# Combination of Symbolic and Statistical Approaches for Grammatical Knowledge Acquisition

**Masaki KIYONO*** and **Jun'ichi TSUJII**
Centre for Computational Linguistics
University of Manchester Institute of Science and Technology
PO Box 88, Manchester, M60 1QD, United Kingdom
kiyono@ccl.umist.ac.uk, tsujii@ccl.umist.ac.uk

## Abstract

The framework we adopted for customizing linguistic knowledge to individual application domains is an integration of symbolic and statistical approaches. In order to acquire domain specific knowledge, we have previously proposed a rule-based mechanism to hypothesize missing knowledge from partial parsing results of unsuccessfully parsed sentences. In this paper, we focus on the statistical process which selects plausible knowledge from a set of hypotheses generated from the whole corpus. In particular, we introduce two statistical measures of hypotheses, *Local Plausibility* and *Global Plausibility*, and describe how these measures are determined iteratively. The proposed method will be incorporated into the tool kit for linguistic knowledge acquisition which we are now developing.

## 1 Introduction

Current technologies in natural language processing are not so mature as to make general purpose systems applicable to any domains; therefore rapid customization of linguistic knowledge to the sublanguage of an application domain is vital for the development of practical systems. In the currently working systems, such customization has been carried out manually by linguists or lexicographers with time-consuming effort.

We have already proposed a mechanism which acquires sublanguage-specific linguistic knowledge from parsing failures and which can be used as a tool for linguistic knowledge customization (Kiyono and Tsujii, 1993; Kiyono and Tsujii, 1994). Our approach is characterized by a mixture of symbolic and statistical approaches to grammatical knowledge acquisition. Unlike probabilistic parsing, proposed by (Fujisaki et al., 1989; Briscoe and Carroll, 1993), which assumes the prior existence of comprehensive linguistic knowledge, our system can suggest new pieces of knowledge including CFG rules, subcategorization frames, and other lexical features. It also differs from previous proposals on lexical acquisition using statistical measures such as (Church et al., 1991; Brent, 1991; Brown et al., 1993) which either deny the prior existence of linguistic knowledge or use linguistic knowledge in ad hoc ways.

Our system consists of two components: (1) the rule-based component, which detects incompleteness of the existing knowledge and generates a set of hypotheses of new knowledge and (2) the corpus-based component which selects plausible hypotheses on the basis of their statistical behaviour. As the rule-based component has been explained in our previous papers, in this paper we focus on the corpus-based component.

After giving a brief explanation of the framework, we describe a data structure called *Hypothesis Graph* which plays a crucial role in the corpus-based process, and then introduce two statistical measures of hypotheses, *Global Plausibility* and *Local Plausibility*, which are iteratively determined to select a set of plausible hypotheses. An experiment which shows the effectiveness of our method is also given.

## 2 The System Organization

### 2.1 Hypothesis Generation

Figure 1 shows the framework of our system. When the parser fails to analyse a sentence, the Hypothesis Generator (HG) produces hypotheses of missing knowledge each of which could rectify the defects of the current grammar. As the parser is a sort of *Chart Parser* and maintains partial parsing results in the form of inactive and active edges, a parsing failure means that no inactive edge of category $S$ spanning the whole sentence exists.

The HG tries to introduce an inactive edge of $S$ by making hypotheses of missing linguistic knowledge. It generates hypotheses of rewriting rules which collect existing sequences of inactive edges into an expected category. It also calls itself recursively to in-

*also a staff member of Matsushita Electric Industrial Co.,Ltd., Shinagawa, Tokyo, JAPAN.

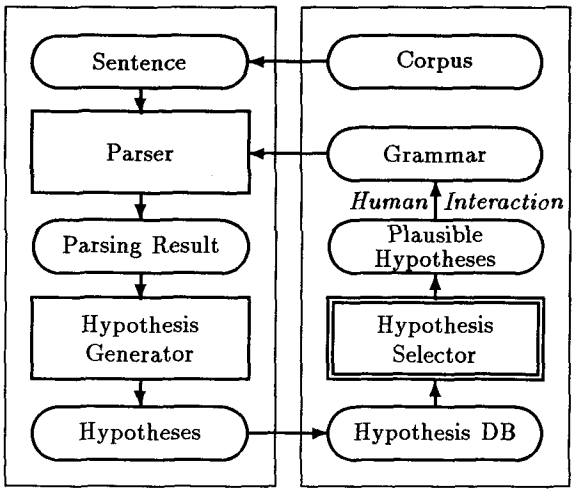


Figure 1: Framework of Grammar Acquisition

troduce necessary inactive edges for each rule of the expected category whose application is prevented due to the lack of necessary inactive edges. The simplest form of the algorithm is shown below.

**[Algorithm]** An inactive edge $[ie(A) : x_0, x_n]$ can be introduced, with label $A$, between word positions $x_0$ and $x_n$ by each of the hypotheses generated from the following two steps.

**[Step 1]** For each sequence of inactive edges, $[ie(B_1) : x_0, x_1]$, ..., $[ie(B_n) : x_{n-1}, x_n]$, spanning from $x_0$ to $x_n$, generates a new rule.

$$A \Rightarrow B_1, \cdots, B_n$$

**[Step 2]** For each existing rule of form $A \Rightarrow A_1, \cdots, A_n$, finds an incomplete sequence of inactive edges, $[ie(A_1) : x_0, x_1]$, ..., $[ie(A_{i-1}) : x_{i-2}, x_{i-1}]$, $[ie(A_{i+1}) : x_i, x_{i+1}]$, ..., $[ie(A_n) : x_{n-1}, x_n]$, and calls this algorithm for $[ie(A_i) : x_{i-1}, x_i]$.

This algorithm has been further augmented in order to treat sentences which contain more than one construction not covered by the current version of the grammar and to generate hypotheses concerning complex features like subcategorization frames.

## 2.2 Hypothesis Filtering

The greater number of the hypotheses generated by the algorithm are linguistically unnatural, because the algorithm does not embody any linguistic principle to judge the appropriateness of hypotheses, and therefore we introduced a set of criteria to filter out unnatural hypotheses (Kiyono and Tsujii, 1993; Kiyono and Tsujii, 1994). This includes, for example,

- The maximum number of daughter constituents of a rule is set to 3.
- Supposing that the current version of the grammar contains all the category conversion rules, a unary rule with one daughter constituent is not generated.
- Using generalizations embodied in the current version of the grammar, a rule containing a sequence of constituents which can be collected into a larger constituent by the current version of grammar is not generated.
- Distinguishing non-lexical categories from lexical categories, a rule whose mother category is a lexical category is not generated.

These criteria significantly reduce the number of hypotheses to be generated.

## 2.3 Hypothesis Graph

As the criteria which the HG uses to filter out unnatural hypotheses are solely based on the forms of hypotheses, they cannot identify the "correct" hypotheses on their own. The correct ones are rather chosen by the Hypothesis Selector (HS), which resorts to examining the statistical behaviour of hypotheses throughout a given corpus.

A straightforward method is to count the frequency of hypotheses, but this simple method does not work, because hypotheses are not independent of each other. A hypothesis is either *competing* with or *complementary* to other hypotheses generated from the same sentence. A group of hypotheses generated for restoring the same inactive edge constitutes a set of competing hypotheses and only one of them contributes to the correct structure of the sentence. On the other hand, two groups of hypotheses which are generated to treat two different parts of the same sentence stand in complementary relationships.

A hypothesis should be recognized as being correct, only when no other competing hypothesis is more plausible. That is, even if a hypothesis is generated frequently, it should not be chosen as the correct one, if more plausible competing hypotheses are always generated together with it. On the other hand, even if a hypothesis is generated only once, it should be chosen as the correct one, if there is no other competing hypothesis.

In order to realize the above conception, the HS maintains mutual relationships among hypotheses as an AND-OR graph. In a graph, AND nodes and OR nodes express complementary relationships and competing relationships, respectively. A node is shared, when different recursion steps in the HG try to restore the same inactive edge. Figure 2 shows the AND-OR graph for the hypotheses generated from the sentence "Failing students looked embarrassed" when the current version of grammar does not contain rules for participles. The top node is an AND node which has two groups of hypotheses that treat two different parts of the sentence, i.e. "failing students" and "looked embarrased".

Sentence: Failing students looked embarrassed.

HP1: $NP \Rightarrow VP, NP$ ("failing students")
HP2: $ADJ \Rightarrow [failing]$
HP3: $VP \Rightarrow VP, VP$ ("looked embarrassed")
HP4: $ADV \Rightarrow [embarrassed]$
HP5: $N \Rightarrow [embarrassed]$
HP6: $ADJ \Rightarrow [embarrassed]$

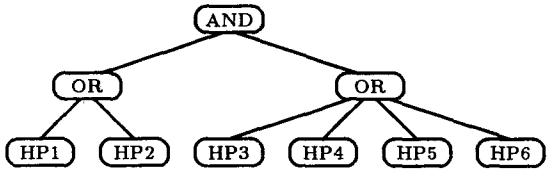



Figure 2: AND-OR Graph of Hypotheses

## 3 Statistical Analysis

### 3.1 Two Measures of Plausibility

The HS uses two measures of plausibility of hypotheses. One is computed for an instance hypothesis and the other is for a generic hypothesis. (See 3.3 for the relationship between the two types of hypotheses.)

**(1) Local Plausibility:** This value shows how plausible an instance hypothesis is as grammatical knowledge to contribute to the correct analysis of a unsuccessfully parsed sentence.

**(2) Global Plausibility:** This value shows how plausible the hypothesis of the generic form is as grammatical knowledge to be acquired.

As we describe in the following section, the Local Plausibility (LP) of an instance hypothesis is computed on the basis of the values of the Global Plausibility (GP) of the generic hyoptheses which are linked to instance hypotheses in the same hypothesis graph. On the other hand, the GP of a generic hypothesis is computed from the LP values of its instance hypotheses across the whole corpus.

Intuitively speaking, the GP of a generic hypothesis is high if its instances are frequently generated and if they receive high LP values, while the LP of a instance hypothesis is high if the GP of the corresponding generic hypothesis is high and if the GP values of the generic hypotheses corresponding to its competing hypotheses are low. Because of this mutual dependence between LP and GP, they cannot be computed in a single step but rather computed iteratively by repeating the following steps until the halt condition is satisfied.

**[Step 1]** Estimates the initial values of LP.

**[Step 2]** Calculates GP values from LP values.

**[Step 3]** Checks the halt condition.

**[Step 4]** Calculates LP values from GP values and `GOTO` **[Step 2]**.

### 3.2 Initial Estimation of Local Plausibility

If the current version of the grammar is reasonably comprehensive, pieces of linguistic knowledge which have to be acquired are likely to be lexical or idiosyncratic. That is, we assume that sublanguage-specificity tends to be manifested by unknown words, new usages of existing words, and syntactic constructions idiosyncratic to the sublanguage. In order to quantify such plausibility, the following value is given to each hypothesis.

$$LP(Hypo_i) = 1 - \frac{W(Hypo_i) \times H(Hypo_i)}{W(S) \times H(S)}$$

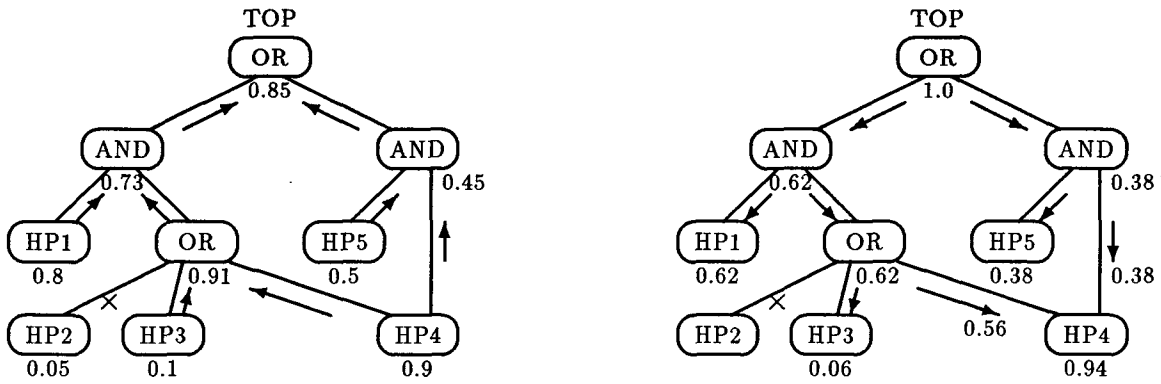
This value shows the proportion of the syntactic structure in the whole sentence which is not covered by the hypothesis. It ranges from 0 to 1 and gets larger if the hypothesis rectifies a smaller part of the sentence. $W(Hypo_i)$, the width of the hypothesis, is defined as the word count of the subtree and $H(Hypo_i)$, the height, is defined as the shortest path from lexical nodes to the top node of the subtree.

### 3.3 Generic Hypothesis and Global Plausibility

The GP of a hypothesis is computed based on the LP values of its instance hypotheses, but the relationship between a generic hypothesis and its instances is not straightforward because we adopted a unification-based grammar formalism. For example, the instance hypothesis of $NP \Rightarrow VP, NP$ in Figure 2 contains not only this CFG skeleton but also further feature descriptions of the three constituents which include specific surface words like "failing" and "students". Unless we generalize them, we cannot obtain the generic form of this instance hypothesis, and therefore cannot judge whether the hypotheses generated from different sentences are identical.

Such generalization of instance hypotheses requires an *inductive* mechanism for judging which parts of the feature specification are common to all instance hypotheses and should be included in a hypothesis of the generic form. This kind of induction is beyond the scope of the current framework, because such induction may need a lot of time and space if it is carried out from scratch. We first gather a set of instance hypotheses which are likely to be instances of the same generic hypothesis which, in turn, is likely to be "correct" linguistic knowledge.

Our current framework uses a simple definition of generic hypotheses and their instances. That is, if two rule hypotheses have the same CFG skeleton, then they are judged to be instances of the same generic hypotheses. As for lexical hypotheses, we use a set of fixed templates of lexical entries in order to acquire detailed knowledge like subcategorization frames. Features which are not included in the templates are ignored in the judgement of whether generic hypotheses are identical.




(a) Collection of Global Plausibility

(b) Distribution of Local Plausibility

Figure 3: Calculation of Local Plausibility

The GP of a generic hypotheses is defined as being the probability of the event that at least one instance hypothesis recovers the true cause of a parsing failure, and it is computed by the following formula when a set of its instance hypotheses is identified. In the formula, $HP$ is a generic hypothesis and $HP_i$ are its instances.

$$GP(HP) = 1 - \prod_{i=1}^{n}(1 - LP(HP_i))$$

The more instance hypotheses are generated, the closer to 1 $GP(HP)$ becomes. If one of the instances is regarded to be recovering the true cause of a parsing failure, the GP of the generic hypothesis is assigned 1, because the hypothesis is indispensable to the analysis of the corpus.

### 3.4 Local Plausibility

The calculation of LP is carried out on each hypothesis graph based on the assumption that an instance hypothesis or a set of instance hypotheses which recovers the true cause(s) of the parsing failure should exist in the graph. This assumption means that the top node of a hypothesis graph is assigned 1 as its LP value.

The LP value assigned to a node is to be distributed to its daughter nodes by considering the GP values of the corresponding generic hypotheses. For example, the daughter nodes of an OR node, which constitute a set of competing hypotheses, receive their LP values which are dividents of the LP value of the mother node proportional to their GP values.

However, as GP is defined only for hypotheses, we first determine the GP values of all nodes in a hypothesis graph in a bottom-up manner, starting from the tip nodes of the graph to which instance hypotheses are attached. Therefore, **[Step 2]** in the statistical analysis is further divided into the following three steps.

**[Step 2-1] Bottom-up Calculation of GP**

The GP value of an intermediate node is determined as follows (See Figure 3(a)).

- The GP value of an OR node is computed by the following formula based on the GP values of the daughter nodes, which corresponds to the probability that at least one of the daughter nodes represents "correct" grammatical knowledge.

$$GP(OR) = 1 - \prod_{i=1}^{m}(1 - GP(Node_i))$$

- The GP value of an AND node is computed by the following formula, which corresponds to the probability that all the daughter nodes represent "correct" grammatical knowledge.

$$GP(AND) = \prod_{i=1}^{m} GP(Node_i)$$

**[Step 2-2] Deletion of Hypotheses**

The nodes which have significantly smaller GP values than the highest one among the daughter nodes of the same mother OR node (less than one tenth, in our current implementation) will be removed from the hypothesis graph. For example, *HP2* in Figure 3 was considered to be much less plausible than *HP4* and removed from the graph.

As a node in a hypothesis graph could have more than one mother nodes, the hypothesis deletion is realized by removing the link between the node representing the hypothesis and one of its mother OR nodes (not removing the node itself). For example, in Figure 3, when *HP4* is removed in comparison with *HP2* or *HP3*, the link between *HP4* and the OR node is removed, while the link between *HP4* and the AND node still remains.

The deletion of less viable nodes accelerates the convergence of the iterative process of computing GP and LP.

**[Step 2-3] Top-down Calculation of LP**

This step distributes the LP assigned to the top node (that is, 1) to the nodes below in a top-down way according to the following rules (See Figure 3(b)).

- The LP value of an OR node is distributed to its daughter nodes proportional to their GP values so that the sum of their LP values is the same as that of the OR node because the daughter nodes of the same OR node represent mutually exclusive hypotheses.

$$LP(Node_i) = \frac{GP(Node_i)}{\sum_{j=1}^{m} GP(Node_j)} LP(OR)$$

- The LP value of an AND node is distributed to its daughter nodes with the same values.

$$LP(Node_i) = LP(AND)$$

If a hypothesis has more than one mother nodes and its LP can be calculated through several paths, the sum of those is given to the hypothesis. For example, the value for *HP4* in Figure 3 is $0.56 + 0.38 = 0.94$.

As we discussed before, these newly computed LP values are used to compute the GP values at **[Step 2]** in the next cycle of iteration.

### 3.5 Halt Condition

The iterative calculation process is regarded to have converged if the GP values of all the generic hypotheses do not change in comparison with the previous cycle, but as it possibly takes a lot of time for the process to reach such a situation, we use an easier condition to stop the process. That is, we count the number of deleted instance hypotheses at each cycle and terminate the iteration when no instance hypothesis is deleted in a number of consecutive iterations. Actually, the process halts after 5 zero-deletion cycles in our current implementation.

When the interative process terminates, the hypotheses with high GP values are presented as the final candidates of new knowledge to be added to the current version of grammar.

## 4 Preliminary Experiment

In order to demonstrate how the HS works, we carried out a preliminary experiment with 1,000 sentences in the UNIX on-line manual (approximately one fifth of the whole manual). As the initial knowledge for the experiment, we prepared a grammar set which contains 120 rules covering English basic expressions and deliberately removed rules for participles in order to check whether the HS can discover adequate rules. The input data to the statistical process is a set of 5,906 instance hypotheses generated from 282 unsuccessfully parsed sentences.

| # | GP | Generic Hypothesis | N |
|---|---|---|---|
| 1 | 1.000000 | `vp => vp,p.` | 20 |
| 1 | 1.000000 | `np => vp,np.` | 26 |
| 1 | 1.000000 | `n => ['double-quote'].` | 1 |
| 1 | 1.000000 | `n => [filename].` | 8 |
| 1 | 1.000000 | `v => [archived].` | 1 |
| 1 | 1.000000 | `n => [directory].` | 4 |
| 1 | 1.000000 | `n => ['EOF'].` | 2 |
| 1 | 1.000000 | `adj => ['non-printing'].` | 1 |
| 1 | 1.000000 | `n => [pathnames].` | 1 |
| 1 | 1.000000 | `n => [cpp].` | 3 |
| 1 | 1.000000 | `n => ['NEWLINE'].` | 3 |
| 1 | 1.000000 | `n => ['.cshrc'].` | 2 |
| 1 | 1.000000 | `n => [backslash].` | 2 |
| 1 | 1.000000 | `n => [aliases].` | 2 |
| 1 | 1.000000 | `adj => [nonseekable].` | 1 |
| 1 | 1.000000 | `n => [wordlist].` | 2 |
| 1 | 1.000000 | `n => [login].` | 3 |
| 138 | 0.925960 | `n => ['TERM'].` | 2 |
| 142 | 0.913933 | `n => [cmdtool].` | 1 |
| 173 | 0.750000 | `n => ['command-line'].` | 2 |
| 189 | 0.683594 | `n => [filenames].` | 1 |
| 232 | 0.500000 | `adj => [backquoted].` | 1 |
| 318 | 0.336694 | `np => np,np.` | 12 |
| 546 | 0.000000 | `adj => [blocking].` | 1 |
| 546 | 0.000000 | `adj => [invisible].` | 1 |

#: Rank, N: Number of instance hypotheses

Table 1: List of "Correct" Hypotheses

The statistical process removed 4,034 instance hypotheses and stopped after 63 cycles of the iterative computation of GP and LP. The instance hypotheses were grouped into 2,876 generic hypotheses and the GP values of 2,331 generic hypotheses were reduced to 0 by the hypothesis deletion.

Table 1 is the list of "correct" hypotheses picked up from the whole list of generic hypotheses sorted by GP values. The hypothesis for participles, `np => vp,np`, is one of the 128 hypotheses whose GP values are 1. This table also shows that quite a few "correct" lexical hypotheses are in higher positions because lexical knowledge for unknown words is indispensable to the successful parsing of the corpus.

The distribution of "correct" hypotheses within the whole list is shown in Table 2. The fact that "correct" hypotheses exist more in higher ranges supports our mechanism. Although some of the "correct" ones have zero GP values, they do not diminish our framework because most of them are the hypotheses treating participles as adjectives, which are the alternative hypotheses of `np => vp,np`.

The parameter which we can adjust to select more plausible hypotheses is the threshold for the hypothesis deletion. Generally speaking, giving a higher threshold causes an increase of the number of deleted hypotheses and therefore accelerates the convergence of the iterative process. In the experiment, however, the use of one fifth as the threshold instead of one tenth did not bring a major difference.

| Range | Rule Hypothesis | Lexical Hypothesis | Correct Hypothesis |
|---|---|---|---|
| 1- 100 | 19 | 81 | 35 |
| 101- 200 | 41 | 59 | 25 |
| 201- 300 | 55 | 45 | 13 |
| 301- 400 | 76 | 24 | 5 |
| 401- 500 | 95 | 5 | 0 |
| 501-1000 | 473 | 27 | 9 |
| 1001-2000 | 770 | 230 | 12 |
| 2001-2876 | 653 | 223 | 9 |
| Total | 2182 | 694 | 108 |

Table 2: Distribution of "Correct" Hypotheses

## 5 Conclusion

The statistical analysis discussed in this paper is based on the assumption that types of linguistic knowledge to be acquired are:

[1] Knowledge for syntactic constructions which is used frequently in the given sublanguage.

[2] Lexical knowledge such as subcategorization frames and number properties, which is often idiosyncratic to the given sublanguage.

[3] Knowledge which belongs neither to [1] nor to [2], but is indispensable to the given corpus.

[1] implies that knowledge for less frequent constructions can be ignored at the initial stage of linguistic knowledge customization. Such knowledge will be discovered after major defects of the current grammar are rectified, because the GP of a generic hypothesis is defined as being sensitive to the frequency of the hypothesis.

[2] means that we assume that the set of initially provided grammar rules has a comprehensive coverage of English basic expressions. This assumption is reflected in the way of the initial estimation of LP values. Also note that only when this assumption is satisfied, can the HG produce a reasonable set of hypotheses. On the other hand, because of this assumption, our framework can learn structurally complex and linguistically meaningful lexical descriptions, like a subcategorization frame.

[3] is reflected in the way of the computation of GP values. A generic hypothesis one of whose instances occurs as a single possible hypothesis that can recover a parsing failure will have the GP value of 1, even though its frequency is very low.

The computation mechanism of GP and LP bears a resemblance to the EM algorithm(Dempster et al., 1977; Brown et al., 1993), which iteratively computes maximum likelihood estimates from incomplete data. As the purpose of our statistical analysis is to choose "correct" hypotheses from a hypothesis set which contains unnatural hypotheses as well, our motivation is different from that of the EM algorithm. However, if we consider that the hypothesis deletion is maxmizing the plausibility of "correct" hypotheses, the computation procedures of both algorithms have a strong similarity.

The grammatical knowledge acquisition method proposed in this paper will be incorporated into the tool kit for linguistic knowledge customization which we are now developing. In the practical use of our method, a grammar maintainer will be shown a list of hypotheses with high GP values and renew the current version of grammatical knowledge. The renewed knowledge will be used in the next cycle of hypothesis generation and selection to achieve the gradual enlargement of linguistic knowledge.

## Acknowledgements

We would like to thank our colleagues in UMIST who gave us many usuful comments. We also want to thank Mr Tsumura and Dr Kawakami of Matsushita, who allowed the first author to study at UMIST.

## References

Michael R. Brent. 1991. Automatic Acquisition of Subcategorization Frames from Untagged Text. In *Proc. of the 29st ACL meeting*, pages 209–214.

Ted Briscoe and John Carroll. 1993. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. *Computational Linguistics*, 19(1):25–59.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using Statistics in Lexical Analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, chapter 6, pages 115–164. Lawrence Erlbaum Associates.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.

T. Fujisaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino. 1989. A Probabilistic Parsing Method for Sentence Disambiguation. In *Proc. of the Int. Workshop on Parsing Technologies*, pages 105–114. Carnegie-Mellon University.

Masaki Kiyono and Jun'ichi Tsujii. 1993. Linguistic Knowledge Acquisition from Parsing Failures. In *Proc. of EACL-93*, pages 222–231.

Masaki Kiyono and Jun'ichi Tsujii. 1994. Hypothesis Selection in Grammar Acquisition. In *Proc. of COLING-94*.