Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 230-237. Association for Computational Linguistics.

Using the Web to Overcome Data Sparseness

Frank Keller and **Maria Lapata**
Division of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
{keller, mlap}@cogsci.ed.ac.uk

Olga Ourioupina
Department of Computational Linguistics
Saarland University
PO Box 15 11 50
66041 Saarbrücken, Germany
ourioupi@coli.uni-sb.de

Abstract

This paper shows that the web can be employed to obtain frequencies for bigrams that are unseen in a given corpus. We describe a method for retrieving counts for adjective-noun, noun-noun, and verb-object bigrams from the web by querying a search engine. We evaluate this method by demonstrating that web frequencies and correlate with frequencies obtained from a carefully edited, balanced corpus. We also perform a task-based evaluation, showing that web frequencies can reliably predict human plausibility judgments.

1 Introduction

In two recent papers, Banko and Brill (2001a; 2001b) criticize the fact that current NLP algorithms are typically optimized, tested, and compared on fairly small data sets (corpora with millions of words), even though data sets several orders of magnitude larger are available, at least for some tasks. Banko and Brill go on to demonstrate that learning algorithms typically used for NLP tasks benefit significantly from larger training sets, and their performance shows no sign of reaching an asymptote as the size of the training set increases.

Arguably, the largest data set that is available for NLP is the web, which currently consists of at least 968 million pages.[1] Data retrieved from the web therefore provides enormous potential for training NLP algorithms, if Banko and Brill's findings generalize. There is a small body of existing research that tries to harness the potential of the web for NLP. Grefenstette and Nioche (2000) and Jones and Ghani (2000) use the web to generate corpora for languages where electronic resources are scarce, while Resnik (1999) describes a method for mining the web for bilingual texts. Mihalcea and Moldovan (1999) and Agirre and Martinez (2000) use the web for word sense disambiguation, and Volk (2001) proposes a method for resolving PP attachment ambiguities based on web data.

A particularly interesting application is proposed by Grefenstette (1998), who uses the web for example-based machine translation. His task is to translate compounds from French into English, with corpus evidence serving as a filter for candidate translations. As an example consider the French compound *groupe de travail*. There are five translation of *groupe* and three translations for *travail* (in the dictionary that Grefenstette (1998) is using), resulting in 15 possible candidate translations. Only one of them, viz., *work group* has a high corpus frequency, which makes it likely that this is the correct translation into English. Grefenstette (1998) observes that this approach suffers from an acute data sparseness problem if the corpus counts are obtained from a conventional corpus such as the British National Corpus (BNC) (Burnard, 1995). However, as Grefenstette (1998) demonstrates, this problem can be overcome by obtaining counts through web searches, instead of relying on the BNC. Grefenstette (1998) therefore effectively uses the web as a way of obtaining counts for compounds that are sparse in the BNC.

[1] This is the number of pages indexed by Google in March 2002, as estimated by Search Engine Showdown (see http://www.searchengineshowdown.com/).

While this is an important initial result, it raises the question of the generality of the proposed approach to overcoming data sparseness. It remains to be shown that web counts are generally useful for approximating data that is sparse or unseen in a given corpus. It seems possible, for instance, that Grefenstette's (1998) results are limited to his particular task (filtering potential translations) or to his particular linguistic phenomenon (noun-noun compounds). Another potential problem is the fact that web counts are far more noisy than counts obtained from a well-edited, carefully balanced corpus such as the BNC. The effect of this noise on the usefulness of the web counts is largely unexplored.

The aim of the present paper is to generalize Grefenstette's (1998) findings by testing the hypothesis that the web can be employed to obtain frequencies for bigrams that are unseen in a given corpus. Instead of having a particular task in mind (which would introduce a sampling bias), we rely on sets of bigrams that are randomly selected from the corpus. We use a web-based approach not only for noun-noun bigrams, but also for adjective-noun and verb-object bigrams, so as to explore whether this approach generalizes to different predicate-argument combinations. We evaluate our web counts in two different ways: (a) comparison with actual corpus frequencies, and (b) task-based evaluation (predicting human plausibility judgments).

2 Obtaining Frequencies from the Web

2.1 Sampling Bigrams

Two types of adjective-noun bigrams were used in the present study: seen bigrams, i.e., bigrams that occur in a given corpus, and unseen bigrams, i.e., bigrams that fail to occur in the corpus. For the seen adjective-noun bigrams, we used the data of Lapata et al. (1999), who compiled a set of 90 bigrams as follows. First, 30 adjectives were randomly chosen from a lemmatized version of the BNC so that each adjective had exactly two senses according to WordNet (Miller et al., 1990) and was unambiguously tagged as "adjective" 98.6% of the time. The 30 adjectives ranged in BNC frequency from 1.9 to 49.1 per million. Gsearch (Corley et al., 2001), a chart parser which detects syntactic patterns in a tagged corpus by exploiting a user-specified context free grammar and a syntactic query, was used to extract all nouns occurring in a head-modifier relationship with one of the 30 adjectives. Bigrams involving proper nouns or low-frequency nouns (less than 10 per million) were discarded. For each adjective, the set of bigrams was divided into three frequency bands based on an equal division of the range of log-transformed co-occurrence frequencies. Then one bigram was chosen at random from each band.

Lapata et al. (2001) compiled a set of 90 unseen adjective-noun bigrams using the same 30 adjectives. For each adjective, the Gsearch chunker was used to compile a list of all nouns that failed to co-occur in a head-modifier relationship with the adjective. Proper nouns and low-frequency nouns were discarded from this list. Then each adjective was paired with three randomly chosen nouns from its list of non-co-occurring nouns.

For the present study, we applied the procedure used by Lapata et al. (1999) and Lapata et al. (2001) to noun-noun bigrams and to verb-object bigrams, creating a set of 90 seen and 90 unseen bigrams for each type of predicate-argument relationship. More specifically, 30 nouns and 30 verbs were chosen according to the same criteria proposed for the adjective study (i.e., minimal sense ambiguity and unambiguous part of speech). All nouns modifying one of the 30 nouns were extracted from the BNC using a heuristic which looks for consecutive pairs of nouns that are neither preceded nor succeeded by another noun (Lauer, 1995). Verb-object bigrams for the 30 preselected verbs were obtained from the BNC using Cass (Abney, 1996), a robust chunk parser designed for the shallow analysis of noisy text. The parser's output was post-processed to remove bracketing errors and errors in identifying chunk categories that could potentially result in bigrams whose members do not stand in a verb-argument relationship (see Lapata (2001) for details on the filtering process). Only nominal heads were retained from the objects returned by the parser. As in the adjective study, noun-noun bigrams and verb-object bigrams with proper nouns or low-frequency nouns (less than 10 per million) were discarded. The sets of noun-noun and verb-object bigrams were divided into three frequency bands and one bigram was chosen at random from each band.

The procedure described by Lapata et al. (2001)

was followed for creating sets of unseen noun-noun and verb-object bigrams: for each of noun or verb, we compiled a list of all nouns with which it failed to co-occur with in a noun-noun or verb-object bigram in the BNC. Again, Lauer's (1995) heuristic and Abney's (1996) partial parser were used to identify bigrams, and proper nouns and low-frequency nouns were excluded. For each noun and verb, three bigrams were randomly selected from the set of their non-co-occurring nouns.

Table 1 lists examples for the seen and unseen noun-noun and verb-object bigrams generated by this procedure.

2.2 Obtaining Web Counts

Web counts for bigrams were obtained using a simple heuristic based on queries to the search engines Altavista and Google. All search terms took into account the inflectional morphology of nouns and verbs.

The search terms for verb-object bigrams matched not only cases in which the object was directly adjacent to the verb (e.g., *fulfill obligation*), but also cases where there was an intervening determiner (e.g., *fulfill the/an obligation*). The following search terms were used for adjective-noun, noun-noun, and verb-object bigrams, respectively:

(1) `"A N"`, where `A` is the adjective and `N` is the singular or plural form of the noun.

(2) `"N1 N2"` where `N1` is the singular form of the first noun and `N2` is the singular or plural form of the second noun.

(3) `"V Det N"` where `V` is the infinitive, singular present, plural present, past, perfect, or gerund for of the verb, `Det` is the determiner *the*, *a* or the empty string, and `N` is the singular or plural form of the noun.

Note that all searches were for exact matches, which means that the search terms were required to be directly adjacent on the matching page. This is encoded using quotation marks to enclose the search term. All our search terms were in lower case.

For Google, the resulting bigram frequencies were obtained by adding up the number of pages that matched the expanded forms of the search terms in (1), (2), and (3). Altavista returns not only the number of matches, but also the number of words that match the search term. We used this count, as it takes multiple matches per page into account, and is thus likely to produce more accurate frequencies.

	adj-noun	noun-noun	verb-object
Altavista	14	10	16
Google	5	3	5

Table 2: Number of zero counts returned by the queries to search engines (unseen bigrams)

The process of obtaining bigram frequencies from the web can be automated straightforwardly using a script that generates all the search terms for a given bigram (from (1)–(3)), issues an Altavista or Google query for each of the search terms, and then adds up the resulting number of matches for each bigram. We applied this process to all the bigrams in our data set, covering seen and unseen adjective-noun, noun-noun, and verb-object bigrams, i.e., 540 bigrams in total.

A small number of bigrams resulted in zero counts, i.e., they failed to yield any matches in the web search. Table 2 lists the number of zero bigrams for both search engines. Note that Google returned fewer zeros than Altavista, which presumably indicates that it indexes a larger proportion of the web. We adjusted the zero counts by setting them to one. This was necessary as all further analyses were carried out on log-transformed frequencies.

Table 3 lists the descriptive statistics for the bigram counts we obtained using Altavista and Google.

From these data, we computed the average factor by which the web counts are larger than the BNC counts. The results are given in Table 4 and indicate that the Altavista counts are between 331 and 467 times larger than the BNC counts, while the Google counts are between 759 and 977 times larger than the BNC counts. As we know the size of the BNC (100 million words), we can use these figures to estimate the number of words on the web: between 33.1 and 46.7 billion words for Altavista, and between 75.9 and 97.7 billion words for Google. These estimates are in the same order of magnitude as Grefenstette and Nioche's (2000) estimate that 48.1 billion words of English are available on the web (based on Altavista counts in February 2000).

noun-noun bigrams										
high		medium		low		unseen		predicate		
process	1.14	user	.95	gala	0	collection, clause, coat		directory		
television	1.53	satellite	.95	edition	0	chain, care, vote		broadcast		
plasma	1.78	nylon	1.20	unit	.60	fund, theology, minute		membrane		
verb-object bigrams										
predicate	high		medium		low		unseen			
fulfill	obligation	3.87	goal	2.20	scripture	.69	participant, muscle, grade			
intensify	problem	1.79	effect	1.10	alarm	0	score, quota, chest			
choose	name	3.74	law	1.61	series	1.10	lift, bride, listener			

Table 1: Example stimuli for seen and unseen noun-noun and verb-object bigrams (with log-transformed BNC counts)

seen bigrams												
	adj-noun				noun-noun				verb-object			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Altavista	0	5.67	3.55	1.06	.67	6.28	3.41	1.21	0	5.46	3.20	1.14
Google	1.26	5.98	3.89	1.00	.90	6.11	3.66	1.20	0	5.85	3.56	1.16
BNC	0	2.19	.90	.69	0	2.14	.74	.64	0	2.55	.68	.58
unseen bigrams												
	adj-noun				noun-noun				verb-object			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Altavista	0	4.04	1.29	.94	0	3.80	1.08	1.12	0	3.72	1.38	1.06
Google	0	3.99	1.68	.96	0	4.00	1.42	1.09	0	4.07	1.76	1.04

Table 3: Descriptive statistics for web counts and BNC counts (log-transformed)

	adj-noun	noun-noun	verb-object
Altavista	447	467	331
Google	977	831	759

Table 4: Average factor by which the web counts are larger than the BNC counts (seen bigrams)

3 Evaluation

3.1 Evaluation Against Corpus Frequencies

While the procedure for obtaining web counts described in Section 2.2 is very straightforward, it also has obvious limitations. Most importantly, it is based on bigrams formed by adjacent words, and fails to take syntactic variants into account (other than intervening determiners for verb-object bigrams). In the case of Google, there is also the problem that the counts are based on the number of matching pages, not the number of matching words. Finally, there is the problem that web data is very noisy and unbalanced compared to a carefully edited corpus like the BNC.

Given these limitations, it is necessary to explore if there is a reliable relationship between web counts and BNC counts. Once this is assured, we can explore the usefulness of web counts for overcoming data sparseness. We carried out a correlation analysis to determine if there is a linear relationship between the BNC counts and Altavista and Google counts. The results of this analysis are listed in Table 5. All correlation coefficients reported in this paper refer to Pearson's r and were computed on log-transformed counts.

A high correlation coefficient was obtained across the board, ranging from .675 to .822 for Altavista counts and from .737 to .849 for Google counts. This indicates that web counts approximate BNC counts for the three types of bigrams under investigation, with Google counts slightly outperforming Altavista counts. We conclude that our simple

	adj-noun	noun-noun	verb-object
Altavista	.821**	.744**	.675**
Google	.849**	.737**	.751**
*$p < .05$ (2-tailed)		**$p < .01$ (2-tailed)	

Table 5: Correlation of BNC counts with web counts (seen bigrams)

heuristics (see (1)–(3)) are sufficient to obtain useful frequencies from the web. It seems that the large amount of data available for web counts outweighs the associated problems (noisy, unbalanced, etc.).

Note that the highest coefficients were obtained for adjective-noun bigrams, which probably indicates that this type of predicate-argument relationship is least subject to syntactic variation and thus least affected by the simplifications of our search heuristics.

3.2 Task-based Evaluation

Previous work has demonstrated that corpus counts correlate with human plausibility judgments for adjective-noun bigrams. This results holds for both seen bigrams (Lapata et al., 1999) and for unseen bigrams whose counts were recreated using smoothing techniques (Lapata et al., 2001). Based on these findings, we decided to evaluate our web counts on the task of predicting plausibility ratings. If the web counts for bigrams correlate with plausibility judgments, then this indicates that the counts are valid, in the sense of being useful for predicting intuitive plausibility.

Lapata et al. (1999) and Lapata et al. (2001) collected plausibility ratings for 90 seen and 90 unseen adjective-noun bigrams (see Section 2.1) using magnitude estimation. Magnitude estimation is an experimental 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. Magnitude estimation requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In the experiments conducted by Lapata et al. (1999) and Lapata et al. (2001), native speakers of English were presented with adjective-noun bigrams and were asked to rate the degree of adjective-noun fit proportional to the modulus item. The resulting judgments were normalized by dividing them by the modulus value and by log-transforming them. Lapata et al. (1999) report a correlation of .570 between mean plausibility judgments and BNC counts for the seen adjective-noun bigrams. For unseen adjective-noun bigrams, Lapata et al. (2001) found a correlation of .356 between mean judgments and frequencies recreated using class-based smoothing (Resnik, 1993).

In the present study, we used the plausibility judgments collected by Lapata et al. (1999) and Lapata et al. (2001) for adjective-noun bigrams and conducted additional experiments to obtain noun-noun and verb-object judgments for the materials described in Section 2.1. We used the same experimental procedure as the original study (see Lapata et al. (1999) and Lapata et al. (2001) for details). Four experiments were carried out, one each for seen and unseen noun-noun bigrams, and for seen and unseen verb-object bigrams. Unlike the adjective-noun and the noun-noun bigrams, the verb-object bigrams were not presented to subjects in isolation, but embedded in a minimal sentence context involving a proper name as the subject (e.g., *Paul fulfilled the obligation*).

The experiments were conducted over the web using the WebExp software package (Keller et al., 1998). A series of previous studies has shown that data obtained using WebExp closely replicates results obtained in a controlled laboratory setting; this was demonstrated for acceptability judgments (Keller and Alexopoulou, 2001), co-reference judgments (Keller and Asudeh, 2001), and sentence completions (Corley and Scheepers, 2002). These references also provide a detailed discussion of the WebExp experimental setup.

Table 6 lists the descriptive statistics for all six judgment experiments: the original experiments by Lapata et al. (1999) and Lapata et al. (2001) for adjective-noun bigrams, and our new ones for noun-noun and verb-object bigrams.

We used correlation analysis to compare web counts with plausibility judgments for seen adjective-noun, noun-noun, and verb-object bigrams. Table 7 (top half) lists the correlation coefficients that were obtained when correlat-

	adj-noun bigrams					noun-noun bigrams					verb-object bigrams				
	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD
Seen	30	−.85	.11	−.13	.22	25	−.15	.69	.40	.21	27	−.52	.45	.12	.24
Unseen	41	−.56	.37	−.07	.20	25	−.49	.52	−.01	.23	21	−.51	.28	−.16	.22

Table 6: Descriptive statistics for plausibility judgments (log-transformed); N is the number of subjects used in each experiment

ing log-transformed web and BNC counts with log-transformed plausibility judgments.

The results show that both Altavista and Google counts correlate with plausibility judgments for seen bigrams. Google slightly outperforms Altavista: the correlation coefficient for Google ranges from .624 to .693, while for Altavista, it ranges from .638 to .685. A surprising result is that the web counts consistently achieve a higher correlation with the judgments than the BNC counts, which range from .488 to .569. We carried out a series of one-tailed t-tests to determine if the differences between the correlation coefficients for the web counts and the correlation coefficients for the BNC counts were significant. For the adjective-noun bigrams, the difference between the BNC coefficient and the Altavista coefficient failed to reach significance ($t(87) = 1.46$, $p > .05$), while the Google coefficient was significantly higher than the BNC coefficient ($t(87) = 1.78$, $p < .05$). For the noun-noun bigrams, both the Altavista and the Google coefficients were significantly higher than the BNC coefficient ($t(87) = 2.94$, $p < .01$ and $t(87) = 3.06$, $p < .01$). Also for the verb-object bigrams, both the Altavista coefficient and the Google coefficient were significantly higher than the BNC coefficient ($t(87) = 2.21$, $p < .05$ and $t(87) = 2.25$, $p < .05$). In sum, for all three types of bigrams, the correlation coefficients achieved with Google were significantly higher than the ones achieved with the BNC. For Altavista, the noun-noun and the verb-object coefficients were higher than the coefficients obtained from the BNC.

Table 7 (bottom half) lists the correlations coefficients obtained by comparing log-transformed judgments with log-transformed web counts for unseen adjective-noun, noun-noun, and verb-object bigrams. We observe that the web counts consistently show a significant correlation with the judgments, the coefficient ranging from .466 to .588 for Altavista counts, and from .446 to .611 for the Google counts. Note that a small number of bigrams produced zero counts even in our web queries; these frequencies were set to one for the correlation analysis (see Section 2.2).

seen bigrams			
	adj-noun	noun-noun	verb-object
Altavista	.642**	.685**	.638**
Google	.650**	.693**	.624**
BNC	.569**	.517**	.488**
unseen bigrams			
Altavista	.466**	.588**	.568**
Google	.446**	.611**	.542**
*$p < .05$ (2-tailed)		**$p < .01$ (2-tailed)	

Table 7: Correlation of plausibility judgments with web counts and BNC counts

To conclude, this evaluation demonstrated that web counts reliably predict human plausibility judgments, both for seen and for unseen predicate-argument bigrams. In the case of Google counts for seen bigrams, we were also able to show that web counts are a better predictor of human judgments than BNC counts. These results show that our heuristic method yields useful frequencies; the simplifications we made in obtaining the counts, as well as the fact that web data are noisy, seem to be outweighed by the fact that the web is up to three orders of magnitude larger than the BNC (see our estimate in Section 2.2).

4 Conclusions

This paper explored a novel approach to overcoming data sparseness. If a bigram is unseen in a given corpus, conventional approaches recreate its frequency using techniques such as back-off, linear interpolation, class-based smoothing or distance-weighted averaging (see Dagan et al. (1999) and Lee (1999)

for overviews). The approach proposed here does not recreate the missing counts, but instead retrieves them from a corpus that is much larger (but also much more noisy) than any existing corpus: it launches queries to a search engine in order to determine how often a bigram occurs on the web.

We systematically investigated the validity of this approach by using it to obtain frequencies for predicate-argument bigrams (adjective-noun, noun-noun, and verb-object bigrams). We first applied the approach to seen bigrams randomly sampled from the BNC. We found that the counts obtained from the web are highly correlated with the counts obtained from the BNC, which indicates that web queries can generate frequencies that are comparable to the ones obtained from a balanced, carefully edited corpus such as the BNC.

Secondly, we performed a tasked-based evaluation that used the web frequencies to predict human plausibility judgments for predicate-argument bigrams. The results show that web counts correlate reliably with judgments, for all three types of predicate-argument bigrams tested, both seen and unseen. For the seen bigrams, we showed that the web frequencies correlate better with judged plausibility than the BNC frequencies.

To summarize, we have proposed a simple heuristic for obtaining bigram counts from the web. Using two different types of evaluation, we demonstrated that this simple heuristic is sufficient to obtain useful frequency estimates. It seems that the large amount of data available outweighs the problems associated with using the web as a corpus (such as the fact that it is noisy and unbalanced).

In future work, we plan to compare web counts for unseen bigrams with counts recreated using standard smoothing algorithms, such as similarity-based smoothing (Dagan et al., 1999) or class-based smoothing (Resnik, 1993). If web counts correlate reliable with smoothed counts, then this provides further evidence for our claim that the web can be used to overcome data sparseness.

References

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

Eneko Agirre and David Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken/Luxembourg/Nancy.

Michele Banko and Eric Brill. 2001a. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In James Allan, editor, *Proceedings of the 1st International Conference on Human Language Technology Research*, San Francisco. Morgan Kaufmann.

Michele Banko and Eric Brill. 2001b. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse.

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.

Martin Corley and Christoph Scheepers. 2002. Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, 9(1).

Steffan Corley, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

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

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.

Gregory Grefenstette and Jean Nioche. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of the RIAO Conference on Content-Based Multimedia Information Access*, pages 237–246, Paris.

Gregory Grefenstette. 1998. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, London.

Rosie Jones and Rayid Ghani. 2000. Automatically building a corpus for a minority language from the web. In *Proceedings of the Student Research Workshop at the 38th Annual Meeting of the Association for Computational Linguistics*, pages 29–36, Hong Kong.

Frank Keller and Theodora Alexopoulou. 2001. Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, 79(3):301–372.

Frank Keller and Ash Asudeh. 2001. Constraints on linguistic coreference: Structural vs. pragmatic factors. In Johanna D. Moore and Keith Stenning, editors, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 483–488, Mahwah, NJ. Lawrence Erlbaum Associates.

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.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen.

Maria Lapata, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgments. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 346–353, Toulouse.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University, Sydney.

Lilian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, University of Maryland, College Park.

Rada Mihalcea and Dan Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, University of Maryland, College Park.

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

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

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, College Park.

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

Martin Volk. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics Conference*, pages 601–606, Lancaster.