A Corpus-based Account of Regular Polysemy: The Case of Context-sensitive Adjectives

Maria Lapata*
Department of Computational Linguistics
Saarland University
PO Box 15 11 50
66041 Saarbrücken, Germany
mlap@coli.uni-sb.de

Abstract

In this paper we investigate polysemous adjectives whose meaning varies depending on the nouns they modify (e.g., *fast*). We acquire the meanings of these adjectives from a large corpus and propose a probabilistic model which provides a ranking on the set of possible interpretations. We identify lexical semantic information automatically by exploiting the consistent correspondences between surface syntactic cues and lexical meaning. We evaluate our results against paraphrase judgments elicited experimentally from humans and show that the model's ranking of meanings correlates reliably with human intuitions: meanings that are found highly probable by the model are also rated as plausible by the subjects.

1 Introduction

Much recent 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. Adjectives, more than other categories, are a striking example of regular polysemy since they are able to take on different meanings depending on their context, viz., the noun or noun class they modify (see Pustejovsky (1995) and the references therein).

The adjective *fast* in (1) receives different interpretations when modifying the nouns *programmer*, *plane* and *scientist*. A *fast programmer* is typically a programmer who programs quickly, a *fast plane* is typically a plane that flies quickly, a *fast scientist* can be a scientist who publishes papers quickly, who performs experiments quickly, who observes something quickly, who reasons, thinks, or runs quickly. Interestingly, adjectives like *fast* are ambiguous across and within the nouns they modify. A *fast plane* is not only a plane that flies quickly, but also a plane that lands, takes off, turns, or travels quickly. Even the more restrictive *fast programmer* allows more than one interpretation. One can easily think of a context where a *fast programmer* thinks, runs or talks quickly.

(1) a. fast programmer
b. fast plane
c. fast scientist

(2) a. easy problem
b. difficult language
c. good cook
d. good soup

Adjectives like *fast* have been extensively studied in the lexical semantics literature and their properties have been known at least since Vendler (1968). The meaning of adjective-noun combinations like those in (1) and (2) 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". In order to account for the meaning of these combinations Vendler (1968, 92) points out that "in most cases not one verb, but a family of verbs is needed". Vendler 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 (1)), *easy* and *difficult* trigger verb-object interpretations (see (2a,b)). An *easy problem* is usually a problem that is easy to solve, whereas a *difficult language* is a language that is difficult to learn, speak, or 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 soup that tastes good or soup that is good to eat.

Pustejovsky (1995) avoids enumerating the various senses for adjectives like *fast* by exploiting the semantics of the nouns they modify. Pustejovsky treats nouns as having a *qualia structure* as part of their lexical entries, which among other things, specifies possible events associated with the entity. For example, the telic (purpose) role of the qualia structure for *problem* has a value equivalent to *solve*. When the adjective *easy* is combined with *problem*, it predicates over the telic role of *problem* and consequently the adjective-noun combination receives the interpretation a problem that is easy to solve.

Pustejovsky (1995) does not give an exhaustive list of the telic roles a given noun may have. Furthermore, in cases where more than one interpretations are provided (see Vendler (1968)), no information is given with respect to the likelihood of these interpretations. Out-of context, the number of interpretations for *fast scientist* is virtually unlimited, yet some interpretations are more likely than others: *fast scientist* is more likely to

*The work reported in this paper was carried out while the author was at the Division of Informatics, University of Edinburgh.

be a scientist who performs experiments quickly or who publishes quickly than a scientist who draws or drinks quickly.

In this paper we focus on polysemous adjective-noun combinations (see (1) and (2)) and attempt to address the following questions: (a) Can the meanings of these adjective-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 combines 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 British National Corpus (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). We evaluate our results by comparing the model's predictions against human judgments and show that the model's ranking of meanings correlates reliably with human intuitions.

2 The Model

2.1 Formalization of Adjective-Noun Polysemy

In order to come up with the meaning 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*, which in turn is modified by the adverbs *fast* or *quickly*. In the general case we want to paraphrase the meaning of an adjective-noun combination by finding the verbs that 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 $P(a,n,v,rel)$ where v is the verbal predicate modified by the adverb a (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. We rewrite $P(a,n,v,rel)$ using the chain rule in (3).

(3) $$P(a,n,v,rel) = P(v) \cdot P(n|v) \cdot P(a|v,n) \cdot P(rel|v,n,a)$$

Although the parameters $P(v)$ and $P(n|v)$ can be straightforwardly estimated from the BNC, the estimation of $P(rel|v,n,a)$ and $P(a|v,n)$ is somewhat problematic. In order to obtain $P(rel|v,n,a)$ we must estimate the frequency $f(v,n,a,rel)$ (see (4)).

(4) $$P(rel|v,n,a) = \frac{f(v,n,a,rel)}{f(v,n,a)}$$

One way to acquire $f(v,n,a,rel)$ 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. Even if we could accurately parse the corpus, it is questionable whether we can find enough data for the estimation of $f(v,n,a,rel)$. There are only six sentences in the entire BNC that can be used to estimate $f(v,n,a,rel)$ for the adjective-noun combination *fast plane* (see (5a)–(5f)). 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. This is rather counter-intuitive since *fast planes* are more likely to fly than swoop in fast. For the adjective-noun combination *fast programmer* there is only one sentence relevant for the estimation of $f(v,n,a,rel)$ in which the modifying adverbial is not *fast* but the semantically related *quickly* (see (6)). The sparse data problem carries over to the estimation of the frequency $f(v,n,a)$.

(5) a. Three planes swooped in, fast and low.
b. The plane was dropping down fast towards Bangkok.
c. The unarmed plane flew very fast and very high.
d. The plane went so fast it left its sound behind.
e. And the plane's going slightly faster than the Hercules or Andover.
f. He is driven by his ambition to build a plane that goes faster than the speed of sound.

(6) It means that programmers will be able to develop new applications more quickly.

We avoid these estimation problems by reducing the parameter space. In particular, we make the following independence assumptions:

(7) $$P(a|v,n) \approx P(a|v)$$

(8) $$P(rel|v,n,a) \approx P(rel|v,n)$$

We assume that the likelihood of an adverb modifying a verb is independent of the verb's arguments (see (7)). Accordingly, we assume that knowing that the adverb a modifying the verb v will contribute little information to the likelihood of the relation *rel* which depends more on the verb and its argument n (see (8)). By substituting (7) and (8) into (3), $P(a,n,v,rel)$ can be written as:

(9) $$P(a,n,v,rel) \approx P(v) \cdot P(n|v) \cdot P(a|v) \cdot P(rel|v,n)$$

We estimate the probabilities $P(v)$, $P(n|v)$, $P(a|v)$, and $P(rel|v,n)$ as follows:

(10) $$P(v) = \frac{f(v)}{\sum_i f(v_i)}$$

(11) $$P(n|v) = \frac{f(n,v)}{f(v)}$$

(12) $$P(a|v) = \frac{f(a,v)}{f(v)}$$

(13) $$P(rel|v,n) = \frac{f(rel,v,n)}{f(v,n)}$$

By substituting equations (10)–(13) into (9) and simplifying the relevant terms, (9) is rewritten as follows:

$$(14)\quad P(a,n,v,rel) \approx \frac{f(rel,v,n)\cdot f(a,v)}{f(v)\cdot \sum_i f(v_i)}$$

Depending on the data (noisy or not) and the task at hand we may choose to estimate the probability $P(v,n,a,rel)$ from reliable corpus frequencies only (e.g., $f(a,v) > 1$ and $f(rel,v,n) > 1$). If we know the interpretation preference of a given adjective (i.e., subject or object), we may vary only the term *v*, keeping the terms *n*, *a* and *rel* constant. Alternatively, as we show in Experiment 1 (see Section 3), we may acquire the interpretation preferences automatically by varying both the terms *rel* and *v*.

2.2 Parameter Estimation

We estimated the parameters described in the previous section from a part-of-speech tagged and lemmatized version of the BNC (100 million words). The estimation of the terms $f(v)$ and $\sum_i f(v_i)$ (see (14)) reduces to the number of times a given verb is attested in the corpus. In order to estimate the terms $f(rel,v,n)$ and $f(a,v)$ the corpus was automatically parsed by Cass (Abney, 1996), a robust chunk parser designed for the shallow analysis of noisy text. We used the parser's built-in function to extract tuples of verb-subjects and verb-objects (see (15)). 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, the verb is missing from (16a) and the noun is missing from (16b).

(15) a. change situation SUBJ
b. come off heroin OBJ
c. deal with situation OBJ

(16) a. isolated people SUBJ
b. smile good SUBJ

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. Non-auxiliary instances of the verb *be* (e.g., OBJ be embassy) were also eliminated since they contribute no semantic information with respect to the events or states that are possibly associated with the noun with which the adjective is combined. Particle verbs (see (15b)) 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 (15c)). 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, upon* were prepositional objects. This resulted in 737,390 distinct types of verb-subject pairs and 1,077,103 distinct types of verb-object pairs.

Generally speaking, the frequency $f(a,v)$ represents not only a verb modified by an adverb derived from the adjective in question (see (17a)) but also constructions like the ones in (17b,c) where the adjective takes an infinitival VP complement whose logical subject can be realized as a *for*-PP (see (17c)). It is relatively straightforward to develop an automatic process which maps an adjective to its corresponding adverb, modulo exceptions and idiosyncrasies. However in the experiments described in the following sections this mapping was manually specified.

(17) 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*

We estimated the frequency $f(a,v)$ by collapsing the counts from cases where the adjective was followed by an infinitival complement (see (17b,c)) and cases where the verb was modified by the adverb corresponding to the related adjective (see (17a)). We focused only on instances where the verb and the adverbial phrase modifying it (AdvP) were adjacent and extracted the verb and the head of the AdvP immediately following or preceding it. From constructions with adjectives immediately followed by infinitival complements with an optionally intervening *for*-PP (see (17c)) we extracted the adjective and the main verb of the infinitival complement.

2.3 Comparison against the Literature

In what follows we explain the properties of the model by applying it to a small number of adjective-noun combinations taken from the lexical semantics literature. Table 1 gives the interpretations of eight adjective-noun combinations discussed in Pustejovsky (1995) and Vendler (1968). Table 2 shows the five most likely interpretations for these combinations as derived by the model discussed in the previous sections (v_1 is the most likely interpretation, v_2 is the second most likely interpretation, etc.).

First notice that our model predicts variation in meaning when the same adjective modifies different nouns by providing different interpretations for *easy problem* and *easy planet* (see Table 2). Our model agrees with Vendler (1968) in the interpretation of *easy problem* (see Tables 1 and 2). Furthermore, it provides the additional meanings "a problem that is easy to deal with, identify, tackle, and handle". Although the model does not derive Vendler's interpretation of *easy planet*, it produces complementary meanings such as "a planet that is easy to predict, identify, plunder, work with". Similarly, although the model does not discover the suggested interpretation for *good umbrella* it comes up with the plausible meaning "an umbrella that covers well". In fact the latter can be considered as a subtype of the meaning suggested by Pustejovsky (1995): an umbrella functions well if it opens well, closes well, covers well, etc. Although Pustejovsky suggests only a subject-related interpretation for *good umbrella*, the model also derives plausible object-related interpretations: "an umbrella that is good to keep, good for waving, good to hold, good to run for, good to leave".

Adjective	Interpretation	
easy problem	a problem that is easy to solve	(Vendler, 1968, 97)
easy planet	a planet that is easy to observe	(Vendler, 1968, 99)
good umbrella	an umbrella that functions well	(Pustejovsky, 1995, 43)
good shoe	a shoe that is good for wearing, for walking	(Vendler, 1968, 99)
fast horse	a horse that runs fast	(Vendler, 1968, 92)
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)

Table 1: Paraphrases for adjective-noun combinations taken from the literature

$P(v,n,a,rel)$	v_1	v_2	v_3	v_4	v_5
$P(v,$*problem*$,$*easy*$,\text{OBJ})$	solve	deal with	identify	tackle	handle
$P(v,$*planet*$,$*easy*$,\text{OBJ})$	predict	identify	plunder	see on	work with
$P(v,$*umbrella*$,$*good*$,\text{SUBJ})$	cover				
$P(v,$*umbrella*$,$*good*$,\text{OBJ})$	keep	wave	hold	run for	leave
$P(v,$*shoe*$,$*good*$,\text{OBJ})$	wear	keep	buy	get	stick
$P(v,$*horse*$,$*fast*$,\text{OBJ})$	run	learn	go	come	rise
$P(v,$*language*$,$*difficult*$,\text{OBJ})$	understand	interpret	learn	use	speak
$P(v,$*careful*$,$*scientist*$,\text{SUBJ})$	calculate	proceed	investigate	study	analyse
$P(v,$*comfortable*$,$*chair*$,\text{OBJ})$	sink into	sit on	lounge in	relax in	nestle in

Table 2: Model-derived paraphrases for adjective-noun combinations, ranked in order of likelihood

The model and Vendler (1968) agree in their interpretation of the pairs *good shoe* and *fast horse*. The model additionally acquires the fairly plausible meanings "a shoe that is good to keep, to buy, and get" for *good shoe* and "a horse that learns, goes, comes and rises fast" for *fast horse*. The model's interpretations for *difficult language* are a superset of the meanings suggested by Vendler (see Table 1). The model's interpretations for *careful scientist* seem intuitively plausible (even though they don't overlap with those suggested by Vendler). Finally, note that the meanings derived for *comfortable chair* are also plausible (the second most likely meaning is the one suggested by Vendler, see Table 1).

The examples in Table 1 may not be entirely representative of the types of polysemous adjective-noun combinations occurring in unrestricted text since they are taken from linguistic texts where emphasis is given on explaining polysemy with examples that straightforwardly illustrate it. In other words, the adjective-noun combinations in Table 1 may be too easy for the model to handle. In Experiment 1 (see Section 3) we test our model on polysemous adjective-noun combinations randomly sampled from the BNC, and formally evaluate our results against human judgments.

3 Experiment 1: Comparison against Human Judgments

3.1 Method

The ideal test of the proposed model of adjective-noun polysemy will be with randomly chosen materials. We evaluate the acquired meanings by comparing the model's rankings against judgments of meaning paraphrases elicited experimentally from human subjects. By comparing the model-derived meanings against human intuitions we are able to explore: (a) whether plausible meanings are ranked higher than implausible ones; (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; (c) whether there is a linear relationship between the model-derived likelihood of a given meaning and its perceived plausibility, using correlation analysis.

3.1.1 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, wrong*). These adjectives had to be unambiguous with respect to their part-of-speech: 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.

We identified adjective-noun pairs using Gsearch (Corley et al., 2000), a chart parser which detects syntactic patterns in a tagged corpus by exploiting a user-specified context free grammar and a syntactic query. Gsearch was run on a lemmatized version of the BNC so as to compile a comprehensive corpus count of all nouns occurring in a modifier-head relationship with each of the nine adjectives. From the syntactic analysis provided by

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

Table 3: Randomly selected example stimuli with log-transformed probabilities derived by the model

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.

We used the model outlined in Section 2 to derive meanings for the 90 adjective-noun combinations. We employed no threshold on the frequencies $f(a,v)$ and $f(rel,v,n)$. In order to obtain the frequency $f(a,v)$ the adjective was mapped to its corresponding adverb. In particular, *good* was mapped to *good* and *well*, *fast* to *fast*, *easy* to *easily*, *hard* to *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, the frequency $f(difficult,v)$ was estimated only on the basis of infinitival constructions (see (17)). We estimated the probability $P(a,n,v,rel)$ for each adjective-noun pair by varying both the terms *v* and *rel*.

In order to generate stimuli covering a wide range of model-derived paraphrases corresponding to different degrees of likelihood, for each adjective-noun combination we divided the set of the derived meanings into three probability "bands" (High, Medium, and Low) of equal size and randomly chose one interpretation from each band. The division ensured that the experimental stimuli represented the model's behavior for likely and unlikely paraphrases and enabled us to test the hypothesis that likely paraphrases correspond to high ratings and unlikely paraphrases correspond to low ratings. 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 3 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 as follows: our initial design created two sets of stimuli, 270 subject-related stimuli and 270 object-related stimuli. For each stimuli set 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 adjectives with both subject-related and object-related interpretations. Each experimental item consisted of an adjective-noun pair and a sentence paraphrasing its meaning. Paraphrases were created by the experimenter by converting the model's output to a simple phrase, usually a noun modified by a relative clause. A native speaker of English was asked to confirm that the paraphrases were syntactically well-formed.

3.1.2 Procedure

The experimental paradigm was Magnitude Estimation (ME), a technique standardly used in psychophysics to measure judgments of sensory stimuli Stevens (1975), which Bard et al. (1996) and Cowart (1997) have applied to the elicitation of linguistic judgments. ME has been shown to provide fine-grained measurements of linguistic acceptability which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers.

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 et al., 1996; Schütze, 1996).

In the present experiment, the subjects were instructed to judge how well a sentence paraphrases an adjective-noun combination proportional to a modulus item. 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 et al., 1998), an interactive software package for administering web-based psychological experiments. Subjects first saw a set of instructions that explained the ME technique and included some examples, and had to fill in a short questionnaire including basic demographic information. Each subject group saw 135 experimental stimuli (i.e., adjective-noun pairs and their paraphrases). Subjects were assigned to subject groups at random, and a random stimulus order was generated for each subject.

3.1.3 Subjects

The experiment was completed by 60 unpaid volunteers, all native speakers of English. Subjects were recruited via postings to local Usenet newsgroups.

3.2 Results

As is standard in magnitude estimation studies (Bard et al., 1996), statistical tests were done using geometric means to normalize the data (the geometric mean is the mean of the logarithms of the ratings).

We first performed an analysis of variance (ANOVA) to determine whether there is a relation between the paraphrases derived by the model and their perceived likelihood. In particular, we tested the hypothesis that meanings assigned high probabilities by the model are perceived as better paraphrases by the subjects and correspondingly that meanings with low probabilities are perceived as worse paraphrases. The descriptive statistics for log-transformed model-derived probabilities are shown in Table 4. The ANOVA revealed that the Probability Band effect was significant, in both by-subjects and by-items analyses: $F_1(2,118) = 101.46$, $p < .01$; $F_2(2,88) = 29.07$, $p < .01$. The geometric mean of the ratings in the High band was $-.0005$, compared to Medium items at $-.1754$ and Low items at $-.2298$ (see Table 5). Post-hoc Tukey tests indicated that the differences between all pairs of conditions were significant at $\alpha = .01$ in the by-subjects analysis. The difference between High and Medium items as well as High and Low items was significant at $\alpha = .01$ in the by-items analysis, whereas the difference between Medium and Low items did not reach significance. These results show that meaning paraphrases derived by the model correspond to human intuitions: paraphrases assigned high probabilities by the model are perceived as better than paraphrases that are assigned low probabilities.

We further explored the linear relationship between the subjects' rankings and the corpus-based model, using correlation analysis. The elicited judgments were compared with the interpretation probabilities which were obtained from the model described in Section 2 to examine the extent to which the proposed interpretations correlate with human intuitions. A comparison between our model and the human judgments yielded a Pearson correlation coefficient of .40 ($p < .01$, $N = 270$). This verifies the Probability Band effect discovered by the ANOVA, in an analysis which compares the individual interpretation likelihood for each item with elicited interpretation preferences, instead of collapsing all the items in three equivalence classes (i.e., High, Medium, Low). In order to evaluate whether the grammatical function has any effect on the relationship between the model-derived meanings and the human judgments, we split the items into those that received a subject interpretation versus those that received an object interpretation. A comparison between our model and the human judgments 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.

Rank	μ	SD	SE	Min	Max
High	-20.5	1.71	.18	-24.0	-15.9
Medium	-22.6	.99	.10	-25.2	-20.2
Low	-23.9	.86	.18	-25.9	-22.5

Table 4: Descriptive statistics for model-derived probabilities

Rank	μ	SD	SE	Min	Max
High	$-.0005$	.2974	.0384	$-.68$	.49
Medium	$-.1754$	.3284	.0424	$-.70$	.31
Low	$-.2298$	.3279	.0423	$-.68$	.37

Table 5: Descriptive statistics for Experiment 1, by subjects

An important question is how well humans agree in their paraphrase judgments for adjective-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. For each subject group we performed correlations on the elicited judgments using leave-one-out resampling (Weiss and Kulikowski, 1991). For the first group, the average inter-subject agreement was .67 (Min = .03, Max = .82, SD = .14), and for the second group .65 (Min = .05, Max = .82, SD = .14). This means that our model performs satisfactorily given that humans do not perfectly agree in their judgments.

The elicited judgments can be further used to derive

Adj	Model	μ	SD	SE	Subjects	μ	SD	SE
difficult	$\surd$ OBJ	−21.6	1.36	.04	$\surd$ OBJ	.07	.36	.07
	SUBJ	−21.8	1.34	.05	SUBJ	−.29	.28	.05
easy	$\surd$ OBJ	−21.6	1.51	.05	$\surd$ OBJ	.10	.34	.06
	SUBJ	−22.1	1.36	.06	SUBJ	−.14	.23	.04
fast	OBJ	−24.2	1.27	.13	OBJ	−.35	.29	.05
	$\surd$ SUBJ	−23.8	1.40	.14	$\surd$ SUBJ	−.15	.45	.08
good	OBJ	−22.1	1.28	.06	OBJ	−.01	.39	.07
	SUBJ	−22.3	1.10	.07	SUBJ	−.16	.30	.05
hard	$\surd$ OBJ	−21.7	1.53	.06	$\surd$ OBJ	.01	.34	.06
	SUBJ	−22.1	1.35	.06	SUBJ	−.25	.24	.04
right	$\surd$ OBJ	−21.7	1.36	.04	$\surd$ OBJ	−.01	.25	.05
	SUBJ	−21.8	1.24	.04	SUBJ	−.24	.44	.08
safe	OBJ	−22.7	1.48	.10	$\surd$ OBJ	.01	.25	.05
	$\surd$ SUBJ	−22.4	1.59	.12	SUBJ	−.34	.43	.08
slow	OBJ	−22.5	1.53	.08	OBJ	−.30	.48	.08
	SUBJ	−22.3	1.50	.07	$\surd$ SUBJ	−.09	.24	.04
wrong	OBJ	−23.2	1.33	.08	$\surd$ OBJ	−.04	.25	.05
	SUBJ	−23.3	1.30	.08	SUBJ	−.24	.37	.08

Table 6: Log-transformed model-derived and subject-based argument preferences for polysemous adjectives

the grammatical function preferences (i.e., subject or object) for a given adjective. In particular, we can determine which is the preferred interpretation for individual adjectives and 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 an 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 judgments for the object interpretation of this particular adjective is 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. Table 6 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. Argument preferences were elicited from human subjects in a similar fashion. For each adjective we gathered the elicited responses pertaining to subject- and object-related interpretations and performed an ANOVA. The biases and the significance of the Grammatical Function effect ($\surd$) are shown in Table 6.

Comparison of the biases derived from the model with ones derived from the elicited judgments shows that the model and the humans are in agreement for all adjectives but *slow*, *wrong* and *safe*. On the basis of human judgments *slow* has a subject bias, whereas *wrong* has an object bias. Although the model could not reproduce this result there is a tendency in the right direction (see Table 6).

Note that in our correlation analysis reported above the elicited judgments 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 behavior by performing separate correlation analyses for preferred and dispreferred biases as determined previously by the ANOVAs conducted for each adjective. 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 judgments 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 judgments, which suggests that the model derives plausible interpretations even in cases where the default argument bias is overridden.

4 Experiment 2: Comparison against Naive Baseline

The probabilistic model described in Section 2 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, setting them, etc.). This simplification is consistent with Pustejovsky's (1995) claim that polysemous adjectives like *easy* are predicates, modifying the events associated with the noun. A "naive" or "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.

4.1 Naive Model

Given an adjective-noun combination we are interested in finding the verbs whose object or subject is the noun appearing in the adjective-noun combination. This can be simply expressed as $P(v|rel,n)$, the conditional probability of a verb v given an argument-noun relation rel,n:

$$(18) \quad P(v|rel,n) = \frac{f(v,rel,n)}{f(rel,n)}$$

The model in (18) assumes that the meaning of an adjective-noun combination is independent of the adjective in question. The model in (18) would 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 frequen-

cies $f(v, rel, n)$ and $f(rel, n)$ from verb-object and verb-subject tuples extracted from the BNC using Cass (Abney, 1996).

4.2 Method

Using the naive model we calculated the meaning probability for each of the 270 stimuli included in Experiment 1. Through correlation analysis we explored the linear relationship between the elicited judgments and the naive baseline model. We further directly compared the two models, our initial, linguistically more informed model, and the naive baseline.

4.3 Results

Using correlation analysis we explored which model performs better at deriving meanings for adjective-noun combinations. A comparison between the naive model's probabilities and the human judgments 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 judgments. 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. Comparison of the two correlation coefficients revealed that their difference was significant ($t(267) = 2.42$, $p < .01$). This means that our original model performs reliably better than a naive baseline at deriving interpretations for polysemous adjective-noun combinations.

We further compared the naive baseline model and the human judgments 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. 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 ($r = .09$, $p = .28$, $N = 135$).

5 Conclusions

In this paper we showed how adjectival meanings can be acquired from a large corpus and provided a probabilistic model which derives a preference ordering on the set of possible interpretations. Our model does not only acquire clusters of meanings (following Vendler's (1968) insight) but furthermore can be used to obtain argument preferences for a given adjective.

We rigorously evaluated the results of our model by eliciting paraphrase judgments from subjects naive to linguistic theory. Comparison between our model and human judgments yielded a reliable correlation of .40 when the upper bound for the task (i.e., inter-subject agreement) is approximately .65. Furthermore, our model performed reliably better than a naive baseline model, which only achieved a correlation of .25. Although adjective-noun polysemy is a well researched phenomenon in the theoretical linguistics literature, the experimental approach advocated here is new to our knowledge.

Furthermore, the proposed model can be viewed as complementary to linguistic theory: it automatically derives 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) it has no means to indicate which ones are likely and which ones are not. Our model fares well on both tasks. It recasts the problem of adjective-noun polysemy in a probabilistic framework deriving 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 a given adjective. These are only implicit in the lexical semantics literature where certain adjectives are exclusively given a verb-subject or verb-object interpretation. We have demonstrated that we can empirically derive argument biases for a given adjective that correspond to human intuitions.

References

Steve Abney. 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.

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

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

Steffan Corley, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2000. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*. To appear.

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

Frank Keller, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998. WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

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

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

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

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

Sholom M. Weiss 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.