for example that Copestake and Briscoe (1995) treat the type coercion as internal to the metonymic word where Pustejovsky (1995) treats it as part of the noun—do not matter for our purposes, because we aim to provide information about metonymic interpretations that is compatible with either account. More specifically, we are concerned with using a real language corpus to automatically acquire the semantic value of the event that is part of the interpretation. We abstract away from theoretical concepts such as semantic type coercion, and instead utilise co-occurrence frequencies in the corpus to predict metonymic interpretations. Very roughly, we acquire a ranked set of interpretations *enjoy V-ing the book* for the construction *enjoy the book* by estimating the probabilities that V is enjoyed and that it is something done to books; and we estimate these probabilities on the basis of the corpus frequencies for V appearing as a (verbal) complement to *enjoy* and for V taking *book* as its object. Similarly, we acquire a ranked set of verb-subject interpretations of *fast plane* by estimating the likelihood of seeing *the plane Vs* and *Vs quickly* in the corpus. See Sections 2 and 3 for more details and motivation of these models.

Our results not only show that we can predict meaning differences when the same adjective or verb is associated with different nouns, but we can also derive—taking into account Vendler's (1968) observation—a cluster of meanings for a single verb/adjective-noun combination. We can also predict meaning differences for a given noun associated with different metonymic verbs and adjectives. We evaluate our results by comparing the model's predictions against human judgements and show that the model's ranking of meanings correlates reliably with human intuitions.

However, the model is limited in its scope. It is suited for the interpretation of well-formed metonymic constructions. But it does not distinguish odd metonymies (see (4)) from acceptable ones: in both cases, paraphrases will be generated, at least in principle (see Section 2.1 for explanation and motivation). In particular, the model does not *learn* conventional constraints, such as *enjoy* must take an event of duration as its argument (Godard and Jayez, 1993). However, such constraints are potentially captured indirectly: if the above conventional constraint is right, *enjoy referring to* should not be attested in the corpus and hence according to our model it won't be part of a possible paraphrase for *enjoy the dictionary* (see Sections 2.4.2 and 2.5.3 for further discussion). Further, since the model abstracts away from semantic type coercion, it does not distinguish between uses of a verb/adjective which are claimed in the lexical semantics literature to be metonymic uses (e.g., *enjoy the book*, *fast programmer*) and those which are claimed to be non-metonymic uses (e.g., *enjoy the marriage*, *fast run-time*). Again, the model of interpretation presented here will generate paraphrases for all these cases (e.g., it will paraphrase *enjoy the marriage* as enjoy going to or participating in the marriage and *fast run-time* as run-time that goes by or passes quickly). The model also does not take discourse context into account; e.g., it will not predict the intuitive interpretation of (5e)). Rather, it determines the most dominant meanings for a given metonymic construction overall, across all of its instances in the corpus.

The remainder of this paper is organised as follows: in the first part (Section 2) we present our probabilistic model of verbal logical metonymy and describe the model parameters. In Experiment 1 we use the model to derive the meaning paraphrases for verb-noun combinations randomly selected from the BNC (see Section 2.3) and formally evaluate our results against human intuitions. Experiment 2 demonstrates that when compared against human judgements our model outperforms a naive baseline in deriving a preference ordering for the meanings of verbal-noun combinations and Experiment 3 evaluates an extension of the basic model. In the second part (Section 3), we focus on adjectival logical metonymy. Section 3.1 introduces our probabilistic formalization for polysemous metonymic adjectives and Sections 3.3–3.6 present our experiments and evaluate our results. Overall, the automatically acquired model of logical metonymy reliably correlates with human intuitions and also predicts the relative degree of ambiguity and acceptability of the various metonymic constructions. In Section 5 we review related work and conclude in Section 6.

2 Metonymic Verbs

2.1 The Model

Consider the verb-noun combinations in (7) and (8). Our task is to come up with (7b) and (8b) as appropriate interpretations for (7a) and (8a). While the interpretations of (7a) and (8a) are relatively straightforward for English speakers given their general knowledge about coffees and films and the activities or events associated with them, a probabilistic model requires detailed information about words and their interdependencies in order to generate the right interpretation. Examples of such interdependencies are verbs co-occurring with *coffee* (e.g., *drink*, *make*, *prepare*) or verbs that are related with *begin* (e.g., *make*, *realise*, *understand*).

(7) a. John began the coffee.
    b. John began drinking the coffee.
(8) a. Mary enjoyed the film.
    b. Mary enjoyed watching the film.

A relatively straightforward approach to the interpretation of (7a) and (8a) would be to extract from the corpus (via parsing) paraphrases for which the additional information (e.g., *drinking* and *watching*), which is absent from (7a) and (8a), is fleshed out. In other words, we would like to find in the corpus sentences whose main verb is *begin* followed either by the progressive VP complement *drinking* or by the infinitive *to drink*, selecting for the NP *coffee* as its object. In the general case we would like to find the activities or events related both to the verb *begin* and the noun *coffee* (e.g., *drinking*, *buying*, *making*, *preparing*). Similarly, in order to paraphrase (8a) we need information about the VP complements which are associated with *enjoy* and can take *film* as their object (e.g., *watching*, *making*, *shooting*).

The above paraphrase-based model is attractive given its simplicity: all we need to do is count the co-occurrences of a verb, its complements, and their objects. The approach is unsupervised, no manual annotation is required, and no corpus-external resources are used. Such a model relies on the assumption that the interpretations of (7a) and (8a) can be approximated by their *usage*, i.e., it assumes that there is an equal likelihood of uttering the metonymic construction as well as its interpretation. However, this assumption is not borne out. Only four sentences in the BNC are relevant for the interpretation of *begin coffee* (see (9)); likewise, four sentences are relevant for the interpretation of *enjoy film* (see (10)).

(9) a. Siegfried bustled in, muttered a greeting and began to **pour** his coffee.
    b. She began to **pour** coffee.
    c. Jenna began to **serve** the coffee.
    d. Victor began **dispensing** coffee.
(10) a. I was given a good speaking part and enjoyed **making** the film.
    b. He's enjoying **making** the film.
    c. Courtenay enjoyed **making** the film.
    d. I enjoy most music and enjoy **watching** good films.
    e. Did you enjoy **acting** alongside Marlon Brando in the recent film The Freshman?

The attested sentences in (9) are misleading if they are taken as the only evidence for the interpretation of *begin coffee*; for on their own they suggest that the most likely interpretation for *begin coffee* is begin to pour coffee, whereas begin to serve coffee and begin dispensing coffee are equally likely as they are attested in the corpus only once. Note that the sentences in (9) fail to capture begin to drink coffee as a potential interpretation for *begin coffee*. On the basis of the sentences in (10), enjoy making the film is the most likely interpretation for (8a), whereas enjoy watching the film and enjoy acting in the film are equally likely.

Table 1
BNC frequencies for *begin*

Examples	*begin* NP	*begin* V-*ing* NP
begin book	35	17
begin sandwich	4	0
begin beer	2	1
begin speech	21	4
begin solo	1	1
begin song	19	8
begin story	31	15

Table 2
BNC frequencies for *enjoy*

Examples	*enjoy* NP	*enjoy* V-*ing* NP
enjoy symphony	34	30
enjoy movie	5	1
enjoy coffee	8	1
enjoy book	23	9
like movie	18	3

Table 3
BNC frequencies for *want*

Examples	*want* NP	*want* V-*ing* NP
want cigarette	18	3
want beer	15	8
want job	116	60

This finding complies with Briscoe et al.'s (1990) results for the LOB corpus: *begin* V NP is very rare when the value of V corresponds to a highly plausible interpretation of *begin* NP. Indeed, one can predict that problems with finding evidence for *begin* V NP will occur on the basis of Gricean principles of language production, where the heuristic *be brief* (which is part of the maxim of Manner) will compel speakers to utter *begin coffee* as opposed to *begin* V *coffee* if V is one of the plausible interpretations of *begin coffee*. Thus on the basis of this Gricean reasoning, one might expect metonymies like (7a) and (8a) to occur with greater frequencies than their respective paraphrases (see (7b) and (8b)). Tables 1–3 show BNC counts of verb-noun metonymies (commonly cited in the lexical semantics literature (Pustejovsky, 1995; Verspoor, 1997)) and their corresponding interpretations when these are attested in the corpus. The data in Tables 1–3 indicates that metonymic expressions are more often attested in the BNC with NPs rather than with VP complements.

The discrepancy between an interpretation and its usage could be circumvented by using a corpus labelled explicitly with interpretation paraphrases. Lacking such a corpus we will sketch below an approach to the interpretation of metonymies which retains the simplicity of the paraphrase-based account but no longer assumes a tight correspondence between a metonymic interpretation and its usage. We present an unsupervised method which generates interpretations for verbal metonymies without recourse to manually annotated data or taxonomic information; it only requires a part-of-speech tagged corpus and a partial parser.

We model the interpretation of a verbal metonymy as the joint distribution $P(e,o,v)$ of three variables: the metonymic verb v (e.g., *enjoy*), its object o (e.g., *film*), and the sought after interpretation e (e.g., *making*, *watching*, *directing*). By choosing the ordering $\langle e,v,o\rangle$ for the variables e, v, and o, we can factor $P(e,o,v)$ as follows:

$$P(e,o,v) = P(e)\cdot P(v|e)\cdot P(o|e,v) \tag{11}$$

The probabilities $P(e)$, $P(v|e)$, and $P(o|e,v)$ can be estimated using maximum likelihood as follows:

$$\hat{P}(e) = \frac{f(e)}{N} \tag{12}$$

$$\hat{P}(v|e) = \frac{f(v,e)}{f(e)} \tag{13}$$

$$\hat{P}(o|e,v) = \frac{f(o,e,v)}{f(e,v)} \tag{14}$$

Although $P(e)$ and $P(v|e)$ can be estimated straightforwardly from a corpus ($f(e)$ amounts to the number of the times a given verb e is attested, N is the sum of all verbs found in the corpus (excluding modals and auxiliaries), and $P(v|e)$ can be obtained through parsing, by counting the number of times a verb v takes e as its complement), the estimation of $P(o|e,v)$ is problematic. It presupposes that co-occurrences of metonymic expressions and their interpretations are to be found in a given corpus but as we've seen previously there is a discrepancy between a metonymic interpretation and its usage. In fact, metonymies occur more frequently than their overt interpretations (expressed by the term $f(o,e,v)$ in (14)) and the interpretations in question are not explicitly marked in our corpus. We will therefore make the following approximation:

$$P(o|e,v) \approx P(o|e) \tag{15}$$

$$\hat{P}(o|e) = \frac{f(o,e)}{f(e)} \tag{16}$$

The rationale behind this approximation is that the likelihood of seeing a noun o as the object of an event e is largely independent of whether e is the complement of another verb. In other words, v is conditionally independent of e, since the likelihood of o is (largely) determined on the basis of e and not of v. Consider again example (8a): *Mary enjoyed the film*. Here, *film*, the object of *enjoy*, is more closely related to the underspecified interpretation e rather than to *enjoy*. For example, watching movies is more likely than eating movies, irrespectively of whether Mary enjoyed or liked watching them. We estimate $P(o|e)$ as shown in (16). The simplification in (15) results in a compact model with a relatively small number of parameters which can be estimated straightforwardly from the corpus in an unsupervised manner. By substituting equations (12), (13), and (16) into (11) and simplifying the relevant terms, (11) is rewritten as follows:

$$P(e,o,v) = \frac{f(v,e) \cdot f(o,e)}{N \cdot f(e)} \tag{17}$$

Assume we want to generate meaning paraphrases for the verb-noun pair *enjoy film* (see (8a)). Table 4 lists the most frequent events related with the verb *enjoy* and the most frequent verbs which take *film* as their object (we describe how the frequencies $f(v,e)$ and $f(o,e)$ were obtained in the following section). We can observe that *seeing*, *watching*, and *making* are all events associated with *enjoy* and with *film*, and will be therefore generated as likely paraphrases for the metonymic expression *enjoy film* (see Table 4 where the underlined verbs indicate common complements between the metonymic verb and its object).

Note that the model in (17) does not represent the fact that the metonymic verb v may have a subject. This in practice means that the model cannot distinguish between the different readings for (18a) and (18b): in (18a) the doctor enjoyed watching the film, whereas in (18b) the director enjoyed making or directing the film. The model in (17) will generate the set of events that are associated with enjoying films (e.g., *watching, making, seeing, going to*) ignoring the contribution of the sentential subject. We present in Section 2.5.1 an extension of the basic model which takes sentential subjects into account.

Table 4
Most frequent complements of *enjoy* and *film*

$f(enjoy,e)$		$f(film,e)$	
play	44	<u>make</u>	176
<u>watch</u>	42	be	154
work with	35	<u>see</u>	89
read	34	<u>watch</u>	65
<u>make</u>	27	show	42
<u>see</u>	24	produce	29
meet	23	have	24
go to	22	<u>use</u>	21
<u>use</u>	17	do	20
take	15	get	18

(18) a. The doctor enjoyed the film.
b. The director enjoyed the film.

It is important to stress that the probabilistic model outlined above is a model of the *interpretation* rather than the *grammaticality* of metonymic expressions. In other words, we do not assume that it can distinguish between well-formed and odd metonymic expressions (see the examples in (4)). In fact, it will generally provide a set of interpretation paraphrases, even for odd formulations. The model in (11) has no component that corresponds to the occurrence of v and o together. Choosing the ordering $\langle o,v,e\rangle$ for the variables o, e, and v would result in the following derivation for $P(e,v,o)$:

$$P(e,o,v) = P(o)\cdot P(v|o)\cdot P(e|o,v) \tag{19}$$

The term $P(v|o)$ in (19) explicitly takes into account the likelihood of occurrence of the metonymic expression. This means that no interpretation will be provided for odd metonymies like *enjoy the highway* as long as they are not attested in the corpus. Such a model penalizes, however, well-formed metonymies which are *not* attested in the corpus. A striking example is *enjoy the ice-cream* which is a plausible metonymy, not attested at all in the BNC and thus by (19) would be incorrectly assigned no interpretations. This is because the maximum likelihood estimate of $P(v|o)$ relies on the co-occurrence frequency $f(v,o)$ which is zero for *enjoy the ice-cream*. But the probabilistic model in (11) will generate meaning paraphrases for metonymic verb-object pairs that have not been attested in the corpus as long as the co-occurrence frequencies $f(v,e)$ and $f(o,e)$ are available.

Finally, note that our model is ignorant with respect to the discourse context within which a given sentence is embedded. This means that it will come up with the same ranked set of meanings for (20b), irrespectively of whether it is preceded by sentence (20a) or (21a). The model thus does not focus on the meaning of individual corpus tokens; instead it determines the most dominant meanings for a given verb/noun combination overall, across all of its instances in the corpus.

(20) a. Who is making the cigarettes for tomorrow's party?
b. John finished three cigarettes.
c. John finished making three cigarettes.

(21) a. Why is the room filled with smoke?
b. John finished three cigarettes.
c. John finished smoking three cigarettes.

2.2 Parameter Estimation

We estimated the parameters of the model outlined in the previous section from a part-of-speech tagged and lemmatised version of the BNC, a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent current British English (Burnard, 1995).

The counts $f(v,e)$ and $f(o,e)$ (see (17)) were obtained automatically from a partially parsed version of the BNC which was created using Cass (Abney, 1996), a robust chunk parser designed for the shallow analysis of noisy text. The parser's built-in function was employed to extract tuples of verb-subjects and verb-objects (see (22)). Although verb-subject relations are not relevant for the present model, they are important for capturing the influence of the sentential subject (see Section 2.5) and modeling the interpretations of polysemous adjectives (which we discuss in the second part of the paper, see Section 3).

(22)	a.	change situation	SUBJ
	b.	come off heroin	OBJ
	c.	deal with situation	OBJ
(23)	a.	isolated people	SUBJ
	b.	smile good	SUBJ

The tuples obtained from the parser's output are an imperfect source of information about argument relations. Bracketing errors, as well as errors in identifying chunk categories accurately, result in tuples whose lexical items do not stand in a verb-argument relationship. For example, inspection of the original BNC sentences from which the tuples in (23) were derived reveal that the verb *be* is missing from (23a) and the noun *smile* is missing from (23b) (see the sentences in (24)).

(24) a. Wenger found that more than half the childless old people in her study of rural Wales saw a relative, a sibling, niece, nephew or cousin at least once a week, though in inner city London there were more isolated old people.

b. I smiled my best smile down the line.

In order to compile a comprehensive count of verb-argument relations we discarded tuples containing verbs or nouns attested in a verb-argument relationship only once. Instances of the verb *be* were also eliminated since they contribute no semantic information with respect to the events or activities that are possibly associated with the noun with which the verb is combined. Particle verbs (see (22b)) were retained only if the particle was adjacent to the verb. Verbs followed by the preposition *by* and a head noun were considered instances of verb-subject relations. The verb-object tuples also included prepositional objects (see (22c)). It was assumed that PPs adjacent to the verb headed by either of the prepositions *in, to, for, with, on, at, from, of, into, through*, and *upon* were prepositional objects.[2] This resulted in 737,390 distinct types of verb-subject pairs and 1,077,103 distinct types of verb-object pairs (see Table 5 which contains information about the tuples extracted from the corpus before and after the filtering).

The frequency $f(v,e)$ represents verbs taking progressive or infinitive VP complements. These were extracted from the parser's output by looking for verbs followed by progressive or infinitival complements (a special tag, VDG, is reserved in the BNC for verbs in the progressive). The latter were detected by looking for verbs followed by infinitives (indicated by the marker *to* (TO0) and a verb in base form (VVI)). The examples below illustrate the information extracted

2 The POS-tagging of the BNC (Leech, Garside, and Bryant, 1994) distinguishes between verb particle constructions like *down* in *climb down the mountain* and *up* in *put up the painting* on the one hand, and prepositions on the other. So this allowed us to distinguish PP complements from NP ones.

Table 5
Tuples extracted from the BNC

	Tokens		Types		
Relation	Parser	Filtering	Tuples	Verbs	Nouns
SUBJ	4,759,950	4,587,762	737,390	14,178	25,900
OBJ	3,723,998	3,660,897	1,078,053	12,026	35,867

Table 6
Paraphrases for verb-noun combinations taken from the literature

John began the story	$\rightarrow$ telling	(Verspoor, 1997, 189)
John began the song	$\rightarrow$ singing	(Verspoor, 1997, 189)
John began the sandwich	$\rightarrow$ eating/making	(Verspoor, 1997, 167)
Mary wants a job	$\rightarrow$ to have	(Pustejovsky, 1995, 45)
John began the book	$\rightarrow$ reading/writing	(Verspoor, 1997, 167)
Bill enjoyed Steven King's last book	$\rightarrow$ reading	(Pustejovsky, 1995, 88)
John began the cigarette	$\rightarrow$ smoking	(Verspoor, 1997, 167)
Harry wants another cigarette	$\rightarrow$ to smoke	(Pustejovsky, 1995, 109)

from the parser's output for obtaining the frequency $f(v,e)$ which collapsed counts for progressive and infinitive complements.

(25)	a.	I had started to write a love-story.	start write
	b.	She started to cook with simplicity.	start cook
	c.	The suspect attempted to run off.	attempt run off
(26)	a.	I am going to start writing a book.	start write
	b.	I've really enjoyed working with you.	enjoy work with
	c.	The phones began ringing off the hook.	begin ring off

Note that some verbs (e.g., *start*) allow both an infinitival and a progressive complement (see (25a) and (26a), respectively), whereas other verbs (e.g., *attempt*) allow only one type of complement (see (25c)). Even for verbs that allow both types of complements there exist syntactic contexts where the two complement types are in complementary distribution: *to start writing* occurs 15 times in the BNC, whereas *to start to write* occurs zero times. The situation is reversed for *starting writing* and *starting to write*, for the former occurs zero times and the latter occurs seven times. Choosing to focus only on one type of complement would result in a lower count for $f(v,e)$ than collapsing the counts observed for both types of complements.

Once we have obtained the frequencies $f(v,e)$ and $f(o,e)$ we can determine the most likely interpretations for metonymic verb-noun combinations. Note that we may choose to impose thresholds on the frequencies $f(v,e)$ and $f(o,e)$ (e.g., $f(v,e) > 1$ and $f(o,e) > 1$) depending on the quality of the parsing data or the type of meaning paraphrases we seek to discover (e.g., likely versus unlikely ones).

To give an example of the paraphrases generated by our model consider the sentences in Table 6. These were cited as examples of logical metonymy in the lexical semantics literature (Pustejovsky, 1995; Verspoor, 1997). The five most likely interpretations for these metonymies (and their respective log-transformed probabilities) are illustrated in Table 7. Note that the model comes up with plausible meanings, some of which overlap with those suggested in the lexical semantics literature (underlined interpretations indicate agreement between the model and the literature). Also, the model derives several meanings, as opposed to the single interpretations provided in most cases in the literature. Consider for example the pair *begin story* in Table 7.

Table 7
Model-derived paraphrases for verbal metonymies, ranked in order of likelihood

begin story		begin song		begin sandwich		want job	
tell	−16.34	sing	−15.14	bite into	−18.12	get	−14.87
write	−17.02	rehearse	−16.15	eat	−18.23	lose	−15.72
read	−17.28	write	−16.86	munch	−19.13	take	−16.40
re-tell	−17.45	hum	−17.45	unpack	−19.14	make	−16.52
recount	−17.80	play	−18.01	make	−19.42	create	−16.62
begin book		**enjoy book**		**begin cigarette**		**want cigarette**	
read	−15.49	read	−16.48	smoke	−16.92	smoke	−16.67
write	−15.52	write	−17.58	roll	−17.63	take	−18.23
appear in	−16.98	browse through	−18.56	light	−17.76	light	−18.45
publish	−17.10	look through	−19.68	take	−18.88	put	−18.51
leaf through	−17.35	publish	−19.93	twitch	−19.17	buy	−18.64

Here, not only the interpretation *tell* is generated but also *write*, *read*, *re-tell*, and *recount*. Another example is *begin song* for which the model generates the interpretations *rehearse*, *write*, *hum*, and *play*, besides *sing*.

The model also exhibits slight variation in the interpretations for a given noun among the different metonymic verbs (compare *begin book* and *enjoy book*; *begin cigarette* and *want cigarette* in Table 7). This is in line with claims made in the lexical semantics literature (Copestake and Briscoe, 1995; Pustejovsky, 1995; Verspoor, 1997), and it ultimately contributes to an improved performance against a 'naive baseline' model (see Experiment 2).

In some cases, the model comes up with counterintuitive interpretations: *bite into* is generated as the most likely interpretation for *begin sandwich* (although the latter interpretation is not so implausible, since eating entails biting into). The model also fails to rank *have* as one of the five most likely interpretations for *want job* (see Table 7). The interpretations *get*, and *take*, are however relatively likely; note that they semantically entail the desired interpretation—namely "have"—as a post-state. The interpretations *make*, and *create* imply the act of hiring rather than finding a job. Our model cannot distinguish between the two types of interpretations. It also cannot discover related meanings. For example that *get* and *take* mean "have" or that *tell*, *re-tell* and *recount* (see Table 7) mean "tell". We return to this issue in Section 4.

In the following section we test our model against verb-noun pairs randomly selected from the BNC and evaluate the meaning paraphrases it generates against human judgements. We explore the linear relationship the subjects' rankings and the model-derived probabilities using correlation analysis.

2.3 Experiment 1: Comparison Against Human Judgements

Although there is no standard way to evaluate the paraphrases generated by the model (there is no gold standard for comparison), a reasonable way to judge the model's performance would seem to be agreement with human paraphrase ratings. This can be roughly done by selecting some metonymic constructions, deriving their paraphrase interpretations using the model outlined in Section 2.1, eliciting human judgements on these paraphrases, and then looking at how well the human ratings correlate with the model probabilities for the same paraphrases.

In the following section we describe our method for assembling the set of experimental materials and eliciting human subject data for the metonymy paraphrasing task. We use correlation analysis to compare the model probabilities against human judgements and explore whether there is a linear relationship between the model-derived likelihood of a given meaning and its perceived plausibility.

In Section 2.4.1 we introduce a naive model of verbal metonymy which does not take the contribution of the metonymic verb into account; metonymic interpretations (i.e., verbs) are simply expressed in terms of their conditional dependence on their objects. We investigate the naive model's performance against the human judgements and the paraphrases generated by our initial model (see Section 2.4).

2.3.1 Method

Materials and Design. From the lexical semantics literature (Pustejovsky, 1995; Verspoor, 1997; McElree et al., 2001) we compiled a list of 20 verbs allowing logical metonymy. From these we randomly selected 12 verbs (*attempt, begin, enjoy, finish, expect, postpone, prefer, resist, start, survive, try*, and *want*). The selected verbs ranged in BNC frequency from 10.9 per million to 905.3 per million. Next, we paired each one of them with five nouns randomly selected from the BNC. The nouns had to be attested in the corpus as the object of the verbs in question. Recall that verb-object pairs were identified using Abney's (1996) chunk parser (see Section 2.2 for details). From the retrieved verb-object pairs, we removed all pairs with BNC frequency of one, as we did not want to include verb-noun combinations that were potentially unfamiliar to the subjects. We used the model outlined in Section 2.1 to derive meaning paraphrases for the 60 verb-noun combinations.

Our materials selection procedure abstracts over semantic distinctions which are made in linguistic analyses. For instance, current models of lexical semantics typically assign verbs such as *enjoy* a non-metonymic sense when it combines with NPs which are purely temporal or eventive in nature, as in *enjoy the marriage* or *enjoy the lecture* (Copestake and Briscoe, 1995; Verspoor, 1997). This is largely because a logical form can be constructed in such cases without the use of semantic type coercion; the event-denoting NP itself is the argument to the predicate *enjoy*. We did not rule out such nouns from our materials, however as our evaluation was conducted on randomly selected verb-noun pairs.

More generally, we abstract over several criteria that Verspoor (1997) used in distinguishing metonymic from non-metonymic uses within the corpus, and we adopt a linguistically 'naive' approach for two reasons. First, while Verspoor (1997) could deploy more refined criteria because she was hand-selecting the materials from the corpus and was focusing only on two metonymic verbs (*begin* and *finish*), our materials were randomly sampled and covered a wider range of metonymic constructions. And secondly, paraphrases for non-metonymic cases (e.g., that *enjoy the lecture* can be paraphrased as *enjoy attending the lecture* or *enjoy listening to the lecture*) may be useful for some potential NLP applications (see the discussion in Section 4.2), since they provide more detailed information about meaning than would be given by a logical form which simply features $enjoy(e,x,e')$, where e' is the (event) variable that denotes the lecture.

Recall from Section 2.2 that thresholding is an option for the counts $f(v,e)$ and $f(o,e)$. We derived model paraphrases without employing any thresholds for these counts. Obtaining $f(v,e)$ from the parsed data was relatively straightforward as there was no structural ambiguity involved. As far as $f(o,e)$ is concerned, the parser's output was post-processed to remove potentially erroneous information, so there was no reason to believe that the frequencies $f(v,e)$ and $f(o,e)$ were noisy. Furthermore, recent work has shown that omitting low-frequency tuples degrades performance for language learning tasks such as PP-attachment (Collins and Brooks, 1995; Daelemans, van den Bosch, and Zavrel, 1999), grapheme-to-phoneme conversion, POS-tagging, and NP-chunking (Daelemans, van den Bosch, and Zavrel, 1999). For our task, employing thresholds for $f(v,e)$ and $f(o,e)$ dramatically decreases the number of derived interpretations. Table 8 shows the decrease in the number of interpretations as the cutoff for $f(v,e)$ and $f(o,e)$ is varied for five verb-object pairs that were included in our experimental study. Note that discarding counts occurring in the corpus only once reduces the number of interpretations by a factor of nearly three. Furthermore, applying frequency cutoffs reduces the range of the obtained probabilities, only likely (but not necessarily plausible) interpretations are obtained with $f(o,e) \geq 4$

Table 8
Number of generated interpretations as frequency cutoff for $f(v,e)$ and $f(o,e)$ is varied

Verb-noun	$f(v,e) \geq 1$ $f(o,e) \geq 1$	$f(v,e) \geq 2$ $f(o,e) \geq 2$	$f(v,e) \geq 3$ $f(o,e) \geq 3$	$f(v,e) \geq 4$ $f(n,e) \geq 4$
finish gig	11	4	3	1
finish novel	31	11	5	3
finish project	65	20	8	6
finish room	79	25	16	10
finish video	44	16	9	6

Table 9
Ten most likely interpretations for *finish room* as frequency threshold is varied

$f(v,e) \geq 1$ $f(o,e) \geq 1$		$f(v,e) \geq 2$ $f(o,e) \geq 2$		$f(v,e) \geq 3$ $f(o,e) \geq 3$		$f(v,e) \geq 4$ $f(o,e) \geq 4$	
decorate	−18.47	decorate	−18.47	fill	−18.88	fill	−18.88
wallpaper	−19.07	fill	−18.89	clean	−19.08	clean	−19.08
clean	−19.09	clean	−19.08	pack	−20.17	pack	−20.17
paper	−19.09	search	−20.13	make	−20.36	make	−20.36
furnish	−19.31	pack	−20.17	view	−20.78	check	−21.24
tidy	−19.92	make	−20.36	check	−21.24	use	−21.78
search	−20.13	dress	−20.55	pay	−21.53	build	−21.96
pack	−20.17	view	−20.78	use	−21.78	give	−22.29
make	−20.36	check	−21.24	build	−21.96	prepare	−22.45
view	−20.78	paint	−21.38	give	−22.29	take	−23.11

and $f(v,e) \geq 4$. However, part of the aims of the experiment outlined below was to explore the quality of interpretations with varied probabilities. Table 9 displays the ten most likely paraphrases (and their log-transformed probabilities) for *finish room* as the cutoff for the frequencies $f(v,e)$ and $f(o,e)$ is varied. Notice that applying a cutoff of three or four eliminates plausible interpretations such as *decorate*, *wallpaper*, *furnish*, and *tidy*. This may be particularly harmful for verb-noun (or adjective-noun) combinations that allow for a wide range of interpretations (as is the case for *finish room*).

We estimated the probability $P(e,o,v)$ for each verb-noun pair by varying the term e. In order to generate stimuli covering a wide range of paraphrases corresponding to different degrees of likelihood, for each verb-noun combination we divided the set of generated meanings into three "probability bands" (High, Medium, and Low) of equal size and randomly chose one interpretation from each band. The division ensured that subjects saw a wide range of paraphrases with different degrees of likelihood.

Our experimental design consisted of two factors: verb-noun pair (*Pair*) and probability band (*Band*). The factor *Pair* included 60 verb-noun combinations and the factor *Band* had three levels, High, Medium, and Low. This yielded a total of *Pair* × *Band* = 60 × 3 = 180 stimuli. The 180 stimuli were administered to two separate groups. The first group saw meaning paraphrases for the verbs *attempt, begin, want, enjoy, try*, and *expect*, whereas the second group saw paraphrases for *finish, prefer, resist, start, postpone*, and *survive*. Example stimuli are shown in Table 10.

Each experimental item consisted of two sentences, a sentence containing a metonymic construction (e.g., *Peter started his dinner*) and a sentence paraphrasing it (e.g., *Peter started eating his dinner*). The metonymic sentences and their paraphrases were created by the authors as follows. The selected verb-noun pairs were converted into a simple sentence by adding a sentential subject and articles or pronouns where appropriate. The sentential subjects were familiar proper

Table 10
Randomly selected example stimuli with log-transformed probabilities derived by the model

Verb-noun	Probability Band					
	High		Medium		Low	
attempt peak	climb	−20.22	claim	−23.53	include	−24.85
begin production	organise	−19.09	influence	−21.98	tax	−22.79
enjoy city	live in	−20.77	come to	−23.50	cut	−24.67
expect reward	collect	−21.91	claim	−23.13	extend	−23.52
finish room	wallpaper	−19.07	construct	−22.49	want	−24.60
postpone payment	make	−21.85	arrange	−23.21	read	−25.92
prefer people	talk to	−20.52	sit with	−22.75	discover	−25.26
resist song	whistle	−22.12	start	−24.47	hold	−26.50
start letter	write	−15.59	study	−22.70	hear	−24.50
survive course	give	−22.88	make	−24.48	write	−26.27
try drug	take	−17.81	grow	−22.09	hate	−23.88
want hat	buy	−17.85	examine	−21.56	land on	−22.38

names (BNC corpus frequency $>$ 30 per million) balanced for gender. All sentences were in the past tense. In the paraphrasing sentences the metonymy was spelled out by converting the model's output to a verb taking either a progressive or infinitive VP complement (e.g., *started to eat* or *started eating*). For verbs allowing both a progressive and an infinitive VP complement we chose the type of complement with which the verb occurred more frequently in the corpus. A native speaker of English other than the authors was asked to confirm that the metonymic sentences and their paraphrases were syntactically well-formed (items found syntactically odd were modified and re-tested). Examples of the experimental stimuli the subjects saw are illustrated in (27) and (28). The complete list of the experimental items is given in Appendix B.

(27) a. **H:** Michael attempted the peak Michael attempted to climb the peak
b. **M:** Michael attempted the peak Michael attempted to claim the peak
c. **L:** Michael attempted the peak Michael attempted to include the peak

(28) a. **H:** Jean enjoyed the city Jean enjoyed living in the city
b. **M:** Jean enjoyed the city Jean enjoyed coming to the city
c. **L:** Jean enjoyed the city Jean enjoyed cutting the city

Procedure. The experimental paradigm was Magnitude Estimation (ME), a technique standardly used in psychophysics to measure judgements of sensory stimuli (Stevens, 1975). The ME procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. Highly reliable judgements can be achieved in this fashion for a wide range of sensory modalities, such as brightness, loudness or tactile stimulation.

The ME paradigm has been extended successfully to the psychosocial domain (Lodge, 1981) and recently Bard, Robertson, and Sorace (1996), and Cowart (1997) showed that linguistic judgements can be elicited in the same way as judgements of sensory or social stimuli. ME requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, to which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In this way, each subject can establish their own rating scale, thus yielding maximally fine-grained data and avoiding the known problems with the conventional ordinal scales for linguistic data (Bard, Robertson, and Sorace, 1996; Cowart, 1997; Schütze, 1996). In particular, ME does not restrict the range of the responses. No matter which modulus a subject chooses, he or she can subsequently assign a higher or lower judgement by using multiples or fractions of the modulus.

In the present experiment, each subject took part in an experimental session that lasted approximately 20 minutes. The experiment was self-paced, and response times were recorded to allow the data to be screened for anomalies. The experiment was conducted remotely over the Internet. Subjects accessed the experiment using their web browser, which established an Internet connection to the experimental server running WebExp 2.1 (Keller, Corley, and Scheepers, 2001), an interactive software package for administering web-based psychological experiments. (For a discussion of WebExp and the validity of web-based data, See Appendix A).

Instructions. Before participating in the actual experiment, subjects were presented with a set of instructions. The instructions explained the concept of numerical magnitude estimation of line length. Subjects were instructed to make estimates of line length relative to the first line they would see, the reference line. Subjects were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numerical estimates were provided to illustrate the concept of proportionality.

The subjects were instructed to judge how well a sentence paraphrases another sentence, using the same technique that they had applied to judging line length. Examples of plausible (see (29a)) and implausible (see (29b)) sentence paraphrases were provided, together with examples of numerical estimates.

(29)	a.	Peter started his dinner	Peter started eating his dinner
	b.	Peter started his dinner	Peter started writing his dinner

Subjects were informed that they would initially have to assign a number to a *reference* paraphrase. For each subsequent paraphrase, subjects were asked to assign a number indicating how good or bad that paraphrase is in proportion to the reference.

Subjects were told that they could use any range of positive numbers for their judgements, including decimals. It was stressed that there was no upper or lower limit to the numbers that could be used (exceptions being zero or negative numbers). Subjects were urged to use a wide range of numbers and to distinguish as many degrees of paraphrase plausibility as possible. It was also emphasized that there were no "correct" answers, and that subjects should base their judgements on first impressions, not spending too much time to think about any one paraphrase.

Demographic Questionnaire. After the instructions, a short demographic questionnaire was administered. The questionnaire included name, email address, age, sex, handedness, academic subject or occupation, and language region. Handedness was defined as "the hand you prefer to use for writing", while language region was defined as "the place (town, federal state, country) where you learned your first language".

Training Phase. The training phase was meant to familiarize subjects with the concept of numeric magnitude estimation using line lengths. Items were presented as horizontal lines, centered in the window of the subject's web browser. After viewing an item, the subject had to provide a numerical judgement over the computer keyboard. After pressing Return, the current item disappeared and the next item was displayed. There was no possibility to revisit previous items or change responses once Return had been pressed. No time limit was set for either the item presentation or for the response, although response times were recorded to allow inspection of the data.

Subjects first judged the modulus item, and then all the items in the training set. The modulus was the same for all subjects, and it remained on the screen all the time to facilitate comparison. Items were presented in random order, with a new randomization being generated for each subject.

The training set contained six horizontal lines. The range of the smallest to largest item was 1:10. The items were distributed evenly over this range, with the largest item covering the maximal window width of the web browser. A modulus item in the middle of the range was provided.

Practice Phase. This phase allowed subjects to practice magnitude estimation of verb-noun paraphrases. Presentation and response procedure was the same as in the training phase, with linguistic stimuli being displayed instead of lines. Each subject judged the whole set of practice items, again in random order.

The practice set consisted of eight paraphrase sentences that were representative of the test materials. The paraphrases were based on the three probability bands and represented a wide probability range. A modulus item selected from the Medium probability band was provided.

Experimental Phase. Presentation and response procedure in the experimental phase were the same as in the practice phase. Each subject group saw 90 experimental stimuli (i.e., metonymic sentences and their paraphrases). As in the practice phase, the paraphrases were representative of the three probability bands (i.e., High, Medium, Low). Again a modulus item from the Medium probability band was provided (see Appendix B). The modulus was the same for all subjects and remained on the screen all the time. Subjects were assigned to groups at random, and a random stimulus order was generated for each subject (for the complete list of experimental stimuli see Appendix B).

Subjects. Sixty-three native speakers of English participated in the experiment. The subjects were recruited over the Internet by advertisements posted to newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be linguistically naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of two subjects were eliminated after inspection of their response times showed that they had not completed the experiment in a realistic time frame (i.e., they provided ratings too quickly with average response time < 1000 ms). The data of one subject was excluded because she was a non-native speaker of English.

This left 60 speakers for analysis. Of these, 53 subjects were right-handed, 7 left-handed; 24 subjects were female, 36 male. The age of subjects ranged from 17 to 62; the mean was 26.4 years.

2.3.2 Results The data were first normalized by dividing each numerical judgement by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgements are normally distributed and is standard practice for magnitude estimation data (Bard, Robertson, and Sorace, 1996; Lodge, 1981). All further analyses were conducted on normalized, log-transformed judgements.

We performed a correlation analysis to determine whether there is a linear relation between the paraphrases generated by the model and their perceived likelihood. This tests the hypothesis that meaning paraphrases assigned high probabilities by the model are perceived as better paraphrases by the subjects than meaning paraphrases with low probabilities. For each experimental item we computed the average of the normalized and log-transformed subject ratings. The mean subject ratings were then compared against the (log-transformed) probabilities assigned by the model for the same items.

The comparison between the absolute model probabilities and the human judgements yielded a Pearson correlation coefficient of .64 ($p < .01$, $N = 174$[3]). The mean subject ratings and the

3 The items are 174 instead of 180; 6 items were discarded due a coding error.

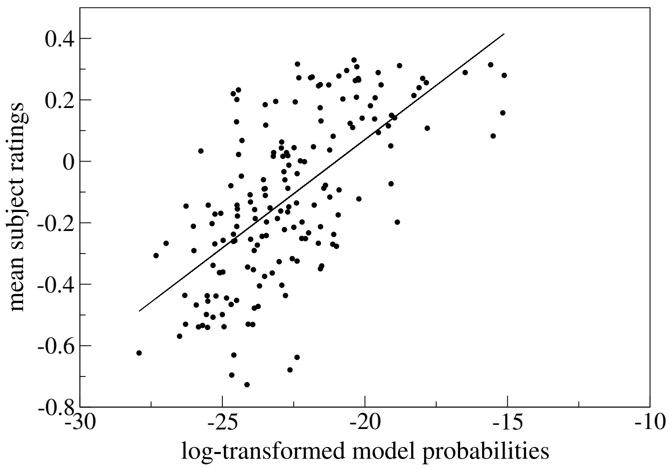


Figure 1
Correlation of elicited judgements and model-derived probabilities for metonymic verb-noun pairs

model probabilities are given in Appendix B. Appendix C contains the descriptive statistics for the model probabilities and the human judgements. The relationship between judgements and probabilities is plotted in Figure 1.

An important question is how well humans agree in their paraphrase judgements for verb-noun combinations. Inter-subject agreement gives an upper bound for the task and allows us to interpret how well the model is doing in relation to humans. To calculate inter-subject agreement we used leave one-out resampling. The technique is a special case of n-fold cross-validation (Weiss and Kulikowski, 1991) and has been previously used for measuring how well humans agree on judging semantic similarity (Resnik and Diab, 2000; Resnik, 1999).

For each subject group we divided the set of the subjects' responses with size m into a set of size $m-1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the mean ratings of the former set with the ratings of the latter. This was repeated m times. Since each group had 30 subjects we performed 30 correlation analyses and report their mean. For the first group of subjects the average inter-subject agreement was .74 (Min = .19, Max = .87, StdDev = .12), and for the second group .73 (Min = .49, Max = .87, StdDev = .09). Our model's agreement with the human data is not far from the average human performance of .74.

In the following section we introduce a naive model of verbal metonymy. We compare the naive model's performance against the human judgements and the paraphrases generated by our initial model. We discuss extensions of the basic model in Section 2.5.1.

2.4 Experiment 2: Comparison against Naive Baseline

2.4.1 Naive Baseline Model In the case of verbal metonymy a naive baseline model can be constructed by simply taking verb-noun co-occurrence data into account, ignoring thus the dependencies between the polysemous verb and its progressive or infinitival VP complements. Consider the sentence *John began the book*. In order to generate appropriate paraphrases for *begin book* we will consider solely the verbs which take book as their object (i.e., *read*, *write*, *buy*, etc.). This can be simply expressed as $P(e|o)$, the conditional probability of a verb e given its object o (i.e., the noun figuring in the metonymic expression) which we estimate as follows:

$$\hat{P}(e|o) = \frac{f(e,o)}{f(o)} \tag{30}$$

The model in (30) treats metonymic verbs as semantically empty, and relies on their object NPs to provide additional semantic information. The counts $f(e,o)$ and $f(o)$ can be easily obtained from the BNC: $f(o)$ amounts to the number of times a noun is attested as an object and $f(e,o)$ are verb-object tuples extracted from the BNC using Cass (Abney, 1996) as described earlier.

2.4.2 Results We used the naive model to calculate the likelihood of the meaning paraphrases that were presented to the subjects (see Experiment 1). Through correlation analysis we explored the linear relationship between the elicited judgements and the naive baseline model. We further directly compared the two models: i.e., our initial, linguistically more informed model and the naive baseline.

Comparison between the probabilities generated by the naive model and the elicited judgements yielded a Pearson correlation coefficient of .42 ($p < .01$, $N = 174$). Recall that our initial model yielded a correlation coefficient of .64. We conducted a one-tailed t-test to determine if the correlation coefficients were significantly different. The comparison revealed that their difference was statistically significant ($t(171) = 1.67$, $p < .05$), indicating that our model performs reliably better than the naive baseline. Comparison between the two models (our initial model introduced in Section 2.1 and the naive baseline model) yielded an intercorrelation of .46 ($p < .01$, $N = 174$). These differences between the 'full' probabilistic model and the naive baseline confirm claims made in the literature: different metonymic verbs have a different *semantic* impact on the resolution of metonymy.

2.5 Experiment 3: Comparison Against Norming Data

Our previous experiments focused on evaluating the plausibility of meaning paraphrases generated by a model which does not take into account the contribution of the sentential subject. However, properties of the subject NP appear to influence the interpretation of the metonymic expression in otherwise neutral contexts. This is illustrated in (31) where the interpretation of *enjoy the book* is influenced by the sentential subject: authors usually write books, whereas critics usually review them.

(31) a. The critic enjoyed the book.
b. The author enjoyed the book.

In this section we present an extension of the basic model outlined in Section 2.1 which takes sentential subjects into account. We evaluate the derived paraphrases and their likelihood again by comparison with human data. This time we compare our model against paraphrase data generated independently by subjects that participated in an experimental study (McElree et al., 2001) that was not designed specifically to test our model.

2.5.1 The Extended Model We model the meaning of sentences like (31) again as the joint distribution of the following variables: the metonymic verb v, its subject s, its object o, and the implicit interpretation e. By choosing the ordering $\langle e, v, s, o \rangle$, we can factor $P(e, o, s, v)$ as follows:

$$P(e,o,s,v) = P(e) \cdot P(v|e) \cdot P(s|e,v) \cdot P(o|e,v,s) \tag{32}$$

The terms $P(e)$ and $P(v|e)$ are easy to estimate from the BNC. For $P(e)$ all we need is a pos-tagged corpus and $P(v|e)$ can be estimated from Cass's output (see equations (12) and (13)).

The estimation of the terms $P(s|e,v)$ and $P(o|e,v,s)$ is however problematic as they rely on the frequencies $f(s,e,v)$ and $f(o,e,v,s)$, respectively. Recall that there is a discrepancy between a metonymic interpretation and its usage. As we've discussed earlier, the underlying interpretation is not overtly expressed in the corpus. Furthermore, the only type of data available to us for the estimation of $P(s|e,v)$ and $P(o|e,v,s)$ is the partially parsed BNC which is not annotated with information regarding the interpretation of metonymies. This means that $P(s|e,v)$ and $P(o|e,v,s)$ need to be somehow approximated. We first assume that the sentential subject s is conditionally independent of the metonymic verb v; secondly we assume that the sentential object o is conditionally independent of v and s:

$$P(s|e,v) \approx P(s|e) \tag{33}$$

$$P(o|e,v,s) \approx P(o|e) \tag{34}$$

The rationale behind the approximation in (33) is that the likelihood of a noun s being a subject of a verb e is largely independent of whether e is the complement of a metonymic verb v. For example, authors usually write, irrespectively of whether they enjoy, dislike, start, or finish doing it. The motivation for the approximation in (33) comes from the observation that an object is more closely related to the verb that selects for it. We are likely to come up with *book* or *letter* for o if we know that o is the object of *read* or *write*. Coming up with an object for o is not so straightforward if all we know is the metonymic verb (e.g., *enjoy*, *finish*) or its sentential subject. It is the verbs that impose semantic restrictions on their arguments rather than other sentential constituents. We estimate $P(s|e)$ and $P(o|e)$ using maximum likelihood:

$$\hat{P}(s|e) = \frac{f(s,e)}{f(e)} \tag{35}$$

$$\hat{P}(o|e) = \frac{f(o,e)}{f(e)} \tag{36}$$

The count $f(s,e)$ amounts to the number of times a noun s is attested as the subject of a verb e; $f(o,e)$ represents nouns attested as objects of e. Verb-argument tuples can be easily extracted from the BNC using Cass (see Section 2.2 for details). Table 11 illustrates the model's performance for metonymic constructions in (37). We only show the five most likely interpretations the model came up with. Interestingly, different interpretations are derived for different subjects. Even though pianists and composers are semantically related, pianists are more likely to begin playing a symphony, whereas composers are more likely to conduct or write a symphony. Similarly, builders tend to renovate houses and architects tend to design them.

(37) a. The composer/pianist began the symphony.
b. The author/student started the book.
c. The builder/architect started the house.
d. The secretary/boss finished the memo.

In the following section we compare the interpretations generated by the model against paraphrases provided by humans. More specifically, we explore whether there is a linear relationship between the frequency of an interpretation as determined in a norming study and the probability of the same interpretation as calculated by the model.

Table 11
Subject-related model interpretations, ranked in order of likelihood

begin symphony				start book			
composer		pianist		author		student	
write	−22.2	play	−24.20	write	−14.87	read	−16.12
conduct	−23.79	hear	−25.38	publish	−16.94	write	−16.48
hear	−25.38	give	−28.44	compile	−17.84	study	−17.59
play	−25.81	do	−29.23	read	−17.98	research	−18.86
create	−25.96	have	−30.13	sign	−18.59	translate	−17.85
start house				finish memo			
builder		architect		secretary		boss	
renovate	−15.43	design	−16.87	write	−19.79	draft	−20.73
build	−17.56	build	−17.20	type	−20.13	send	−21.97
demolish	−18.37	restore	−19.08	send	−20.87	sign	−22.04
dismantle	−19.69	purchase	−19.32	sign	−22.01	hand	−22.12
erect	−19.81	site	−19.73	make	−22.74	write	−22.74

2.5.2 Method For the experiment described in this section we used the norming data reported in McElree et al. (2001). In McElree et al.'s study subjects were given sentence fragments such as (38) and were asked to complete them. Potential completions for fragment (38a) include *writing* or *reading*. The study consisted of 142 different sentences similar to those shown in (38) and included 15 metonymic verbs. 30 sentences were constructed for each of the metonymic verbs *start*, *begin*, *complete*, and *finish* and a total of 22 sentences for *attempt, endure, expect, enjoy, fear, master, prefer, resist, savor, survive*, and *try*.

(38) a. The writer finished ___ the novel.
b. The soldier attempted ___ the mountain.
c. The teenager finished ___ the novel.

The completions can be used to determine interpretation preferences for the metonymic constructions simply by counting the verbs that human subjects use to complete sentences like (38). For example, five completions were provided by the subjects for fragment (38b): *climb*, *hike*, *scale*, *walk* and *take*. Of these climb was by far the most likely with 78 (out of 88) subjects generating this interpretation.[4] The most likely interpretations for (38a) and (38c) were respectively *write* (13 out of 28 subjects) and *read* (18 out of 22).

For each of the sentences included in McElree et al.'s (2001) study we derived interpretation paraphrases using the model presented in Section 2.5.1. We next compared the interpretations common in the model and the human data.

2.5.3 Results In Experiment 1 we evaluated the paraphrases generated by the model by eliciting plausibility judgements from subjects and showed that our model produces an intuitively plausible ranking of meanings. Here, we evaluate the quality of the produced paraphrases by directly comparing them to norming data acquired independently of our model and the particular corpus we are using.

The comparison between (log-transformed) model probabilities and (log-transformed) completion frequencies yielded a Pearson correlation coefficient of .422 ($p < .01$, $N = 341$). We also

4 McElree et al.'s (2001) linguistic materials were manually constructed and not controlled for frequency. For example one would expect (38b) to be relatively rare, even in a large corpus. This is true for the BNC where the combination *attempt mountain* is not attested at all.

compared the completion frequencies against interpretation probabilities derived using the model presented in Section 2.3 which does not take subject-related information into account. The comparison yielded a correlation coefficient of .216 ($p < .01$, $N = 341$). We carried out a one-tailed t-test to determine if the difference between the two correlation coefficients is significant. The comparison revealed that the difference was statistically significant ($t(338) = 2.18$, $p < .05$). This means that the fit between the norming data and the model is better when the model explicitly incorporates information about the sentential subject. The two models are as expected intercorrelated ($r = .264$, $p < .01$, $N = 341$).

2.6 Discussion

We have demonstrated that the meanings acquired by our probabilistic model correlate reliably with human intuitions. These meanings go beyond the examples found in the theoretical linguistics literature. The verb-noun combinations we interpret were randomly sampled from a large balanced corpus providing a rich inventory for their meanings. We have shown that the model has four defining features: (a) it is able to derive intuitive meanings for verb-noun combinations, (b) it generates clusters of meanings (following Vendler's (1968) insight), (c) it predicts variation in interpretation among the different nouns: the same verb may carry different meanings depending on its subject or object (compare *begin book* vs. *begin house* and *the author began the house* vs. *the architect began the house*), and (d) it represents variation in interpretation among the different metonymic verbs (e.g., *begin book* vs. *enjoy book*). This latter property demonstrates that although the model does not explicitly encode linguistic constraints for resolving metonymies, it generates interpretations that broadly capture linguistic differences (e.g., *attempt* imposes different constraints on interpretation from *begin* or *enjoy*). Furthermore, these interpretations for metonymic verb-noun pairs are discovered automatically, without presupposing the existence of a pre-defined taxonomy or a knowledge base.

Note that the evaluation procedure to which we subject our model is rather strict. The derived verb-noun combinations were evaluated by subjects naive to linguistic theory. Although verbal logical metonymy is a well researched phenomenon in the theoretical linguistics literature, the experimental approach advocated here is new to our knowledge. Comparison between our model and human judgements yielded a reliable correlation of .64 when the upper bound for the task (i.e., inter-subject agreement) is on average .74. Furthermore, our model performed reliably better than a naive baseline model, which only achieved a correlation of .42. When compared against norming data, an extended version of our model which takes subject information into account reached a correlation of .42. Comparison against norming data is a strict test on unseen data that was not constructed explicitly to evaluate our model but is independently motivated and does not take our corpus (i.e., the BNC) or our particular task into account.

We next investigate whether such an approach generalizes to other instances of logical metonymy by looking at adjective-noun combinations. Adjectives pose a greater challenge for our modeling task as they can potentially allow for a wider range of interpretations and can exhibit preferences for a verb-subject or verb-object paraphrase (see Section 1). Following the approach we adopted for verbal metonymy, we define the interpretation of polysemous adjective-noun combinations as a paraphrase generation task. We provide a probabilistic model which not only paraphrases adjective-noun pairs (e.g., *fast plane*) with a related verb (e.g., *fly*) but also predicts whether the noun modified by the adjective (e.g., *plane*) is likely to be the verbal object or subject. The model achieves this by combining distributional information about how likely it is for any verb to be modified by the adjective in the adjective-noun combination or its corresponding adverb with information about how likely it is for any verb to take the modified noun as its object or subject. We obtain quantitative information about verb-adjective modification and verb-argument relations from the BNC and evaluate our results by comparing the model's predictions against human judgements. Consistent with our results on verbal metonymy, we show that the model's ranking of meanings correlates reliably with human intuitions.

3 Metonymic Adjectives

3.1 The Model

Consider again the adjective-noun combinations in (39). In order to come up with the interpretation of plane that flies quickly for *fast plane* we would like to find in the corpus a sentence whose subject is the noun *plane* or *planes* and whose main verb is *fly*. We would also expect *fly* to be modified by the adverb *fast* or *quickly*. In the general case, we would like to gather from the corpus sentences indicating what planes do fast. Similarly, for the adjective-noun combination *fast scientist* we would like to find in the corpus information indicating what the activities that scientists perform fast are, whereas for *easy problem* we need information about what one can do with problems easily (e.g., one can solve problems easily) or about what problems are (e.g., easy to solve or easy to set).

(39) a. fast plane
b. fast scientist
c. fast programmer
d. easy problem

In sum, in order to come up with a paraphrase of the meaning of an adjective-noun combination we need to know which verbs take the head noun as their subject or object and are modified by an adverb corresponding to the modifying adjective. This can be expressed as the joint probability $P(a,e,n,rel)$ where e is the verbal predicate modified by the adverb a (directly derived from the adjective present in the adjective-noun combination) bearing the argument relation *rel* (i.e., subject or object) to the head noun n. By choosing the ordering $\langle e,n,a,rel\rangle$ for the variables a,e,n, and *rel* we can rewrite $P(a,e,n,rel)$, using the chain rule, as follows:

$$P(a,e,n,rel) = P(e)\cdot P(n|e)\cdot P(a|e,n)\cdot P(rel|e,n,a) \qquad (40)$$

Although the terms $P(e)$ and $P(n|e)$ can be straightforwardly estimated from the BNC (see (12) for $P(e)$; $P(n|e)$ can be obtained by counting the number of times a noun n co-occurs with a verb e either as its subject or object), the estimation of $P(a|e,n)$ and $P(rel|e,n,a)$ faces problems similar to those for metonymic verbs. Let us consider more closely the term $P(rel|e,n,a)$ which can be estimated as shown in (41) below.

$$\hat{P}(rel|e,n,a) = \frac{f(rel,e,n,a)}{f(e,n,a)} \qquad (41)$$

One way to obtain $f(rel,e,n,a)$ would be to fully parse the corpus so as to identify the verbs which take the head noun n as their subject or object and are modified by the adverb a, assuming it is equally likely to find in a corpus the metonymic expression (e.g., *fast plane*) and its paraphrase interpretation (i.e., plane that flies quickly). As in the case of verb-noun metonymies this assumption is unjustified: for the adjective-noun combination *fast plane* there are only six sentences in the entire BNC that correspond to $f(rel,e,n,a)$. According to the sentences in (42) the most likely interpretation for *fast plane* is plane that goes fast (see examples (42a)–(42c)). The interpretations plane that swoops in fast, plane that drops down fast and plane that flies fast are all equally likely, since they are attested in the corpus only once (see examples (42d)–(42f)). This is rather unintuitive since *fast planes* are more likely to fly than swoop in fast. Similar problems affect the frequency $f(e,n,a)$.

(42) a. The plane **went** so fast it left its sound behind.
b. And the plane's **going** slightly faster than the Hercules or Andover.
c. He is driven by his ambition to build a plane that **goes** faster than the speed of sound.

d. Three planes **swooped in**, fast and low.
e. The plane was **dropping down** fast towards Bangkok.
f. The unarmed plane **flew** very fast and very high.

In default of a corpus explicitly annotated with interpretations for metonymic adjectives, we will make the following independence assumptions:

$$P(a|e,n) \approx P(a|e) \tag{43}$$

$$P(rel|e,n,a) \approx P(rel|e,n) \tag{44}$$

The rationale behind the approximation in (43) is that the likelihood of seeing an adverbial a modifying a verb e bearing an argument relation to a noun n is largely independent of that specific noun. For example, flying can be carried out fast or slowly or beautifully irrespectively of whether it is a pilot or a bird who is doing the flying. Similarly, the adverb peacefully is more related to dying than killing or injuring irrespectively of who the agent of these actions is. Accordingly, we assume that the argument relation rel is independent of whether the verb e (standing in relation rel with noun n) is modified by an adverb a (see (44)). In other words, it is the verb e and its argument n that determine the relation rel rather than the adjective/adverb a. Knowing that flying is conducted slowly will not affect the likelihood of inferring a subject relation for *plane* and *fly*. Yet, we are likely to infer an object relation for *plane* and *construct* irrespectively of whether the constructing is done slowly, quickly, or automatically. We estimate the probabilities $P(e)$, $P(n|e)$, $P(a|e)$, and $P(rel|e,n)$ as follows:

$$\hat{P}(e) = \frac{f(e)}{N} \tag{45}$$

$$\hat{P}(n|e) = \frac{f(n,e)}{f(e)} \tag{46}$$

$$\hat{P}(a|e) = \frac{f(a,e)}{f(e)} \tag{47}$$

$$\hat{P}(rel|e,n) = \frac{f(rel,e,n)}{f(e,n)} \tag{48}$$

By substituting equations (45)–(48) into (41) and simplifying the relevant terms, (41) is rewritten as follows:

$$P(a,e,n,rel) = \frac{f(rel,e,n) \cdot f(a,e)}{f(e) \cdot N} \tag{49}$$

Assume we want to discover a meaning paraphrase for the adjective-noun combination *fast plane*. We need to find the verb e and the relation rel (i.e., subject or object) that maximize the term $P(fast,e,plane,rel)$. Table 12 gives a list of the most frequent verbs modified by the adverb *fast* in the BNC (see the term $f(a,e)$ in equation (49)) and Table 13 lists the verbs for which the noun *plane* is the most likely object or subject (see the term $f(rel,e,n)$ in equation (49)). We describe how the frequencies $f(rel,e,n)$, $f(a,e)$, and $f(e)$ were estimated from a lemmatised version of the BNC in the following section.

Table 12 can be thought of as a list of the activities that can be fast (i.e., going, growing, flying), whereas Table 13 specifies the events associated with the noun *plane*. Despite our simplifying assumptions the model given in (49) will come up with plausible meanings for adjective-noun combinations like *fast plane*. Note that the verbs *fly*, *come*, and *go* are most likely to take

Table 12
Most frequent verbs modified by the adverb *fast*

$f(fast,e)$		$f(fast,e)$	
go	29	work	6
grow	28	grow in	6
beat	27	learn	5
run	16	happen	5
rise	14	walk	4
travel	13	think	4
move	12	keep up	4
come	11	fly	4
drive	8	fall	4
get	7	disappear	4

Table 13
Most frequent verbs taking as an argument the noun *plane*

$f(\text{SUBJ},e,plane)$		$f(\text{OBJ},e,plane)$	
fly	20	catch	24
come	17	board	15
go	15	take	14
take	14	fly	13
land	9	get	12
touch	8	have	11
make	6	buy	10
arrive	6	use	8
leave	5	shoot	8
begin	5	see	7

the noun *plane* as their subject (see Table 13). These verbs also denote activities that are fast (see Table 12 where the underlined verbs are events that are associated both with the adverb *fast* and the noun *plane*). Further note that a subject interpretation is more likely than an object interpretation for *fast plane* since none of the verbs likely to have *plane* as their object are modified by the adverb *fast* (compare Tables 12 and 13).

As in the case of metonymic verbs, the probabilistic model outlined above acquires meanings for polysemous adjective-noun combinations in an unsupervised manner without presupposing annotated corpora or taxonomic information. The obtained meanings are not discourse-sensitive; they can be thought of as default semantic information associated with a particular adjective-noun combination. This means that our model is unable to predict that programmer that runs fast is a likely interpretation for *fast programmer* when the latter is in a context like the one given in (5) repeated here as (50).

(50) a. All the office personnel took part in the company sports day last week.
b. One of the programmers was a good athlete, but the other was struggling to finish the courses.
c. The fast programmer came first in the 100m.

3.2 Parameter Estimation

As in the case of verbs the parameters of the model were estimated using a part-of-speech tagged and lemmatised version of the BNC. The counts $f(e)$ and N (see (49)) reduce to the number of times a given verb is attested in the corpus. The frequency $f(rel,e,n)$ was obtained using Abney's (1996) chunk parser (see Section 2.2 for details).

Generally speaking, the frequency $f(a,e)$ represents not only a verb modified by an adverb derived from the adjective in question (see example (51a)) but also constructions like the ones shown in (51b,c), where the adjective takes an infinitival VP complement whose logical subject can be realized as a *for*-PP (see example (51c)). In cases of verb-adverb modification we assume access to morphological information which specifies what counts as a valid adverb for a given adjective. In most cases adverbs are formed by adding the suffix *-ly* to the base of the adjective (e.g., *slow-ly*, *easy-ly*). Some adjectives have identical adverbs (e.g., *fast*, *right*). Others have idiosyncratic adverbs (e.g., the adverb of *good* is *well*). It is relatively straightforward to develop an automatic process which maps an adjective to its corresponding adverb, modulo exceptions and indiosyncracies, however in the experiments described in the following sections this mapping was manually specified.

(51) a. comfortable chair → a chair *on* which one *sits comfortably*
b. comfortable chair → a chair that is *comfortable* to *sit on*
c. comfortable chair → a chair that is *comfortable* for me to *sit on*

In cases where the adverb does not immediately succeed the verb, the parser is not guaranteed to produce a correct analysis since it does not resolve structural ambiguities. So we adopted a conservative strategy, where to obtain the frequency $f(a,e)$ we only looked at instances where the verb and the adverbial phrase (AdvP) modifying it were adjacent. More specifically, in cases where the parser identified an AdvP following a VP, we extracted the verb and the head of the AdvP (see the examples in (52)). In cases where the AdvP was not explicitly identified we extracted the verb and the adverb immediately following or preceding it (see the examples in (53)) assuming that the verb and the adverb stand in a modification relation. The examples below illustrate the parser's output and the information that was extracted for the frequency $f(a,e)$.

(52) a. [$_{NP}$ Oriental art] [$_{VP}$ came] [$_{AdvP}$ more slowly.] come slowly
b. [$_{NP}$ The issues] [$_{VP}$ will not be resolved] [$_{AdvP}$ easily.] resolve easily
c. [$_{NP}$ Arsenal] [$_{VP}$ had been pushed] [$_{AdvP}$ too hard.] push hard
(53) a. [$_{NP}$ Some art historians] [$_{VP}$ write] well [$_{PP}$ about the present.] write well
b. [$_{NP}$ The accidents] [$_{VP}$ could have been easily avoided.] avoid easily
c. [$_{NP}$ A system of molecules] [$_{VP}$ is easily shown] [$_{VP}$ to stay constant.] show easily
d. [$_{NP}$ Their economy] [$_{VP}$ was so well run.] run well

Adjectives with infinitival complements (see (51b,c)) were extracted from the parser's output. We concentrated solely on adjectives immediately followed by infinitival complements with an optionally intervening *for*-PP (see (51c)). The adjective and the main verb of the infinitival complement were counted as instances of the quantity $f(a,e)$. The examples in (54) illustrate the process.

(54) a. [$_{NP}$ These early experiments] [$_{VP}$ were easy] [$_{VP}$ to interpret.] easy interpret
b. [$_{NP}$ It] [$_{VP}$ is easy] [$_{PP}$ for an artist] [$_{VP}$ to show work independently.] easy show
c. [$_{NP}$ It] [$_{VP}$ is easy] [$_{VP}$ to show] [$_{VP}$ how the components interact.] easy show

Finally, the frequency $f(a,e)$ collapsed the counts from cases where the adjective was followed by an infinitival complement (see the examples in (54)) and cases where the verb was modified by the adverb corresponding to the related adjective (see the examples in (52)–(53)). For example, assume that we are interested in the frequency $f(easy, show)$. In this case, we will take into account not only sentences (54b,c) but also sentence (53b). Assuming this was the only evidence in the corpus, the frequency $f(easy, show)$ would be three.

Once we have obtained the frequencies $f(a,e)$ and $f(rel,e,n)$ we can determine what the most likely interpretations for a given adjective-noun combination are. If we know the interpretation preference of a given adjective (i.e., subject or object), we may vary only the term e in $P(a,n,rel,e)$, keeping the terms n, a and rel constant. Alternatively, we could acquire the interpretation preferences automatically by varying both the terms rel and e. In Experiment 4 (see Section 3.3) we acquire both meaning paraphrases and argument preferences for polysemous adjective-noun combinations.

In what follows we illustrate the properties of the model by applying it to a small number of adjective-noun combinations displayed in Table 14. The adjective-noun combinations and their respective interpretations are taken from the lexical semantics literature (i.e., Pustejovsky 1995 and Vendler 1968). The five most likely model-derived paraphrases for these combinations are shown in Tables 15 and 16.

The model comes up with plausible meanings, some of which overlap with those suggested in the lexical semantics literature (underlined interpretations indicate agreement between the

Table 14
Paraphrases for adjective-noun combinations taken from the literature

easy problem	→ a problem that is easy to solve	(Vendler, 1968, 97)
easy text	→ text that reads easily	(Vendler, 1968, 99)
difficult language	→ a language that is difficult to speak, learn, write, understand	(Vendler, 1968, 99)
careful scientist	→ a scientist who observes, performs, runs experiments carefully	(Vendler, 1968, 92)
comfortable chair	→ a chair on which one sits comfortably	(Vendler, 1968, 98)
good umbrella	→ an umbrella that functions well	(Pustejovsky, 1995, 43)

Table 15
Object-related interpretations for adjective-noun combinations, ranked in order of likelihood

easy problem		easy text		difficult language		comfortable chair		good umbrella	
solve	−15.14	read	−17.42	understand	−17.15	sink into	−18.66	keep	−21.59
deal with	−16.12	handle	−18.79	interpret	−17.59	sit on	−19.13	wave	−21.61
identify	−16.83	use	−18.83	learn	−17.67	lounge in	−19.15	hold	−21.73
tackle	−16.92	interpret	−19.05	use	−17.79	relax in	−19.33	run for	−21.73
handle	−16.97	understand	−19.15	speak	−18.21	nestle in	−20.51	leave	−22.28

Table 16
Subject-related interpretations for adjective-noun combinations, ranked in order of likelihood

easy text		good umbrella		careful scientist	
see	−19.22	cover	−23.05	calculate	−22.31
read	−19.50			proceed	−22.67
understand	−19.66			investigate	−22.78
achieve	−19.71			study	−22.90
explain	−20.40			analyse	−22.92

model and the literature). Observe that the model predicts different meanings when the same adjective modifies different nouns and derives a cluster of meanings for a single adjective-noun combination. An *easy problem* is not only a problem that is easy to solve (see Vendler's (1968) identical interpretation in Table 14) but also a problem that is easy to deal with, identify, tackle, and handle. The meaning of *easy problem* is different from the meaning of *easy text* which in turn is easy to read, handle, interpret, and understand. The interpretations the model arrives at for *difficult language* are a superset of the interpretations suggested by Vendler (1968). The model comes up with the additional meanings language that is difficult to interpret and language that is difficult to use. Although the meanings acquired by the model for *careful scientist* do not overlap with the ones suggested by Vendler (1968) they seem intuitively plausible: a *careful scientist* is a scientist who calculates, proceeds, investigates, studies, and analyses carefully. These are all possible events associated with scientists.

The model derives subject and object-related interpretations for *good umbrella* which covers well and is good to keep, good for waving, good to hold, good to run for, and good to leave. A subject interpretation can be also derived for *easy text*. Our parser treats inchoative and non-inchoative uses of the same verb as distinct surface structures (e.g., *text that one reads easily* vs. *text that reads easily*); as a result, *read* is generated as a subject- and object-paraphrase for *easy text* (compare Table 15 and 16). The object-related interpretation is nevertheless given a higher probability in comparison to the subject-related one, which seems intuitively correct (there is an understood but unexpressed agent in structures like *text that reads easily*). In general, subject and object interpretations are derived on the basis of verb-subject and verb-object constructions which have been extracted from the corpus heuristically without taking into account information about theta-roles, syntactic transformations (with the exception of passivization), or diathesis alternations such as the middle or causative/inchoative alternation.

Although the model can be used to provide several interpretations for a given adjective-noun combination, not all of these interpretations are useful or plausible (see the subject interpretations for *easy text*). Also, the meanings acquired by our model are a simplified version of the ones provided in the lexical semantics literature. An adjective-noun combination may be paraphrased with another adjective-noun combination (e.g., a *good meal* is a tasty meal) or with an NP instead of an adverb (e.g., a *fast decision* is a decision that takes a short amount of time). We are making the simplifying assumption that a polysemous adjective-noun combination can be paraphrased by a sentence consisting of a verb whose argument is the noun the adjective is in construction with (cf. earlier discussion concerning non-metonymic uses of verbs like *enjoy*).

In the following section we evaluate the meaning paraphrases generated by the model against human judgements. As in the case of verbs, the model is tested on examples randomly sampled from the BNC and the linear relationship between the subjects' rankings and the model-derived probabilities is explored using correlation analysis. In Section 3.5.1 we assess whether our model outperforms a naive baseline in deriving interpretations for metonymic adjectives.

3.3 Experiment 4: Comparison against Human Judgements

The experimental method was the same as in Experiment 1. Meaning paraphrases for adjective-noun combinations were obtained using the model introduced in Section 3.1. The model's rankings were compared against paraphrase judgements elicited experimentally from human subjects. The comparison between model probabilities and their perceived likelihood enabled us to explore (a) whether there is a linear relationship between the likelihood of a given meaning as derived by the model and its perceived plausibility and (b) whether the model can be used to derive the argument preferences for a given adjective, i.e., whether the adjective is biased towards a subject or object interpretation or whether it is equi-biased.

3.3.1 Method

Materials and Design. We chose nine adjectives according to a set of minimal criteria and paired each adjective with 10 nouns randomly selected from the BNC. We chose the adjectives as follows: we first compiled a list of all the polysemous adjectives mentioned in the lexical semantics literature (Vendler, 1968; Pustejovsky, 1995). From these we randomly sampled nine adjectives (*difficult, easy, fast, good, hard, right, safe, slow*, and *wrong*). These adjectives had to be relatively unambiguous syntactically: in fact, each adjective was unambiguously tagged as "adjective" 98.6% of the time, measured as the number of different part-of-speech tags assigned to the word in the BNC. The nine selected adjectives ranged in BNC frequency from 57.6 per million to 1,245 per million.

Adjective-noun pairs were extracted from the parser's output. Recall that the BNC was parsed using Abney's (1996) chunker (see Sections 2.2 and 3.2 for details). From the syntactic analysis provided by the parser we extracted a table containing the adjective and the head of the noun phrase following it. In the case of compound nouns, we only included sequences of two nouns, and considered the rightmost occurring noun as the head. From the retrieved adjective-noun pairs, we removed all pairs with BNC frequency of one, as we wanted to reduce the risk of paraphrase ratings being influenced by adjective-noun combinations unfamiliar to the subjects. Furthermore, we excluded pairs with deverbal nouns (i.e., nouns derived from a verb) such as *fast programmer* since an interpretation can be easily arrived at for these pairs by mapping the deverbal noun to its corresponding verb. A list of deverbal nouns was obtained from two dictionaries, CELEX (Burnage, 1990) and NOMLEX (Macleod et al., 1998).

Table 17
Randomly selected example stimuli with log-transformed probabilities derived by the model

Adjective-noun	Probability Band					
	High		Medium		Low	
difficult customer	satisfy	−20.27	help	−22.20	drive	−22.64
easy food	cook	−18.94	introduce	−21.95	finish	−23.15
fast pig	catch	−23.98	stop	−24.30	use	−25.66
good postcard	send	−20.17	draw	−22.71	look at	−23.34
hard number	remember	−20.30	use	−21.15	create	−22.69
right school	apply to	−19.92	complain to	−21.48	reach	−22.90
safe drug	release	−22.24	try	−23.38	start	−25.56
slow child	adopt	−19.90	find	−22.50	forget	−22.79
wrong colour	use	−21.78	look for	−22.78	look at	−24.89

We used the model outlined in Section 3.1 to derive meaning paraphrases for the 90 adjective-noun combinations. We employed no threshold on the frequencies $f(e,a)$ and $f(rel,e,n)$. The frequency $f(e,a)$ was obtained by mapping the adjective to its corresponding adverb: the adjective *good* was mapped to the adverbs *good* and *well*, the adjective *fast* to the adverb *fast*, *easy* was mapped to *easily*, *hard* was mapped to the adverb *hard*, *right* to *rightly* and *right*, *safe* to *safely* and *safe*, *slow* to *slowly* and *slow*, and *wrong* to *wrongly* and *wrong*. The adverbial function of the adjective *difficult* is expressed only periphrastically (i.e., *in a difficult manner*, *with difficulty*). As a result we obtained the frequency $f(difficult,e)$ only on the basis of infinitival constructions (see the examples in (54)). We estimated the probability $P(a,n,rel,e)$ for each adjective-noun pair by varying both the terms e and rel. We thus derived both subject-related and object-related paraphrases for each adjective-noun pair.

For each adjective-noun combination the set of the derived meanings was again divided into three "probability bands" (High, Medium, and Low) of equal size and one interpretation was selected from each band. The division ensured that the experimental stimuli represented the model's behaviour for likely and unlikely paraphrases. We performed separate divisions for object-related and subject-related paraphrases resulting in a total of six interpretations for each adjective-noun combination, as we wanted to determine whether there are differences in the model's predictions with respect to the argument function (i.e., object or subject) and also because we wanted to compare experimentally-derived adjective biases against model-derived biases. Example stimuli (with object-related interpretations only) are shown in Table 17 for each of the nine adjectives.

Our experimental design consisted of the factors adjective-noun pair (*Pair*), grammatical function (*Func*) and probability band (*Band*). The factor *Pair* included 90 adjective-noun combinations. The factor *Func* had two levels (subject and object), whereas the factor *Band* had three levels (High, Medium, and Low). This yielded a total of $Pair \times Func \times Band = 90 \times 2 \times 3 = 540$ stimuli. The number of the stimuli was too large for subjects to judge in one experimental session. We limited the size of the design by selecting a total of 270 stimuli according to the following criteria: our initial design created two sets of stimuli, 270 subject-related stimuli and 270 object-related stimuli. For each set of stimuli (i.e., object- and subject-related) we randomly selected five nouns for each of the nine adjectives together with their corresponding interpretations in the three probability bands (High, Medium, Low). This yielded a total of $Pair \times Func \times Band = 45 \times 2 \times 3 = 270$ stimuli. This way, stimuli were created for each adjective in both subject-related and object-related interpretations.

We administered the 270 stimuli to two separate subject groups. Each group saw 135 stimuli consisting of interpretations for all adjective-noun pairs. For the first group five adjectives were represented by object-related meanings only (*difficult*, *easy*, *good*, *hard*, *slow*); these adjectives were presented to the second group with subject-related interpretations only. Correspondingly, for the first group four adjectives were represented by subject-related meanings only (*safe*, *right*, *wrong*, *fast*); the second group saw these adjectives with object-related interpretations.

Each experimental item consisted of an adjective-noun pair and a sentence paraphrasing its meaning. Paraphrases were created by the experimenters by converting the model's output to a simple phrase, usually a noun modified by a relative clause. A native speaker of English other than the authors was asked to confirm that the paraphrases were syntactically well-formed (items found syntactically odd were modified and re-tested). Example stimuli are shown in (55). The complete list of the experimental items is given in Appendix B.

(55) a. **H:** difficult customer a customer who is difficult to satisfy
b. **M:** difficult customer a customer who is difficult to help
c. **L:** difficult customer a customer who is difficult to drive

(56) a. **H:** fast horse a horse that runs fast
b. **M:** fast horse a horse that works fast
c. **L:** fast horse a horse that sees quickly

Procedure. The method used was magnitude estimation, with the same experimental protocol as in Experiment 4.

Instructions, demographic questionnaire, and training phase. The instructions were the same as in Experiment 1 with the exception that this time the subjects were asked to judge how well a sentence paraphrases an adjective-noun combination. The demographic questionnaire and the training phase were the same as in Experiment 1.

Experimental Phase. Each subject group saw 135 metonymic sentences and their paraphrases. A modulus item from the Medium probability band was provided (see Appendix B). The modulus was the same for all subjects and remained on the screen all the time. Subjects were assigned to groups at random, and a random stimulus order was generated for each subject.

Subjects. Sixty-five native speakers of English participated in the experiment.[5] The subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists. Participation was voluntary and unpaid.

The data of one subject were eliminated after inspection of his response times showed that he had not completed the experiment in a realistic time frame (average response time < 1000ms). The data of four subjects were excluded because they were non-native speakers of English.

This left 60 subjects for analysis. Of these, 54 subjects were right-handed, six left-handed; 22 subjects were female, 38 male. The age of the subjects ranged from 18 to 54 years, the mean was 27.4 years.

3.4 Results

The data were normalized as in Experiment 1. We tested the hypothesis that paraphrases with high probabilities are perceived as better paraphrases than paraphrases assigned low probabilities by examining the degree to which the elicited judgements correlate with the probabilities derived by the model. Again the judgements refer to the average of log-transformed and normalized subject ratings per experimental item. The comparison between model probabilities and the human judgements yielded a Pearson correlation coefficient of .40 ($p < .01$, $N = 270$). Figure 2 plots the relationship between judgements and model probabilities. The descriptive statistics for model probabilities and the judgements are given in Appendix C.

5 None of the participants of Experiment 1 took part in Experiment 4.

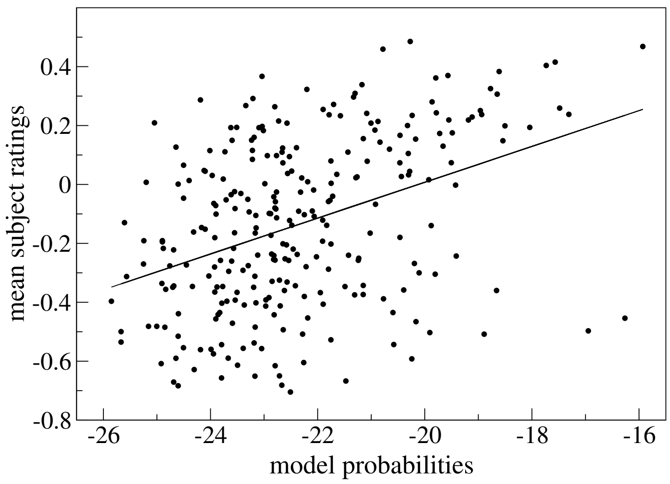


Figure 2
Correlation of elicited judgements and model-derived probabilities for metonymic adjective-noun pairs

In order to evaluate whether the grammatical function has any effect on the relationship between the model-derived meaning paraphrases and the human judgements, we split the items into those that received a subject interpretation and those that received an object interpretation. A comparison between our model and the human judgements yielded a correlation of $r = .53$ ($p < .01$, $N = 135$) for object-related items and a correlation of $r = .21$ ($p < .05$, $N = 135$) for subject-related items. Note that a weaker correlation is obtained for subject-related interpretations. One explanation for that could be the parser's performance, i.e., the parser is better at extracting verb-object tuples than verb-subject tuples. Another hypothesis (which we test below) is that most adjectives included in the experimental stimuli have an object-bias, and therefore subject-related interpretations are generally less preferred than object-related ones.

Using leave-one-out resampling (see Section 2.3.2 for details) we calculated how well humans agree in their judgements for metonymic adjective-noun combinations. For the first group, the average inter-subject agreement was .67 (Min = .03, Max = .82, StdDev = .14), and for the second group .65 (Min = .05, Max = .82, StdDev = .14).

The elicited judgements can be further used to derive the grammatical function preferences (i.e., subject or object) for a given adjective. In particular, we can determine the preferred interpretation for individual adjectives on the basis of the human data and then compare these preferences against the ones produced by our model. Argument preferences can be easily derived from the model's output by comparing subject-related and object-related paraphrases. For each adjective we gathered the subject- and object-related interpretations derived by the model and performed a one-way ANOVA in order to determine the significance of the Grammatical Function effect.

We interpret a significant effect as bias towards a particular grammatical function. We classify an adjective as object-biased if the mean of the model-derived probabilities for the object interpretation of this particular adjective is significantly larger than the mean for the subject interpretation; subject-biased adjectives are classified accordingly, whereas adjectives for which no effect of Grammatical Function is found are classified as equi-biased. The effect of Grammatical Function was significant for the adjectives *difficult* ($F(1,1806) = 8.06$, $p < .01$), *easy* ($F(1,1511) = 41.16$, $p < .01$), *hard* ($F(1,1310) = 57.67$, $p < .01$), *safe* ($F(1,382) = 5.42$, $p < .05$), *right* ($F(1,2114) = 9.85$, $p < .01$), and *fast* ($F(1,92) = 4.38$, $p < .05$). The effect of

Table 18
Log-transformed model-derived and subject-based argument preferences for polysemous adjectives

Adjective	Model	Mean	StdDev	StdEr	Subjects	Mean	StdDev	StdEr
difficult	$\surd$ OBJ	−21.62	1.36	.04	$\surd$ OBJ	.0745	.3753	.0685
	SUBJ	−21.80	1.34	.05	SUBJ	−.2870	.2777	.0507
easy	$\surd$ OBJ	−21.60	1.51	.05	$\surd$ OBJ	.1033	.3364	.0614
	SUBJ	−22.11	1.36	.06	SUBJ	−.1437	.2308	.0421
fast	OBJ	−24.20	1.27	.13	OBJ	−.3544	.2914	.0532
	$\surd$ SUBJ	−23.80	1.40	.14	$\surd$ SUBJ	−.1543	.4459	.0814
good	OBJ	−22.12	1.28	.06	OBJ	−.0136	.3898	.0712
	SUBJ	−22.27	1.10	.07	SUBJ	−.1563	.2965	.0541
hard	$\surd$ OBJ	−21.69	1.53	.06	$\surd$ OBJ	.0030	.3381	.0617
	SUBJ	−22.12	1.35	.06	SUBJ	−.2543	.2436	.0445
right	$\surd$ OBJ	−21.65	1.36	.04	$\surd$ OBJ	−.0054	.2462	.0450
	SUBJ	−21.84	1.24	.04	SUBJ	−.2413	.4424	.0808
safe	OBJ	−22.75	1.48	.10	$\surd$ OBJ	.0037	.2524	.0461
	$\surd$ SUBJ	−22.39	1.59	.12	SUBJ	−.3399	.4269	.0779
slow	OBJ	−22.49	1.53	.08	OBJ	−.3030	.4797	.0876
	SUBJ	−22.32	1.50	.07	$\surd$ SUBJ	−.0946	.2357	.0430
wrong	OBJ	−23.15	1.33	.08	$\surd$ OBJ	−.0358	.2477	.0452
	SUBJ	−23.29	1.30	.08	SUBJ	−.2356	.3721	.0679

Grammatical Function was not significant for the adjectives *good* ($F(1,741) = 3.95$, $p = .10$), *slow* ($F(1,759) = 5.30$, $p = .13$), and *wrong* ($F(1,593) = 1.66$, $p = .19$). Table 18 shows the biases for the nine adjectives as derived by our model. The presence of the symbol $\surd$ indicates significance of the Grammatical Function effect as well as the direction of the bias.

Ideally, we would like to elicit argument preferences from human subjects in a similar fashion. However, since it is unpractical to experimentally elicit judgements for all paraphrases derived by the model, we will obtain argument preferences from the judgements based on the restricted set of experimental stimuli, under the assumption that they correspond to a wide range of model paraphrases (i.e., they correspond to a wide range of probabilities) and therefore they are representative of the entire set of model-derived paraphrases. This assumption is justified by the fact that items were randomly chosen from the three probability bands (i.e., High, Medium, Low). For each adjective we gathered the elicited responses pertaining to subject- and object-related interpretations and performed an Analysis of Variance (ANOVA).

The ANOVA indicated that the Grammatical Function effect was significant for the adjective *difficult* in both by-subjects and by-items analyses ($F_1(1,58) = 17.98$, $p < .01$; $F_2(1,4) = 53.72$, $p < .01$), and for the adjective *easy* in the by-subjects analysis only ($F_1(1,58) = 10$, $p < .01$; $F_2(1,4) = 8.48$, $p = .44$). No effect of Grammatical Function was found for *good* ($F_1(1,58) = 2.55$, $p = .12$; $F_2(1,4) = 1.01$, $p = .37$). The effect of Grammatical Function was significant for the adjective *hard* in the by-subjects analysis only ($F_1(1,58) = 11.436$, $p < .01$; $F_2(1,4) = 2.84$, $p = .17$), whereas for the adjective *slow* the effect was significant by subjects and by items ($F_1(1,58) = 4.56$, $p < .05$; $F_2(1,4) = 6.94$, $p = .058$). For *safe* and *right* the main effect was significant in both by-subjects and by-items analyses ($F_1(1,58) = 14.4$, $p < .0005$; $F_2(1,4) = 17.76$, $p < .05$ and $F_1(1,58) = 6.51$, $p < .05$; $F_2(1,4) = 15.22$, $p = .018$, respectively). The effect of Grammatical Function was significant for *wrong* and *fast* only by subjects ($F_1(1,58) = 5.99$, $p = .05$; $F_2(1,4) = 4.54$, $p = .10$ and $F_1(1,58) = 4.23$, $p = .05$; $F_2(1,4) = 4.43$, $p = .10$). The biases for these adjectives are shown in Table 18. The presence of the symbol $\surd$ indicates significance of the Grammatical Function effect (as determined from the by-subjects analyses) as well as the direction of the bias.

We expect a correct model to assign on average higher probabilities to object-related interpretations and lower probabilities to subject-related interpretations for an object-biased adjective; accordingly, we expect the model to assign on average higher probabilities to subject-related interpretations for subject-biased adjectives. Comparison of the biases derived from the model with ones derived from the elicited judgements shows that the model and the humans are in agreement for all adjectives but *slow*, *wrong*, and *safe*. On the basis of human judgements *slow* has a subject bias and *wrong* has an object bias (see Table 18). Although the model could not reproduce this result there is a tendency in the right direction.

Note that in our correlation analysis reported above the elicited judgements were compared against model-derived paraphrases without taking argument preferences into account. We would expect a correct model to produce intuitive meanings at least for the interpretation a given adjective favors. We further examined the model's behaviour by performing separate correlation analyses for preferred and dispreferred biases as determined previously by the ANOVAs conducted for each adjective (see Table 18). Since the adjective *good* was equi-biased we included both biases (i.e., object-related and subject-related) in both correlation analyses. The comparison between our model and the human judgements yielded a Pearson correlation coefficient of .52 ($p < .01$, $N = 150$) for the preferred interpretations and a correlation of .23 ($p < .01$, $N = 150$) for the dispreferred interpretations. The result indicates that our model is particularly good at deriving meanings corresponding to the argument-bias for a given adjective. However, the dispreferred interpretations also correlate significantly with human judgements, which suggests that the model derives plausible interpretations even in cases where the argument bias is overridden.

In sum, the correlation analysis supports the claim that adjective-noun paraphrases with high probability are judged more plausible than pairs with low probability. It also suggests that the meaning preference ordering produced by the model is intuitively correct since subjects' perception of likely and unlikely meanings correlates with the probabilities assigned by the model.

The probabilistic model evaluated here explicitly takes adjective/adverb and verb co-occurrences into account. However, one could derive meanings for polysemous adjective-noun combinations by solely concentrating on verb-noun relations, ignoring thus the adjective/adverb and verb dependencies. For example, in order to interpret the combination *easy problem* we could simply take into account the types of activities which are related with problems (i.e., solving them, giving them, etc.). This simplification is consistent with Pustejovsky's (1995) claim that polysemous adjectives like *easy* are predicates, modifying some aspect of the head noun and more specifically the events associated with the noun. A 'naive baseline' model would be one which simply takes into account the number of times the noun in the adjective-noun pair acts as the subject or object of a given verb, ignoring the adjective completely. This raises the question of how well would such a naive model perform at deriving meaning paraphrases for polysemous adjective-noun combinations.

In the following section we present such a naive model of adjective-noun polysemy. We compare the model's predictions against the elicited judgements. Using correlation analysis we attempt to determine whether the naive model can provide an intuitively plausible ranking of meanings (i.e., whether perceived likely/unlikely meanings are given high/low probabilities). We further compare the naive model to our initial model (see Section 3.1) and discuss their differences.

3.5 Experiment 5: Comparison against Naive Baseline

3.5.1 Naive Baseline Model Given an adjective-noun combination we are interested in finding the events most closely associated with the noun modified by the adjective. In other words we are interested in the verbs whose object or subject is the noun appearing in the adjective-noun combination. This can be simply expressed as $P(e|rel,n)$, the conditional probability of a verb e

given an argument-noun relation rel, n:

$$P(e|rel,n) = \frac{f(e,rel,n)}{f(rel,n)} \tag{57}$$

The model in (57) assumes that the meaning of an adjective-noun combination is independent of the adjective in question. Consider for example the adjective-noun pair *fast plane*. We need to find the verbs e and the argument relation *rel* that maximize the probability $P(e|rel, plane)$. In the case of *fast plane* the verb that is most frequently associated with planes is *fly* (see Table 13 in Section 3.1). Note that this model will come up with the same probabilities for *fast plane* and *wrong plane* since it does not take the identity of the modifying adjective into account. We estimated the frequencies $f(e,rel,n)$ and $f(rel,n)$ from verb-object and verb-subject tuples extracted from the BNC using Cass (Abney, 1996) (see Section 2.2 for details on the extraction and filtering of the argument tuples).

3.6 Results

Using the naive model we calculated the meaning probability for each of the 270 stimuli included in Experiment 4 explored the linear relationship between the elicited judgements and the naive baseline model through correlation analysis. The comparison yielded a Pearson correlation coefficient of .25 ($p < .01$, $N = 270$). Recall that we obtained a correlation of .40 ($p < .01$, $N = 270$) when comparing our original model to the human judgements. Not surprisingly the two models are intercorrelated ($r = .38$, $p < .01$, $N = 270$). An important question is whether the difference between the two correlation coefficients ($r = .40$ and $r = .25$) is due to chance. A one-tailed t-test revealed that their difference was significant ($t(267) = 2.42$, $p < .01$). This means that our original model (see Section 3.1) performs reliably better than the naive baseline at deriving interpretations for metonymic adjective-noun combinations.

We further compared the naive baseline model and the human judgements separately for subject-related and object-related items. The comparison yielded a correlation of $r = .29$ ($p < .01$, $N = 135$) for object interpretations. Recall that our original model yielded a correlation coefficient of .53 (see Section 3.4). A one-tailed t-test revealed that the two correlation coefficients were significantly different ($t(132) = 3.03$, $p < .01$). No correlation was found for the naive model when compared against elicited subject interpretations ($r = .09$, $p = .28$, $N = 135$).

3.7 Discussion

We have demonstrated that the meanings acquired by our probabilistic model correlate reliably with human intuitions. Our model does not only acquire clusters of meanings (following Vendler's 1968 insight) but furthermore can be used to obtain a tripartite distinction of adjectives depending on the type of paraphrase they prefer: subject-biased adjectives tend to modify nouns which act as subjects of the paraphrasing verb, object-biased adjectives tend to modify nouns which act as objects of the paraphrasing verb, whereas equi-biased adjectives display no preference for either argument role. A comparison between the argument preferences produced by the model and human intuitions revealed that most adjectives (six out of nine) display a preference for an object interpretation (see Table 18), two adjectives are subject-biased (i.e., *fast*, *slow*) and one adjective is equi-biased (i.e., *good*).

We rigorously evaluated[6] the results of our model by eliciting paraphrase judgements from subjects naive to linguistic theory. Comparison between our model and human judgements yielded

6 We have not compared the model's predictions against norming data for adjectival metonymies primarily because such data was not available to us. To our knowledge McElree et al.'s (2001) study is the only example of a norming study on logical metonymy; however, it concentrates on verbs.

a reliable correlation of .40 when the upper bound for the task (i.e., inter-subject agreement) is approximately .65. We have demonstrated that a naive baseline model which interprets adjective-noun combinations by focusing solely on the events associated with the noun is outperformed by a more detailed model which not only considers verb-argument relations but also adjective-verb and adverb-verb dependencies.

Although the events associated with the different nouns are crucially important for the meaning of polysemous adjective-noun combinations, it seems that more detailed linguistic knowledge is needed in order to produce intuitively plausible interpretations. This is by no means surprising. To give a simple example consider the adjective-noun pair *fast horse*. There is a variety of events associated with the noun *horse*, yet only a subset of those are likely to occur fast. The three most likely interpretations for *fast horse* according to the naive model are a horse that needs something fast, a horse that gets something fast, horse that does something fast. A model which uses information about verb-adjective or verb-adverb dependencies provides a more plausible ranking: a *fast horse* is a horse that runs, learns, or goes fast. A similar situation arises when one considers the pair *careful scientist*. According to the naive model a *careful scientist* is more likely to believe, say, or make something carefully. However, none of these events are particularly associated with the adjective *careful*.

Although our experiments on adjectival logical metonymy revealed a reliable correlation between the model's ranking and human intuitions, the fit between model probabilities and elicited judgements was lower for metonymic adjectives ($r = .40$) than for metonymic verbs ($r = .64$). One explanation is that the semantic restrictions that adjectives impose on the nouns with which they combine appear to be less strict than the ones imposed by verbs (consider the adjective *good* which can combine with nearly any noun). A consequence of this is that metonymic adjectives seem to allow a wider range of interpretations than verbs. This means that there will be a larger variation in subjects' responses to the generated model paraphrases when it comes to adjectives, affecting thus the linear relationship between our model and the elicited judgements. Our hypothesis is further supported by the inter-subject agreement which is lower for metonymic adjectives than for verbs (.65 versus .74). As explained in the previous sections, our model does not take the wider context within which the adjective-noun combination is found into account. Precisely because of the ease with which some adjectives combine with practically any noun, it may be the case that more information (i.e., context) is needed in order to obtain intuitively plausible interpretations for metonymic adjectives. The model presented here can be extended to incorporate contextual information (e.g., intra- and extra-sentential information), but we leave this to future work.

4 General Discussion

In this paper we focused on the automatic interpretation of logical metonymy. We showed how meaning paraphrases for metonymic expressions can be acquired from a large corpus and provided a probabilistic model which derives a preference ordering on the set of possible meanings in an unsupervised manner, without relying on the availability of a disambiguated corpus. The proposed approach utilises surface syntactic analysis and distributional information that can be gleaned from a corpus while exploiting correspondences between surface cues and meaning.

Our probabilistic model reflects linguistic observations about the nature of metonymic constructions: it predicts variation in interpretation for different verbs/adjectives with respect to the noun they select for and is faithful to Vendler's (1968) claim that metonymic expressions are usually interpreted by a cluster of meanings instead of a single meaning. This contrasts with Pustejovsky's (1995) approach which typically assigns a single reading to the metonymic construction, and with the account put forward by Copestake and Briscoe (1995) which assigns one interpretation conventionally, albeit a default interpretation—in fact this interpretation might also be of quite a general type (e.g., the event argument to *enjoy* in *enjoy the pebble* can be assigned the general type *act-on*). Our model provides plausible alternatives to the default, and it augments the general semantic type in the interpretation that's assigned by these theories with a plausible range of more specific values, although, in contrast to the hybrid model of interpretation described in Copestake and Lascarides (1997), it does not predict the contexts in which a statistically dispreferred interpretation is the correct one.

Our approach can be viewed as complementary to linguistic theory: while our model does not identify odd metonymies in the way that a rule-based model might (we return to this in Section 4.1), our model *does* automatically derive a *ranking* of meanings, thus distinguishing likely from unlikely interpretations. Even if linguistic theory was able to enumerate all possible interpretations for a given adjective (note that in the case of polysemous adjectives we would have to take into account all nouns or noun classes the adjective could possibly modify), in most cases it does not indicate which ones are likely and which ones are not. Our model fares well on both tasks. It recasts the problem of logical metonymy in a probabilistic framework and derives a large number of interpretations not readily available from linguistic introspection. The information acquired from the corpus can be also used to quantify the argument preferences of metonymic adjectives. These are only implicit in the lexical semantics literature where certain adjectives are exclusively given a verb-subject or verb-object interpretation.

4.1 Limitations and Extensions

We chose to model metonymic constructions and their meanings as joint distributions of interdependent linguistic events (e.g., verb-argument relations, verb-adverb modification). Although linguistically-informed, our approach relies on approximations and simplifying assumptions, partly motivated by the absence of corpora explicitly annotated with metonymic interpretations. We generate meaning paraphrases for metonymic constructions solely from co-occurrence data without taking advantage of taxonomic information. Despite the simplicity of this approach and its portability to languages for which lexical semantic resources may not be available, there are certain regularities about the derived interpretations which our model fails to detect.

Consider the interpretations produced for *begin song* repeated here from Table 7: *sing*, *rehearse*, *write*, *hum*, and *play*. Our model fails to capture the close correspondence for some of these meanings. For example, *hum* and *sing* are sound emission verbs; they additionally entail the performance or execution of the song. The verbs *rehearse* and *play* capture only the performance aspect and can be thus considered supertypes of *hum* and *sing* (one can play/rehearse a song by humming it, singing it, drumming it, whistling it, etc.). The verb *write* on the other hand is neither a performance nor a sound emission verb, it has to do with communication and creation. Another example is *comfortable chair* for which the model generates: *sink into*, *sit on*, *lounge in*, *relax in*, *nestle in*. The verbs *sink into*, *sit on*, *lounge in*, and *nestle in* describe the position one assumes when sitting in the chair, whereas *relax in* refers to the state of the person in the chair. There is no notion of semantic proximity built in the model and correspondences among semantically related interpretations are not automatically recognized.

An alternative to the knowledge-free approach advocated here is to use taxonomic information for obtaining some degree of generalisation over the acquired interpretations. Semantic classifications such as WordNet (Miller et al., 1990) or Levin (1993) can be used to group the obtained verbs into semantically coherent classes. Furthermore, the WordNet semantic hierarchy can be used to directly estimate probabilities involving either nouns (e.g., $P(rel|e,n)$, $P(o|e)$) or verbs (e.g., $P(v|e)$, $P(a|e)$). Probabilities can be defined in terms of senses from a semantic hierarchy by exploiting the fact that the senses can be grouped into classes consisting of semantically similar senses (Resnik, 1993; Clark and Weir, 2001; McCarthy, 2000; Li and Abe, 1998). So, the probability $P(book|read)$ can be estimated by taking into account nouns which belong to the same semantic class as *book* and can be read (e.g., journals, novels, scripts) or by focusing on verbs that are semantically related to *read* (e.g., interpret, communicate, understand). Note that estimation of probabilities over classes rather than words can effectively overcome data sparseness and potentially lead to better probability estimates. Currently our models cannot estimate probabilities for unseen word combinations and WordNet could be used for recreating the frequencies of these combinations (Lapata, Keller, and McDonald, 2001; Clark and Weir, 2001). However, part of our aim here was to investigate whether it is at all possible to generate interpretations for metonymic constructions without the use of prior knowledge bases which might bias the acquisition process in uncontrolled and idiosyncratic ways.

A related issue is the fact that our models are ignorant about the potentially different senses of the noun in the metonymic construction. For example, the combination *fast plane* may be a fast aircraft, or a fast tool, or a fast geometrical plane. Our model derives meanings related to all three senses of the noun *plane*. For example, a *fast plane* is not only a plane (i.e., an aircraft) which flies, lands, or travels quickly, but also a plane (i.e., a surface) which transposes or rotates quickly and a plane (i.e., a tool) which smoothes something quickly. However, more paraphrases are derived for the 'aircraft' sense of plane; these paraphrases also receive a higher ranking. This is not surprising since the number of verbs related with the 'aircraft' sense of plane are more frequent than the verbs related with the other two senses. In contrast to *fast plane*, however, *efficient plane* should probably bias towards the 'tool' sense of plane, even though the 'aircraft' sense is more frequent in the corpus. One could make the model sensitive to this by investigating the synonyms to the various senses of plane; moreover the 'tool' sense bias of *efficient plane* could also be inferred on the basis that *efficient plane* co-occurs with different verbs from *fast plane*. There are also cases where a model-derived paraphrase does not provide disambiguation clues with respect to the meaning of the noun. Consider the adjective-noun combination *fast game*. The model comes up with the paraphrases game that runs fast or game that goes fast. Both paraphrases may well refer to either the 'contest', 'activity', or 'prey' sense of *game*. Note finally that our model can be made sensitive to word sense distinctions by taking into account noun classes rather than word forms; this modification would allow us to apply the model to word sense disambiguated metonymic expressions.

In this paper we have focused on the automatic *interpretation* of logical metonymy without explicitly dealing with the *recognition* of verbs or adjectives undergoing logical metonymy. In default of the latter study which we plan for the future we sketch here briefly how our proposal can be extended to recognizing logical metonymies. A very simple approach would be to use the proposed model to generate interpretations for metonymic and non-metonymic constructions. The derived paraphrases and the range of their probabilities can be then used to quantify the degree of "metonymic-ness" of a given verb or adjective. One would expect that a larger number of paraphrases is generated for verbs or adjectives for which logical metonymy is possible. We tested this hypothesis using a metonymic verb (i.e., *enjoy*) and a non-metonymic one (i.e., *play*). *Enjoy* is attested 5,344 times in the BNC in a verb-object relation, whereas *play* is attested 12,597 times (again these numbers are based on information extracted using Cass (Abney, 1996)). Using the model presented in Section 2.1 we generated meaning paraphrases for all verb-object tuples found for *enjoy* and *play*. The model generated 44,701 paraphrases for *enjoy* and 9,741 for *play*. Comparison between the probabilities assigned to the interpretations for *enjoy* and *play* revealed that the paraphrases obtained for *enjoy* were on average more likely (Mean $= -23.53$, Min $= -27.02$, Max $= -15.32$) than those discovered for *play* (Mean $= -24.67$, Min $= -28.25$, Max $= -17.24$). The difference was statistically significant (using an independent samples t-test; $t(54,440) = 2.505$, $p < .01$). This result indicates that *enjoy* is more likely to undergo logical metonymy than *play*. Another potential indicator of metonymic use is the likelihood of a given verb or adjective to be found in a certain syntactic construction. Consider the adjective *blue* for which our model (see Section 3.1) does not generate any meaning paraphrases. The reason for this is presumably the fact that *blue* is not attested as a verb modifier (in contrast to adjectives like *easy* or *fast*).

Such an approach could potentially predict differences in productivity among metonymic verbs. One would expect the metonymic uses of *attempt* for example to be much less productive

Table 19
Model-derived paraphrases for odd metonymies, ranked in order of likelihood

begin dictionary		begin rock		begin keyboard		begin highway	
compile	−19.32	crunch across	−18.09	use	−20.11	obstruct	−20.40
flick through	−19.59	climb	−18.39	play	−20.44	regain	−20.79
use	−19.80	run towards	−18.70	operate	−20.56	build	−20.80
publish	−20.34	percolate through	−18.78	assemble	−20.78	use	−20.81
advance	−20.39	dissolve	−19.37	tune	20.80	detach	−20.82

than the metonymic uses of *enjoy* and *begin*. One could conceivably predict this on the basis of the frequency and *diversity* of NPs in *attempt* NP constructions that are attested in the corpus, compared with those for *enjoy* NP and *begin* NP. Furthermore, the model generates paraphrases for *attempt* that are on average less likely in comparison to those generated for *enjoy* or *begin*. However, we leave this for future work.

As argued in Section 2.1 our model cannot distinguish between well-formed and odd metonymic constructions; in fact, it will generally provide meaning paraphrases even for combinations that are deemed to be odd. Consider the examples in (58) and their interpretations in Table 19. In general the paraphrases generated for problematic data are of worse quality than those produced for well-formed metonymies. In most cases the model will generate unavailable interpretations (see *begin highway*, *begin keyboard*). Consider, however the pair *begin rock*. Pustejovsky (1995) observes that although there is no generally available interpretation for a sentence like *Mary began the rock* because of what we understand *begin* to require of its argument and our knowledge of what rocks are and what you can do to them, as speakers/hearers we tend to accommodate information into the context so as to interpret otherwise ill-formed expressions. Our model generates meaning paraphrases that are relatively plausible assuming different pragmatic contexts for *begin rock*. One can begin *climbing* or *running towards* a rock. Someone's footsteps can *crunch across* a frozen rock, a material can *percolate through* a rock, and rain water can *dissolve* a rock.

(58) a. ?John began the dictionary.
b. ?Mary began the rock.
c. *John began a keyboard.
d. *John began the highway.

Despite the fact that the model does not recognize odd metonymies, one would expect that low probabilities are assigned to ungrammatical constructions. Table 20 reports some descriptive statistics on well-formed (top half) and odd metonymies (bottom half) taken from the lexical semantics literature (Verspoor, 1997; Pustejovsky, 1995). A higher number of interpretations is generated for well-formed metonymies. Using an independent samples t-test we can compare the differences in the probabilities assigned to the two types of metonymies. Take for example *begin dictionary*: the average probability of its interpretations is lower than *begin book* ($t(608) = 5.07$ $p < .01$), *begin cigarette* ($t(178) = 2.77$, $p < .01$), *begin coffee* ($t(178) = 2.1$, $p < .05$), and *begin story* ($t(455) = 4.1$, $p < .01$). Similar results are obtained when comparing *begin keyboard* against the well-formed metonymies in Table 19: the probability of its interpretations is on average lower than those assigned to *begin book*, ($t(582) = 3.87$, $p < .01$), *begin cigarette* ($t(152) = 2.15$, $p < .01$), and *begin story* ($t(429) = 3.02$, $p < .01$). The difference between *begin coffee* and *begin keyboard* is not statistically significant ($t(152) = 1.39$, $p = 0.167$). However, the mean for *begin coffee* is slightly higher than *begin keyboard*. Although here we focus on verbs, similar comparisons can be applied to adjectives. As shown in Table 20 the probabilities for *easy programmer* are on average lower than those for *easy problem* ($t(405) = 8.74$, $p < .01$).

Table 20
Descriptives for odd and well-formed metonymies

Well-formed	*N*	Mean	StdDev	StdErr
begin book	534	−21.54	1.515	.066
begin cigarette	104	−21.83	1.613	.158
begin coffee	104	−22.03	1.626	.159
begin story	381	−21.73	1.493	.076
easy problem	358	−21.14	1.606	.085
Odd	***N***	**Mean**	**StdDev**	**StdErr**
begin dictionary	76	−22.48	1.440	.165
begin keyboard	50	−22.40	1.337	.189
begin rock	50	−23.55	1.376	.193
begin highway	37	−22.52	1.337	.219
easy programmer	49	−23.23	1.289	.184

Table 21
BNC frequencies for odd metonymic expressions

Odd	*begin* NP	*begin* V-*ing* NP
begin chair	0	9
begin tunnel	0	4
begin keyboard	0	0
begin tree	1	13
begin highway	0	2
begin film	0	7
begin nail	0	4
begin door	0	18
begin dictionary	0	3
begin rock	0	17

Finally, recall from Section 2.1 that on the basis of Gricean reasoning one would expect to find in a corpus well-formed metonymies more often than their paraphrases (see Tables 1–3). Following this line of reasoning one might expect for conventionally odd metonymies the opposite situation, i.e., to find the paraphrases more often than the metonymies proper. We tested this hypothesis for some examples cited in the literature (Verspoor, 1997; Pustejovsky, 1995) by examining whether paraphrases corresponding to odd metonymies are attested in the BNC as VP complements. We found plausible paraphrases in the BNC for almost all verb-noun pairs illustrated in Table 21. This suggests that the corpus data relating to the odd and well-formed examples are largely compliant with the Gricean predictions. Corpus co-occurrences of verb-noun combinations and their paraphrases could be exploited in creating a system aimed at quantifying the grammaticality of metonymic expressions. However, it is beyond the scope of the present study to develop such a system.

4.2 Relevance for NLP Applications

The meaning paraphrases discovered by our model could be potentially useful for a variety of NLP tasks. One obvious application is Natural Language Generation. For example, a generator that has knowledge of the fact that *fast plane* corresponds to a plane that flies fast can exploit this information either to render the text shorter (in cases where the input representation is a sentence) or longer (in cases where the input representation is an adjective-noun pair). Information retrieval is another relevant application. Consider a search engine faced with the query *fast plane*. Presumably one would not like to obtain information about planes in general or about planes that

Table 22
Five most likely interpretations for *good author* and *good language*

good author			good language		
write	SUBJ	−21.81	use	OBJ	−17.39
work	SUBJ	−21.97	verse in	OBJ	−17.89
describe	SUBJ	−22.03	speak	OBJ	−18.32
know	OBJ	−22.05	know	OBJ	−19.18
engage with	OBJ	−22.23	learn	OBJ	−19.36

go down or burn fast but rather about planes that fly or travel fast. So knowledge about the most likely interpretations of *fast plane* could help rank relevant documents before non-relevant ones or restrict the number of retrieved documents.

Note that in the case of adjectives, it is not just the paraphrase but also the grammatical function that needs to be determined. Rendering an adjective-noun combination with an object- or subject-related paraphrase can be worked out by computing the biases discussed in Section 3.4. So, if we know that *fast* has a subject bias, we can concentrate only on the subject-related interpretations. The choice of grammatical function is less straightforward in the case of equi-biased adjectives. In fact, it is possible that the interpretation for the same adjective varies depending on the noun it modifies. For example, a *good author* writes well, whereas a *good language* is good to learn, hear, study. A simple way to address this is to select the interpretations with the highest probability. For *good author* and *good language* the five most likely interpretations (and their grammatical functions) according to the model (see Section 3.3) are given in Table 22. As can be seen, subject-related interpretations are ranked higher for *good author*; the opposite is true for *good language*. Another possibility for determining the grammatical function for equi-biased adjectives is to directly compare verb-object and verb-subject interpretations for a particular adjective-noun combination. To illustrate with an example, consider the following. For *good author* the model produces 107 object-related paraphrases and 199 subject-related ones. Furthermore, the subject-related probabilities are on average higher than the object-related ones and the difference is statistically significant (using a one-tailed t-test, $t(304) = 3.26$, $p < .01$). For *good language* the object-related paraphrases are 253 and the subject-related ones are 180. The former are assigned higher probabilities than the latter and the difference is statistically significant ($t(431) = 3.80$, $p < .01$).

Machine translation is another related application. A logical metonymy may be acceptable in a source language but unacceptable in the target language. Consider the example in (59): its direct translation into German produces a semantically unacceptable sentence (see (60a)). In this case we need to spell out the metonymy in order to obtain an acceptable translation for German and our model can be used to provide the missing information by generating meaning paraphrases. Under such an approach, we would not translate (59) directly but one of its paraphrases (see (60b,c)).

(59) Peter attempted the peak.

(60) a. * Peter hat den Gipfel versucht.
Peter has the peak attempted
'Peter attempted the peak.'
b. Peter hat den Gipfel zu besteigen versucht.
Peter has the peak to climb attempted
'Peter attempted to climb the peak.'
c. Peter hat den Gipfel zu erreichen versucht.
Peter has the peak to reach attempted
'Peter attempted to reach the peak.'

5 Related Work

In contrast to the extensive theoretical literature on the topic of logical metonymy, little attention has been paid to the phenomenon from an empirical perspective. Briscoe et al. (1990) and Verspoor (1997) undertake a manual analysis of logical metonymies found in naturally occurring text. Their results show that logical metonymy is a relatively widespread phenomenon and that most metonymic examples can be interpreted on the basis of the head noun's qualia structure assuming a theoretical framework similar to Pustejovsky's (1991). Verspoor's (1997) analysis of the metonymic verbs *begin* and *finish* demonstrates that context plays a relatively small role for the interpretation of these verbs: 95.0% of the logical metonymies for *begin* and 95.6% the logical metonymies for *finish* can resolved on the basis of information provided by the noun the verb selects for. Briscoe et al.'s (1990) work further suggests ways of acquiring qualia structures for nouns by combining information extracted from machine-readable dictionaries and corpora. Our probabilistic formulation of logical metonymy allows us to discover interpretations for metonymic constructions without presupposing the existence of qualia-structures. In fact, we show that a simple statistical learner in combination with a shallow syntactic analyzer yields relatively intuitive results, considering the simplifications and approximations in the system.

Perhaps more relevant to the work presented here are previous approaches to the automatic interpretation of general metonymy (Lakoff and Johnson, 1980; Nunberg, 1995). This is slightly different from logical metonymy, in that the examples aren't usually analysed in terms of semantic type coercion. But the phenomena are closely related. Generally speaking an expression A is considered a metonymy, if A deviates from its literal denotation in that it stands for an entity B which is not expressed explicitly but is conceptually related to A via a contiguity relation r (Markert and Hahn, 1997). A typical example of general metonymy is given in (61): in (61a) the *bottle* stands for its content (i.e., the liquid in the bottle) and in (61b) *Shakespeare* stands for his works. The contiguity relation r between the bottle and its liquid is *Container for Contents*; for Shakespeare and his works the contiguity relation is *Producer for Product*.

(61) a. Denise drank the bottle.
b. Peter read Shakespeare.

Previous approaches to processing metonymy typically rely heavily on the availability of manually constructed knowledge-bases or semantic networks (Fass, 1991; Inverson and Helmreich, 1992; Bouaud, Bachimont, and Zweigenbaum, 1996; Hobbs et al., 1993). Furthermore, most implementations either contain no evaluation (Fass, 1991; Inverson and Helmreich, 1992; Hobbs et al., 1993) or report results on the development data (Bouaud et al. 1996).

The approach put forward by Utiyama, Murata, and Isahara (2000) is perhaps the most comparable to our own work. Utiyama et al. describe a statistical approach to the interpretation of general metonymies for Japanese. Utiyama et al.'s algorithm interprets verb-object metonymies by generating the entities for which the object stands for. These entities are ranked using a statistical measure. Given an expression like (62), nouns related to *Shakespeare* are extracted from the corpus (e.g., *si* 'poem', *tyosyo* 'writings', *sakuhin* 'works') and ranked according to their likelihood. Two types of syntactic relations are used as cues for the interpretation of metonymic expressions: (a) the noun phrase *A no B* roughly corresponding to the English *B of A* where A is the noun figuring in the metonymic expression (e.g., *Shakespeare* in (62)) and B is the noun it stands for (e.g., *sakuhin* 'works') and (b) nouns co-occurring with the target noun (e.g., *Shakespeare*) within the target sentence.

(62) Shakespeare wo yomu
Shakespeare ACC read
'read Shakespeare'

Given a metonymy of the form $A\ R\ V$ the appropriateness of a noun B as an interpretation of

A is defined as:

$$L_Q(B|A,R,V) = \frac{P(B|A,Q)P(R,V|B)}{P(R,V)} \tag{63}$$

where V is the verb in the metonymic expression, A is its object, R is A's case marker (e.g., *wo* (accusative)), B is the noun A stands for, and Q is the relation Q bears to A (e.g., *no*). The probabilities in (63) are estimated from a large morphologically analysed Japanese corpus of newspaper texts (approximately 153 million words). A Japanese thesaurus is used for the estimation of the term $P(R,V|B)$ when the frequency $f(R,V,B)$ is zero (see Utiyama et al. 2000 for the derivation and estimation of (63)). Utiyama et al.'s approach is tested on 75 metonymies taken from the literature. It achieves a precision of 70.6% as measured by one of the authors according to the following criterion: a metonymic interpretation was considered correct if it made sense in some context.

Our approach is conceptually similar to Utiyama et al. (2000). Metonymies are interpreted using corpora as the inventory of the missing information. In contrast to Utiyama et al., we use no information external to the corpus (e.g., a thesaurus); sparse data problems are tackled via independence assumptions and syntactic information is obtained through shallow text analysis (Utiyama et al. 2000 rely on morphological analysis to provide cues for syntactic information). The most striking difference, however, between our work and Utiyama et al.'s is methodological. Their evaluation is subjective and limited to examples taken from the literature. The appropriateness of their statistical measure (see (63)) is not explored and it is not clear whether it can derive an intuitively plausible ranking of interpretations or whether it can extend to examples found in naturally occurring text. We test our probabilistic formulation of logical metonymy against a variety of examples taken from the corpus and the derived interpretations are evaluated objectively using standard experimental methodology. Furthermore, the appropriateness of the proposed model is evaluated via comparisons to a naive baseline.

6 Conclusions

In this paper we proposed a statistical approach to logical metonymy. We acquired the meanings of metonymic constructions from a large corpus and introduced a probabilistic model which provides a ranking on the set of possible interpretations. We identified semantic information automatically by exploiting the consistent correspondences between surface syntactic cues and meaning.

We evaluated our results against paraphrase judgements elicited experimentally from subjects naive to linguistic theory and showed that the model's ranking of meanings correlates reliably with human intuitions. Comparison between our model and human judgements yields a reliable correlation of .64 for verb-noun combinations and .40 for adjective-noun pairs. Furthermore, our model performs reliably better than a naive baseline model, which only achieves a correlation of .42 in the case of verbs and .25 in the case of adjectives.

Our approach combined insights from linguistic theory (i.e., Pustejovsky's 1995 theory of Qualia Structure and Vendler's 1968 observations) with corpus-based acquisition techniques, probabilistic modeling, and experimental evaluation. Our results empirically test the validity of linguistic generalisations and extend their coverage. Furthermore, in agreement with other lexical acquisition studies (Merlo and Stevenson, 2001; Barzilay and McKeown, 2001; Siegel and McKeown, 2000; Light, 1996; McCarthy, 2000; Rooth et al., 1999), we show that that it is possible to extract semantic information from corpora even if they are not semantically annotated in any way.

Acknowledgments

This work was supported by ESRC grant number R000237772 (Data Intensive Semantics and Pragmatics) and the DFG (Gottfried Wilhelm Leibniz Award to Manfred Pinkal). Alex Lascarides is supported by an ESRC research fellowship. Thanks to Brian McElree and Matt Traxler for making available to us the results of their norming study and to Ann Copestake, Frank Keller, Scott McDonald, Manfred Pinkal, Owen Rambow and three anonymous reviewers for valuable comments.

References

2001. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Abney, Steve. 1996. Partial parsing via finite-state cascades. In John Carroll, editor, *Workshop on Robust Parsing*, pages 8–15, Prague. European Summer School in Logic, Language and Information.

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

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Barzilay, Regina and Kathy McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (ACL, 2001), pages 50–57.

Bouaud, Jacques, Bruno Bachimont, and Pierre Zweigenbaum. 1996. Processsing metonymy: A domain-model heuristic graph traversal approach. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 137–142, Copenhagen, Denmark.

Briscoe, Ted, Ann Copestake, and Bran Boguraev. 1990. Enjoy the paper: Lexical semantics via lexicology. In *Proceedings of 13th International Conference on Computational Linguistics*, pages 42–47, Helsinki, Finland.

Burnage, Gavin. 1990. Celex – a guide for users. Technical report, Centre for Lexical Information, University of Nijmegen.

Burnard, Lou, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Clark, Stephen and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Collins, Michael and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. In David Yarowsky and Kenneth W. Church, editors, *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 27–38, Cambridge, MA.

Copestake, A. and E. J. Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1):15–67.

Copestake, A. and A. Lascarides. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter for the Association for Computational Linguistics*, pages 136–143, Madrid.

Copestake, Ann. 2001. The semi-generative lexicon: Limits on lexical productivity. In *Proceedings of the 1st International Workshop on Generative Approaches to the Lexicon*, Geneva.

Corley, Martin, Frank Keller, and Christoph Scheepers. 2000. Conducting psycholinguistic experiments over the world wide web. Unpubl. ms., University of Edinburgh and Saarland University.

Corley, Martin and Christoph Scheepers. 2002. Syntactic priming in English: Evidence from response latencies. *Psychonomic Bulletin and Review*, 9(1):126–131.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, Thousand Oaks, CA.

Daelemans, Walter, Antal van den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34(1–3):11–43.

Fass, Dan. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Godard, Danièle and Jacques Jayez. 1993. Towards a proper treatment of coercion phenomena. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 168–177, Columbus, OH.

Hobbs, Jerry R., Martin Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Inverson, Eric and Stephen Helmreich. 1992. An integrated approach to non-literal phrase interpretation. *Computational Intelligence*, 8(3):477–493.

Keller, Frank and Theodora Alexopoulou. 2001. Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, 79(3):301–371.

Keller, Frank and Ash Asudeh. 2002. Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry*, 33(2).

Keller, Frank, Martin Corley, and Christoph Scheepers. 2001. Conducting psycholinguistic experiments over the world wide web. Unpublished Manuscript, University of Edinburgh and Saarland University.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Lapata, Maria, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgments. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (ACL, 2001), pages 346–353.

Lascarides, Alex and Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics*, 34(2):387–414.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. The tagging of the British national corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 622–628, Kyoto, Japan.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Li, Hang and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Light, Marc. 1996. *Morphological Cues for Lexical Semantics*. Ph.D. thesis, University of Rochester.

Lodge, Milton. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions*. Sage Publications, Beverley Hills, CA.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. Comlex: A lexicon of nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography*, pages 187–193, Liège, Belgium.

Markert, Katja and Udo Hahn. 1997. On the interaction of metonymies and anaphora. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 1010–1015, Nagoya, Japan.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pages 256–263, Seattle, WA.

McElree, Brian, Matthew J. Traxler, Martin J. Pickering, Rachel E. Seely, and Ray Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 1(78):17–25.

Merlo, Paola and Susanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Nunberg, Geoffrey. 1995. Transfers of meaning. *Journal of Semantics*, 12(1):109–132.

Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

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

Pustejovsky, James and Pierrette Bouillon. 1995. Logical polysemy and aspectual coercion. *Journal of Semantics*, 12:133–162.

Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, pages 95–130.

Resnik, Philip and Mona Diab. 2000. Measuring verb similarity. In Lila R. Gleitman and Aravid K. Joshi, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 399–404, Mahwah, NJ. Lawrence Erlbaum Associates.

Resnik, Philip Stuart. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, College Park, MD.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago.

Siegel, Eric and Kathleen McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–629.

Stevens, S. Smith. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley, New York.

Utiyama, Masao, Masaki Murata, and Hitoshi Isahara. 2000. A statistical approach to the processing of metonymy. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 885–891, Saarbrücken, Germany.

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

Verspoor, Cornelia Maria. 1997. *Contextually-Dependent Lexical Semantics*. Ph.D. thesis, University of Edinburgh.

Weiss, Sholom M. and Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

Appendix A. The WebExp Software Package

As part of the evaluation of the probabilistic models presented in this paper we conducted two psycholinguistic experiments. These experiments were administered using WebExp (Corley, Keller, and Scheepers, 2000), a software package designed for conducting psycholinguistic studies over the web.[7]

WebExp is a set of Java classes for conducting psycholinguistic experiments over the world wide web. The software consists of two modules: the WebExp server, which is a stand-alone Java application, and the WebExp client, which is implemented as a Java applet. The server application runs on the web server that hosts the experiment, and waits for client applets to connect to it. It issues experimental materials to clients, and records participants' responses. The client applet is typically embedded into a web page that contains the instructions for the experiment. When a participant starts the experiment, the WebExp client will download the experimental materials from the WebExp server and administer them to the participant. After the experiment is completed, it will send the participants' responses to the server, along with other participant-specific data.

As Java is a full-fledged programming language, it gives the web designer maximal control over the interactive features of a web site. WebExp makes use of this flexibility to keep the experimental procedure as constant as possible across participants. An important aspect is that the sequence in which the experimental items are administered is fixed for each participant: the participant does not have the possibility to go back to previous stimuli and to inspect or change previous responses. (If the participant hits the 'back' button on the browser, the experiment will terminate.) WebExp also provides timings of participant responses by measuring the response onset time and the completion time for each answer. The studies reported in this paper make no direct use of these timings. Nevertheless, the timings are useful to screen the responses for anomalies, i.e., to eliminate the subjects who responded too quickly (and thus probably did not complete the experiment in a serious fashion), or those who responded too slowly (and thus were probably distracted while doing the experiment). WebExp automatically tests the response timings against upper and lower limits provided by the experimenter and excludes participants whose timings are anomalous. Further manual checks can be carried out on the response timings later on.

Apart from providing response timing, WebExp also offers a set of safeguards that are meant to ensure the authenticity of the participants taking part, and exclude participants from participating more than once.

Email Address Each participant has to provide their email address. An automatic plausibility check is conducted on the address to ensure that it is syntactically valid. If the address is valid, then WebExp sends an email to this address at the end of the experiment (the email typically contains a message thanking the participant for taking part). If the email bounces, the experimenter is effectively informed that a participant is likely to have used a fake identity, and has the option of excluding their responses from further analysis.

Personal Data Before the experiment proper commences, each participant has to fill in a short questionnaire supplying name, age, sex, handedness, and language background. These data allow manual plausibility checks such that participants who provide implausible answers can be

7 The WebExp software package is distributed free of charge for non-commercial purposes. For information on how to obtain the latest version, please visit the site `http://www.hcrc.ed.ac.uk/web_exp/`. A central entry page for all experiments using WebExp can be found at `http://www.language-experiments.org/`.

eliminated from the data set.

Responses A manual inspection of the responses allows the experimenter to detect participants who have misunderstood the instructions or who responded in an anomalous fashion, e.g., by giving the same response to every item.

Connection Data The software also logs data related to the participant's web connection. This includes the internet address of their machine and the operating system and browser they use. This information (in addition to the email address) is valuable in detecting participants who take part more than once.

Apart from making it possible to administer experiments over the web, WebExp can also be used in a conventional laboratory setting. Here, WebExp has the advantage of being platform-independent (as it is implemented in Java), i.e., it will run on any computer that is connected to the internet and runs a web browser. Comparisons of experimental data obtained from internet experiments (using WebExp) and their laboratory-based counterparts (Keller and Asudeh, 2002; Keller and Alexopoulou, 2001; Corley and Scheepers, 2002) revealed high correlations between the two types of data sets; the comparisons also demonstrated that the same main effects are obtained from web-based and laboratory-based experiments.

Appendix B. Materials

B.1. Experiment 1

The following is a list of the materials used in Experiment 1.

(64) David finished a course David finished writing a course

The modulus is shown in (64). The verb-noun pairs and their selected interpretations (Interpr) are illustrated in Table 23. In addition, we show the mean ratings (Rtg) for each paraphrase according to the subjects' responses and their probability (Prob) according to the model.

Table 23
Materials for Experiment 1, with Mean Ratings and model probabilities

Verb-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Rtg	Prob
attempt definition	analyse	−0.087	−21.44	recall	−0.0338	−22.84	support	−0.1571	−23.87
attempt peak	climb	0.2646	−20.22	claim	−0.0900	−23.53	include	−0.4450	−24.85
attempt question	reply to	0.1416	−18.96	set	−0.2666	−21.63	stick to	−0.3168	−22.55
attempt smile	give	0.2490	−19.43	rehearse	−0.1976	−22.21	look at	−0.4722	−23.74
attempt walk	take	0.1807	−19.81	schedule	−0.1649	−22.71	lead	−0.1863	−23.85
begin game	play	0.2798	−15.11	modify	−0.3403	−21.52	command	−0.2415	−23.46
begin photograph	develop	0.0816	−21.11	test	−0.2147	−22.49	spot	−0.4054	−23.69
begin production	organise	0.0502	−19.09	influence	−0.2329	−21.98	tax	−0.4367	−22.78
begin test	take	0.2699	−17.97	examine	−0.1424	−21.78	assist	−0.3440	−24.11
begin theory	formulate	0.2142	−18.28	present	0.1314	−21.54	assess	−0.1356	−22.40
enjoy book	read	0.2891	−16.48	discuss	−0.1515	−23.33	build	−0.5404	−25.52
enjoy city	live in	0.2028	−20.77	come to	0.1842	−23.50	cut	−0.6957	−24.67
enjoy concert	listen to	0.2779	−20.91	throw	−0.2442	−23.61	make	−0.2571	−24.97
enjoy dish	cook	−0.1223	−20.21	choose	−0.2373	−24.61	bring	−0.3385	−25.33
enjoy story	write	−0.0731	−19.08	learn	−0.0887	−23.50	choose	−0.2607	−24.61
expect order	hear	0.2087	−20.29	read	0.0628	−22.92	prepare	−0.3634	−23.25
expect poetry	see	0.1100	−20.43	learn	−0.0601	−22.81	prove	−0.4985	−25.00
expect reply	get	0.2696	−20.23	listen to	0.1178	−23.48	share	−0.2725	−23.77
expect reward	collect	0.2721	−21.91	claim	0.1950	−23.13	extend	−0.3743	−23.52
expect supper	eat	0.2487	−21.27	start	0.0285	−23.20	seek	−0.3526	−23.91
finish gig	play	0.2628	−20.34	plan	−0.1780	−24.47	use	−0.5341	−25.69
finish novel	translate	0.0474	−21.81	examine	−0.1323	−24.01	take	−0.4375	−25.53
finish project	work on	0.3113	−18.79	study	0.0679	−24.31	sell	−0.1692	−25.05
finish room	wallpaper	0.1497	−19.07	construct	0.0444	−22.48	show	−0.6305	−24.59
finish video	watch	0.3165	−22.37	analyse	−0.0482	−24.33	describe	−0.1718	−25.26
postpone bill	debate	−0.0401	−22.38	give	−0.2028	−25.36	think	−0.6238	−27.92
postpone decision	make	0.3297	−20.38	publish	−0.1087	−24.03	live	−0.4984	−25.56
postpone payment	make	0.2745	−21.85	arrange	0.0166	−23.21	read	−0.4675	−25.91
postpone question	hear	−0.1974	−23.45	assess	−0.0795	−24.70	go to	−0.5383	−24.94
postpone trial	go to	−0.1110	−23.49	make	−0.1425	−25.50	hear	0.0336	−25.75
prefer bike	ride	0.2956	−20.64	mount	−0.1617	−22.95	go for	−0.2119	−24.49
prefer film	go to	0.2456	−21.63	develop	−0.1861	−23.07	identify	−0.4657	−24.69
prefer gas	use	0.3078	−20.28	measure	−0.2903	−23.88	encourage	−0.5074	−25.33
prefer people	talk with	0.1235	−20.52	sit with	0.0283	−22.75	discover	−0.2687	−25.26
prefer river	swim in	0.0936	−19.53	sail	0.0433	−22.93	marry	−0.7269	−24.13
resist argument	contest	−0.0126	−22.66	continue	0.0224	−24.43	draw	−0.4551	−25.51
resist invitation	accept	0.2497	−21.56	leave	−0.3620	−25.10	offer	−0.5301	−26.29
resist pressure	take	−0.1481	−22.67	make	−0.3602	−24.98	see	−0.4382	−25.22
resist proposal	work on	−0.0597	−23.56	take on	0.1286	−24.50	call	−0.2906	−25.99
resist song	whistle	−0.0013	−22.11	start	−0.1551	−24.47	hold	−0.5691	−26.50
start experiment	implement	0.1744	−21.57	study	0.0184	−22.70	need	−0.5299	−24.09
start letter	write	0.3142	−15.59	study	−0.0877	−22.70	hear	−0.4526	−24.50
start treatment	receive	0.2888	−19.53	follow	0.1933	−22.45	assess	−0.2536	−24.01
survive course	give	0.0164	−22.87	make	−0.1426	−24.48	write	−0.1458	−26.27
survive journey	make	0.2719	−22.31	take	0.2324	−24.43	claim	−0.5388	−25.84
survive problem	create	−0.2697	−21.12	indicate	−0.3267	−23.01	confirm	−0.3625	−25.08
survive scandal	experience	0.2200	−24.61	create	−0.2115	−26.02	take	−0.3068	−27.33
survive wound	receive	0.2012	−24.49	produce	−0.4361	−26.32	see	−0.2668	−26.97

Table 23
(*continued*)

Verb-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Rtg	Prob
try drug	take	0.1077	−17.81	grow	−0.2516	−22.09	hate	−0.4777	−23.88
try light	turn	0.2397	−18.10	reach for	−0.1163	−21.23	come with	−0.5310	−23.93
try shampoo	use	0.1404	−20.09	pack	−0.3496	−21.56	like	−0.2581	−24.56
try sport	get into	0.1379	−19.65	encourage	−0.2378	−21.09	consider	−0.2223	−22.82
try vegetable	eat	0.2068	−19.64	chop	−0.0780	−21.38	compare	−0.3248	−22.38
want bed	lay on	0.1154	−19.17	reserve	0.0369	−21.24	settle in	0.0015	−22.26
want hat	buy	0.2560	−17.84	examine	−0.2127	−21.56	land on	−0.6379	−22.38
want man	marry	0.0826	−15.50	torment	−0.2764	−20.99	assess	−0.2510	−22.22
want money	make	0.1578	−15.16	handle	−0.0929	−20.91	represent	−0.4031	−22.92
want programme	produce	−0.1980	−18.86	teach	−0.1743	−20.94	hate	−0.6788	−22.63

B.2. Experiment 4

The experimental item in (65) was presented as the modulus in Experiment 4. The experimental materials are shown in Tables 24 and 25.

(65) hard substance a substance that is hard to alter

Table 24
Materials for Experiment 4, with Mean Ratings (object interpretations)

Adjective-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Rtg	Prob
difficult consequence	cope with	.3834	−18.61	analyse	.0270	−20.43	refer to	−.3444	−24.68
difficult customer	satisfy	.4854	−20.27	help	.3228	−22.20	drive	−.4932	−22.64
difficult friend	live with	.2291	−19.11	approach	.0798	−21.75	miss	−.5572	−23.04
difficult group	work with	.3066	−18.64	teach	.2081	−21.00	respond to	.1097	−22.66
difficult hour	endure	.3387	−21.17	complete	−.1386	−21.83	enjoy	.1600	−23.18
easy comparison	make	.4041	−17.73	discuss	−.0901	−22.09	come to	.3670	−23.03
easy food	cook	.2375	−18.93	introduce	−.3673	−21.94	finish	−.1052	−23.15
easy habit	get into	.2592	−17.48	explain	−.2877	−21.79	support	−.0523	−23.70
easy point	score	.3255	−18.77	answer	.1198	−20.65	know	−.0307	−23.43
easy task	perform	.4154	−17.56	manage	.3094	−21.30	begin	.0455	−22.48
fast device	drive	−.0908	−22.35	make	.2948	−23.65	see	−.4817	−25.00
fast launch	stop	−.5438	−23.79	make	.0075	−25.19	see	−.3963	−25.84
fast pig	catch	−.5596	−23.98	stop	.6285	−24.30	use	−.5350	−25.66
fast rhythm	beat	.0736	−20.46	feel	−.1911	−25.24	make	−.1296	−25.60
fast town	protect	−.5896	−23.66	make	−.4564	−23.90	use	−.4996	−25.66

Table 24
(*continued*)

Adjective-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Rtg	Prob
good climate	grow up in	.2343	−19.65	play in	−.6498	−22.18	experience	−.4842	−23.20
good documentation	use	.2549	−21.89	produce	−.2374	−22.38	include	.1110	−23.73
good garment	wear	.2343	−20.23	draw	−.6498	−22.71	measure	−.4842	−23.16
good language	know	.2188	−19.18	reinforce	−.1383	−22.48	encourage	−.0418	−22.81
good postcard	send	.1540	−20.17	draw	−.3248	−22.71	look at	.2677	−23.34
hard logic	understand	.2508	−18.96	express	.0980	−22.76	impose	−.2398	−23.11
hard number	remember	.0326	−20.30	use	−.3428	−21.14	create	−.4122	−22.69
hard path	walk	.2414	−21.08	maintain	.0343	−21.64	explore	.1830	−23.01
hard problem	solve	.4683	−15.92	express	.0257	−21.26	admit	−.2913	−23.39
hard war	fight	.2380	−17.31	get through	.2968	−21.32	enjoy	−.5381	−23.18
slow child	adopt	−.5028	−19.72	find	−.7045	−22.52	forget	−.6153	−23.93
slow hand	grasp	−.5082	−18.03	win	.2524	−22.07	produce	−.3360	−22.52
slow meal	provide	−.0540	−19.55	begin	.2546	−21.14	bring	−.3965	−23.29
slow minute	take	−.1396	−19.48	fill	.1131	−22.06	meet	−.6083	−23.30
slow progress	make	.3617	−18.50	bring	−.1519	−22.64	give	−.2700	−24.89
safe building	use	.1436	−20.83	arrive at	−.1640	−23.55	come in	.0306	−23.96
safe drug	release	.1503	−23.23	try	.1930	−23.62	start	.1614	−24.31
safe house	go to	.2139	−20.87	get	−.3438	−22.41	make	−.3490	−23.19
safe speed	arrive at	−.0242	−23.55	keep	.1498	−23.59	allow	.2093	−25.04
safe system	operate	.2431	−19.78	move	−.2363	−22.85	start	.0013	−24.60
right accent	speak in	.1732	−19.90	know	−.1223	−22.50	hear	.0946	−22.79
right book	read	.1938	−18.89	lend	−.0188	−22.60	suggest	.0946	−24.90
right school	apply to	.2189	−21.76	complain to	−.3736	−22.82	reach	−.2756	−23.69
right structure	build	−.1084	−19.88	teach	−.1084	−22.76	support	−.0505	−24.92
right uniform	wear	.1990	−19.79	provide	−.1084	−24.09	look at	−.0505	−25.24
wrong author	accuse	−.1925	−21.90	read	.0450	−24.09	consider	.0653	−24.50
wrong colour	use	.2366	−21.78	look for	.0587	−22.78	look at	−.1907	−24.89
wrong note	give	.0222	−22.29	keep	.2014	−22.64	accept	−.1462	−24.16
wrong post	assume	−.3000	−20.10	make	.2579	−23.81	consider	−.0466	−24.50
wrong strategy	adopt	.2804	−19.86	encourage	.1937	−23.51	look for	.0135	−24.39

Table 25
Materials for Experiment 4, with Mean Ratings (subject interpretations)

Adjective-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Interpr	Prob
difficult customer	buy	−.2682	−20.19	pick	−.2050	−22.58	begin	−.3560	−24.83
difficult friend	explain	−.4658	−20.16	neglect	−.5274	−21.75	enjoy	−.4711	−23.59
difficult passage	read	.1668	−20.46	speak	−.3600	−22.62	appear	−.4030	−23.78
difficult piece	read	.1052	−20.30	survive	−.5080	−22.28	continue	−.1006	−23.89
difficult spell	break	−.3047	−19.80	create	−.2412	−22.81	start	−.3661	−23.50

Table 25
(*continued*)

Adjective-noun	High			Medium			Low		
	Interpr	Rtg	Prob	Interpr	Rtg	Prob	Interpr	Interpr	Prob
easy car	start	−.1652	−21.01	move	−.2401	−21.42	close	−.5750	−23.94
easy change	occur	.1999	−20.32	prove	.2332	−21.57	sit	−.0932	−23.27
easy food	cook	.0443	−20.28	change	−.6046	−22.25	form	−.3918	−22.96
easy habit	develop	.1099	−21.43	start	.1156	−23.21	appear	−.3490	−24.70
easy task	fit	−.3882	−20.77	end	−.0982	−22.90	continue	.0474	−24.11
fast device	go	.2638	−22.80	come	−.3652	−23.91	add	−.2219	−24.68
fast horse	run	.4594	−20.78	work	.0025	−22.98	add	−.5901	−24.64
fast lady	walk	−.0261	−22.31	work	−.0716	−23.90	see	−.4816	−25.15
fast pig	run	.2081	−22.57	come	−.1807	−23.91	get	−.2764	−24.75
fast town	grow	−.3601	−18.66	spread	−.3289	−22.84	sell	−.3462	−24.33
good ad	read	.1248	−22.39	sell	.2154	−22.72	run	−.0832	−22.78
good climate	change	−.3748	−21.30	improve	−.3312	−22.57	begin	−.4093	−23.36
good egg	look	−.0581	−21.79	develop	−.2457	−22.02	appear	.1149	−24.01
good light	work	−.0022	−19.42	spread	−.2023	−21.75	increase	−.4349	−23.82
good show	run	.0787	−21.07	continue	−.0798	−22.79	die	−.6569	−23.79
hard fish	bite	−.3583	−20.39	pull	−.2579	−22.53	appear	−.2568	−22.79
hard logic	get	−.1211	−21.90	sell	−.4533	−22.18	go	−.4388	−24.59
hard substance	keep	.0227	−21.28	remain	.0978	−22.93	seem	.1971	−23.03
hard toilet	flush	−.1796	−20.46	look	−.3465	−23.77	start	−.6835	−24.60
hard war	break out	−.4969	−16.94	grow	−.2792	−22.21	increase	−.2602	−23.60
safe building	approach	−.6815	−22.87	stay	.0852	−23.09	start	−.5152	−23.53
safe drug	come	−.3802	−19.41	play	−.5562	−22.24	try	−.3126	−22.45
safe man	eat	−.5434	−21.23	ignore	−.6673	−21.89	agree	−.6509	−23.16
safe speed	go	−.3116	−16.26	leave	−.6136	−20.24	remain	.1267	−22.87
safe system	operate	.3697	−19.92	continue	.0374	−21.48	think	−.4845	−22.90
slow child	react	.1485	−22.66	adapt	.1556	−23.21	express	−.0256	−24.60
slow hand	move	.0738	−22.24	draw	.0039	−23.38	work	−.0346	−25.56
slow meal	go	.1237	−20.57	run	−.1474	−21.47	become	−.2802	−23.17
slow minute	pass	.2717	−23.17	start	−.4423	−23.49	win	−.6709	−24.64
slow sleep	come	−.0671	−19.56	follow	−.3108	−22.56	seem	−.2169	−24.85
right accent	go	−.1727	−18.54	sound	.1928	−21.14	fall	−.3926	−22.76
right book	read	−.2429	−19.50	feel	−.1027	−21.74	discuss	−.2195	−23.61
right character	live	−.2505	−22.65	set	−.4063	−23.14	feel	−.1651	−23.92
right people	vote	−.4541	−21.70	eat	−.5921	−22.81	answer	−.0992	−24.68
right school	teach	.0159	−20.91	start	−.3466	−24.03	stand	−.3839	−24.88
wrong author	go	−.4348	−20.59	think	−.4128	−22.96	read	−.5542	−24.50
wrong business	think	−.4018	−23.15	spend	−.4416	−23.85	hope	−.5608	−24.18
wrong colour	go	.1846	−20.93	show	−.0819	−23.54	seem	.2869	−24.18
wrong note	conclude	−.2575	−21.24	show	−.3480	−23.87	tell	−.2732	−24.45
wrong policy	encourage	−.0401	−21.71	identify	−.2167	−23.56	accept	.0183	−23.76

Appendix C. Descriptive Statistics

Table 26 displays the descriptive statistics for the model probabilities and the subject ratings for Experiments 1 and 4.

Table 26
Descriptives for model probabilities and subject ratings

Model Probabilities, Experiment 1					
Rank	Mean	StdDev	StdEr	Min	Max
High	−21.62	2.59	.26	−24.62	−15.11
Medium	−22.65	2.19	.22	−26.32	−20.90
Low	−23.23	2.28	.22	−27.92	−22.22
Subject Ratings, Experiment 1					
Rank	Mean	StdDev	StdEr	Min	Max
High	0.1449	0.2355	0.0309	−0.2697	0.3297
Medium	−0.1055	0.2447	0.0321	−0.4361	0.2324
Low	−0.3848	0.2728	0.0358	−0.7269	0.0336
Model Probabilities, Experiment 4					
Rank	Mean	StdDev	StdEr	Min	Max
High	−20.49	1.71	.18	−23.99	−15.93
Medium	−22.62	.99	.10	−25.24	−20.24
Low	−23.91	.86	.18	−25.85	−22.46
Subject Ratings, Experiment 4					
Rank	Mean	StdDev	StdEr	Min	Max
High	−.0005	.2974	.0384	−.68	.49
Medium	−.1754	.3284	.0424	−.70	.31
Low	−.2298	.3279	.0423	−.68	.37

Figure 1

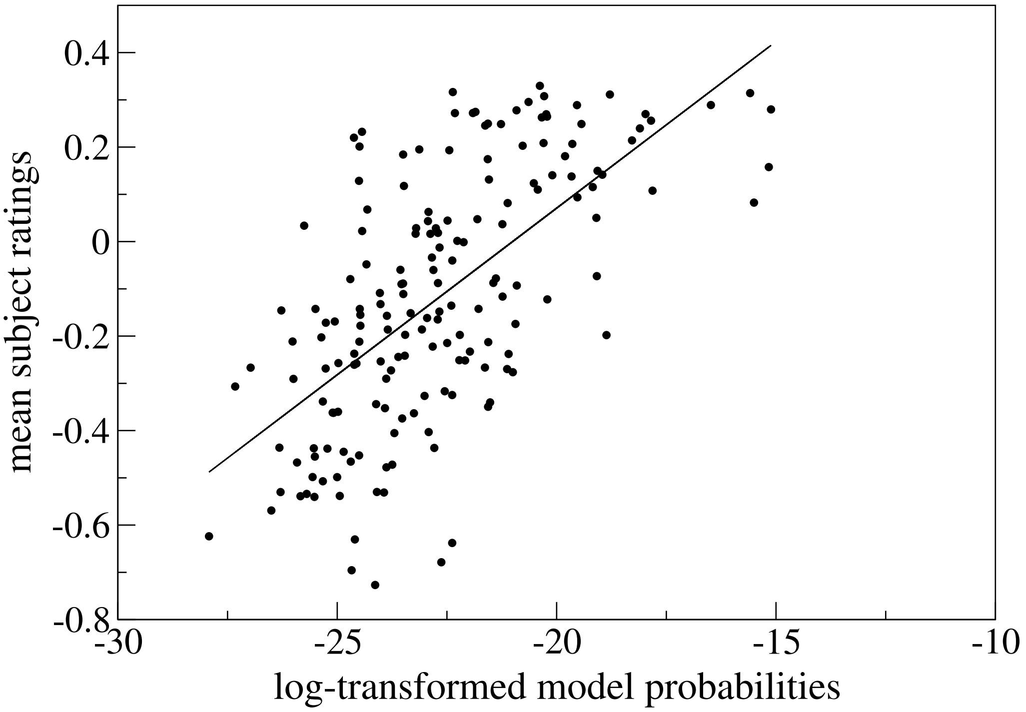
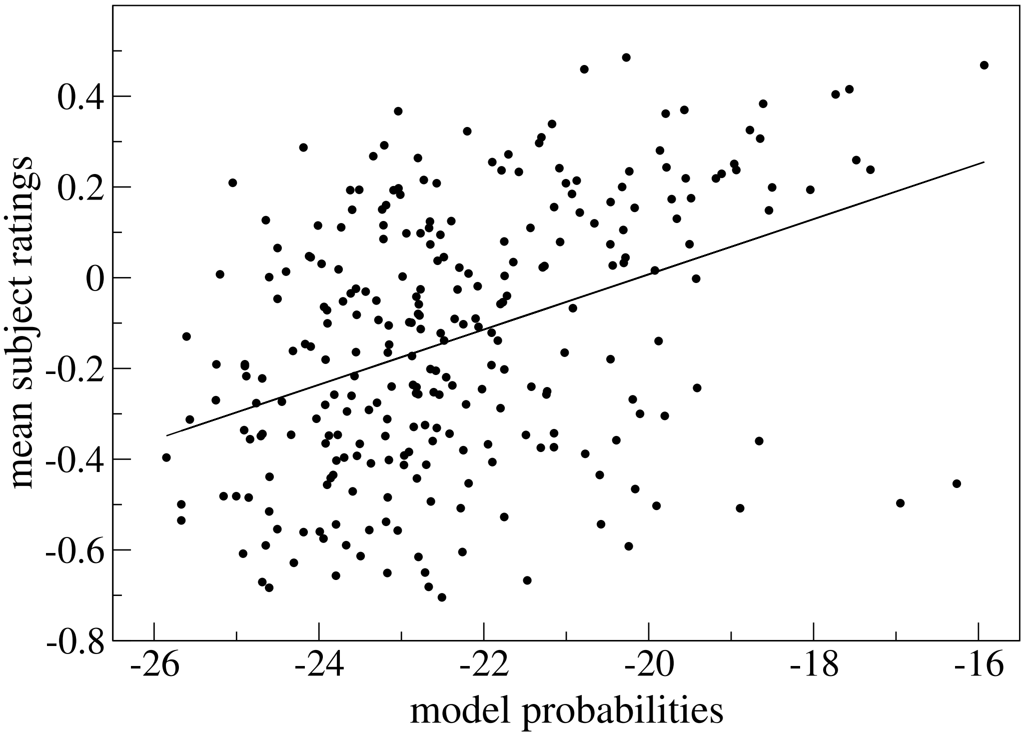

0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
mean subject ratings
-30
-25
-20
-15
-10
log-transformed model probabilities

Figure 2


0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
mean subject ratings
-26
-24
-22
-20
-18
-16
model probabilities