Automatic Selection of Collocations for Instruction

Adam Skory, Maxine Eskenazi

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

askory@cs.cmu.edu, max@cs.cmu.edu

Abstract

For teaching of collocations no resource exists that comprehensively ranks collocations in terms of usefulness for learners. Towards developing a method to produce such a resource, we define a collocation's utility in terms of its unpredictability; the inability of a student to derive the meaning of the collocation from her semantic knowledge of its constituent words. We conduct an experiment comparing knowledge of phrasal verb collocations to familiarity with each collocation's verb constituent in order to have empirical measures of predictability. We then investigate corpus-based methods to approximate collocation predictability and find statistically significant correlations between a subset of these methods and the experimental data. This demonstrates that automated statistical approaches can significantly approximate the predictability of phrasal verbs according to our measures. We intend for this research to lead to development of resources for automated content selection in CALL.

Index Terms: collocations, language learning

1. Introduction

1.1. Defining Collocation

Collocations are broadly described as frequently occurring word sequences with meaning or usage patterns beyond the regular generative rules of syntax or semantics. [1]. The statistical description of collocations defines them as sequences of words that occur more often than chance would predict. A linguistic description of collocation involves separating productive phrases from lexical phrases. Nattinger & DeCarrico [2] describe lexical phrases as

> 'chunks' of language of varying lengths, phrases like *as it were, on the other hand, as X would have us believe*, and so on. As such, they are multi-word lexical phenomena that exist somewhere between the traditional poles of lexicon and syntax, conventionalized form/function composites that occur more frequently and have more idiomatically determined meaning than language that is put together each time.

Traditionally, the deciding factor for inclusion of a linguistic item in a lexicon was considered to be whether or not it can be derived from other elements of the lexicon. Items such as lexical phrases, however, can have rich and variable meanings with cultural significance and usage patterns that are not always derivable. As students become more familiar with a phrase, they are more motivated to store that phrase, and it's non-derivable attributes, in their lexicon. [2]

This same concept applies to collocations, which are distinguished in several ways from the more general category of lexical phrases. For one, lexical phrases can be arbitrarily long and complex. Collocations are rarely longer than three words, and can be used in place of atomic parts of speech such as nouns and verbs. For example, the idiomatic expression *hit the nail on the head* is a strongly fixed lexical phrase, but would not generally be considered a collocation. The lexical phrases *hit list* and *hit on* are both collocations and both take the role of simple parts of speech.

Evert [3] defines collocations as semi-compositional word pairs, with one 'free' element (the *base*) and the other element lexically determined (the *collocate*). This definition provides conceptual roles for a collocation’s constituents; the verb *hit* is thus the base of *hit on*, and the particle *on* is the collocate.

1.2. Collocation in Language Learning

Extensive instruction of collocations reflects a lexical approach to language teaching [4]. Research in language development and psycholinguistics has shown that the role of the lexicon is much larger than was previously thought. There is evidence that phrases that can be syntactically decomposed, if frequent enough, may be stored as units in the lexicon [2].

This evidence has led to a resurgence of the lexical approach, and many classroom and paper-based activities exist to teach collocations. Reading exercises, games, and matching tasks are popular with English teachers at many different levels. [5] One activity of particular interest to developers of language learning software is the cloze task, also known as fill-in-the-blank sentences. Cloze tasks are appealing tools for CALL because they can be produced automatically from authentic sources, such as news text and web pages, and can be automatically graded.

Woolard [6] gives a survey of available resources for teaching collocations, including uses of traditional dictionaries and a list of collocation-specific dictionaries. These resources cannot be used directly to discriminate between collocations for teaching, but will be of value if used in conjunction with methods to evaluate the relative utility of collocations for explicit instruction.

Educators [4,6] suggest the use of *predictability* as a method to determine which collocations should be included in lessons and activities. Instructors can observe students' difficulties predicting the meaning of a collocation from its constituents. They can use this classroom experience qualitatively to select instructional content.

2. Empirical Ranking Methods

In order to give an empirical and quantitative assessment of collocation predictability we compared English learners' performance with phrasal verbs to their familiarity with constituents of the same. Rather than relying on manual assessment of instructional utility, we engage the concept of semantic predictability and develop candidate measures of real student data.

To reduce hidden variables that might arise from constituents of multiple parts of speech, the study was limited to two-word phrasal verbs. We intend for future studies to investigate generalization of these results to other types of collocations. To find a superset of candidate phrasal verbs we selected 10 of the most frequent English prepositions to serve as particles. For each preposition we collected up to 100 verbs that give high mutual information scores when immediately preceding that preposition. We used (MI) for this step because it is one of the simplest but most standard association measures used to identify collocations [3]. Calculations of MI were made using the English Gigaword corpus [7].

For each collocation in this superset we found rankings using the statistical measures detailed in Section (3). We then chose 60 phrasal verbs (6 verbs for each of the 10 prepositions) according to the variance of that collocation's respective rankings by each measure. We propose that each measure represents a distribution along which collocations are spread, so our goal was to find a subset of collocations that would be spread as evenly as possible according to as many of these measures as possible.

2.1. Design of the task

Sentences containing these 60 collocations were extracted from the Gigaword corpus. For each collocation one sentence of 20 words or less was chosen manually. These sentences were used to create multiple-choice cloze tasks. The preposition was removed and three manually created distractors were added. The correct answer plus the distractors were presented in randomized order.

Each cloze task was completed by at least 11 people. The 34 participants were college or graduate students and non-native speakers of English. 14 native languages were represented.

Each participant saw 20 cloze tasks, corresponding to 20 of the 60 collocations. After completing these cloze tasks, the participants were presented just the verb of each phrasal verb and asked to rank their familiarity with that verb on a 5-point scale ranging from “1. I have never seen this verb.” to “5. I know every meaning of this verb, when to use it, and when not to use it.” These ratings record cases where participants predict the meaning of a phrasal verb without full knowledge of the verb's meaning and cases where they do not predict its meaning despite a strong familiarity with the verb.

2.2. Evaluation of the task

Different measures of the experimental results may correspond to dissimilar qualitative opinions of collocation difficulty and predictability. Thus we must avoid selecting any one such interpretation as most representative of student performance. We calculate a broad range of measures to capture distinct patterns in the data.

The simplest of these measures ranks collocations based on the proportion of correct responses to the corresponding tasks. We expect this measure to represent the overall difficulty of a collocation. We denote a task-completion event as *e*, and the set of all events as *E*. An event *e* represents a tuple *e:=((x,y),p,c,r)*, where *(x,y)* denotes the two words of the collocation, *p* denotes the participant, *c* is a Boolean representing success or failure at completing the task, and *r* is *p*'s familiarity rating for *x* (the verb).

The proportion of correct responses to a task is defined as:

$$\mathrm{PC}(x_i, y_i) := \frac{|\{e \in E : e[(x,y)] = (x_i, y_i) \wedge e[c] = \mathrm{True}\}|}{|\{e \in E : e[(x,y)] = (x_i, y_i)\}|} \qquad (1)$$

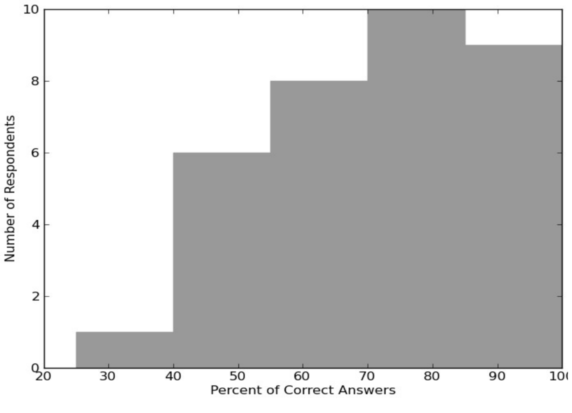

Figure (1): Distribution of respondents by score.

This method assumes an even distribution in terms of the levels of the participants, however our sample included a larger proportion of high-performing speakers. To address this skew we score each participant according to her success rate:

$$\mathrm{score}(p_i) := \frac{|\{e \in E : e[p] = p_i \wedge e[c] = \mathrm{True}\}|}{|\{e \in E : e[p] = p_i\}|} \qquad (2)$$

Figure (1) shows the rightwards skew in the distribution of respondents by score. A measurement of the difficulty of the tasks that is less sensitive to this skew considers the scoring characteristics of the subset of successful respondents (P_T) for a given collocation (*x,y*). We consider the minimum, mean, and median of those scores.

$$P_\mathrm{T}(x_i, y_i) := \{p_i \in P : (\forall e \in E : e[p] = p_i \Rightarrow e[c] = \mathrm{True})\}$$

$$\mathrm{PS}_\mathrm{MIN}(x_i, y_i) := \underset{p_i \in P_\mathrm{T}(x_i, y_i)}{\mathrm{ARGMIN}} \mathrm{score}(p_i) \qquad (3)$$

$$\mathrm{PS}_\mathrm{AVG}(x_i, y_i) := \frac{\sum_{p_i \in P_{corect}(x_i, y_i)} \mathrm{score}(p_i)}{|P_\mathrm{T}(x_i, y_i)|} \qquad (4)$$

$$P_h := \{p_h \in P : p_h < p_i\};\ P_j := \{p_j \in P : p_j > p_i\}$$

$$p_\mathrm{MDN} := \underset{P_h, P_j}{\mathrm{ARGMIN}} |\ (|P_h| - |P_j|)\ |$$

$$\mathrm{PS}_\mathrm{MDN}(x_i, y_i) := \mathrm{score}(p_\mathrm{MDN}) \qquad (5)$$

Next we take the verb familiarity ratings into account and implement a measure that represents *un*predictability of collocations. The measure gives most weight to cases when the verb is well known, but the phrasal verb was not identified. We define this measure as the sum of the corresponding verb ratings for each cloze task that was not completed correctly.

$$E_\mathrm{F}(x_i, y_i) := \{e \in E : e[(x,y)] = (x_i, y_i) \wedge e[c] = \mathrm{False}\}$$

$$\mathrm{SUMR}_\mathrm{F}(x_i, y_i) := \sum_{e \in E_\mathrm{F}(x_i, y_i)} e[r] \qquad (6)$$

Research has shown that people more reliably self-assess what they do *not* know [8], so we implement an alternative measure that ranks the cloze tasks in inverse proportion to the number of times each was completed correctly but rated as not well known.

$$E_\mathrm{T}(x_i, y_i) := \{e \in E : e[(x,y)] = (x_i, y_i) \wedge e[c] = \mathrm{True}\}$$

$$\mathrm{SUMR}_\mathrm{-T}(x_i, y_i) := -\sum_{e \in E_\mathrm{T}(x_i, y_i)} e[r] \qquad (7)$$

Finally, for completeness, we include one last measure that is the inverse sum of the ratings of the correct answers plus the sum of the ratings of the incorrect answers.

$$\mathrm{SUMR}_\mathrm{T+F} := \mathrm{SUMR}_\mathrm{T} + \mathrm{SUMR}_\mathrm{+F} \qquad (8)$$

To demonstrate each measure's output, we manually categorize collocations as (4) least difficult, (3) less difficult but predictable, (2) more difficult and less predictable, (1) and most difficult and *un*predictable. Table (1) shows examples of these categories ranked by each measure on a scale of 1 (most difficult) to 60 (least difficult).

Category	**Example**	**PC**	**PS$_\mathrm{MIN}$**	**PS$_\mathrm{AVG}$**	**PS$_\mathrm{MDN}$**	**SUMR$_\mathrm{F}$**	**SUMR$_\mathrm{T}$**	**SUMR$_\mathrm{F+T}$**
(4)	*stare at*	55	22	27	34	48	59	57
(3)	*dine on*	32	32	28	16	19	27	26
(2)	*plead with*	7	9	31	10	4	7	8
(1)	*bustle about*	6	4	5	5	6	13	5

Table (1): example phrasal verb ranks by each empirical measure.

3. Statistical Ranking Methods

Many statistical methods exist for collocation identification (as opposed ranking collocations that have already been identified as such). They are evaluated in binary terms; each collocation identified by an algorithm is compared to human evaluation and determined to be correct if it matches. Collocation identification is primarily based on statistical association measures. They extract and quantify relationships between words based on relative frequencies in a corpus. Rather than deciding if a word pair is or is not a collocation by using a threshold on an association measure, we propose using values of association measures directly as approximate ranks for predictability of the collocation. We implement several common association measures and correlate each with our empirical measures of collocation predictability.

One very common association measure is mutual information (MI), which expresses the information that presence of one word gives about the likelihood of another word occurring [9]. For two words, *x* and *y*, MI is calculated as follows:

$$\mathrm{MI}(x,y) := \log_2 \frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\mathrm{P}(y)} \qquad (1)$$

Another measure conceptually similar to MI is the log odds ratio:

$$\mathrm{LOR}(x,y) := \log \frac{\mathrm{P}(y|x)}{\mathrm{P}(y|\neg x)} \qquad (2)$$

In this case a word pair is scored as a collocation according to the ratio of (a), the probability of the word *y* occurring given word *x*, to (b), the probability of word *y* occurring given the absence of word *x* [10]. If any of these probabilities are zero the log-odds ratio will be negative infinity. To avoid this, discounting can be used. Evert [3] discounts by calculating a modified probability function, $\mathrm{P}_{\mathrm{DISC}}$, as if a ½ count were added to all zero and non-zero event counts from the corpus.

$$\mathrm{LOD}(x,y) = \mathrm{LOR}_{\mathrm{DISC}}(x,y) := \log \frac{\mathrm{P}_{\mathrm{DISC}}(y|x)}{\mathrm{P}_{\mathrm{DISC}}(y|\neg x)} \qquad (3)$$

Synonymy relations from WordNet [11] are used by Pearce [12] to develop a heuristic measure of collocation strength. A collocation is a combination of "lexical realizations" of two concepts that is significantly more frequent than other possible realizations for the same concepts. Sets of synonyms ("SynSets") from WordNet make it possible to count the frequency of these alternate lexical realizations. Pearce's method uses a comparison of raw frequency of a collocation and raw frequencies of synonym substitutions. We modify this method to allow investigation of other association measures. Formally, we represent WordNet as a function that, given a word *x*, returns a set *ß* of SynSets for that word:

$$\mathrm{WordNet}(x) = ß_x = \{S_{x1}, S_{x2}, \ldots, S_{xn}\} \qquad (7a)$$

Each SynSet *S* is composed of synonyms of *x* grouped by concept:

$$\forall S_x \in ß_x : S = \{s_{x1}, s_{x2}, \ldots, s_{xn}\} \qquad (7b)$$

Given a synset *S*, a second word *y*, and a bigram association measure *f*, we define the association measure for *S* as:

$$f(S_x, y) := \frac{\sum_{s_x \in S_x} f(s_x, y)}{|S_x|} \qquad (7c)$$

We can now define two similar synonym-based heuristics, *SynSets$_{MAX}$(f,x,y)* and *SynSets$_{AVG}$(f,x,y)* that rank collocations inversely to the association measure of synonym-substituted word-pairs - the stronger the associations of *x*'s synonyms with *y*, the more predictable *x* is likely to be. [12]

$$\mathrm{SS}_{\mathrm{MAX}}(f,x,y) := -\underset{S_x \in ß_x}{\mathrm{ARGMAX}} f(S_x, y) \qquad (8)$$

$$\mathrm{SS}_{\mathrm{AVG}}(f,x,y) := \frac{-\sum_{S_x \in ß_x} f(S_x, y)}{|ß_x|} \qquad (9)$$

In addition to association measures applied both to collocations directly and to their SynSets, we also implement ranking according to raw frequency of the base of the collocation (the verb), and raw frequency of the collocation itself:

$$\mathrm{RFV} := \mathrm{Count}(W_1); \mathrm{RFC} := \mathrm{Count}(W_{1,} W_2) \qquad (10;11)$$

As with the empirical measures, we demonstrate the ranking output of statistical measures with examples. $\mathrm{SS}_{\mathrm{MAX}}$ and $\mathrm{SS}_{\mathrm{AVG}}$ did not differ greatly for any association measure; from hereon we omit discussion of $\mathrm{SS}_{\mathrm{AVG}}$ rankings for brevity.

Category	**Example**	**MI**	**LOR**	**LOR$_{\mathrm{DISC}}$**	**RFV**	**RFC**	**SS$_{\mathrm{MAX}}$(MI)**	**SS$_{\mathrm{MAX}}$(LOR)**	**SS$_{\mathrm{MAX}}$(LOR$_{\mathrm{DISC}}$)**
(4)	*stare at*	27	42	40	40	49	43	17	46
(3)	*dine on*	20	28	39	35	38	10	42	9
(2)	*plead with*	12	39	22	45	39	8	38	8
(1)	*bustle about*	5	3	4	25	1	9	41	22

Table (2): Examples of approximate phrasal verb rankings based on each statistical measure.

4. Results

We proposed a definition of instructional utility of collocations in terms of their semantic predictability, and conducted an experiment to obtain empirical rankings of phrasal verbs according to candidate measures of difficulty and predictability. To judge if statistical methods can approximate these rankings, we calculated rankings of the same 60 phrasal verbs using each statistical measure on the Gigaword corpus. We then made pair-wise comparisons of rankings produced by empirical measures (rows in Table 3) to each corpus-based statistical method (columns in Table 3). The value in each cell of Table (3) is the Pearson Correlation Coefficient (PCC) for that pair of ranked lists. The statistical significance of each PCC value was calculated as a two-tailed p-value [13]. P-values of less than 0.05 are statistically significant, and significant pairs are shown in bold.

	PC	**PS$_{\mathrm{MIN}}$**	**PS$_{\mathrm{AVG}}$**	**PS$_{\mathrm{MDN}}$**	**SUM R$_{\mathrm{F}}$**	**SUM R$_{\mathrm{T}}$**	**SUM R$_{\mathrm{F+T}}$**
MI	0.0552	-0.044	-0.052	-0.150	0.2427	0.0136	0.1197
LOR	**0.3100**	0.1275	0.2032	0.2475	0.0885	**0.4772**	**0.3535**
LOR DISC	0.0567	-0.092	-0.115	-0.124	0.2028	-0.094	0.0142
RFV	**0.2861**	0.1729	0.2327	**0.2935**	-0.008	**0.4533**	**0.2976**
RFC	**0.2930**	0.0142	0.0725	0.0959	0.1796	**0.3740**	**0.3149**
SS$_{\mathrm{MAX}}$ (MI)	0.1491	0.0190	0.1693	0.1921	0.0877	0.1656	0.1667
SS$_{\mathrm{MAX}}$ (LOR)	0.0462	0.0943	0.0044	0.0627	0.0009	0.0404	0.0178
SS$_{\mathrm{MAX}}$ (LOR DISC)	**0.2784**	0.0338	0.1692	0.2247	0.2273	0.2202	**0.2584**

Table (3): PCC for comparison of phrasal verb rankings based on empirical measures (colums) and statistical approximations (rows).

The strongest correlations were between the inverse sum of rankings for correct responses ($SUMR_T$) and log-odds ratio (LOR), raw frequency of the verb (RFV), and raw frequency of the collocation (RFC), in that order with correlation coefficients (p-values) of 0.4772 (0.0001), 0.4533 (0.0032), and 0.3740 (0.0003) respectively. The combination of $SUMR_F$ and the positive sum of ratings for incorrect responses ($SUMR_{T+F}$) also correlated significantly with these statistical measures, as did the proportion of correct responses (PC). PC and $SUMR_{T+F}$ were the only empirical measures to correlate with a WordNet-based ranking which was $SS_{MAX}(LOR_{DISC})$. The correlation of PC with the discounted log-odds ratio (LOR_{DISC}) was the highest and most statistically significant correlation made without using the verbal rating data. No statistically significant correlations were found for any rankings based on mutual information (MI).

5. Discussion

Correlations of approximate and empirical rankings of phrasal verb predictability reveal that the frequency of the verb is the best statistical feature for approximation. The log-odds ratio is directly related to verb frequency, and collocation frequency is also indirectly influenced by it, indicating that phrasal verbs with frequent verb constituents have a strong tendency to be more frequent themselves. This suggests that the predictability of a collocation for language learners cannot be separated from the knowledge of each of its constituents. If the best statistical estimate of a student's familiarity with a verb is its frequency, and if predictability of a collocation is directly related to knowledge of its constituents, then it follows that measures influenced by the frequency of a verb will be the best indicators of phrasal verb predictability.

The failure of mutual information to correlate significantly with any measure of the experimental results is notable. The success of ranking based on the raw frequency of the verb suggests the cause of this failure. For binary collocation identification tasks, MI is favored as less sensitive to the raw frequencies of constituent words. We have found that approximation of phrasal verb predictability for language learners requires sensitivity to these frequencies.

For the practical scope of this study we chose one definition for the utility of teaching a collocation; predictability. We found that several methods of quantifying this can be approximated statistically. We did not select *a priori* which empirical measure best represents predictability. Future work will compare these measures to experts' qualitative assessments.

The measures that we found possible to approximate statistically are strongly influenced by verb frequency. Our results show that students are more likely to both know and predict very frequent collocations, but explicit instruction of only the rarest collocations in language is not necessarily most useful. Broader pedagogical goals may suggest additional criteria for utility. The application of corpus-based statistical methods for collocation selection may then be meaningfully applied when selecting content *within* categories defined by other criteria such as task and student model.

6. Conclusion

We have developed corpus-based ranking methods for one form of collocation, phrasal verbs, that significantly approximate empirical measures of collocation difficulty and predictability. These methods can be integrated as features in CALL systems, teaching resources, and automatic content selection.

We collected empirical data on the difficulty and the predictability of 60 phrasal verbs. Several measures were used to rank these collocations according to distinct patterns in the data. We then implemented common statistical association measures and applied them to the same set of phrasal verbs, using the English Gigaword corpus. Pairwise correlations revealed that statistical measures can be used to significantly approximate the rankings of empirical measures. The strongest correlations involved statistical measures strongly influenced by the frequency of the verb constituent of each phrasal verb. This is evidence that predictability of a collocation for language learners is directly proportional to knowledge of its constituents, and frequency is the best general estimator for familiarity with a word. Defining instructional utility as semantic predictability, then, could lead to an over-emphasis on rare collocations. Accordingly, statistical approximations of predictability will be best applied in conjunction with other task-based criteria.

We have shown that automated, corpus-based methods can be meaningfully applied to the task of developing ranked content resources for instruction of phrasal verbs. Future work will concern the use of machine learning approaches to combine these and other sources of information for the integration of other criteria in automatic rankings of collocations. We can use these improved rankings in the development of CALL applications. Furthermore, given more data these methods may be extended to create ranked instructional resources of other collocation types and of complex lexical phrases.

7. Acknowledgements

We would like to thank Alan Black for his advice. This project is supported by REAP-PT, and funded by the Carnegie Mellon | Portugal Information and Communication Technologies Institute.

8. References

[1] K.R. Mckeown and D.R. Radev, "Collocations," A Handbook of Natural Language Processing, R. Dale, H. Moisl, and H. Somers, CRC Press, 2000, pp. 507-523.

[2] J. Nattinger and J. DeCarrico, Lexical Phrases and Language Teaching, Oxford, UK: Oxford University Press, 1992.

[3] S. Evert, "The Statistics of Word Cooccurrences Word Pairs and Collocations," 2005.

[4] J. Hill, "Revising priorities: from grammatical failure to collocational success," Teaching Collocation, M. Lewis, Boston, MA: Thomson Heinle, 2000, pp. 47-69.

[5] J. Conzett, "Integrating collocation into a reading and writing course," Teaching Collocation, M. Lewis, Boston, MA: Thomson Heinle, 2000, pp. 70-87.

[6] G. Woolard, "Collocation - Encouraging Learner Independence," Teaching Collocation, M. Lewis, Boston, MA: Thomson Heinle, 2000, pp. 28-46.

[7] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda, English Gigaword Fourth Edition, Philadelphia: Linguistic Data Consortium, 2009.

[8] M. Heilman and M. Eskenazi, "Self-Assessment in Vocabulary Tutoring," Ninth International Conference on Intelligent Tutoring Systems, 2008.

[9] K.W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," Computational Linguistics, vol. 16, 1990, pp. 22-29.

[10] D. Pearce, "A comparative evaluation of collocation extraction techniques," Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), 2002, p. 1530–1536.

[11] WordNet, Cambridge, Massachusetts: MIT Press, 1998.

[12] D. Pearce, "Synonymy in Collocation Extraction," Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, 2001.

[13] C.D. Manning and H. Schuetze, Foundations of Statistical Natural Language Processing, Cambridge, Massachusetts: The MIT Press, 1999.