Using the Web to Obtain Frequencies for Unseen Bigrams

Frank Keller*
University of Edinburgh

Mirella Lapata†
University of Sheffield

This article 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: (a) a high correlation between Web frequencies and corpus frequencies; (b) a reliable correlation between Web frequencies and plausibility judgments; (c) a reliable correlation between Web frequencies and frequencies recreated using class-based smoothing; (d) a good performance of Web frequencies in a pseudodisambiguation task.

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 NLP tasks. Banko and Brill (2001a, 2001b) experiment with context-sensitive spelling correction, a task for which large amounts of data can be obtained straightforwardly, as no manual annotation is required. They demonstrate that the learning algorithms typically used for spelling correction benefit significantly from larger training sets, and that 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,[1] which currently consists of at least 3,033 million pages.[2] Data retrieved from the Web therefore provide enormous potential for training NLP algorithms, if Banko and Brill's (2001a, 2001b) findings for spelling corrections generalize; potential applications include tasks that involve word *n*-grams and simple surface syntax. 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 for which electronic resources are scarce, and Resnik (1999) describes a method for mining the Web in order to obtain bilingual texts. Mihalcea and Moldovan (1999) and Agirre and Martinez (2000) use the Web for word sense disambiguation, Volk (2001) proposes a method for resolving PP attachment ambiguities based on Web data, Markert, Nissim, and Modjeska (2003) use the Web for the resolution of nominal anaphora,

* School of Informatics, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: keller@inf.ed.ac.uk
† Department of Computer Science, 211 Portobello Street, Sheffield S1 4DP, UK. E-mail: mlap@dcs.shef.ac.uk

1 A reviewer points out that information providers such as Lexis Nexis ⟨http://www.lexisnexis.com/⟩ might have databases that are even larger than the Web. Lexis Nexis provides full-text access to news sources (including newspapers, wire services, and broadcast transcripts) and legal data (including case law, codes, regulations, legal news, and law reviews).

2 This is the number of pages indexed by Google in December 2002, as estimated by Search Engine Showdown ⟨http://www.searchengineshowdown.com/⟩.

© 2003 Association for Computational Linguistics

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

and Zhu and Rosenfeld (2001) use Web-based *n*-gram counts to improve language modeling.

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. An example is the French compound *groupe de travail*. There are five translations 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, namely, *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 counts are obtained from a conventional corpus. However, as Grefenstette (1998) demonstrates, this problem can be overcome by obtaining counts through Web searches, instead of relying on a corpus. Grefenstette (1998) therefore effectively uses the Web as a way of obtaining counts for compounds that are sparse in a given corpus.

Although 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 are 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. The effect of this noise on the usefulness of the Web counts is largely unexplored.

Zhu and Rosenfeld (2001) use Web-based *n*-gram counts for language modeling. They obtain a standard language model from a 103-million-word corpus and employ Web-based counts to interpolate unreliable trigram estimates. They compare their interpolated model against a baseline trigram language model (without interpolation) and show that the interpolated model yields an absolute reduction in word error rate of .93% over the baseline. Zhu and Rosenfeld's (2001) results demonstrate that the Web can be a source of data for language modeling. It is not clear, however, whether their result carries over to tasks that employ linguistically meaningful word sequences (e.g., head-modifier pairs or predicate-argument tuples) rather than simply adjacent words. Furthermore, Zhu and Rosenfeld (2001) do not undertake any studies that evaluate Web frequencies directly (i.e., without a task such as language modeling). This could be done, for instance, by comparing Web frequencies to corpus frequencies, or to frequencies re-created by smoothing techniques.

The aim of the present article is to generalize Grefenstette's (1998) and Zhu and Rosenfeld's (2001) 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 a corpus. We use a Web-based approach for bigrams that encode meaningful syntactic relations and obtain Web frequencies not only for noun-noun bigrams, but also for adjective-noun and verb-object bigrams. We thus explore whether this approach generalizes to different predicate-argument combinations. We evaluate our Web counts in four ways: (a) comparison with actual corpus frequencies from two different corpora, (b) comparison with human plausibility judgments, (c) comparison with frequencies re-created using class-based smoothing, and (d) performance in a pseudodisambiguation task on data sets from the literature.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

2. Obtaining Frequencies from the Web

2.1 Sampling Bigrams from the BNC

The data sets used in the present experiment were obtained from the British National Corpus (BNC) (see Burnard [1995]). The BNC is a large, synchronic corpus, 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. The corpus represents many different styles and varieties and is not limited to any particular subject field, genre, or register.

For the present study, the BNC was used to extract data for three types of predicate-argument relations. The first type is adjective-noun bigrams, in which we assume that the noun is the predicate that takes the adjective as its argument.[3] The second predicate-argument type we investigated is noun-noun compounds. For these, we assume that the rightmost noun is the predicate that selects the leftmost noun as its argument (as compound nouns are generally right-headed in English). Third, we included verb-object bigrams, in which the verb is the predicate that selects the object as its argument. We considered only direct NP objects; the bigram consists of the verb and the head noun of the object. For each of the three predicate-argument relations, we gathered two data sets, one containing seen bigrams (i.e., bigrams that occur in the BNC) and one with unseen bigrams (i.e., bigrams that do not occur in the BNC).

For the seen adjective-noun bigrams, we used the data of Lapata, McDonald, and Keller (1999), who compiled a set of 90 bigrams as follows. First, 30 adjectives were randomly chosen from a part-of-speech-tagged and 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. Lapata, McDonald, and Keller used the part-of-speech-tagged version that is made available with the BNC and was tagged using CLAWS4 (Leech, Garside, and Bryant 1994), a probabilistic part-of-speech tagger, with error rate ranging from 3% to 4%. The lemmatized version of the corpus was obtained using Karp et al.'s (1992) morphological analyzer.

The 30 adjectives ranged in BNC frequency from 1.9 to 49.1 per million words; that is, they covered the whole range from fairly infrequent to highly frequent items. Gsearch (Corley et al. 2001), a chart parser that 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. Examples of the syntactic patterns the parser identified are given in Table 1. In the case of adjectives modifying compound nouns, only sequences of two nouns were included, and the rightmost-occurring noun was considered the head. Bigrams involving proper nouns or low-frequency nouns (less than 10 per million words) were discarded. This was necessary because the bigrams were used in experiments involving native speakers (see Section 3.2), and we wanted to reduce the risk of including words unfamiliar to the experimental subjects. For each adjective, the set of bigrams was divided into three frequency bands based on an equal division of the

3 This assumption is disputed in the theoretical linguistics literature. For instance, Pollard and Sag (1994) present an analysis in which there is mutual selection between the noun and the adjective.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Table 1
Example of patterns used for the extraction of adjective-noun bigrams.

Pattern	Example
`A N`	educational material
`A Adv N`	usual weekly classes
`A N N`	environmental health officers

range of log-transformed co-occurrence frequencies. Then one bigram was chosen at random from each band. This procedure ensures that the whole range of frequencies is represented in our sample.

Lapata, Keller, and McDonald (2001) compiled a set of 90 unseen adjective-noun bigrams using the same 30 adjectives. For each adjective, Gsearch was used to compile a list of all nouns that did not co-occur in a head-modifier relationship with the adjective. Again, 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. Examples of seen and unseen adjective-noun bigrams are shown in Table 2.

For the present study, we applied the procedure used by Lapata, McDonald, and Keller (1999) and Lapata, Keller, and McDonald (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 from Lauer (1995) that looks for consecutive pairs of nouns that are neither preceded nor succeeded by another

Table 2
Example stimuli for seen and unseen adjective-noun, noun-noun, and verb-object bigrams (with log-transformed BNC counts).

Adjective-Noun Bigrams

Adjective	High		Medium		Low		Unseen
hungry	animal	1.79	pleasure	1.38	application	0	tradition, innovation, prey
guilty	verdict	3.91	secret	2.56	cat	0	system, wisdom, wartime
naughty	girl	2.94	dog	1.6	lunch	.69	regime, rival, protocol

Noun-Noun Bigrams

High		Medium		Low		Unseen	Head Noun
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

Verb	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

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

noun. Lauer's heuristic (see (1)) effectively avoids identifying as two-word compounds noun sequences that are part of a larger compound.

$$(1) \qquad C = \{(w_2, w_3) \mid w_1 w_2 w_3 w_4;\ w_1, w_4 \notin N;\ w_2, w_3 \in N\}$$

Here, $w_1\ w_2\ w_3\ w_4$ denotes the occurrence of a sequence of four words and N is the set of words tagged as nouns in the corpus. C is the set of compounds identified by Lauer's (1995) heuristic.

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 recognizes chunks and simplex clauses (i.e., sequences of nonrecursive clauses) using a regular expression grammar and a part-of-speech-tagged corpus, without attempting to resolve attachment ambiguities. It comes with a large-scale grammar for English and a built-in tool that extracts predicate-argument tuples out of the parse trees that Cass produces.

The parser's output was postprocessed 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. Tuples containing verbs or nouns attested in a verb-argument relationship only once were eliminated. Particle verbs were retained only if the particle was adjacent to the verb (e.g., *come off heroin*). Verbs followed by the preposition *by* and a head noun were considered instances of verb-subject relations. It was assumed that PPs adjacent to the verb headed by any of the prepositions *in, to, for, with, on, at, from, of, into, through*, and *upon* were prepositional objects (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 words) 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, Keller, and McDonald (2001) was followed for creating sets of unseen noun-noun and verb-object bigrams: for each noun or verb, we compiled a list of all nouns with which it did not co-occur within 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 formed by pairing it with a noun randomly selected from the set of the non-co-occurring nouns for that noun or verb. Table 2 lists examples for the seen and unseen noun-noun and verb-object bigrams generated by this procedure.

The extracted bigrams are in several respects an imperfect source of information about adjective-noun or noun-noun modification and verb-object relations. First notice that both Gsearch and Cass detect syntactic patterns on part-of-speech-tagged corpora. This means that parsing errors are likely to result because of tagging mistakes. Second, even if one assumes perfect tagging, the heuristic nature of our extraction procedures may introduce additional noise or miss bigrams for which detailed structural information would be needed.

For instance, our method for extracting adjective-noun pairs ignores cases in which the adjective modifies noun sequences of length greater than two. The heuristic in (1) considers only two-word noun sequences. Abney's (1996) chunker recognizes basic syntactic units without resolving attachment ambiguities or recovering missing information (such as traces resulting from the movement of constituents). Although parsing is robust and fast (since unlike in traditional parsers, no global optimization takes

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

place), the identified verb-argument relations are undoubtedly somewhat noisy, given the errors inherent in the part-of-speech tagging and chunk recognition procedure. When evaluated against manually annotated data, Abney's (1996) parser identified chunks with 87.9% precision and 87.1% recall. The parser further achieved a per-word accuracy of 92.1% (where per-word accuracy includes the chunk category and chunk length identified correctly).

Despite their imperfect output, heuristic methods for the extraction of syntactic relations are relatively common in statistical NLP. Several statistical models employ frequencies obtained from the output of partial parsers and other heuristic methods; these include models for disambiguating the attachment site of prepositional phrases (Hindle and Rooth 1993; Ratnaparkhi 1998), models for interpreting compound nouns (Lauer 1995; Lapata 2002) and polysemous adjectives (Lapata 2001), models for the induction of selectional preferences (Abney and Light 1999), methods for automatically clustering words according to their distribution in particular syntactic contexts (Pereira, Tishby, and Lee 1993), automatic thesaurus extraction (Grefenstette 1994; Curran 2002), and similarity-based models of word co-occurrence probabilities (Lee 1999; Dagan, Lee, and Pereira 1999). In this article we investigate alternative ways for obtaining bigram frequencies that are potentially useful for such models despite the fact that some of these bigrams are identified in a heuristic manner and may be noisy.

2.2 Sampling Bigrams from the NANTC

We also obtained corpus counts from a second corpus, the North American News Text Corpus (NANTC). This corpus differs in several important respects from the BNC. It is substantially larger, as it contains 350 million words of text. Also, it is not a balanced corpus, as it contains material from only one genre, namely, news text. However, the text originates from a variety of sources (*Los Angeles Times, Washington Post, New York Times* News Syndicate, Reuters News Service, and *Wall Street Journal*). Whereas the BNC covers British English, the NANTC covers American English. All these differences mean that the NANTC provides a second, independent standard against which to compare Web counts. At the same time the correlation found between the counts obtained from the two corpora can serve as an upper limit for the correlation that we can expect between corpus counts and Web counts.

The NANTC corpus was parsed using MINIPAR (Lin 1994, 2001), a broad-coverage parser for English. MINIPAR employs a manually constructed grammar and a lexicon derived from WordNet with the addition of proper names (130,000 entries in total). Lexicon entries contain part-of-speech and subcategorization information. The grammar is represented as a network of 35 nodes (i.e., grammatical categories) and 59 edges (i.e., types of syntactic [dependency] relationships). MINIPAR employs a distributed-chart parsing algorithm. Instead of a single chart, each node in the grammar network maintains a chart containing partially built structures belonging to the grammatical category represented by the node. Grammar rules are implemented as constraints associated with the nodes and edges.

The output of MINIPAR is a dependency tree that represents the dependency relations between words in a sentence. Table 3 shows a subset of the dependencies MINIPAR outputs for the sentence *The fat cat ate the door mat*. In contrast to Gsearch and Cass, MINIPAR produces all possible parses for a given sentence. The parses are ranked according to the product of the probabilities of their edges, and the most likely parse is returned. Lin (1998) evaluated the parser on the SUSANNE corpus (Sampson 1995), a domain-independent corpus of British English, and achieved a recall of 79% and precision of 89% on the dependency relations.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227l1604.pdf by guest on 24 March 2025

Table 3
Examples of dependencies generated by MINIPAR for *The fat cat ate the door mat*.

Head	Relation	Modifier	Description
cat	N:det:Det	the	determiner of noun
cat	N:mod:A	fat	adjective modifier of noun
eat	V:subj:N	cat	subject of verb
eat	V:obj:N	mat	object of verb
mat	N:det:Det	the	determiner of noun
mat	N:nn:N	door	prenominal modifier of noun

For our experiments, we concentrated solely on adjective-noun, noun-noun, and verb object relations (denoted as N:mod:A, N:nn:N, and V:obj:N in Table 3). From the syntactic analysis provided by the parser, we extracted all occurrences of bigrams that were attested both in the BNC and the NANTC corpus. In this way, we obtained NANTC frequency counts for the bigrams that we had randomly selected from the BNC. Table 4 shows the NANTC counts for the set of seen bigrams from Table 2.

Because of the differences in the extraction methodology (chunking versus full parsing) and the text genre (balanced corpus versus news text), we expected that some BNC bigrams would not be attested in the NANTC corpus. More precisely, zero frequencies were returned for 23 adjective-noun, 16 verb-noun, and 37 noun-noun bigrams. The fact that more zero frequencies were observed for noun-noun bigrams than for the other two types is perhaps not surprising considering the ease with which novel compounds are created (Levi 1978). We adjusted the zero counts by setting them to .5. This was necessary because all further analyses were carried out on log-transformed frequencies (see Table 4).

Table 4
Log-transformed NANTC counts for seen adjective-noun, noun-noun, and verb-object bigrams.

Adjective-Noun Bigrams						
Adjective	High		Medium		Low	
hungry	animal	.90	pleasure	-.30	application	.60
guilty	verdict	2.82	secret	.95	cat	-.30
naughty	girl	.69	dog	-.30	lunch	-.30

Noun-Noun Bigrams						
High		Medium		Low		Head Noun
process	-.30	user	-.30	gala	-.30	directory
television	2.70	satellite	-.30	edition	-.30	broadcast
plasma	-.30	nylon	0	unit	0	membrane

Verb-Object Bigrams						
Verb	High		Medium		Low	
fulfill	obligation	2.38	goal	2.04	scripture	-.30
intensify	problem	1.20	effect	.60	alarm	-.30
choose	name	2.25	law	.90	series	.48

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

2.3 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 in which 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:

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

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

(4) `"V Det N"`, where `V` is the infinitive, singular present, plural present, past, perfect, or gerund form of the verb, `Det` is the determiner *the*, the determiner *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 words in the search terms had to be directly adjacent to score a match. This is encoded by enclosing the search term in quotation marks. All our search terms were in lower case. We searched the whole Web (as indexed by AltaVista and Google); that is, the queries were not restricted to pages in English.

Based on the Web searches, we obtained bigram frequencies by adding up the number of pages that matched the morphologically expanded forms of the search terms (see the patterns in (2)–(4)). This process can be automated straightforwardly using a script that generates all the search terms for a given bigram, 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., a set of 540 bigrams in total). The queries were carried out in January 2003 (and thus the counts are higher than those reported in Keller, Lapata, and Ourioupina [2002], which were generated about a year earlier).

For some bigrams that were unseen in the BNC, our Web-based procedure returned zero counts; that is, there were no matches for those bigrams in the Web searches. It is interesting to compare the Web and NANTC with respect to zero counts: Both data sources are larger than the BNC and hence should be able to mitigate the data sparseness problem to a certain extent. Table 5 provides the number of zero counts for both Web search engines and compares them to the number of bigrams that yielded no matches in the NANTC. We observe that the Web counts are substantially less sparse than the NANTC counts: In the worst case, there are nine bigrams for which our Web queries returned no matches (10% of the data), whereas up to 82 bigrams were unseen in the NANTC (91% of the data). Recall that the NANTC is 3.5 times larger than the BNC, which does not seem to be enough to substantially mitigate data sparseness. All further analyses were carried out on log-transformed frequencies; hence we adjusted zero counts by setting them to .5.

Table 6 shows descriptive statistics for the bigram counts we obtained using AltaVista and Google. For comparison, this table also provides descriptive statistics for the BNC and NANTC counts (for seen bigrams only) and for the counts

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

Table 5
Number of zero counts returned by queries to search engines and in the NANTC (for bigrams unseen in BNC).

	Adjective-Noun	Noun-Noun	Verb-Object
AltaVista	2	9	1
Google	2	5	0
NANTC	76	82	78

Table 6
Descriptive statistics for Web counts, corpus counts, and counts re-created using class-based smoothing (log-transformed).

	Adjective-Noun				Noun-Noun				Verb-Object			
	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*
						Seen Bigrams						
AltaVista	1.15	5.84	3.72	1.02	.60	6.16	3.52	1.22	.48	5.86	3.42	1.13
Google	1.54	6.11	4.01	1.01	.90	6.30	3.80	1.23	.60	5.96	3.70	1.11
BNC	0	2.19	.89	.69	0	2.14	.74	.64	0	2.55	.68	.58
NANTC	-.30	2.84	.84	.96	-.30	3.02	.56	.94	-.30	3.73	1.90	.98
Smoothing	-.06	2.32	1.28	.51	-.70	1.71	.30	.61	-.51	2.07	.53	.57
						Unseen Bigrams						
AltaVista	-.30	5.00	1.50	.99	-.30	3.97	1.20	1.14	-.30	3.88	1.55	1.06
Google	-.30	4.11	1.79	.95	-.30	4.15	1.60	1.12	0	4.19	1.90	1.04
Smoothing	-.03	2.10	1.25	.46	-1.01	1.93	.28	.66	-.70	1.95	.53	.58

Table 7
Average factor by which Web counts are larger than BNC counts (seen bigrams).

	Adjective-Noun	Noun-Noun	Verb-Object
AltaVista	665	691	550
Google	1,306	1,151	1,064

re-created using class-based smoothing (see Section 3.3 for details on the re-created frequencies).

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 7 and indicate that the AltaVista counts are between 550 and 691 times larger than the BNC counts, and that the Google counts are between 1,064 and 1,306 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 available on the Web: between 55.0 and 69.1 billion words for AltaVista, and between 106.4 and 139.6 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 compiled in February 2000). They also agree with Zhu and Rosenfeld's (2001) estimate that the effective size of the Web is between 79 and 108 billion words (based on AltaVista, Lycos, and FAST counts; no date given).

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

2.4 Potential Sources of Noise in Web Counts

The method we used to retrieve Web counts is based on very simple heuristics; it is thus inevitable that the counts generated will contain a certain amount of noise. In this section we discuss a number of potential sources of such noise.

An obvious limitation of our method is that it relies on the page counts returned by the search engines; we do not download the pages themselves for further processing. Note that many of the bigrams in our sample are very frequent (up to 10^6 matches; see the "Max" columns in Table 6), hence the effort involved in downloading all pages would be immense (though methods for downloading a representative sample could probably be devised).

Our approach estimates Web frequencies based not on bigram counts directly, but on page counts. In other words, it ignores the fact that a bigram can occur more than once on a given Web page. This approximation is justified, as Zhu and Rosenfeld (2001) demonstrated for unigrams, bigrams, and trigrams: Page counts and *n*-gram counts are highly correlated on a log-log scale. This result is based on Zhu and Rosenfeld's queries to AltaVista, a search engine that at the time of their research returned both the number of pages and the overall number of matches for a given query.[4]

Another important limitation of our approach arises from the fact that both Google and AltaVista disregard punctuation and capitalization, even if the search term is placed within quotation marks. This can lead to false positives, for instance, if the match crosses a phrase boundary, such as in (5), which matches *hungry prey*. Other false positives can be generated by page titles and links, such as the examples (6) and (7) which match *edition broadcast*.[5]

(5) The lion will kill only when it's **hungry. Prey** can usually sense when lions are hunting.

(6) 10th **Edition Broadcast** Products Catalog (as a page title)

(7) Issue/**Edition/Broadcast** (as a link)

The fact that our method does not download Web pages means that no tagging, chunking, or parsing can be carried out to ensure that the matches are correct. Instead we rely on the simple adjacency of the search terms, which is enforced by using queries enclosed within quotation marks (see Section 2.3 for details). This means that we miss any nonadjacent matches, even though a chunker or parser (such as the one used for extracting BNC or NANTC bigrams) would find them. An example is an adjective-noun bigram in which an adverbial intervenes between the adjective and the noun (see Table 1).

Furthermore, the absence of tagging, chunking, and parsing can also generate false positives, in particular for queries containing words with part-of-speech ambiguity. (Recall that our bigram selection procedure ensures that the predicate word, but not the argument word, is unambiguous in terms of its POS tagging in the BNC.) As an example, consider *process directory*, which in our data set is a noun-noun bigram (see Table 2). One of the matches returned by Google is (8), in which *process* is a verb. Another example is *fund membrane*, which is a noun-noun bigram in our data set but matches (9) in Google.

4 Note that this feature of AltaVista has since been discontinued; hence in the present article we had no option but to use page counts. However, Keller, Lapata, and Ourioupina (2002) used AltaVista match counts (instead of page counts) on the same data sets; their results agree with the ones reported in the present article very closely.

5 Some of the examples in (5)–(9) were kindly provided by a reviewer.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

(8) The global catalog server's function is to **process directory** searches for the entire forest.

(9) Green grants **fund membrane** technology.

Another source of noise is the fact that Google (but not AltaVista) will sometimes return pages that do not include the search term at all. This can happen if the search term is contained in a link to the page (but not on the page itself).

As we did not limit our Web searches to English (even though many search engines now allow the target language for a search to be set), there is also a risk that false positives are generated by cross-linguistic homonyms, that is, by words of other languages that are spelled in the same way as the English words in our data sets. However, this problem is mitigated by the fact that English is by far the most common language on the Web, as shown by Grefenstette and Nioche (2000). Also, the chance of two such homonyms forming a valid bigram in another language is probably fairly small.

To summarize, Web counts are certainly less sparse than the counts in a corpus of a fixed size (see Section 2.3). However, Web counts are also likely to be significantly more noisy than counts obtained from a carefully tagged and chunked or parsed corpus, as the examples in this section show. It is therefore essential to carry out a comprehensive evaluation of the Web counts generated by our method. This is the topic of the next section.

3. Evaluation

3.1 Evaluation against Corpus Frequencies

Since Web counts can be relatively noisy, as discussed in the previous section, it is crucial to determine whether there is a reliable relationship between Web counts and corpus 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 whether there is a linear relationship between BNC and NANTC counts and AltaVista and Google counts. All correlation coefficients reported in this article refer to Pearson's r.[6] All results were obtained on log-transformed counts.[7]

Table 8 shows the results of correlating Web counts with corpus counts from the BNC, the corpus from which our bigrams were sampled (see Section 2.1). A high correlation coefficient was obtained across the board, ranging from .720 to .847 for AltaVista counts and from .720 to .850 for Google counts. This indicates that Web counts approximate BNC counts for the three types of bigrams under investigation. Note that there is almost no difference between the correlations achieved using Google and AltaVista counts.

It is important to check that these results are also valid for counts obtained from other corpora. We therefore correlated our Web counts with the counts obtained from NANTC, a corpus that is larger than the BNC but is drawn from a single genre, namely, news text (see Section 2.2). The results are shown in Table 9. We find that

6 Correlation analysis is a way of measuring the degree of linear association between two variables. Effectively, we are fitting a linear equation $y = ax + b$ to the data; this means that the two variables x and y (which in our case represent frequencies or judgments) can still differ by a multiplicative constant a and an additive constant b, even if they are highly correlated.

7 It is well-known that corpus frequencies have a Zipfian distribution. Log-transforming them is a way of normalizing the counts before applying statistical tests. We apply correlation analysis on the log-transformed data, which is equivalent to computing a log-linear regression coefficient on the untransformed data.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Table 8
Correlation of BNC counts with Web counts (seen bigrams).

	Adjective-Noun	Noun-Noun	Verb-Object
AltaVista	.847**	.720**	.762**
Google	.850**	.720**	.766**
Smoothing	.248*	.277**	.342**

*$p < .05$ (one-tailed). **$p < .01$ (one-tailed).

Table 9
Correlation of NANTC counts with Web counts (seen bigrams).

	Adjective-Noun	Noun-Noun	Verb-Object
AltaVista	.712**	.667**	.788**
Google	.712**	.662**	.787**
BNC	.710**	.672**	.814**
Smoothing	.338**	.317**	.263*

*$p < .05$ (one-tailed). **$p < .01$ (one-tailed).

Google and AltaVista counts also correlate significantly with NANTC counts. The correlation coefficients range from .667 to .788 for AltaVista and from .662 to .787 for Google. Again, there is virtually no difference between the correlations for the two search engines. We also observe that the correlation between Web counts and BNC is generally slightly higher than the correlation between Web counts and NANTC counts. We carried out one-tailed *t*-tests to determine whether the differences in the correlation coefficients were significant. We found that both AltaVista counts ($t(87) = 3.11, p < .01$) and Google counts ($t(87) = 3.21$, $p < .01$) were significantly better correlated with BNC counts than with NANTC counts for adjective-noun bigrams. The difference in correlation coefficients was not significant for noun-noun and verb-object bigrams, for either search engine.

Table 9 also shows the correlations between BNC counts and NANTC counts. The intercorpus correlation can be regarded as an upper limit for the correlations we can expect between counts from two corpora that differ in size and genre and that have been obtained using different extraction methods. The correlation between AltaVista and Google counts and NANTC counts reached the upper limit for all three bigram types (one-tailed *t*-tests found no significant differences between the correlation coefficients). The correlation between BNC counts and Web counts reached the upper limit for noun-noun and verb-object bigrams (no significant differences for either search engine) and significantly exceeded it for adjective-noun bigrams for AltaVista ($t(87) = 3.16, p < .01$) and Google ($t(87) = 3.26, p < .01$).

We conclude that simple heuristics (see Section 2.3) 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.). We found that Web counts were highly correlated with frequencies from two different corpora. Furthermore, Web counts and corpus counts are as highly correlated as counts from two different corpora (which can be regarded as an upper bound).

Note that Tables 8 and 9 also provide the correlation coefficients obtained when corpus frequencies are compared with frequencies that were re-created through class-

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

based smoothing, using the BNC as a training corpus (after removing the seen bigrams). This will be discussed in more detail in Section 3.3.

3.2 Evaluation against Plausibility Judgments

Previous work has demonstrated that corpus counts correlate with human plausibility judgments for adjective-noun bigrams. This result holds both for seen bigrams (Lapata, McDonald, and Keller 1999) and for unseen bigrams whose counts have been re-created using smoothing techniques (Lapata, Keller, and McDonald 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 the intuitive plausibility of predicate-argument pairs. The degree of correlation between Web counts and plausibility judgments is an indicator of the quality of the Web counts (compared to corpus counts or counts re-created using smoothing techniques).

3.2.1 Method. For seen and unseen adjective-noun bigrams, we used the two sets of plausibility judgments collected by Lapata, McDonald, and Keller (1999) and Lapata, Keller, and McDonald (2001), respectively. We conducted four additional experiments to collect judgments for noun-noun and verb-object bigrams, both seen and unseen. The experimental method was the same for all six experiments.

Materials. The experimental stimuli were based on the six sets of seen or unseen bigrams extracted from the BNC as described in Section 2.1 (adjective-noun, noun-noun, and verb-object bigrams). In the adjective-noun and noun-noun cases, the stimuli consisted simply of the bigrams. In the verb-object case, the bigrams were embedded in a short sentence to make them more natural: A proper-noun subject was added.

Procedure. The experimental paradigm was magnitude estimation (ME), a technique standardly used in psychophysics to measure judgments of sensory stimuli (Stevens 1975), which Bard, Robertson, and Sorace (1996) and Cowart (1997) have applied to the elicitation of linguistic judgments. The ME procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. In contrast to the five- or seven-point scale conventionally used to measure human intuitions, ME employs an interval scale and therefore produces data for which parametric inferential statistics are valid.

ME requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, to which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In this way, each subject can establish his or her own rating scale, thus yielding maximally fine-graded data and avoiding the known problems with the conventional ordinal scales for linguistic data (Bard, Robertson, and Sorace 1996; Cowart 1997; Schütze 1996).

The experiments reported in this article were carried out using the *WebExp* software package (Keller et al. 1998). A series of previous studies has shown that data obtained using *WebExp* closely replicate results obtained in a controlled laboratory setting; this has been demonstrated for acceptability judgments (Keller and Alexopoulou 2001), coreference judgments (Keller and Asudeh 2001), and sentence completions (Corley and Scheepers 2002).

In the present experiments, subjects were presented with bigram pairs and were asked to rate the degree of plausibility proportional to a modulus item. They first saw a set of instructions that explained the ME technique and the judgment task. The concept of plausibility was not defined, but examples of plausible and implausible bigrams were given (different examples for each stimulus set). Then subjects were asked to fill in a questionnaire with basic demographic information. The experiment proper

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Table 10
Descriptive statistics for plausibility judgments (log-transformed). *N* is the number of subjects used in each experiment.

	Adjective-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

consisted of three phases: (1) a calibration phase, designed to familiarize subjects with the task, in which they had to estimate the length of five horizontal lines; (2) a practice phase, in which subjects judged the plausibility of eight bigrams (similar to the ones in the stimulus set); (3) the main experiment, in which each subject judged one of the six stimulus sets (90 bigrams). The stimuli were presented in random order, with a new randomization being generated for each subject.

Subjects. A separate experiment was conducted for each set of stimuli. The number of subjects per experiment is shown in Table 10 (in the column labeled *N*). All subjects were self-reported native speakers of English; they were recruited by postings to newsgroups and mailing lists. Participation was voluntary and unpaid.

WebExp collects by-item response time data; subjects whose response times were very short or very long were excluded from the sample, as they are unlikely to have completed the experiment adequately. We also excluded the data of subjects who had participated more than once in the same experiment, based on their demographic data and on their Internet connection data, which is logged by *WebExp*.

3.2.2 Results and Discussion. The experimental data were normalized by dividing each numerical judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Bard, Robertson, and Sorace 1996; Cowart 1997; Stevens 1975). All further analyses were conducted on the normalized, log-transformed judgments.

Table 10 shows the descriptive statistics for all six judgment experiments: the original experiments by Lapata, McDonald, and Keller (1999) and Lapata, Keller, and McDonald (2001) for adjective-noun bigrams, and our new ones for noun-noun and verb-object bigrams.

We used correlation analysis to compare corpus counts and Web counts with plausibility judgments. Table 11 (top half) lists the correlation coefficients that were obtained when correlating log-transformed Web counts (AltaVista and Google) and corpus counts (BNC and NANTC) with mean plausibility judgments for seen adjective-noun, noun-noun, and verb-object bigrams. The results show that both AltaVista and Google counts correlate well with plausibility judgments for seen bigrams. The correlation coefficient for AltaVista ranges from .641 to .700; for Google, it ranges from .624 to .692. The correlations for the two search engines are very similar, which is also what we found in Section 3.1 for the correlations between Web counts and corpus counts.

Note 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 whether 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 AltaVista coefficient was significantly

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

higher than the BNC coefficient ($t(87) = 1.76$, $p < .05$), whereas the difference between the Google coefficient and the BNC coefficient failed to reach significance. For the noun-noun bigrams, both the AltaVista and the Google coefficients were significantly higher than the BNC coefficient ($t(87) = 3.11$, $p < .01$ and $t(87) = 2.95$, $p < .01$). Also, for the verb-object bigrams, both the AltaVista coefficient and the Google coefficient were significantly higher than the BNC coefficient ($t(87) = 2.64$, $p < .01$ and $t(87) = 2.32$, $p < .05$).

A similar picture was observed for the NANTC counts. Again, the Web counts outperformed the corpus counts in predicting plausibility. For the adjective-noun bigrams, both the AltaVista and the Google coefficient were significantly higher than the NANTC coefficient ($t(87) = 1.97$, $p < .05$; $t(87) = 1.81$, $p < .05$). For the noun-noun bigrams, the AltaVista coefficient was higher than the NANTC coefficient ($t(87) = 1.64$, $p < .05$), but the Google coefficient was not significantly different from the NANTC coefficient. For verb-object bigrams, the difference was significant for both search engines ($t(87) = 2.74$, $p < .01$; $t(87) = 2.38$, $p < .01$).

In sum, for all three types of bigrams, the correlation coefficients achieved with AltaVista were significantly higher than the ones achieved by either the BNC or the NANTC. Google counts outperformed corpus counts for all bigrams with the exception of adjective-noun counts from the BNC and noun-noun counts from the NANTC.

The bottom panel of Table 11 shows the correlation 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, with the coefficient ranging from .480 to .578 for AltaVista counts and from .473 to .595 for the Google counts. Table 11 also provides the correlations between plausibility judgments and counts re-created using class-based smoothing, which we will discuss in Section 3.3.

An important question is how well humans agree when judging the plausibility of adjective-noun, noun-noun, and verb-noun bigrams. Intersubject agreement gives an upper bound for the task and allows us to interpret how well our Web-based method performs in relation to humans. To calculate intersubject agreement we used leave-

Table 11
Correlation of plausibility judgments with Web counts, corpus counts, and counts re-created using class-based smoothing. "Agreement" refers to the intersubject agreement on the judgment task.

	Adjective-Noun	Noun-Noun	Verb-Object
	Seen Bigrams		
AltaVista	.650**	.700**	.641**
Google	.641**	.692**	.624**
BNC	.569**	.517**	.488**
NANTC	.526**	.597**	.491**
Smoothing	.329**	.318**	.223*
Agreement	.630**	.641**	.604**
	Unseen Bigrams		
AltaVista	.480**	.578**	.551**
Google	.473**	.595**	.520**
Smoothing	.342**	.372**	.298**
Agreement	.550**	.570**	.640**

*$p < .05$ (one-tailed). **$p < .01$ (one-tailed).

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

one-out resampling. This technique is a special case of n-fold cross-validation (Weiss and Kulikowski 1991) and has been previously used for measuring how well humans agree in judging semantic similarity (Resnik 1999, 2000).

For each subject group, we divided the set of the subjects' responses with size n into a set of size $n-1$ (i.e., the response data of all but one subject) and a set of size 1 (i.e., the response data of a single subject). We then correlated the mean ratings of the former set with the ratings of the latter. This was repeated n times (see the number of participants in Table 6); the mean of the correlation coefficients for the seen and unseen bigrams is shown in Table 11 in the rows labeled "Agreement."

For both seen and unseen bigrams, we found no significant difference between the upper bound (intersubject agreement) and the correlation coefficients obtained using either AltaVista or Google counts. This finding holds for all three types of bigrams. The same picture emerged for the BNC and NANTC counts: These correlation coefficients were not significantly different from the upper limit, for all three types of bigrams, both for seen and for unseen bigrams.

To conclude, our evaluation demonstrated that Web counts reliably predict human plausibility judgments, both for seen and for unseen predicate-argument bigrams. AltaVista counts for seen bigrams are a better predictor of human judgments than BNC and NANTC counts. These results show that our heuristic method yields valid frequencies; the simplifications we made in obtaining the Web counts (see Section 2.3), as well as the fact that Web data are noisy (see Section 2.4), seem to be outweighed by the fact that the Web is up to a thousand times larger than the BNC.

3.3 Evaluation against Class-Based Smoothing

The evaluation in the last two sections established that Web counts are useful for approximating corpus counts and for predicting plausibility judgments. As a further step in our evaluation, we correlated Web counts with counts re-created by applying a class-based smoothing method to the BNC.

We re-created co-occurrence frequencies for predicate-argument bigrams using a simplified version of Resnik's (1993) selectional association measure proposed by Lapata, Keller, and McDonald (2001). In a nutshell, this measure replaces Resnik's (1993) information-theoretic approach with a simpler measure that makes no assumptions with respect to the contribution of a semantic class to the total quantity of information provided by the predicate about the semantic classes of its argument. It simply substitutes the argument occurring in the predicate-argument bigram with the concept by which it is represented in the WordNet taxonomy. Predicate-argument co-occurrence frequency is estimated by counting the number of times the concept corresponding to the argument is observed to co-occur with the predicate in the corpus. Because a given word is not always represented by a single class in the taxonomy (i.e., the argument co-occurring with a predicate can generally be the realization of one of several conceptual classes), Lapata, Keller, and McDonald (2001) constructed the frequency counts for a predicate-argument bigram for each conceptual class by dividing the contribution from the argument by the number of classes to which it belongs. They demonstrate that the counts re-created using this smoothing technique correlate significantly with plausibility judgments for adjective-noun bigrams. They also show that this class-based approach outperforms distance-weighted averaging (Dagan, Lee, and Pereira 1999), a smoothing method that re-creates unseen word co-occurrences on the basis of distributional similarity (without relying on a predefined taxonomy), in predicting plausibility.

In the current study, we used the smoothing technique of Lapata, Keller, and McDonald (2001) to re-create not only adjective-noun bigrams, but also noun-noun

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/0891201033227116O4.pdf by guest on 24 March 2025

Table 12
Correlation of counts re-created using class-based smoothing with Web counts.

	Adjective-Noun	Noun-Noun	Verb-Object
	Seen Bigrams		
AltaVista	.344**	.362**	.361**
Google	.330**	.343**	.349**
	Unseen Bigrams		
AltaVista	.439**	.386**	.412**
Google	.444**	.421**	.397**

*$p < .05$ (one-tailed). **$p < .01$ (one-tailed).

and verb-object bigrams. As already mentioned in Section 2.1, it was assumed that the noun is the predicate in adjective-noun bigrams; for noun-noun bigrams, we treated the right noun as the predicate, and for verb-object bigrams, we treated the verb as the predicate. We applied Lapata, Keller, and McDonald's (2001) technique to the unseen bigrams for all three bigram types. We also used it on the seen bigrams, which we were able to treat as unseen by removing all instances of the bigrams from the training corpus.

To test the claim that Web frequencies can be used to overcome data sparseness, we correlated the frequencies re-created using class-based smoothing on the BNC with the frequencies obtained from the Web. The correlation coefficients for both seen and unseen bigrams are shown in Table 12. In all cases, a significant correlation between Web counts and re-created counts is obtained. For seen bigrams, the correlation coefficient ranged from .344 to .362 for AltaVista counts and from .330 to .349 for Google counts. For unseen bigrams, the correlations were somewhat higher, ranging from .386 to .439 for AltaVista counts and from .397 to .444 for Google counts. For both seen and unseen bigrams, there was only a very small difference between the correlation coefficients obtained with the two search engines.

It is also interesting to compare the performance of class-based smoothing and Web counts on the task of predicting plausibility judgments. The correlation coefficients are listed in Table 11. The re-created frequencies are correlated significantly with all three types of bigrams, both for seen and unseen bigrams. For the seen bigrams, we found that the correlation coefficients obtained using smoothed counts were significantly lower than the upper bound for all three types of bigrams ($t(87) = 3.01$, $p < .01$; $t(87) = 3.23$, $p < .01$; $t(87) = 3.43$, $p < .01$). This result also held for the unseen bigrams: The correlations obtained using smoothing were significantly lower than the upper bound for all three types of bigrams ($t(87) = 1.86$, $p < .05$; $t(87) = 1.97$, $p < .05$; $t(87) = 3.36$, $p < .01$).

Recall that the correlation coefficients obtained using the Web counts were not found to be significantly different from the upper bound, which indicates that Web counts are better predictors of plausibility than smoothed counts. This fact was confirmed by further significance testing: For seen bigrams, we found that the AltaVista correlation coefficients were significantly higher than correlation coefficients obtained using smoothing, for all three types of bigrams ($t(87) = 3.31$, $p < .01$; $t(87) = 4.11$, $p < .01$; $t(87) = 4.32$, $p < .01$). This also held for Google counts ($t(87) = 3.16$, $p < .01$; $t(87) = 4.02$, $p < .01$; $t(87) = 4.03$, $p < .01$). For unseen bigrams, the AltaVista coefficients and the coefficients obtained using smoothing were not significantly different

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

for adjective-noun bigrams, but the difference reached significance for noun-noun and verb-object bigrams ($t(87) = 2.08$, $p < .05$; $t(87) = 2.53$, $p < .01$). For Google counts, the difference was again not significant for adjective-noun bigrams, but it reached significance for noun-noun and verb-object bigrams ($t(87) = 2.34$, $p < .05$; $t(87) = 2.15$, $p < .05$).

Finally, we conducted a small study to investigate the validity of the counts that were re-created using class-based smoothing. We correlated the re-created counts for the seen bigrams with their actual BNC and NANTC frequencies. The correlation coefficients are reported in Tables 8 and 9. We found that the correlation between re-created counts and corpus counts was significant for all three types of bigrams, for both corpora. This demonstrates that the smoothing technique we employed generates realistic corpus counts, in the sense that the re-created counts are correlated with the actual counts. However, the correlation coefficients obtained using Web counts were always substantially higher than those obtained using smoothed counts. These differences were significant for the BNC counts for AltaVista ($t(87) = 8.38$, $p < .01$; $t(87) = 5.00$, $p < .01$; $t(87) = 5.03$, $p < .01$) and Google ($t(87) = 8.35$, $p < .01$; $t(87) = 5.00$, $p < .01$; $t(87) = 5.03$, $p < .01$). They were also significant for the NANTC counts for AltaVista ($t(87) = 4.12$, $p < .01$; $t(87) = 3.72$, $p < .01$; $t(87) = 6.58$, $p < .01$) and Google ($t(87) = 4.08$, $p < .01$; $t(87) = 3.06$, $p < .01$; $t(87) = 6.47$, $p < .01$).

To summarize, the results presented in this section indicate that Web counts are indeed a valid way of obtaining counts for bigrams that are unseen in a given corpus: They correlate reliably with counts re-created using class-based smoothing. For seen bigrams, we found that Web counts correlate with counts that were re-created using smoothing techniques (after removing the seen bigrams from the training corpus). For the task of predicting plausibility judgments, we were able to show that Web counts outperform re-created counts, both for seen and for unseen bigrams. Finally, we found that Web counts for seen bigrams correlate better than re-created counts with the real corpus counts.

It is beyond the scope of the present study to undertake a full comparison between Web counts and frequencies re-created using all available smoothing techniques (and all available taxonomies that might be used for class-based smoothing). The smoothing method discussed above is simply one type of class-based smoothing. Other, more sophisticated class-based methods do away with the simplifying assumption that the argument co-occurring with a given predicate (adjective, noun, verb) is distributed evenly across its conceptual classes and attempt to find the right level of generalization in a concept hierarchy, by discounting, for example, the contribution of very general classes (Clark and Weir 2001; McCarthy 2000; Li and Abe 1998). Other smoothing approaches such as discounting (Katz 1987) and distance-weighted averaging (Grishman and Sterling 1994; Dagan, Lee, and Pereira 1999) re-create counts of unseen word combinations by exploiting only corpus-internal evidence, without relying on taxonomic information. Our goal was to demonstrate that frequencies retrieved from the Web are a viable alternative to conventional smoothing methods when data are sparse; we do not claim that our Web-based method is necessarily superior to smoothing or that it should be generally preferred over smoothing methods. However, the next section will present a small-scale study that compares the performance of several smoothing techniques with the performance of Web counts on a standard task from the literature.

3.4 Pseudodisambiguation

In the smoothing literature, re-created frequencies are typically evaluated using pseudodisambiguation (Clark and Weir 2001; Dagan, Lee, and Pereira 1999; Lee 1999; Pereira, Tishby, and Lee 1993; Prescher, Riezler, and Rooth 2000; Rooth et al. 1999).

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

The aim of the pseudodisambiguation task is to decide whether a given algorithm re-creates frequencies that make it possible to distinguish between seen and unseen bigrams in a given corpus. A set of pseudobigrams is constructed according to a set of criteria (detailed below) that ensure that they are unattested in the training corpus. Then the seen bigrams are removed from the training data, and the smoothing method is used to re-create the frequencies of both the seen bigrams and the pseudobigrams. The smoothing method is then evaluated by comparing the frequencies it re-creates for both types of bigrams.

We evaluated our Web counts by applying the pseudodisambiguation procedure that Rooth et al. (1999), Prescher, Riezler, and Rooth (2000), and Clark and Weir (2001) employed for evaluating re-created verb-object bigram counts. In this procedure, the noun n from a verb-object bigram (v, n) that is seen in a given corpus is paired with a randomly chosen verb v' that does not take n as its object within the corpus. This results in an unseen verb-object bigram (v', n). The seen bigram is now treated as unseen (i.e., all of its occurrences are removed from the training corpus), and the frequencies of both the seen and the unseen bigram are re-created (using smoothing, or Web counts, in our case). The task is then to decide which of the two verbs v and v' takes the noun n as its object. For this, the re-created bigram frequency is used: The bigram with the higher re-created frequency (or probability) is taken to be the seen bigram. If this bigram is really the seen one, then the disambiguation is correct. The overall percentage of correct disambiguations is a measure of the quality of the re-created frequencies (or probabilities). In the following, we will first describe in some detail the experiments that Rooth et al. (1999) and Clark and Weir (2001) conducted. We will then discuss how we replicated their experiments using the Web as an alternative smoothing method.

Rooth et al. (1999) used pseudodisambiguation to evaluate a class-based model that is derived from unlabeled data using the expectation maximization (EM) algorithm. From a data set of 1,280,712 (v, n) pairs (obtained from the BNC using Carroll and Rooth's [1998] parser), they randomly selected 3,000 pairs, with each pair containing a fairly frequent verb and noun (only verbs and nouns that occurred between 30 and 3,000 times in the data were considered). For each pair (v, n) a fairly frequent verb v' was randomly chosen such that the pair (v', n) did not occur in the data set. Given the set of (v, n, v') triples (a total of 1,337), the task was to decide whether (v, n) or (v', n) was the correct (i.e., seen) pair by comparing the probabilities $P(n|v)$ and $P(n|v')$. The probabilities were re-created using Rooth et al.'s (1999) EM-based clustering model on a training set from which all seen pairs (v, n) had been removed. An accuracy of 80% on the pseudodisambiguation task was achieved (see Table 13).

Prescher, Riezler, and Rooth (2000) evaluated Rooth et al.'s (1999) EM-based clustering model again using pseudodisambiguation, but on a separate data set using a

Table 13
Percentage of correct disambiguations on the pseudodisambiguation task using Web counts and counts re-created using EM-based clustering (Rooth et al. 1999).

Data Set	N	AltaVista Conditional Probability	AltaVista Joint Probability	Rooth et al.
Subject	717	71.2	68.5	—
Objects	620	85.2	77.5	—
Subjects and objects	1,337	77.7	72.7	80.0

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Table 14
Percentage of correct disambiguations on the pseudodisambiguation task using Web counts and counts re-created using EM-based clustering (Prescher, Riezler, and Rooth 2000).

Data Set	*N*	AltaVista Conditional Probability	AltaVista Joint Probability	VA Model	VO Model
Subjects	159	66.7	59.1	—	—
Objects	139	70.5	66.2	—	—
Subjects and objects	298	68.5	62.4	79.0	88.3

slightly different method for constructing the pseudobigrams. They used a set of 298 (v, n, n') BNC triples in which (v, n) was chosen as in Rooth et al. (1999) but paired with a randomly chosen noun n'. Given the set of (v, n, n') triples, the task was to decide whether (v, n) or (v, n') was the correct pair in each triple. Prescher, Riezler, and Rooth (2000) reported pseudodisambiguation results with two clustering models: (1) Rooth et al.'s (1999) clustering approach, which models the semantic fit between a verb and its argument (VA model), and (2) a refined version of this approach that models only the fit between a verb and its object (VO model), disregarding other arguments of the verb. The results of the two models on the pseudodisambiguation task are shown in Table 14.

At this point, it is important to note that neither Rooth et al. (1999) nor Prescher, Riezler, and Rooth (2000) used pseudodisambiguation for the final evaluation of their models. Rather, the performance on the pseudodisambiguation task was used to optimize the model parameters. The results in Tables 13 and 14 show the pseudodisambiguation performance achieved for the best parameter settings. In other words, these results were obtained on the development set (i.e., on the same data set that was used to optimize the parameters), not on a completely unseen test set. This procedure is well-justified in the context of Rooth et al.'s (1999) and Prescher, Riezler, and Rooth's (2000) work, which aimed at building models of lexical semantics, not of pseudodisambiguation. Therefore, they carried out their final evaluations on unseen test sets for the tasks of lexicon induction (Rooth et al. 1999) and target language disambiguation (Prescher, Riezler, and Rooth 2000), once the model parameters had been fixed using the pseudodisambiguation development set.[8]

Clark and Weir (2002) use a setting similar to that of Rooth et al. (1999) and Prescher, Riezler, and Rooth (2000); here pseudodisambiguation is employed to evaluate the performance of a class-based probability estimation method. In order to address the problem of estimating conditional probabilities in the face of sparse data, Clark and Weir (2002) define probabilities in terms of classes in a semantic hierarchy and propose hypothesis testing as a means of determining a suitable level of generalization in the hierarchy. Clark and Weir (2002) report pseudodisambiguation results on two data sets, with an experimental setup similar to that of Rooth et al. (1999). For the first data set, 3,000 pairs were randomly chosen from 1.3 million (v, n) tuples extracted from the BNC (using the parser of Briscoe and Carroll [1997]). The selected pairs con-

8 Stefan Riezler (personal communication, 2003) points out that the main variance in Rooth et al.'s (1999) pseudodisambiguation results comes from the class cardinality parameter (start values account for only 2% of the performance, and iterations do not seem to make a difference at all). Figure 3 of Rooth et al. (1999) shows that a performance of more than 75% is obtained for every reasonable choice of classes. This indicates that a "proper" pseudodisambiguation setting with separate development and test data would have resulted in a similar choice of class cardinality and thus achieved the same 80% performance that is cited in Table 13.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Table 15
Percentage of correct disambiguations on the pseudodisambiguation task using Web counts and counts re-created using class-based smoothing (Clark and Weir 2002).

Data Set	N	AltaVista Conditional Probability	AltaVista Joint Probability	Clark and Weir	Li and Abe	Resnik
Objects (low frequency)	3000	83.9	81.1	72.4	62.9	62.6
Objects (high frequency)	3000	87.7	85.3	73.9	68.3	63.9

tained relatively frequent verbs (occurring between 500 and 5,000 times in the data). The data sets were constructed as proposed by Rooth et al. (1999). The procedure for creating the second data set was identical, but this time only verbs that occurred between 100 and 1,000 times were considered. Clark and Weir (2002) further compared their approach with Resnik's (1993) selectional association model and Li and Abe's (1998) tree cut model on the same data sets. These methods are directly comparable, as they can be used for class-based probability estimation and address the question of how to find a suitable level of generalization in a hierarchy (i.e., WordNet). The results of the three methods used on the two data sets are shown in Table 15.

We employed the same pseudodisambiguation method to test whether Web-based frequencies can be used for distinguishing between seen and artificially constructed unseen bigrams. We obtained the data sets of Rooth et al. (1999), Prescher, Riezler, and Rooth (2000), and Clark and Weir (2002) described above. Given a set of (v, n, v') triples, the task was to decide whether (v, n) or (v', n) was the correct pair. We obtained AltaVista counts for $f(v, n)$, $f(v', n)$, $f(v)$, and $f(v')$ as described in Section 2.3.[9] Then we used two models for pseudodisambiguation: the joint probability model compared the joint probability estimates $f(v, n)$ and $f(v', n)$ and predicted that the bigram with the highest estimate is the seen one. The conditional probability model compared the conditional probability estimates $f(v, n)/f(v)$ and $f(v', n)/f(v')$ and again selected as the seen bigram the one with the highest estimate (in both cases, ties were resolved by choosing at random).[10] The same two models were used to perform pseudodisambiguation for the (v, n, n') triples, where we have to choose between (v, n) and (v, n'). Here, the probability estimates $f(v, n)$ and $f(v, n')$ were used for the joint probability model, and $f(v, n)/f(n)$ and $f(v, n')/f(n')$ for the conditional probability model.

The results for Rooth et al.'s (1999) data set are given in Table 13. The conditional probability model achieves a performance of 71.2% correct for subjects and 85.2% correct for objects. The performance on the whole data set is 77.7%, which is below the performance of 80.0% reported by Rooth et al. (1999). However, the difference is not found to be significant using a chi-square test comparing the number of correct and incorrect classifications ($\chi^2(1) = 2.02$, $p = .16$). The joint probability model performs consistently worse than the conditional probability model: It achieves an overall accuracy of 72.7%, which is significantly lower than the accuracy of the Rooth et al. (1999) model ($\chi^2(1) = 19.50$, $p < .01$).

9 We used only AltaVista counts, as there was virtually no difference between AltaVista and Google counts in our previous evaluations (see Sections 3.1–3.3). Google allows only 1,000 queries per day (for registered users), which makes it time-consuming to obtain large numbers of Google counts. AltaVista has no such restriction.

10 The probability estimates are $P(a, b) = f(a, b)/N$ and $P(a|b) = f(a, b)/f(a)$ for the joint probability and the conditional probability, respectively. However, the corpus size N can be ignored, as it is constant.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

A similar picture emerges with regard to Prescher, Riezler, and Rooth's (2000) data set (see Table 14). The conditional probability model achieves an accuracy of 66.7% for subjects and 70.5% for objects. The combined performance of 68.5% is significantly lower than the performance of both the VA model ($\chi^2(1) = 7.78$, $p < .01$) and the VO model ($\chi^2(1) = 33.28$, $p < .01$) reported by Prescher, Riezler, and Rooth (2000). Again, the joint probability model performs worse than the conditional probability model, achieving an overall accuracy of 62.4%.

We also applied our Web-based method to the pseudodisambiguation data set of Clark and Weir (2002). Here, the conditional probability model reached a performance of 83.9% correct on the low-frequency data set. This is significantly higher than the highest performance of 72.4% reported by Clark and Weir (2002) on the same data set ($\chi^2(1) = 115.50$, $p < .01$). The joint probability model performs worse than the conditional model, at 81.1%. However, this is still significantly better than the best result of Clark and Weir (2002) ($\chi^2(1) = 63.14$, $p < .01$). The same pattern is observed for the high-frequency data set, on which the conditional probability model achieves 87.7% correct and thus significantly outperforms Clark and Weir (2002), who obtained 73.9% ($\chi^2(1) = 283.73$, $p < .01$). The joint probability model achieved 85.3% on this data set, also significantly outperforming Clark and Weir (2002) ($\chi^2(1) = 119.35$, $p < .01$).

To summarize, we demonstrated that the simple Web-based approach proposed in this article yields results for pseudodisambiguation that outperform class-based smoothing techniques, such as the ones proposed by Resnik (1993), Li and Abe (1998), and Clark and Weir (2002). We were also able to show that a Web-based approach is able to achieve the same performance as an EM-based smoothing model proposed by Rooth et al. (1999). However, the Web-based approach was not able to outperform the more sophisticated EM-based model of Prescher, Riezler, and Rooth (2000). Another result we obtained is that Web-based models that use conditional probabilities (where unigram frequencies are used to normalize the bigram frequencies) generally outperform a more simple-minded approach that relies directly on bigram frequencies for pseudodisambiguation.

There are a number of reasons why our results regarding pseudodisambiguation have to be treated with some caution. First of all, the two smoothing methods (i.e., EM-based clustering and class-based probability estimation using WordNet) were not evaluated on the same data set, and therefore the two results are not directly comparable. For instance, Clark and Weir's (2002) data set is substantially less noisy than Rooth et al.'s (1999) and Prescher, Riezler, and Rooth's (2000), as it contains only words and nouns that occur in WordNet. Furthermore, Stephen Clark (personal communication, 2003) points out that WordNet-based approaches are at a disadvantage when it comes to pseudodisambiguation. Pseudodisambiguation assumes that the correct pair is unseen in the training data; this makes the task deliberately hard, because some of the pairs might be frequent enough that reliable corpus counts can be obtained without having to use WordNet (using WordNet is likely to be more noisy than using the actual counts). Another problem with WordNet-based approaches is that they offer no systematic treatment of word sense ambiguity, which puts them at a disadvantage with respect to approaches that do not rely on a predefined inventory of word senses.

Finally, recall that the results for the EM-based approaches in Tables 13 and 14 were obtained on the development set (as pseudodisambiguation was used as a means of parameter tuning by Rooth et al. [1999] and Prescher, Riezler, and Rooth [2000]). It is possible that this fact inflates the performance values for the EM-based approaches (but see note 8).

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

4. Conclusions

This article explored a novel approach to overcoming data sparseness. If a bigram is unseen in a given corpus, conventional approaches re-create its frequency using techniques such as back-off, linear interpolation, class-based smoothing or distance-weighted averaging (see Dagan, Lee, and Pereira [1999] and Lee [1999] for overviews). The approach proposed here does not re-create 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 the 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. We then obtained bigram counts from NANTC, a corpus that is substantially larger than the BNC. Again, we found that Web counts are highly correlated with corpus counts. This indicates that Web queries can generate frequencies that are comparable to the ones obtained from a balanced, carefully edited corpus such as the BNC, but also from a large news text corpus such as NANTC.

Secondly, we performed an 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 corpus frequencies.

To substantiate the claim that the Web counts can be used to overcome data sparseness, we compared bigram counts obtained from the Web with bigram counts re-created using a class-based smoothing technique (a variant of the one proposed by Resnik [1993]). We found that Web frequencies and re-created frequencies are reliably correlated, and that Web frequencies are better at predicting plausibility judgments than smoothed frequencies. This holds both for unseen bigrams and for seen bigrams that are treated as unseen by omitting them from the training corpus.

Finally, we tested the performance of our frequencies in a standard pseudodisambiguation task. We applied our method to three data sets from the literature. The results show that Web counts outperform counts re-created using a number of class-based smoothing techniques. However, counts re-created using an EM-based smoothing approach yielded better pseudodisambiguation performance than Web counts.

To summarize, we have proposed a simple heuristic method for obtaining bigram counts from the Web. Using four different types of evaluation, we demonstrated that this simple heuristic method is sufficient to obtain valid 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).

A number of questions arise for future research: (1) Are Web frequencies suitable for probabilistic modeling, in particular since Web counts are not perfectly normalized, as Zhu and Rosenfeld (2001) have shown? (2) How can existing smoothing methods benefit from Web counts? (3) How do the results reported in this article carry over to languages other than English (for which a much smaller amount of Web data are available)? (4) What is the effect of the noise introduced by our heuristic approach? The last question could be assessed by reproducing our results using a snapshot of the Web, from which argument relations can be extracted more accurately using POS tagging and chunking techniques.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Finally, it will be crucial to test the usefulness of Web-based frequencies for realistic NLP tasks. Preliminary results are reported by Lapata and Keller (2003), who use Web counts successfully for a range of NLP tasks, including candidate selection for machine translation, context-sensitive spelling correction, bracketing and interpretation of compounds, adjective ordering, and PP attachment.

Acknowledgments

This work was conducted while both authors were at the Department of Computational Linguistics, Saarland University, Saarbrücken. The work was inspired by a talk that Gregory Grefenstette gave in Saarbrücken in 2001 about his research on using the Web as a corpus. The present article is an extended and revised version of Keller, Lapata, and Ourioupina (2002). Stephen Clark and Stefan Riezler provided valuable comments on this research. We are also grateful to four anonymous reviewers for *Computational Linguistics*; their feedback helped to substantially improve the present article. Special thanks are due to Stephen Clark and Detlef Prescher for making their pseudodisambiguation data sets available.

References

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

Abney, Steve and Marc Light. 1999. Hiding a semantic class hierarchy in a Markov model. In Andrew Kehler and Andreas Stolcke, editors, *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8, College Park, MD.

Agirre, Eneko 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*, pages 11–19, Saarbrücken, Germany.

Banko, Michele 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 First International Conference on Human Language Technology Research*. Morgan Kaufmann, San Francisco.

Banko, Michele 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*, pages 26–33, Toulouse, France.

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

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Burnard, Lou, editor, 1995. *Users Reference Guide, British National Corpus*. British National Corpus Consortium, Oxford University Computing Services, Oxford, England.

Carroll, Glenn and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In Nancy Ide and Atro Voutilainen, editors, *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, pages 36–45, Granada, Spain.

Clark, Stephen and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics*, pages 95–102, Pittsburgh, PA.

Clark, Stephen and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Corley, Martin 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):126–131.

Corley, Steffan, 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.

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

Curran, James. 2002. Scaling context space. In *Proceedings of the 40th Annual Meeting of*

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

the Association for Computational Linguistics, pages 231–238, Philadelphia.
Dagan, Ido, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.
Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic, Boston.
Grefenstette, Gregory. 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.
Grefenstette, Gregory 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.
Grishman, Ralph and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 742–747, Kyoto, Japan.
Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.
Jones, Rosie 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.
Karp, Daniel, Yves Schabes, Martin Zaidel, and Dania Egedi. 1992. A freely available wide coverage morphological analyzer for English. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 950–954, Nantes, France.
Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 33(3):400–401.
Keller, Frank and Theodora Alexopoulou. 2001. Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, 79(3):301–372.
Keller, Frank 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. Erlbaum, Mahwah, NJ.
Keller, Frank, 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.
Keller, Frank, Maria Lapata, and Olga Ourioupina. 2002. Using the Web to overcome data sparseness. In Jan Hajič and Yuji Matsumoto, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Philadelphia.
Lapata, Maria. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics*, pages 63–70, Pittsburgh, PA.
Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.
Lapata, Maria and Frank Keller. 2003. Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. Unpublished manuscript, University of Sheffield, Sheffield, England, and University of Edinburgh, Edinburgh, Scotland.
Lapata, Maria, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgments. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 346–353, Toulouse, France.
Lapata, Maria, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen, Norway.
Lauer, Mark. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University, Sydney, Australia.
Lee, Lilian. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, College Park, MD.
Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. The tagging of the British national corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 622–628, Kyoto, Japan.
Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025

Press, New York.
Li, Hang and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.
Lin, Dekang. 1994. PRINCIPAR—An efficient broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 482–488, Kyoto, Japan.
Lin, Dekang. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of the LREC Workshop on the Evaluation of Parsing Systems*, pages 48–56, Granada, Spain.
Lin, Dekang. 2001. LaTaT: Language and text analysis tools. In *Proceedings of the First International Conference on Human Language Technology Research*. Morgan Kaufmann, San Francisco.
Markert, Katja, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the Web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, Budapest, pages 39–46.
McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*, pages 256–263, Seattle, WA.
Mihalcea, Rada 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, College Park, MD.
Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.
Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.
Prescher, Detlef, Stefan Riezler, and Mats Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany.
Ratnaparkhi, Adwait. 1998. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 1079–1085, Montréal.
Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia.
Resnik, Philip. 1999. Mining the Web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, MD.
Resnik, Philip. 2000. Measuring verb similarity. In Lila R. Gleitman and Aravid K. Joshi, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 399–404. Erlbaum, Mahwah, NJ.
Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, College Park, MD.
Sampson, Geoffrey. 1995. *English for the Computer*: *The SUSANNE Corpus and Analytic Scheme*. Oxford University Press, Oxford.
Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago.
Stevens, S. S. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley, New York.
Volk, Martin. 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, England.
Weiss, Sholom M. and Casimir A. Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.
Zhu, Xiaojin and Ronald Rosenfeld. 2001. Improving trigram language modeling with the World Wide Web. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, UT.

Downloaded from http://direct.mit.edu/coli/article-pdf/29/3/459/1798134/089120103322711604.pdf by guest on 24 March 2025