To appear in Suzanne Stevenson and Paola Merlo's volume of papers from the 1998 CUNY Sentence Processing Conference (Benjamins)

Verb Sense and Verb Subcategorization Probabilities

Douglas Roland
University of Colorado
Department of Linguistics
Boulder, CO 80309-0295
Douglas.Roland@colorado.edu

Daniel Jurafsky
University of Colorado
Dept. of Linguistics, Computer Science, and Inst. of Cognitive Science
Boulder, CO 80309-0295
jurafsky@colorado.edu

1 Introduction

The probabilistic relation between verbs and their arguments plays an important role in psychological theories of human language processing. For example, Ford, Bresnan and Kaplan (1982) proposed that verbs like *position* have two lexical forms: a more preferred form that subcategorizes for three arguments (SUBJ, OBJ, PCOMP) and a less preferred form that subcategorizes for two arguments (SUBJ, OBJ). Many recent psychological experiments suggest that humans use these kinds of verb-argument preferences as an essential part of the process of sentence interpretation. (Clifton et al. 1984, Ferreira & McClure 1997, Garnsey et al. 1997, MacDonald 1994, Mitchell & Holmes 1985, Boland et al. 1990, Trueswell et al. 1993). It is not completely understood how these preferences are realized, but one possible model proposes that each lexical entry for a verb expresses a conditional probability for each potential subcategorization frame (Jurafsky 1996, Narayanan and Jurafsky 1998).

Unfortunately, different methods of calculating verb subcategorization probabilities yield different results. Recent studies (Merlo 1994, Gibson et al. 1996, Roland & Jurafsky 1997) have found differences between syntactic and subcategorization frequencies computed from corpora and those computed from psychological experiments. Merlo (1994) showed that the subcategorization frequencies derived from corpus data were different from the subcategorization data derived from a variety of psychological protocols. Gibson et al. showed that experimental PP attachment preferences did not correspond with corpus frequencies for the same attachments. In addition, different genres of corpora have been found to have different properties (Biber 1988, 1993).

In an attempt to understand this variation in subcategorization frequencies, we studied five different corpora and found two broad classes of differences.

1) **Context-based Variation**: We found that much of the subcategorization frequency variation could be accounted for by differing contexts. For example the production of sentences in isolation differs from the production of sentences in connected discourse. We show how these contextual differences (particularly differences in the use of anaphora and other syntactic devices for cohesion) directly affect the observed subcategorization frequencies.

2) **Word-sense Variation**: Even after controlling for the above context effects, we found variation in subcategorization frequencies. We show that much of this remaining variation is due to the use of different senses of the same verb. Different verb senses (i.e. different *lemmas*) tend to have different subcategorization probabilities. Furthermore, when context-based variation is controlled for, each verb sense tends towards having unified subcategorization probability across sources.

These two sources of variation have important implications. One important class of implications is for cognitive models of human language processing. Our results suggest that the verb sense or *lemma* is the proper locus of probabilistic expectations. The *lemma* (our definition follows Levelt (1989) and others) is the locus of semantic information in the lexical entry. Thus we assume that the verb *hear* meaning 'to try a legal case' and *hear* meaning 'to perceive auditorily' are distinct lemmas. Also following Levelt, we assume that a lemma expresses expectations for syntactic and semantic arguments. Unlike Levelt and many others, our *Lemma Argument Probability* hypothesis assumes that each verb lemma contains a vector of probabilistic expectations for its possible argument frames. For simplicity, in the experiments reported in this paper we measure these probabilities only for syntactic argument frames, but the *Lemma Argument Probability* hypothesis bears equally on the semantic/thematic expectations shown by studies such as Ferreira and Clifton (1986) and Trueswell et al. (1994).

Our results also suggest that the subcategorization frequencies that are observed in a corpus result from the probabilistic combination of the lemma's expectations and the probabilistic effects of context.

The other important implication of these two sources of variation is methodological. Our results suggest that, because of the inherent differences between isolated sentence production and connected discourse, probabilities from one genre should not be used to normalize experiments from the other. In other words, 'test-tube' sentences are not the same as 'wild' sentences. We also show that seemingly innocuous methodological devices, such as beginning sentences-to-be-completed with proper nouns (*Debbie remembered…*) can have a strong effect on resulting probabilities. Finally, we show that such frequency norms need to be based on the lemma or semantics, and not merely on shared orthographic form.

2 Methodology

We compared five different sources of subcategorization information. Two of these are psychological sources; corpora derived from psychological experiments in which subjects are asked to produce single isolated sentences. We chose two widely-cited studies, Connine et al. (1984) (CFJCF) and Garnsey et al. (1997) (Garnsey). The three non-experimental corpora we used are all on-line corpora which have been tagged and parsed as part of the Penn Treebank project (Marcus et al. 1993): the Brown corpus (BC), the Wall Street Journal corpus (WSJ), and the Switchboard corpus (SWBD). These three all consist of connected discourse and are available from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

Although both sets of psychological data consist of single sentence productions, there are differences. In the study by Connine et al. (1984), subjects were given a list of words (e.g. *charge*) and asked to write sentences using them, based on a given topic or setting (e.g. *downtown*). We used the frequencies published in Connine et al. (1984) as well as the sentences from the subject response sheets, provided by Charles Clifton. In the sentence completion methodology used by Garnsey et al. (1997), subjects are given a sentence fragment and asked to complete it. These fragments consisted of a proper name followed by the verb in the preterite form (i.e. *Debbie remembered* _________). We used the frequency data published for 48 verbs as well as the sentences from the subject response sheets, provided by Sue Garnsey.

We used three different sets of connected discourse data. The Brown corpus is a 1-million-word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc). The texts had all been published in 1961, and the corpus was assembled at Brown University in 1963-1964 (Francis and Kucera 1982). Because the Brown corpus is the only one of our five corpora which was explicitly balanced, and because it has become a standard for on-line corpora, we often use it as a benchmark to compare with the other corpora. The Wall Street Journal corpus is a 1-million word collection of Dow Jones Newswire stories. Switchboard is a corpus of telephone conversations between strangers, collected in the early 1990's (Godfrey et al. 1992). We used only the half of the corpus that was processed by the Penn

Treebank project; this half consists of 1155 conversations averaging 6 minutes each, for a total of 1.4 million words in 205,000 utterances.

We studied the 127 verbs used in the Connine et al. study and the 48 verbs published from the Garnsey et al. study. The Connine et al. and Garnsey et al. data sets have nine verbs in common. Table 1 shows the number of tokens of the relevant verbs that were available in each corpus. It also shows whether the sample size for each verb was fixed or frequency dependent. We controlled for verb frequency in all cross-corpus comparisons.

Corpus	*Token/Type*	*examples per verb*
CFJCF	5,400 (127 CFJCF verbs)	n ≅ either 29, 39, or 68
Garnsey	5,200 (48 Garnsey verbs)	n ≅ 108
BC	21,000 (127 CFJCF verbs) 6,600 (48 Garnsey verbs)	$0 \preceq n \preceq 2{,}644$
WSJ	25,000 (127 CFJCF verbs) 5,700 (48 Garnsey verbs)	$0 \preceq n \preceq 11{,}411$
SWBD	10,000 (127 CFJCF verbs) 4,400 (48 Garnsey verbs)	$0 \preceq n \preceq 3{,}169$

Table 1: Approximate size of each corpus

Deriving subcategorization probabilities from the five corpora involved both automatic scripts and some hand re-coding. Our set of complementation patterns is based in part on our collaboration with the FrameNet project (Baker et al. 1998, Lowe et al. 1997). Our 17 major categories were 0, PP, VPto, Sforto, Swh, Sfin, VPing, VPbrst, NP, [NP NP], [NP PP], [NP Vpto], [NP Swh], [NP Sfin], Quo, Passives, and Other. These categories include only true syntactic arguments and exclude adjuncts, following the distinction made in Treebank (Marcus et al. 1993). We used a series of regular expression searches and *tgrep* scripts[1] to compute probabilities for these subcategorization frames from the three syntactically parsed Treebank corpora (BC, WSJ, SWBD). Some categories (in particular the quotation category Quo) were difficult to code automatically and so were re-coded by hand. Since the Garnsey et al. data used a more limited set of subcategorizations, we re-coded portions of this data into the 17 categories. The Connine et al. data had an additional confound; 4 of the 17 categories did

[1] We evaluated the error rate of our search strings by hand-checking a random sample of our data. The error rate in our data is between 3% and 7%. The error rate is given as a range due to the subjectivity of some types of errors. 2-6% of the error rate was due to mis-parsed sentences in Treebank, including PP attachment errors, argument/adjunct errors, etc. 1% of the error rate was due to inadequacies in our search strings, primarily in locating displaced arguments via the Treebank 1 style notation used in the Brown Corpus data.

not distinguish arguments from adjuncts. Thus we re-coded portions of the Connine et al. data to include only true syntactic arguments and not adjuncts.

We also hand tagged the data from seven verbs for semantic sense. We used the semantic senses provided in Wordnet (Miller et al. 1993). We collapsed across senses in the few cases where we could not reliably distinguish between the Wordnet senses. When there were more than 100 tokens of a verb in a single corpus, we coded the first 100 randomly selected examples. This sample size was chosen to match the maximum sample size in the psychological corpora.

The subcategorization frequencies for a verb can be treated as a vector in multidimensional space. This allowed us to use the cosine of the angle between the vectors (Salton & McGill 1983) as a measure of the agreement between the subcategorization frequencies of verbs in different corpora. Table 2 shows the vectors for the verb *hear* in the Brown corpus and in the Wall Street Journal corpus. Using Formula 1, the cosine of the two vectors shown in Table 2 is 0.98. For non-negative vectors, the cosine ranges from 0 (complementary distribution) to 1 (complete agreement).

hear	*0*	*PP*	*Swh*	*Sfin*	*VPbrst*	*NP*	*NP PP*	*passive*
BC	4	12	3	1	15	47	4	14
WSJ	0	17	3	5	13	56	10	10

Table 2: Raw subcategorization vectors for *hear* from BC and WSJ.

$$\boxed{\text{Cosine} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}}$$

Formula 1: Cosine of two vectors, x and y.

To measure whether the differences shown in the cosine were significant, we performed a chi-squared test on the same vectors, collapsing low frequency categories into an *other* category.

3 Isolated sentence versus connected-discourse corpora

A portion of the subcategorization frequency differences are the result of the inherently different nature of single sentence production and connected discourse sentence production. This section will show that the single sentence / connected discourse opposition affects subcategorization through two general mechanisms: the use of discourse cohesion in connected discourse and the use of default referents in null context (isolated sentence production).

Discourse cohesion

The first difference between single sentence production and connected discourse involves discourse cohesion. Unlike isolated sentences, a sentence in connected discourse must cohere with rest of the discourse. Halliday and Hasan (1976) use the notion of cohesion to show why sentences such as "So we pushed him under the other one" sound odd as the start of a conversation. Because a large number of syntactic phenomena such as pronominalization, fronting, deixis, and passivization play a role in discourse coherence, we would expect these syntactic devices to be used differently in connected discourse than in single sentence production. In addition, to the extent that these syntactic phenomena affect subcategorization, we would expect sentences produced in isolation (such as in the Connine et al. and Garnsey et al. experiments) to have different subcategorization probabilities than sentences found in connected discourse, such as in the Brown corpus, the Wall Street Journal corpus, and the Switchboard corpus. Because we counted dislocated arguments and pronominalized arguments in the same categories as their non-dislocated and full NP counterparts, pronominalization and most kinds of movement do not affect our subcategorization frequencies. Two syntactic devices that do affect our subcategorization frequencies are passivization and zero anaphora.

The passive in English is generally described as having one of two broad functions: (1) de-emphasizing the identity of the agent and (2) keeping an undergoer topic in subject position. (Thompson 1987). Because both of these functions are more relevant for multi-sentence discourse, one would expect that sentences produced in isolation would make less use of passivization. As shown in Table 3, we found a much greater use of the passive in all of the connected discourse corpora than in the isolated sentences from Connine et al.[2]

[2] We also found that there were more passives in the written than in the spoken corpora, supporting Chafe (1992).

Data Source	% passive sentences
Garnsey	—
CFJCF	0.6%
Switchboard	2.2%
Wall Street Journal	6.7%
Brown corpus	7.8%

Table 3: Use of passives in each corpus.

Zero anaphora also plays a role in discourse cohesion. Whether an argument of a verb may be omitted depends on factors such as the semantics of the verb, what kind of omission the verb lexically licenses, the definiteness of the argument, and the nature of the context (Fillmore 1969, 1986; Fraser and Ross 1970; Resnik 1996 *inter alia*). In one common case of zero anaphora, Definite Null Complementation (DNC), "the speaker's authority to omit a complement exists only within an ongoing discourse in which the missing information can be immediately retrieved from the context" (Fillmore, 1986). For example the verb *follow* licenses DNC only if the 'thing followed' can be recovered from the context, as shown in example (1). Because the referent must be recoverable from the context, this type of zero anaphora is unlikely to occur in single sentence production, where the context is limited at best.

(1) The shot reverberated in diminishing whiplashes of sound. Hush **followed**. (Brown corpus)

The lack of Definite Null Complementation in single sentence production results in single sentence corpora having a lower occurrence of the [0] subcategorization frame. For example the direct object of the verb *follow* is often omitted in the connected discourse corpora, but never omitted in the Connine et al. data set. By hand-counting every instance of *follow* in all four corpora, we found that every case of omission was caused by definite null complementation. The referent is usually in a preceding sentence or a preceding clause of the same sentence.

Data Source	*% [0] subcat frame*
Garnsey	—
CFJCF	0%
Wall Street Journal	5%
Switchboard	11%
Brown	22%

Table 4: The object of *follow* is only omitted in connected-discourse corpora (numbers are hand-counted, and indicate % of omitted objects out of all instances of *follow*)

Default referents
In connected discourse, the context controls which referents are used as arguments of the verb. In single sentence production tasks, there is no larger context to provide this influence. In the absence of such demands, one might expect the subjects to use a wider variety of arguments with the verbs. On the contrary, we observe that the subjects favor a set of default referents – those which are accessible in the experimental context, or which are prototypical arguments of the verb. We found three kinds of biases toward these default referents.

First, non-zero subjects of single sentence productions were more likely to be *I* or *we* than subjects in connected discourse. Presumably the participants tended to use themselves as the topic of the sentence since in a null context there was no topic under discussion. Table 5 shows that the single sentence production data has a higher use of first person subjects than the written connected discourse data. Note that the Switchboard corpus also has a higher use of first person subjects. This could reflect a tendency for the participants, who are talking to strangers, to use themselves as a topic, given the absence of shared background.

Data Source	% first person subject
Garnsey	—
CFJCF	40%
Switchboard	39%
Brown corpus	18%
Wall Street Journal	7%

Table 5: Greater use of first person subject in isolated-sentences.

Second, VP internal NPs (e.g. NPs which are c-commanded by the subject of the verb) are more likely to be anaphorically related to the subject of the verb. This includes cases such as (2) where the embedded NP is co-referential with the subject, and cases such as (3) where the embedded NP and the subject are related by a possession or part-whole relationship. To simplify judgement of relatedness, we only counted co-referential pronouns and traces. We did not count inferentially related NPs.

(2) Tom_i noticed that he_i was getting taller. (Garnsey et al. data)

(3) Alice_i prayed that her_i daughter wouldn't die. (Garnsey et al. data)

By contrast, VP-internal NPs in the natural corpora were more likely to refer to referents other than the subject of the verb. This additional sentence-internal anaphora in the isolated sentences is presumably a strategy for avoiding sentences like (4) which require the creation of an additional referent that is not already present in the context.

(4) Alice prayed that Bob's daughter wouldn't die. (made up example)

Table 6 shows how often the subject was anaphorically related to a VP internal NP in a hand-counted random sample of 100 examples from each corpus.

Data Source	*% related subject/NP*
Garnsey	41%
CFJCF	26%
Wall Street Journal	15%
Brown corpus	12%
Switchboard	8%

Table 6: Use of VP-internal NPs which are anaphorically related to the subject.

Third, the objects in the single sentence production data were more likely to be *prototypical* objects. That is, subjects tended to use default, relatively predictable head nouns for the direct objects of verbs. For example, of the 107 Garnsey sentences with the verb *accept,* 12 (11%) had a direct object whose head nouns was *award*. In fact 33% of the 107 sentences had a direct object whose head was one of the most common four words *award*, *fact*, *job*, or *invitation.* By contrast, the 112 Brown corpus sentences used a far greater variety of objects; it would take 12 different object nouns to account for 33% of the 112 sentences. Furthermore, the most common Brown corpus objects were pronouns (*it*, *them*); no common noun occurred more than 3 times in the 112 sentences. A formal metric of argument prototypicality is the token/type ratio. The ratio of the number of object noun tokens to object noun types will be high when a small number of types account for a greater percentage of the tokens. Table 7 shows that the token/type ratio is much higher for Garnsey data set than for the Brown corpus.

Data Source	*token count*	*type count*	*Argument token/type ratio*
Garnsey	107	54	2.0
CFJCF	—	—	—
Wall Street Journal	138	105	1.3
Brown corpus	112	86	1.3
Switchboard	15	14	1.1

Table 7: Token/Type ratio for arguments of *accept*

These uses of default references can all be seen as a device that experimental participants use to avoid introducing multiple new referential expressions into the single sentences. Natural sentences are known to generally contain only one new (inactive) piece of information per intonation contour (Chafe 1987) or clause (Givon 1979, 1984, 1987).

This section has shown several different ways in which discourse context affects observed subcategorization frequencies. These effects suggest that a psychological model of subcategorization probabilities will need to control for such discourse context effects. These contextual effects also have a methodological implication. Because of the biases inherent in isolated sentence production, we should not expect results from such psychological experiments to directly match natural language use.

4 Other experimental factors

The previous section discussed context effects that distinguish isolated sentence corpora from connected discourse corpora. This section discusses a further experimental bias that is specific to the sentence completion task. In sentence completion, the participants are given a prompt consisting of a syntactic subject as well as a verb. The nature of this syntactic subject can influence the verb subcategorization of the resulting sentence. Indeed this fact explains the single largest mismatch between the Garnsey data set and Brown corpus data. The verb *worry* was the only verb in these two corpora with an opposite preference between direct object and sentential complement; in Brown *worry* was more likely to take a direct object, while in the Garnsey data set *worry* was more likely to take a sentential complement.

Subcategorizations of *worry*	% Direct Object	% Sentential Complement
Garnsey	1%	24%
BC	14%	4%

Table 8: Subcategorization of *worry* affected by sentence-completion paradigm.

This reversal in preference was caused by the properties of two of the subcategorization frames of *worry*. In frame 1 below, *worry* takes an experiencer as a subject, and subcategories for a finite sentence [Sfin]. In frame 2 below, *worry* takes a stimulus as a subject, and subcategorizes for an [NP].

#	*frame*	*example*
1	[experiencer] worries [stimulus]	Samantha worried that trouble was coming in waves. (Garnsey)
2	[stimulus] worries [experiencer]	Her words remained with him, worrying him for hours. (BC)

Table 9: Uses of *worry*.

In the Garnsey protocol, proper names (highly animate) were provided. This provides a bias towards the first use, since animate subjects are more likely to be experiencers than stimuli. All of the sentential complement uses in the the Brown corpus data had a human/animate subject. In the direct object uses, only 30% of the subjects were animate. It is uncontroversial that the nature of the prompt in a sentence completion experiment affects factors such as whether the sentence will be active or passive. This analysis shows that the nature of the prompt has more subtle but equally important effect on how subjects will use a verb.

5 Different verb senses have different subcategorization frequencies

Much work on subcategorization frequencies assumes implicitly that these frequencies were indexed by the orthographic word. Presumably this is because in many cases (e.g. Connine et al. (1984) and Garnsey et al. (1997)) these frequencies were collected to use in norming reading studies. Since we are making a psychological claim about the locus of frequency effects in the mental lexicon, the orthographic word assumption may not be a good one. Indeed, linguists have long suggested that the *lemma* or *sense* of a word is the locus of subcategorization; for example Green (1974) showed that two different senses of the verb *run* had different subcategorizations. Indeed, since Gruber (1965) and Fillmore (1968), linguists have been trying to show that the syntactic subcategorization of a verb is related to the semantics of its arguments. Thus one might expect a verb meaning *accuse* to have a different set of syntactic properties than a verb meaning *bill*. Similarly, if two senses of a single verb mean *accuse* and *bill*, these two senses should have different syntactic properties. The notion of a semantic base for subcategorization probabilities is consistent with work such as Argmaman et al. (1998), which shows that verbs and their nominalizations have similar subcategorization preferences.

We propose that this fact about possible subcategorizations is also a fact about subcategorization *probabilities*, as the *Lemma Argument Probability* hypothesis:

> *Lemma Argument Probability* hypothesis: The lemma or word sense is the locus of argument expectations. Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.

We give a four-step argument for the *Lemma Argument Probability* hypothesis. In this section we start by showing that different corpora can yield different subcategorization probabilities. We show that different corpora contain different senses of verbs. We then show that it is this different distribution of lemmas or senses that accounts for much of the inter-corpus variability in subcategorization frequencies. Finally, in section 6, we show a specific example of how when context-based variation is controlled for, each verb sense has a unified subcategorization probability vector across sources.

In order to investigate the relationship between verb sense and verb subcategorization, we hand coded the data for six verbs for sense/lemma. We primarily compare the data from the Brown corpus and the Wall Street Journal corpus since these two corpora had the largest amount of data. Although the data from the other corpora was less plentiful, it still provided useful insights.

First, we analyze three verbs, *pass*, *charge*, and *jump*, which were chosen because they had large differences in subcategorization frequencies between the Wall Street Journal corpus and the Brown corpus. Table 10 shows that all three verbs have significant differences in subcategorization frequencies between the Brown corpus and the Wall Street Journal corpus.

Verb	*Cosine (all senses combined)*	*Do BC and WSJ have different subcategorization probabilities?*
pass	0.75	Yes ($X^2 = 22.2$, $p < .001$)
charge	0.65	Yes ($X^2 = 46.8$, $p < .001$)
jump	0.50	Yes ($X^2 = 49.6$, $p < .001$)

Table 10: Agreement between WSJ and BC data.

Next, we measured how often each sense occurred in each corpus. We found that each of the verbs showed a significant difference in the distribution of senses between the Brown corpus and the Wall Street Journal corpus, as shown in Table 11. This is consistent with Biber et al. (1998), who note that different genres have different distributions of word senses.

Verb	*Do BC and WSJ have different distributions of verb sense?*
pass	Yes ($X^2 = 59.4$, $p < .001$)
charge	Yes ($X^2 = 35.1$, $p < .001$)
jump	Yes ($X^2 = 103$, $p < .001$)

Table 11: Differences in distribution of verb senses between BC and WSJ.

Table 12 uses the verb *charge* to show how the sense distributions are different for a particular verb. The types of topics contained in a corpus influence which senses of a verb are used. Since Brown corpus contains a balanced variety of topics, while the Wall Street Journal corpus is strongly biased towards business related discussion, we expect to see more of the business-related senses in the Wall Street Journal corpus. Indeed we found that the two business related senses of *charge* (*accuse* and *bill*) are used more frequently in the Wall Street Journal corpus, although they also occur commonly in the Brown corpus, while the *attack* sense of charge is used only in the Brown corpus. The *credit card* sense is probably more common in corpora that are more recent than the Brown corpus.

Senses of *charge*	BC %	WSJ %	Example of the senses of **charge**.
attack	23%	0%	His followers shouted the old battle cry after him and **charged** the hill, firing as they ran. (BC)
run	8%	0%	She **charged** off to the bedrooms. (BC)
appoint	6%	4%	The commission is **charged** with designing a ten year recovery program. (WSJ)
accuse	39%	58%	Separately, a Campeau shareholder filed suit, **charging** Campeau, Chairman Robert Campeau and other officers with violating securities law. (WSJ)
bill	24%	36%	Currently the government **charges** nothing for such filings. (WSJ)
credit card	0%	2%	Many auto dealers now let buyers **charge** part or all of their purchase on the American Express card….(WSJ)
TOTAL	100%	100%	

Table 12: Examples of common senses of *charge*.

We also found this effect of corpus topic on verb sense in the isolated sentence corpora. When topics such as *home*, *school*, and *downtown* were provided to the subjects in the Connine et al. sentence production study, subjects used different senses of the verbs. For example the school setting caused 5 out of 9 subjects to use the *test* sense of the verb pass. By contrast, the *test* sense was used only 2 times in 230 examples in the Brown corpus.

	movement	test	pass the buck
home	6	1	1
downtown	5	1	0
school	4	5	0

Table 13: Uses of *pass* in different settings in the CFJCF sentence production study

For each of these three verbs, we then examined the subcategorization frequencies for each sense. In each case, the relative frequency of the verb senses in each corpus resulted in a difference in the overall subcategorization frequency for that verb. This is due to each of the senses having separate subcategorization probabilities. Table 14 illustrates that different senses of the verb *charge* have different subcategorizations (examples of each sense are given in Table 12).

Senses of *charge*	that-S	NP	NP PP[3]	passive	Other
appoint	0%	0%	0%	4%	0%
accuse	18%	0%	12% (with)	24%	2%
bill	0%	9%	24% (for)	1%	1%
credit card	0%	0%	2% (on)	0%	0%

Table 14: Different senses of *charge* in WSJ have different subcategorization probabilities. Dominant prepositions are listed in parentheses after the frequency.

Further evidence that subcategorization probabilities are based on verb sense is provided by the fact that for two of the verbs, *pass* and *charge*, the agreement for the most common sense was better than the agreement for all senses combined. The third verb, jump, also shows improvement, but the single sense value is not significant. This is because the nearly complementary distribution of senses between the corpora results in low sample sizes for one of the corpora whenever only a single sense is taken into consideration. Table 15 shows that the

[3] The set of subcategorization frames that we use does not take the identity of the preposition into account.

agreement for the most common sense is better than the agreement for all senses combined. We attribute the remaining disagreement between the corpora to context and discourse based subcategorization differences.

Verb	*Cosine (all senses combined)*	*Cosine (most common sense)*
pass	0.75	0.95
charge	0.65	0.80
jump	0.50	0.59

Table 15: Improvement in agreement when after controlling for verb sense.

We also examined three verbs with good agreement (*kill*, *stay*, and *jump* - Table 16) in overall subcategorization between the Wall Street Journal corpus and the Brown corpus data as a preliminary effort to see what factors might prevent subcategorization frequencies from changing between corpora.

Verb	*Cosine (all senses combined)*	*Do BC and WSJ have different subcategorization probabilities? (X^2)*
kill	1.00	No
stay	1.00	No
try	1.00	No

Table 16: Agreement between BC and WSJ data.

We would expect no changes in subcategorization (beyond context/discourse changes) in cases where 1) the verb only had one common sense, or 2) the multiple senses of a verb had similar subcategorizations. We found that all three verbs with high agreement did in fact have different distributions of sense between the corpora, as shown in Table 17. These verbs showed equally high agreement for their most frequent senses.

Verb	*Do BC and WSJ have different distributions of verb sense?*
kill	Yes ($X^2 = 26.9$, $p < .001$)
stay	Yes ($X^2 = 26.1$, $p < .001$)
try	Yes ($X^2 = 8.74$, $p < .025$)

Table 17: Differences in distribution of verb sense between BC and WSJ.

Why do certain sense differences not cause subcategorization differences? One factor is that senses that are very closely (polysemously or metaphorically) related, like the senses of *kill* and *stay*, tend to have similar subcategorization probabilities across corpora. However, contextual factors may combine with the subcategorization probabilities for the similar senses, resulting in different observed probabilities. For example, the verb *jump* has two senses related by metonymy, *leap* and *rise in price*. While these have similar possible subcategorizations, the actual distribution of these subcategorizations was very different in the Brown corpus and the Wall Street Journal corpus data, due to the discourse circumstances under which each of the senses was used. The information demands in the Wall Street Journal resulted in stock price jumps being given with a distance and stopping point (jumped five eighths to five dollars a share).

This section has shown that different verb senses can have different subcategorization probabilities. It also showed that different corpora tend to have a different distribution of verb senses, and that this different distribution can result in overall subcategorization differences between the corpora. Showing that different senses have different subcategorizations is only part of the argument for the Lemma Argument Probability hypothesis. Section 6 will complete the argument by investigating one verb in detail and showing that a given sense/lemma has the same subcategorization probability vector across sources when we control for context-based variation.

This relationship between verb sense and subcategorization leads to an important methodological caveat as well: our psychological models and experimental protocols which rely on verb subcategorization frequencies must also take verb sense into account.

6 Evidence for the Lemma Argument Probability Hypothesis

The previous section showed that different senses of a verb could have different subcategorizations. In this section we show preliminary evidence that a single sense tends to have a single subcategorization probability vector, when we control for other factors. We use data for the verb *hear*, which is one of the few verbs that appeared on all five corpora.

Our procedure is to show that the agreement between subcategorization vectors iteratively improves as we control for more factors, from .88 for agreement between uncontrolled vectors, to .99 for agreement between vectors controlled for verb sense as well as discourse context effects.

We began by calculating the average agreement between each of the 10 possible pairs of corpora. For example we compared the Brown corpus and the Wall Street Journal corpus, the Brown corpus and the Connine data set, the Brown corpus and the Garnsey data set, the Brown corpus and the Switchboard corpus, the Wall Street Journal corpus and the Switchboard corpus, and so on. The average agreement was .88.

We then controlled for the 'isolated-sentence' effect by *only* comparing pairs of corpora if they were *both* isolated-sentences or *both* connected sentences. Thus we compared the Garnsey data set to the Connine data set, the Brown corpus to the Wall Street Journal corpus, the Wall Street Journal corpus to the Switchboard corpus, and the Brown corpus to the Switchboard corpus. The average agreement improved to .93. We then controlled for spoken versus written effects by comparing only the Brown corpus and the Wall Street Journal corpus. The average agreement improved to .98. Finally, instead of comparing all sentences with *hear* in the Brown corpus to all sentences with *hear* in the Wall Street Journal corpus, we compared only sentences which used the single most frequent sense of *hear*. The average agreement improved to .99. Table 18 shows a schematic of our comparisons. Note that although verb sense is controlled for only in the final step, controlling for sense results in improvement at any point in the chart. For example, the average agreement for all corpora also improves to .89 when we control for sense.

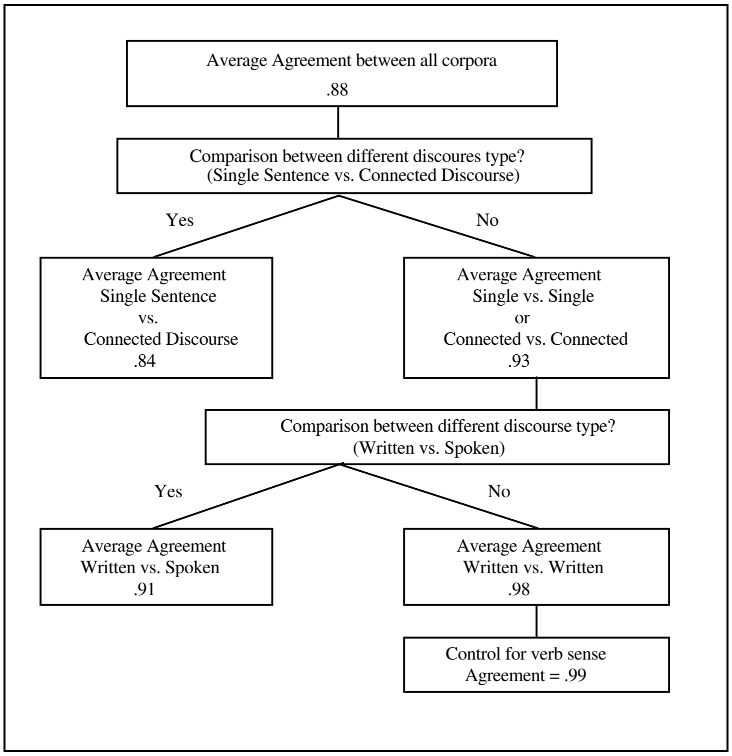

Table 18: Improvements in agreement for 'hear'.

Unfortunately, this methodology does not allow us to assign factor weights to the relative contributions of verb sense and discourse context. While we had hoped to establish such weights, it now seems to us that such factor weights would be extremely dependent on the verb and the idiosyncrasies of the context.

7 Conclusion

We have shown that subcategorization frequency variation is caused by factors including the discourse cohesion effects of natural corpora, the default referent effects of isolated-sentence experiments, the prompt given in sentence production experiment, the effects of different genres on verb sense, and the effect of verb sense on subcategorization. Our evidence shows clearly that in clear cases of polysemy, such as the *accuse* and *bill* senses of *charge*, each sense has a different set of subcategorization probabilities. We have not investigated subtler differences in meaning, such as in *load the wagon with hay* and *load hay into the wagon*. Such alternations are usually modeled by one of two theories. Our data is currently unable to distinguish between them. For example, a Lexical Rule account (Levin and Rappaport Hovov 1995) might consider each valence possibility as a distinct lemma; our results merely show that these lemmas would have to be associated with lemma probabilities. An alternative constructional account (Goldberg 1995) would include both valence possibilities as part of a single lemma for load, with separate valence probabilities. In the constructional account, the shadings in sense being determined by the combination of lexical meaning and constructional meaning.

Our experiments do have a number of implications both for cognitive modeling and for psycholinguistic methodology. The *Lemma Argument Probability* hypothesis makes a psychological claim about mental representation: that each lemma contains a vector of probabilistic expectations for its arguments. While we have only explored verbal lemmas, we assume this claim also holds of other predicates such as adjectives and nouns. Furthermore, our results suggest that the observed subcategorization probabilities can be explained by a probabilistic combination of these lemma probabilities with other probabilistic factors. That is, the probability of linguistic events occurring “in the world” can be accounted for by probabilistic combinations of mentally represented linguistic knowledge. If this is true, it supports models of human language interpretation such as Narayanan and Jurafsky (1998) which similarly rely on the Bayesian combination of different probabilistic sources of lexical and non-lexical knowledge.

Acknowledgments

This project was supported by the generosity of the NSF via NSF IIS-9733067, NSF IRI-9704046, NSF IRI-9618838 and the Committee on Research and Creative Work at the graduate school of the University of Colorado, Boulder. Many thanks to Giulia Bencini, Charles Clifton, Charles Fillmore, Susanne Gahl, Sue Garnsey, Adele Goldberg, Michelle Gregory, Uli Heid, Paola Merlo, Laura Michaelis, Neal Perlmutter, Bill Raymond, Philip Resnik, and two anonymous reviewers.

References

Argmaman, Pearlmutter, and Garnsey (1998). *Lexical Semantics as a Basis for Argument Structure Frequency Biases*. Poster presented at CUNY Sentence Processing Conference.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). *The Berkeley FrameNet Project*. Proceedings of the 1998 COLING-ACL Conference, Montreal, Canada. 86-90.
Biber, D. (1988) *Variation across speech and writing*. Cambridge University Press, Cambridge.
Biber, D. (1993) *Using Register-Diversified Corpora for General Language Studies*. Computational Linguistics, 19(2), 219-241.
Biber, D, Conrad, S., & Reppen, R. (1998) *Corpus Linguistics*. Cambridge University Press, Cambridge.
Boland, J. E., Tanenhaus, M. K., Garnsey, S. M. (1990) *Evidence for the immediate use of verb control information in sentence processing*. Journal of Memory & Language, 29(4), 413-432.
Chafe, W. (1982) *Integration and involvement in speaking, writing, and oral literature*. In Tannen, D. ed. Spoken and Written Language, Norwood, New Jersey:Ablex.
Chafe, W. (1987) *Cognitive constraints on information flow*. In Tomlin, R. S. (ed). Coherence and grounding in discourse. Amsterdam Benjamins, 1-16.
Clifton, C., Frazier, L., & Connine, C. (1984) *Lexical expectations in sentence comprehension*. Journal of Verbal Learning and Verbal Behavior, 23, 696-708.
Connine, C., Ferreira, F., Jones, C., Clifton, C., and Frazier, L. (1984*) Verb Frame Preference: Descriptive Norms*. Journal of Psycholinguistic Research 13, 307-319.
Dowty, D. (1979) *Word meaning and Montague grammar*. Dordrecht: Reidel.
Ferreira, F. and Clifton, C. (1986). *The independence of syntactic processing*. Journal of Memory and Language 25, 348-368.
Ferreira, F., and McClure, K.K. (1997). *Parsing of Garden-path Sentences with Reciprocal Verbs*. Language and Cognitive Processes 12, 273-306.
Fillmore, C. J. (1968) *The Case for Case*. In Bach, E. W. and Harms, R. T. eds. Universals in Linguistic Theory. Holt, Rinehart & Winston, New York: 1-88.
Fillmore, C. J. (1969). *Types of lexical information*. In Ferenc Kiefer (ed.) Studies in Syntax and Semantics. Dordrecht: Reidel, 109-137.
Fillmore, C. J. (1986). *Pragmatically Controlled Zero Anaphora*. Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society, Berkeley, CA. 95-107.
Ford, M.; Bresnan, J., Kaplan, R. M. (1982) *A Competence-Based Theory of Syntactic Closure*. In Bresnan, Joan (ed..) The Mental Representation of Grammatical Relations. Cambridge: MIT Press, 1982. 727-796.
Francis, W. and Kucera, H. (1982) *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin
Fraser, B., and Ross, J. R. (1970). *Idioms and unspecified NP deletion*. Linguistic Inquiry 1. 264-265.

Gahl, S. (1998). *Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus.* Proceedings of ACL-98, Montreal.

Garnsey, S. M., Pearlmutter, N. J., Myers, E. & Lotocky, M. A. (1997). *The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences.* Journal of Memory and Language 37, 58-93.

Gibson, E., Schutze, C., & Salomon, A. (1996). *The relationship between the frequency and the processing complexity of linguistic structure.* Journal of Psycholinguistic Research 25(1), 59-92.

Givon, T. (1979) *On understanding grammar.* NY: Academic Press.

Givon, T. (1984) *Syntax: a functional/typological introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Givon, T. (1987) *Beyond foreground and background.* In Tomlin, R.S. (ed). Coherence and grounding in discourse. Amsterdam: Benjamins.

Godfrey, J., E. Holliman, J. McDaniel. (1992) *SWITCHBOARD : Telephone speech corpus for research and development*. Proceedings of ICASSP-92, 517-520, San Francisco.

Goldberg, A.E. (1995) *Constructions*. Chicago: University of Chicago Press.

Green, G. (1974) *Semantics and Syntactic Regularity*. Bloomington: Indiana University Press.

Gruber, J. (1965). *Studies in lexical relations.* Bloomington: Indiana University Linguistics Club. [MIT Dissertation, 1965]

Halliday, M. A. K., and Hasan, R. (1976) Cohesion in English. London/New York Longman.

Juliano, C., and Tanenhaus, M.K. *Contingent frequency effects in syntactic ambiguity resolution.* In proceedings of the 15th annual conference of the cognitive science society, LEA: Hillsdale, NJ.

Jurafsky, D. (1996) *A probabilistic model of lexical and syntactic access and disambiguation*. Cognitive Science, 20, 137-194.

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge: MIT Press.

Levin, B. and Hovav, M. R. (1995). *Unaccusativity at the syntax-lexical semantics interface*. Cambridge: MIT Press.

Lowe, J. B., Baker, C.F., and Fillmore, C.J. (1997). *A frame-semantic approach to semantic annotation.* Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics: Why, What, and How?" in conjunction with ANLP-97. Washington, D.C., USA.

MacDonald, M. C. (1994) *Probabilistic constraints and syntactic ambiguity resolution.* Language and Cognitive Processes 9, 157-201.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A.. (1993) *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics 19(2), 313-330.

Marcus, M. P., Kim, G. Marcinkiewicz, M.A., MacIntyre, R., Ann Bies, Ferguson, M., Katz, K., and Schasberger, B.. (1994) *The Penn Treebank: Annotating predicate argument structure*. ARPA Human Language Technology Workshop, Plainsboro, NJ, 114-119.

Merlo, P. (1994). *A Corpus-Based Analysis of Verb Continuation Frequencies for Syntactic Processing*. Journal of Pyscholinguistic Research 23(6), 435-457.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993) *Introduction to WordNet: an on-line lexical database*.

Mitchell, D. C. and V. M. Holmes. (1985) *The role of specific information about the verb in parsing sentences with local structural ambiguity.* Journal of Memory and Language 24, 542--559.

Narayanan, S. and Jurafsky, D. (1998) Bayesian models of human sentencing processing. Procedings of 20th annual conference of the Cognitive Science Society. 752-757.

Resnik, Philip (1996). Selectional constraints: an information-theoretic model and its computational realization. Cognition 61(1-2), 127-159

Roland, D. and Jurafsky, D. *Computing verbal valence frequencies: corpora versus norming studies*. Poster session presented at the CUNY sentence processing conference, Santa Monica, CA.

Salton, G. and McGill, M.J. (1983), *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Thompson, S. A. (1987) *The Passive in English: A Discourse Perspective*. In Channon, Robert & Shockey, Linda (Eds.) In Honor of Ilse Lehiste/Ilse Lehiste Puhendusteos. Dordrecht: Foris, 497-511.

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. (1994). *Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution.* Journal of Memory and Language 33, 285-318.

Trueswell, J., Tanenhaus, M. K., and Kello, C. (1993) *Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference from Garden-Paths.* Journal of Experimental Psychology: Learning, Memory and Cognition 19(3), 528-553.