Verb Frame Frequency as a Predictor of Verb Bias

Maria Lapata and Frank Keller
Institute for Communicating and Collaborative Systems
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
phone: +44-131-650-4436, fax: +44-131-650-6626
email: {mlap, keller}@cogsci.ed.ac.uk

Sabine Schulte im Walde
Institute for Natural Language Processing
University of Stuttgart
Azenbergstraße 12, 70174 Stuttgart, Germany
phone: +49-711-121-1386, fax: +49-711-121-1366
email: schulte@ims.uni-stuttgart.de

Final Version, December 5, 2000

Abstract

There is considerable evidence showing that the human sentence processor is guided by lexical preferences in resolving syntactic ambiguities. Several types of preferences have been identified, including morphological, syntactic, and semantic ones. However, the literature fails to provide a uniform account of what lexical preferences are and how they should be measured. The present paper provides evidence for the view that lexical preferences are records of prior linguistic experience. We show that a type of lexial syntactic preference, viz., verb biases as measured by norming experiments, can be approximated by verb frame frequencies extracted from a large, balanced corpus by using computational learning techniques.

1. Introduction

1.1. Lexical Preferences and Sentence Processing

A number of researchers have argued that lexical preferences guide the human parser in resolving syntactic ambiguities. Such lexical preferences include morphological preferences (e.g., the tendency of a verb to occur in a particular tense, Trueswell, 1996), syntactic preferences (e.g., the

---

We would like to thank Chris Brew, Matt Crocker, Scott McDonald, Mats Rooth, and Patrick Sturt for comments regarding this work. Special thanks go to Susan Garnsey, Patrick Sturt, and Martin Traxler for making their data available to us. The financial support of the Alexander S. Onassis Foundation (Lapata) and the Economic and Social Research Council (Keller, Lapata) is gratefully acknowledged.

To appear in *Journal of Psycholinguistic Research*, 2001.

tendency of a verb to occur with a particular subcategorization frame, Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Kello, 1993), and semantic preferences (e.g., the tendency of a noun to occur as the object of a particular verb, Garnsey et al., 1997; Pickering, Traxler, & Crocker, 2000).

While there is agreement on the importance of lexical preferences for ambiguity resolution, the literature fails to provide a uniform account of what lexical preferences are and how they should be measured. The notion of preference has been operationalized in a number of ways; while morphological preferences can be obtained straightforwardly through corpus frequencies, syntactic and semantic preferences are typically determined on the basis of norming experiments.

The present paper provides evidence for a unified view of lexical preferences as records of prior linguistic experience. In this view, the human parser keeps track of the frequency of certain aspects of the lexical items it is exposed to; such frequencies then guide the parser in resolving syntactic ambiguities. The parser's linguistic experience can be approximated through large, balanced language corpora. We will demonstrate that lexical frequencies obtained from such corpora are highly correlated with the preferences obtained from norming experiments, thus providing evidence for an experience-based view of lexical preferences.

The present research focuses on verb bias, a type of lexical syntactic preference. The relevance of verb bias can be illustrated with respect to the direct object/sentential complement (NP/S) ambiguity and the transitive/intransitive (NP/0) ambiguity, illustrated in (1) and (2), respectively. In these examples, the verb frame (and thus the syntactic structure of the sentence it is part of) can only be disambiguated once the processor has read past the post-verbal NP. A garden path effect is expected if the parser has initially postulated the wrong frame.

(1) a. **NP frame:** The teacher knew the answer to the question.
b. **S frame:** The teacher knew the answer was false.
(2) a. **NP frame:** As the professor lectured the students the fire-alarm went off.
b. **0 frame:** As the professor lectured the students fell asleep.

A given verb is biased towards a frame if it tends to prefer this frame over the other frames it allows. The verb *know*, for instance, is S biased, i.e., it prefers the S frame over the NP frame. A verb that shows approximately equal preference for both frames is referred to as equi-biased.

1.2. Norming Studies

In the sentence processing literature, verb biases are typically obtained by conducting norming studies, either involving sentence completion tasks (Garnsey, Lotocky, Pearlmutter, & Myers, 1997) or production tasks (Connine, Ferreira, Jones, Clifton, & Frazier, 1984; Pickering et al., 2000). Occasionally, manually collected corpus counts have been used (Sturt, Pickering, & Crocker, 1999).

In a completion task, subjects are given sentence fragments such as (3) and (4) and are asked to complete the sentence. Possible completions for fragment (3) include *the answer* (NP frame) and *the student was ill* (S frame), or *that the student was ill* (S′ frame). Potential completions for fragment (4) include *the students the fire-alarm went off* (transitive frame) and *everyone was chatting* (intransitive frame).

(3) The teacher knew ___ .
(4) As the professor lectured ___ .

In a production task, subjects are given an isolated verb and are asked to produce a sentence that contains this verb. In some studies, the experimenter also provides a general topic (e.g., sports, travel) for the sentence to be produced (Connine et al., 1984).

In a manual collection task, verb biases are obtained through sampling from a corpus. The corpus occurrences for the verbs in the sample are inspected by hand in order to determine the frequency with which a verb is attested in a given frame.

Manual counting is not feasible if a large number of verbs or frames are to be investigated. This makes large-scale testing of the experience-based view of lexical preferences difficult. The present study demonstrates how such difficulties can be overcome with the help of chunking and parsing techniques that are standard in computational linguistics. Using such techniques, verb biases for a large number of verbs and frames can be obtained automatically. Experiment 1 focuses on NP/S ambiguous verbs, whereas Experiment 2 deals with the transitive/intransitive ambiguity.

1.3. Previous Work

Merlo (1994) conducted a study on a subset of the Penn Treebank corpus (Marcus, Santorini, & Marcinkiewicz, 1993). This corpus is annotated with part-of-speech and phrase structure information. Merlo (1994) focused on NP/S ambiguous verbs, which she extracted automatically from the Treebank. Merlo (1994) failed to find a strong correspondence between corpus frequencies and completion frequencies for NP/S ambiguous verbs. However, this study was based on a small, unbalanced corpus sample (ca. 340,000 words). It seems likely that better results can be achieved with a large corpus that delivers more reliable frame frequencies.

Gibson, Schütze, and Salomon (1996) and Gibson and Schütze (1999) also used a subset of the Penn Treebank, viz., the Brown and Wall Street Journal corpora (ca. two million words) and obtained relative frequencies of the attachment sites of conjoined NPs. They failed to find a correlation between the corpus-derived attachment preferences and measurements of syntactic complexity which were obtained from an off-line study and two self-paced reading experiments.

Roland and Jurafsky (2001) compared subcategorization frequencies for 17 frames which were derived from completion studies and a number of different corpora. Roland and Jurafsky's (2001) study demonstrated that corpus type (written vs. spoken) and discourse type (single sentences vs. connected discourse) influenced the verb frame frequencies. This result points to the need for using a balanced corpus that incorporates samples from several discourse types and from both spoken and written language.

2. Experiment 1: NP/S Ambiguity

2.1. Method

2.1.1. Materials

The corpus used for our frame extraction experiments was a part-of-speech annotated, lemmatized version of the British National Corpus (BNC). The BNC (Burnard, 1995) is a large, synchronic corpus of British English, consisting of 90 million words of text and 10 million words of speech. The BNC is a balanced corpus, i.e., it was compiled so as to represent a wide range of present day British English. The written part includes samples from newspapers, magazines, books (both academic and fiction), letters, and school and university essays, among other kinds of text. The spoken part consists of spontaneous conversations, recorded from volunteers balanced by age,

region, and social class. Other samples of spoken language are also included, ranging from business or government meetings to radio shows and phone-ins.[1]

The fact that the BNC is a balanced corpus makes it particularly attractive for psycholinguistic research: frequencies obtained from the BNC should be more representative of the language experience of native speakers than the ones obtained from unbalanced corpora such as the Penn Treebank typically used in earlier studies of subcategorization preferences. Note that a large part of the Penn Treebank represents only one genre, viz., newspaper text (taken from the Wall Street Journal).

2.1.2. Procedure

Instead of hand-counting subcategorization frequencies or relying on syntactically annotated corpora (such as the Penn Treebank used by Merlo (1994), for instance), we employed computational learning techniques. These techniques form the basis for well-established methods used in computational linguistics to extract subcategorization information from corpora (Manning, 1993; Brent, 1993; Briscoe & Carroll, 1997).

We compared the performance of two methods for extracting verb frames. The first method (the chunking method) is knowledge-poor and relies on a chunk grammar to produce a partially parsed version of the corpus. The second method (the parsing method) is more knowledge-intensive and makes use of a stochastic grammar to obtain a fully parsed version of the corpus. For validation purposes, the outputs of both methods were compared to frequencies obtained from a hand-annotated sample of the BNC.

The chunking method identified surface syntactic structure using Gsearch (Corley, Corley, Keller, Crocker, & Trewin, 2001), a tool which allows the search of arbitrary part-of-speech tagged corpora for shallow syntactic patterns based on a user-specified grammar and a syntactic query. Depending on the grammar specification (i.e., recursive or not) Gsearch can be used as a full context-free parser or as a chunk parser. A chunk is a non-recursive syntactic unit (Abney, 1997); chunking typically leaves attachment decisions unresolved.

We used Gsearch to extract tokens matching the syntactic patterns ‘V NP’, ‘V NP VP’, and ‘V *that* NP VP’ by specifying a chunk grammar for recognizing the verbal complex and NPs. We discarded all frames with a frequency smaller than 10, as they are likely to be unreliable given our heuristic approach. As a result, we extracted 674 verb types for the NP frame, 128 verb types for the S′ frame and 80 verb types for the S frame.

The parsing method identified syntactic structure using a robust statistical parser (Carroll & Rooth, 1998). The parser utilizes a probabilistic context-free grammar (PCFG) for English. In a PCFG, context-free rules are annotated with probabilities, and the probability of a parse is computed as the product of the probabilities of the relevant rules. The grammar used in the present experiment is an extension of the standard PCFG model in that it incorporates information about the lexical heads of constituents. Such a head-lexicalized PCFG provides a grammar model that combines lexicalized rules and lexical coherence relations between constituents. The parameters of this model were iteratively trained on a tagged version of the BNC by applying the Expectation Maximization algorithm, an unsupervised machine learning technique (Baum, 1972).

The head-lexicalized PCFG obtained this way was used to parse the whole BNC. This produced a set of parse trees for each sentence, where each tree was annotated with information about

[1]For additional information, see the BNC web page at `http://info.ox.ac.uk/bnc/`.

	GPML			TTK			CFJCF			BNC man			BNC chunk		
	N	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*
TTK	49	.91*	76%												
CFJCF	15	.87*	73%	12	.78†	67%									
BNC manual	50	.75*	62%	24	.69*	50%	12	.96*	100%						
BNC chunking	90	.69*	58%	44	.64*	52%	21	.54†	71%	52	.89*	77%			
BNC parsing	90	.81*	74%	44	.69*	61%	24	.74*	50%	55	.92*	78%	658	.84*	93%

$\dagger p < .01$; $*p < .001$; *N* number of verbs; *r* correlation; *b* agreement

Table 1: Correlations between frame frequencies and completion norms for the NP/S ambiguity

the lexical head and the probability of each subtree. The most probable parse for each sentence was determined, and the verb and all verbal arguments extracted. Based on this information, the subcategorization frame of the verb was identified. This procedure resulted 3,108 verb types for the NP frame, 1,818 verb types for the S′ frame and 861 verb types for the S frame.

2.2. Results

To estimate verb biases for the NP/S ambiguity, we first computed the relative frequencies of the NP frame and the S frame for a given verb. As is standard in the literature, we conflated the frequencies for the S and S′ complements (as both complements represent the same frame). Based on the relative frequencies, we then computed biases using the metric proposed by Garnsey et al. (1997): verbs were classified as NP-biased if the NP frequency was at least twice the S frequency; S-biased verbs were classified accordingly, and the remainder was classified as equi-biased.

Before comparing the biases estimated from corpus frame frequencies to those estimated from psycholinguistic norms, we carried out a validation task to determine the accuracy of our corpus frame frequencies. This was done by comparing our results to frame frequencies manually derived from the BNC. Sturt et al. (1999) randomly sampled 100 tokens each of 106 verbs and annotated five frame categories: NP, S, S′, intransitive, and other. We only included verbs that are truly ambiguous, i.e., that were found in both the NP and the S frame in Sturt et al.'s (1999) data.

The results of the validation study are reported in Table 1. The comparison with the relative frequencies obtained using the chunking method yielded a Pearson correlation coefficient of .89. The agreement on the classification task was 77%, i.e., the chunking method and manual collected data agreed on the bias for 77% of the verbs that they had in common. Using the frequencies obtained by the computationally more intensive parsing method, we obtained a somewhat higher correlation of .92. The agreement on the bias was 78%. Comparison of the two computational methods also yielded a high correlation of .84 and a verb bias agreement of 93%.

These results show that both techniques produce reliable verb frame frequencies when compared to manually obtained data. The shallow chunking technique performs almost as well as the computationally much more intensive parsing approach. Also, the results obtained using both techniques were highly correlated. Note that the baseline for predicting the biases is 33%, assuming that we just classify the verbs at random. Clearly both methods performed considerably better than this baseline.

The next step was to compute the correlation between the frame frequencies extracted from the corpus and frame frequencies reported in norming studies using human subjects. We corre-

lated the corpus-derived frame frequencies with the norming data gathered by Garnsey et al. (1997) (henceforth GLPM), Trueswell et al. (1993) (henceforth TTK) and Connine et al. (1984) (henceforth CFJCF). The first two studies employed a sentence completion task and collected norms specifically for NP/S ambiguous verbs. GLPM collected completion norms for 100 verbs, and TTK for 50 verbs. CFJCF employed a free production task. Their study included 127 verbs, and 19 verb frames were counted, including the NP frame and the S frame (the S′ frame was not counted separately).

The results of the correlation analysis are displayed in Table 1. Using the shallow parsing method, we obtained a .69 correlation between the extracted frequencies and the GLPM frequencies, and a correlation of .64 with the TTK frequencies. The correlation between GLPM and TTK was .91. This figure can serve as an upper bound; it indicates the maximum performance we can expect from an automatic method when used to obtain frame frequencies. The shallow parsing method also performed fairly well in estimating verb biases. The GLPM completion study and the chunking data agreed in estimating the bias for 58% of the verbs that they had in common (90 verbs overlap). The bias agreement was slightly lower for the TTK norms, viz., 52% (44 verbs overlap). This means that the chunking method produced considerable better results than the chance baseline of 33%, but worse than the upper bound of 76%, which was the agreement between GLPM and TTK (49 verbs overlap).

We expected better results for the parsing method, which produces more accurate frame frequencies. This prediction was confirmed by the substantially higher correlation of .81 with GLPM; the GLPM completion norms and the parsing method reached a 74% agreement in estimating the bias for 90 verbs, which is close to the upper bound of 76%, measured as the agreement between the GLPM and the TTK norms. For the TTK data, we observed a slight increase in the correlation compared with the chunking method ($r = .69$); the bias agreement was 61%.

Let us now consider the CFJCF data. These data are less suitable for comparison with the extracted frame frequencies, since only a small number of the verbs are ambiguous between the NP and the S frame, i.e., occur with both frames in the CFJCF study. (All non-ambiguous verbs were excluded from the comparison, because the notion of verb bias does not apply; including them would artificially inflate our correlation and agreement figures.) We found a high correlation ($r = .96$) with the manually annotated sample of the BNC; a high correlation of .74 was also obtained for parsing method. The correlation for the chunking data was lower at .54, but still significant. The manually obtained frame frequencies and the CFJCF production data reached a 100% agreement in estimating the bias for the 12 verbs they shared, while the chunking method achieved an agreement of 71% for 21 verbs. The agreement of the parsing method and the CFJCF data was unexpectedly low at 50% (on 24 verbs). The comparison of the CFJCF frequencies with the GLPM and TTK frequencies yielded high correlation ($r = .87$ and $r = .78$) and agreement figures (73% and 67%).

2.3. Discussion

This experiment demonstrated a high correlation between verb frame frequencies extracted from a large, balanced corpus and frame frequencies reported in norming studies. Particularly good results were achieved with the frame frequencies extracted using full parsing; the verb biases obtained from these frequencies reached an agreement of 74% with the biases obtained from the norming experiment reported in GLPM. This percentage is only slightly lower than the upper bound of 76%, measured as the agreement between the GLPM norms and the TTK norms.

A significant correlation was also found between the frame frequencies in the corpus and the

frequencies obtained from CFJCF's free production study. However, this result has to be interpreted with caution due to the small number of verbs that exhibited the NP/S ambiguity in this norming study (correlations and biases were computed on only 12 to 24 verbs). We return to this issue in the next experiment.

To summarize, our findings demonstrate that even a knowledge-poor method based on a shallow syntactic processor such as Gsearch provides frequencies data that correlates well with completion and production frequencies. The correlation can be improved by using a fully parsed version of the corpus. This result provides initial evidence for an experience-based view of lexical preferences.

An obvious question concerns the generality of this finding, i.e., if it extends to other ambiguities. Experiment 2 addresses this question with respect to the NP/0 ambiguity. Note that the NP/0 ambiguity poses a challenge to our corpus-based techniques, since the intransitive frame is considerably harder to detect automatically than the NP frame or the S frame.

3. Experiment 2: NP/0 Ambiguity

3.1. Method

3.1.1. Materials

The present experiment used the same corpus as Experiment 1.

3.1.2. Procedure

Again, we compared the performance of our chunking method to the computationally intensive full parsing approach. To validate the frequencies delivered by these two methods, we compared them to the data from Sturt et al. (1999), i.e., to frequencies obtained by manually annotating a sample from the BNC.

As before, the chunking method made use of the Gsearch corpus query system. This time we used a more sophisticated grammar in an attempt to achieve a reasonable accuracy in recognizing the intransitive frame. Given that a query of the form 'NP V' could potentially match all verbs in the corpus, we extracted tokens matching the patterns exemplified in (5). For example, pattern (5a) matches corpus tokens containing verbs optionally followed by adverbials or PPs and punctuation. In order to avoid extracting transitive verbs that subcategorize for PPs we excluded PPs headed by prepositions *for* and *to*. Pattern (5c) looks for instances of verbs followed by subordinating conjunctions (e.g., *although, until, so that*).

(5)
a.	V (ADV|PP) PUN	Peter dances beautifully.
b.	V (ADV) PP	Peter dances on the table.
c.	V (ADV|PP) SUBJ	Peter dances whenever he likes.

Note that this approach is guaranteed to overgenerate. In particular, it is prone to recognize as intransitive instances of transitive verbs that are part of a relative clause with a gapped object (see (6a)) and miss intransitive verbs followed by NPs that are not arguments but modifiers (see (6b)).[2]

(6)
a. The man upstairs, the first husband **whom Maria is to leave**, is not self-aware.
b. After these visits, I **slept the whole night**.

[2]The examples in (6) are taken from the BNC.

	CFJCF			PTC prod			PTC compl			BNC man			BNC chunk		
	N	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*	*N*	*r*	*b*
PTC production	19	.80*	68%												
PTC completion	16	.62	38%	69	.70*	62%									
BNC manual	28	.25	11%	22	.54†	68%	16	.55	4%						
BNC chunking	12	.11	17%	85	.42*	52%	64	.23	39%	67	.26	37%			
BNC parsing	64	.61*	56%	102	.66*	67%	70	.42*	40%	66	.54*	63%	1862	.43*	56%

$\dagger p < .01$; $*p < .001$; *N* number of verbs; *r* correlation; *b* agreement

Table 2: Correlations between frame frequencies and completion norms for the NP/0 ambiguity

It follows that the chunking method will only deliver very approximate frame frequencies for the NP/0 ambiguity. As in Experiment 1 verbs with frame frequencies smaller than 10 were discarded. After applying this frequency cutoff we obtained 2,371 verb types for the intransitive frame and 2,402 verb types for the transitive frame.

The parsing method, on the other hand, computes the most likely syntactic analysis for each sentence, and thus should be able to cope with problems such as the ones illustrated by the examples in (6). We used the procedure described in Experiment 1 for the parsing method and obtained 3,108 verb types for the transitive frame and 3,080 verb types for the transitive frame.

3.2. Results

As in Experiment 1, all results were based on relative frame frequencies. Again, we used GLPM's metric to estimate verb biases: a verb was classified as NP-biased if the NP frequency was at least twice the intransitive frequency; intransitive biased verbs were classified accordingly; the remainder was classified as equi-biased.[3]

We first carried out a validation study to determine how accurate the two parsing methods were in extracting subcategorization frequencies and estimating verb biases. Again, the standard of comparison was the manual corpus frequencies of Sturt et al. (1999). We only included ambiguous verbs, i.e., verbs that were attested both in the NP frame and in the intransitive frame in the Sturt et al. (1999) data.

The results of the validation study are reported in Table 2. The chunking method yielded a non-significant correlation of .26 with the manually collected data. The bias was predicted correctly for 37% of the verbs, which is close to the 33% agreement expected by chance. The parsing method obtained a significant correlation of .54 and an agreement of 63%. The correlation between the chunking method and the parsing method was significant but relatively low ($r = .43$), as was the agreement (56%). This demonstrates that identifying the NP/0 ambiguity automatically is considerably harder than detecting the NP/S ambiguity, especially for the knowledge-poor chunking method.

To test the psycholinguistic validity of our frame counts, we compared the verb biases predicted by the corpus frequencies with the ones reported in CFJCF. We also correlated the corpus frequencies with norming data gathered by Pickering et al. (2000) (henceforth PTC). PTC collected frame frequencies specifically for the NP/0 ambiguity by conducting a free production study, where

[3]See Pickering et al. (2000) for an alternative method of computing verb biases based on only two classes, NP biased and intransitive biased.

subjects were given a verb and were asked to provide a sentence, and a completion study, where subjects were asked to complete a sentence fragment (see example (4)). We correlated our corpus frame frequencies both with the production and completion data provided by PTC (henceforth PTC production and PTC completion, respectively). No comparisons involving the GLPM and TTK data could be carried out, since these data sets do not include norms for the intransitive frame. The results of all pairwise correlations and agreement figures are presented in Table 2.

For the frequencies obtained by the parsing method, a significant correlation of .61 was achieved with the CFJCF data; the agreement in predicting the verb bias was 56%. The correlation between the chunking data and the CFJCF data failed to reach significance ($r = .11$), and there as only 17% agreement on the bias, which is below chance level. This confirmed our observation that the chunker yields unreliable frame frequencies for the NP/0 ambiguity.

For the PTC data, we obtained a significant correlation of .66 between the parsed corpus frequencies and the PTC production frequencies; the agreement on the verb bias was 67%, which is only slightly lower than the upper bound of 68%, measured as the agreement between the PTC and CFJCF production data. A lower, yet significant, correlation of .42 was obtained when comparing the parsing data and the PTC completion data, with 40% agreement on the bias. Expectedly, the results for the chunking method were poor. The comparison between the chunking frequencies and the PTC production data yielded a correlation of .42 and a bias agreement of .51%. The correlation between the chunking data and the PTC completion data failed to be significant ($r = .23$) and agreement was low at 39%.

Another interesting finding is that the corpus frequencies correlate better with the PTC production data than with the PTC completion data (see Table 2). This result can be explained by the fact that the unrestricted text provided by a corpus is fairly similar to the unrestricted text produced in an unconstrained production task. In a completion task, on the other hand, the sentential context will constrain the subject's responses in a way that is not easily approximated by corpus samples. The hypothesis that completion and production tasks yield different types of data is confirmed by the fact that we found a high correlation ($r = .80$) between the two production studies (CFJCF and PTC production), while the correlation between the two production studies and the completion study (PTC completion) was lower ($r = .62$ and $r = .70$, respectively).

Surprisingly, the correlation between the manual frequencies and the CFJCF frequencies was non-significant ($r = .25$). This might be due to the small number of verbs that were shared by CFJCF and the manual data (28 verbs). However, the correlation between the manual frequencies and the PTC production data was significant ($r = .54$), with a bias agreement of 68%.

3.3. Discussion

The second experiment shed light on the limitations of a shallow approach to detecting syntactic structure. The Gsearch method yielded somewhat unreliable frequencies for the NP/0 ambiguity, and failed to estimate the verb biases in agreement with CFJCF's and PTC's completion data. However, a full parsing approach based on a stochastic grammar model yielded a significant correlation with the CFJCF and PTC production frequencies, and reached an agreement of 67% with the latter. The agreement is only slightly lower than the upper bound of 68% measured as the agreement between the CFJCF and PTC production data. While this result is lower than the 74% agreement achieved for the NP/S ambiguity, it is still well above the chance baseline of 33%.

There are two reasons for why we failed to find a higher agreement between the corpus frequencies and the frequencies in the norming studies. One is that the task of identifying the frames

correctly in the corpus is hard, especially for the intransitive frame. This problem is particularly grave for the chunking method, which seriously underestimates the frequencies for the intransitive frame. Reasons for this include tagging and parsing errors as well as problems with the extraction method, which misses certain syntactic constructions (e.g., NP ellipsis, which is difficult even for the parsing method).

Second, a perfect match between production or completion frequencies and corpus frequencies cannot be expected for a number of theoretical reasons, even if we managed to identify the frames in the corpus perfectly. Roland and Jurafsky (2001) investigate this issue in some detail. They identify discourse type as one of the factors that influence verb use and hence subcategorization. Text and speech corpora contain narrative and connected discourse. Verbs in connected discourse can be expected to behave differently from the verbs used in production studies, which are typically presented in isolation, perhaps with a very general discourse topic provided. In completion studies, subjects have to fill in a gap in an isolated sentence; this means that most of the discourse context that is part of a corpus occurrence of a verb is missing.

Roland and Jurafsky (2001) identify semantic factors as another potential explanation for the difference between the corpus frequencies and production norms. Different senses of a verb have different subcategorization frames, therefore frame frequencies are likely to depend on sense frequencies. The present study failed to distinguish different verb senses, and hence produced potentially unrealistic frame frequencies compared to completion studies. In a completion study, the sense of a verb is disambiguated by the sentential context, and subjects will produce frames for this sense. Note however, that the same observation does not apply to free production studies, where subjects are only presented with an isolated verb. This is reflected in our results by the fact that we achieved a high correlation with the production frequencies reported by Pickering et al. (2000), but only a relatively poor correlation with Pickering et al.'s (2000) completion frequencies.

Both of the problems pointed out by Roland and Jurafsky (2001) are a consequence of how norming studies are typically conducted. All the completion data referred to in the present paper were obtained using sentence fragments that were manually designed by the experimenters and presented to subjects in isolation. Such hand-crafted stimuli are bound to yield results that differ from data obtained from realistic samples of text or speech as they occur in corpora. Therefore, a more adequate approach to collecting norming data would be to use materials that are designed on the basis of corpus samples, and presented to subjects in the context in which they occur in the corpus. In this experimental setting, which is much closer the actual language experience of native speakers, subjects can be expected to generate completions that are in accordance with the discourse type and the verb sense of a given stimulus. We expect that such realistic completion norms will yield a high correlation with the frame frequencies obtained from corpora.

4. Conclusions

We demonstrated that verb frame frequencies obtained from a large, balanced corpus make it possible to predict verb biases for the NP/S and the NP/0 ambiguity, as determined using sentence completion and free production tasks. We estimated corpus frame frequencies using a probabilistic parsing approach (Schulte im Walde, 2000) and found that this method predicts the verb bias correctly 74% of the time for the NP/S ambiguity and 67% of the time for the NP/0 ambiguity. These results approach the upper bounds of 76% and 68%, respectively (measured as the agreement between two norming experiments).

Our findings counteract the pessimism expressed by Connine et al. (1984) regarding the relevance of corpus data and its correlation with experimentally obtained norms. It also disconfirms the results of an earlier corpus study by Merlo (1994), which failed to find a significant correspondence between completion norms corpus-derived frame frequencies. Merlo's (1994) study was based on a small, unbalanced corpus sample; it should not be expected to yield the same reliability as a study carried out on a balanced 100 million word corpus.

By demonstrating a correlation between verb biases obtained from corpora and norming experiments, we provided evidence for a view of lexical preferences as records of prior linguistic experience. Such a view is compatible with the tuning hypothesis (Mitchell, Cuetos, Corley, & Brysbaert, 1996), which states that the human parser deals with ambiguity by initially selecting the syntactic analysis that has worked most frequently in the past. Note however that our results only concern lexical frequencies; the tuning hypothesis predicts that the human parser also keeps track of coarse-grained syntactic frequencies (e.g., PP attachment frequencies). Further studies on parsed corpora are required to test this prediction.

We also discussed the reasons as to why we failed to obtain a perfect match between experimental norms and corpus frequencies. We argued that the norming studies reported in the literature are unrealistic because they do not provide discourse context for their materials and fail to control for verb sense ambiguities (the latter point applies to production studies only). Both factors were shown by Roland and Jurafsky (2001) to influence verb frame frequencies. This problem could be addressed by conducting realistic completion experiments that use materials extracted from corpora and present them in a discourse context.

Note that data collected by the present study are of practical importance for psycholinguistic research. Through the use of automatic extraction techniques, we obtained frame frequencies for 3,172 verbs (excluding particle verbs). These frequencies can serve as norms for the construction of materials in psycholinguistic experiments that manipulate verb bias. Deriving norms for such a large number of verbs would be very costly if done experimentally or through manual inspection of corpus samples. Our parsing method yielded frame frequencies not only for the NP, S, S′, and intransitive frame, but for a total of 88 frames (Schulte im Walde, 2000). These frequencies are useful for the psycholinguistic investigation of ambiguities other than the NP/S or NP/0 ambiguity. (The frequency data are publicly available; please contact the authors for details.)

References

Abney, S. (1997). Part-of-speech tagging and partial parsing. In S. Young & G. Bloothooft (Eds.), *Corpus-based methods in language and speech* (pp. 118–136). Dordrecht: Kluwer.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities, 3*(1), 1–8.

Brent, M. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics, 19*(3), 243–262.

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing* (pp. 46–55). Washington, DC.

Burnard, L. (1995). *Users guide for the British National Corpus.* British National Corpus Consortium, Oxford University Computing Service.

Carroll, G., & Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing* (pp. 36–45). Granada.

Connine, C. M., Ferreira, F., Jones, C., Clifton, C., & Frazier, L. (1984). Verb frame preferences: Descriptive norms. *Journal of Psycholinguistic Research*, *13*(4), 307–319.

Corley, S., Corley, M., Keller, F., Crocker, M. W., & Trewin, S. (2001). Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*. (to appear)

Garnsey, S. M., Lotocky, M. A., Pearlmutter, N. J., & Myers, E. M. (1997). *Argument structure frequency biases for 100 sentence-complement-taking verbs.* (Unpublished manuscript, University of Illinois at Urbana-Champaign)

Garnsey, S. M., Pearlmutter, N. J., Myers, E. M., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.

Gibson, E., & Schütze, C. T. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language*, *40*(2), 263–279.

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

Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (pp. 235–242). Columbus, OH.

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.

Merlo, P. (1994). A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research*, *23*(6), 435–457.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1996). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, *24*(6), 469–488.

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, *43*(3), 447–475.

Roland, D., & Jurafsky, D. (2001). Verb sense and verb subcategorization probabilities. In S. Stevenson & P. Merlo (Eds.), *The lexical basis of sentence processing: Formal, computational, and experimental issues.* Amsterdam: John Bejamins. (to appear)

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics* (pp. 747–753). Saarbrücken/Luxembourg/Nancy.

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, *40*(1), 136–150.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*(4), 566–585.

Trueswell, J. C., Tanenhaus, M. K., & 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.