Descriptive and Empirical Approaches to Capturing Underlying Dependencies among Parsing Errors

Tadayoshi Hara1 Yusuke Miyao1 Jun’ichi Tsujii1,2,3

1Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, JAPAN
2School of Computer Science, University of Manchester
3NaCTeM (National Center for Text Mining)
{harasan,yusuke,tsujii}@is.s.u-tokyo.ac.jp

Abstract

In this paper, we provide descriptive and empirical approaches to effectively extracting underlying dependencies among parsing errors. In the descriptive approach, we define some combinations of error patterns and extract them from given errors. In the empirical approach, on the other hand, we re-parse a sentence with a target error corrected and observe errors corrected together. Experiments on an HPSG parser show that each of these approaches can clarify the dependencies among individual errors from each point of view. Moreover, the comparison between the results of the two approaches shows that combining these approaches can achieve a more detailed error analysis.

1 Introduction

For any kind of technology, analyzing causes of errors given by a system is a very helpful process for improving its performance. In recent sophisticated parsing technologies, the process has taken on more and more important roles since critical ideas for parsing performance have already been introduced and the researches are now focusing on exploring the rest of the pieces for making additional improvements.

In most cases for parsers’ error analysis, researchers associate output errors with failures in handling certain linguistic phenomena and attempt to avoid them by adding or modifying corresponding settings of their parsers. However, such an analysis cannot been done so smoothly since parsing errors sometimes depend on each other and the underlying dependencies behind superficial phenomena cannot be captured easily.

In this paper, we propose descriptive and empirical approaches to effective extraction of dependencies among parsing errors and engage in a deeper error analysis with them. In our descriptive approach, we define various combinations of error patterns as organized error phenomena on the basis of linguistic knowledge, and then extract such combinations from given errors. In our empirical approach, on the other hand, we re-parse a sentence under the condition where a target error is corrected, and errors which are additionally corrected are regarded as dependent errors. By capturing dependencies among parsing errors through systematic approaches, we can effectively collect errors which are related to the same linguistic properties.

In the experiments, we applied both of our approaches to an HPSG parser Enju (Miyao and Tsujii, 2005; Ninomiya et al., 2006), and then evaluated the obtained error classes. After examining the individual approaches, we explored the combination of them.

2 Parser and its evaluation

A parser is a system which interprets structures of given sentences from some grammatical or in some cases semantical viewpoints, and interpreted structures are utilized as essential information for various natural language tasks such as information extraction, machine translation, and so on. In most cases, an output structure of a parser is based on a certain grammatical framework such as CFG, CCG (Steedman, 2000), LFG (Kaplan and Bresnan, 1995) or HPSG (Pollard and Sag, 1994).

Since such a framework can usually produce more than one probable structure for a sentence, a parser
While various parsers take different manners in capturing linguistic phenomena based on their frameworks, they are at least required to obtain some kinds of relations between the words in sentences. On the basis of the requirements, a parser is usually evaluated on how correctly it gives intended linguistic relations. “Predicate argument relation” is one of the most common evaluation measurements for a parser since it is a very fundamental linguistic behavior and is less dependent on parser systems. This measure divides linguistic structural phenomena in a sentence into minimal predicative events. In one predicate argument relation, a word which represents an event (predicate) takes some words as participants (arguments). Although no fixed formulation exists for the relations, there are to a large extent common conceptions for them based on linguistic knowledge among researchers.

Figure 1 shows an example of predicate argument relations given by Enju. In the sentence “John has come,” “has” is a predicate of type “aux_arg12” and takes “John” and “come” as the first and second arguments. “come” is also a predicate of the type “verb_arg1” and takes “John” as the first and the only argument. In this formalism, each predicate type is represented as a combination of “the grammatical nature of a word” and “the arguments which it takes,” which are represented by the descriptions in Table 1. “aux_arg12” in Figure 1 indicates that it is an auxiliary word and takes two arguments “ARG1” and “ARG2.”

In order to improve the performance of a parser, analyzing parsing errors is very much worth the effort. Since the errors are output according to a given evaluation measurement such as “predicate argument relation,” we researchers carefully explore them and infer the linguistic phenomena which cause the erroneous outputs. Figure 2 shows an example of parsing errors for sentence “I watched the girl on TV.” Note that the errors are based on predicate argument relations as shown above and that the predicate types are abbreviated in this figure. When we focus on the error output, we can observe that “ARG1” of predicate “on” was mistaken by the parser. In this case, “ARG1” represents a modifiee of the preposition, and we then conclude that the ill attachment of a prepositional phrase caused this error. By continuing such error analysis, weak points of the parser are revealed and can be useful clues for further improvements.

However, in most researches on parsing technologies, error analysis has been limited to narrow and shallow explorations since there are various dependencies behind erroneous outputs. In Figure 3, for example, two errors were given: wrong outputs for “ARG1” of “which” and “ARG2” of “read.” Both of these two errors originated from the fact that the relative clause took a wrong antecedent “the shelf.” In this sentence, the former

Table 1: Descriptions for predicate types

| Abbr. | Full |
|-------|------|
| aux   | auxiliary |
| verb  | verb |
| prep  | prepositional |
| det   | determiner |
| adj   | adjunction |
| app   | apposition |
| relative | relative |
| lgs   | logical subject |
| coord | coordination |
| conj  | conjunction |
| _argN1,.. | take argument(s) |
| _mod  | modify a word |

Figure 1: Predicate argument relations

Figure 2: An example of parsing errors

Figure 3: Co-occurring parsing errors
“ARG1” directly corresponds to the antecedent while the latter “ARG2” indirectly referred to the same antecedent as the object of the verb “read.” The two predicate argument relations thus took the same word as their common arguments, and therefore the two errors co-occurred.

On the other hand, one-way inductive relations also exist among errors. In Figure 4, “ARG1” of “for” and “to” were mistaken by a parser. We can know that each of the errors was caused by an ill attachment of a prepositional phrase with the same analysis as shown in Figure 2. What is important in this example is the manner in their occurrences. The former error can appear by itself (Analysis 1) while the latter cannot because of the structural conflict with the former error (Analysis 2). The appearance of the latter error thus induces that of the former error. In error analysis, we have to correctly capture such various relations, which leads us to a costly and less rewarding analysis.

In order to make advancements on this problem, we propose two types of approaches to realizing a deeper error analysis on parsing. In the experiments, we examine our approaches for actual errors which are given by the HPSG parser Enju (Miyao and Tsujii, 2005; Ninomiya et al., 2006). Enju was developed for capturing detailed syntactic or semantic properties and relations for a sentence with an HPSG framework (Pollard and Sag, 1994). In this research, we focus on error analysis based on predicate argument relations, and in the experiments with Enju, utilize the relations which are represented in parsed tree structures.

3 Two approaches for error analysis

In this section, we propose two approaches for error analysis which enable us to capture underlying dependencies among parsing errors. Our descriptive approach matches the patterns of error combinations with given parsing errors and collects matched erroneous participants. Our empirical approach, on the other hand, detects co-occurring errors by re-parsing a sentence under a situation where each of the errors is forcibly corrected.

3.1 Descriptive approach

Our descriptive approach for capturing dependencies among parsing errors is to extract certain representative structures of errors and collect the errors which involve them. Parsing errors have a tendency to occur with certain patterns of structures representing linguistic phenomena. We first define such patterns through observations with a part of error outputs, and then match them with the rest.

Table 2 summarizes the patterns for erroneous phenomena which we defined for matching in the experiments. In the table, the patterns for 14 phenomena are given and classified into four types according to their matching manners. Each of the patterns for “Argument selection” examine whether a focused argument for a certain predicate type is erroneous or not. Figure 5 shows the pattern for “Prepositional attachment,” which col-

![Figure 4: Sketch of error propagation](image-url)
They completed the sale of it to him for $1,000.

Figure 5: Pattern for “Prepositional attachment”

Figure 6: Pattern for “Gerund acts as modifier or not”

Figure 7: Pattern for “To-infinitive for modifier/argument of verb”

Figure 8: Pattern for “Subject for passive sentence or not”

selects wrong ARG1 for predicate type “prep_arg”. From the sentence in the figure, we can obtain two errors for “Prepositional attachment” around prepositions “to” and “for.” On the other hand, each “Predicate type selection” pattern collects errors around a word whose predicate type is erroneous. Figure 6 shows the pattern for “Gerund acts as modifier or not,” which collects errors around gerunds whose predicate types are erroneous. From the example sentence in the figure, we can obtain an erroneous predicate type for “expecting” and collect errors around it for “Gerund acts as modifier or not.”

We can implement more structural errors than simple argument or predicate type selections. Figures 7 and 8 show the patterns for “To-infinitive for modifier/argument of verb” and “Subject for passive sentence or not” respectively. The pattern for the latter phenomenon collects errors on recognitions of prepositional phrases which behave as subjects for passive expressions. The pattern collects errors not only around prepositions but also around the verbs which take the prepositional phrases as a subject.

Since these patterns are based on linguistic knowledge given by a human, the process could provide a relatively precise analysis with a lower cost than a totally manual analysis.

3.2 Empirical approach

Our empirical approach, on the other hand, briefly traces the parsing process which results in each of the target errors. We collect co-occurring errors as strongly relevant ones, and then extract dependencies among the obtained groups. Parsing errors could originate from wrong processing at certain stages in the parsing, and errors with a common origin would by necessity appear together. We re-parse a target sentence under the condition where a certain error is forcibly corrected and then collect errors which are corrected together as the “relative” ones. An error group where all errors are relative to each other can be regarded as a “co-occurring error group.” Errors in the same co-
Figure 9: Pattern for “Relative clause attachment”

Example:
Correct answer:
The book on the shelf which: relative_arg1 I read yesterday

Parser output:
Re-parse a sentence under the condition where each error is forcibly corrected

Correct Error 1 Error 2 Error 3 Error 4 corrected together
Correct Error 2 Error 1 Error 3 Error 4 corrected together
Correct Error 3 Error 1 Error 2 Error 4 corrected together
Correct Error 4 Error 1 Error 2 Error 5 corrected together
Correct Error 5 Error 1 Error 2 Error 3 Error 4 corrected together

Extract co-occurring error groups and inductive relations

Co-occurring error group Co-occurring error group
Error 1 Error 2 Error 3 Error 4

Figure 10: An image of our empirical approach

Co-occurring error group

Error 5

Correct answer:
It has no bearing on our work force today

Parser output:
It has no bearing on our work force today

Re-parse a sentence under the condition where each error is forcibly corrected

Correct Error 1 Error 2 Error 3 Error 4 Error 5 corrected together
Correct Error 2 Error 1 Error 3 Error 4 Error 5 corrected together
Correct Error 3 Error 1 Error 2 Error 4 Error 5 corrected together
Correct Error 4 Error 1 Error 2 Error 3 Error 5 corrected together
Correct Error 5 Error 1 Error 2 Error 3 Error 4 corrected together

Co-occurring error group

Extract co-occurring error groups and inductive relations

Error 5

Table 3: Errors extracted with descriptive analysis

| Error types                                | # of Errors | Patterns |
|--------------------------------------------|-------------|----------|
| Analyzed                                   | 2,078       | 1,671    |
| [Argument selection]                      |             |          |
| Prepositional attachment                   | 579         |          |
| Adjunction attachment                      | 261         |          |
| Conjunction attachment                     | 43          | 40       |
| Head selection for noun phrase             | 30          | 30       |
| Coordination                               | 202         | 184      |
| [Predicate type selection]                 |             |          |
| Preposition/Adjunction                     | 108         | 54       |
| Gerund acts as modifier or not             | 84          | 31       |
| Coordination/conjunction                    | 54          | 27       |
| # of arguments for preposition             | 51          | 17       |
| Adjunction/adjunctive noun                 | 13          | 13       |
| [More structural errors]                   |             |          |
| To-infinitive for modifier/argument of verb| 120         | 22       |
| Subject for passive sentence or not        | 8           | 3        |
| [Others]                                   |             |          |
| Comma                                      | 444         | 372      |
| Relative clause attachment                 | 102         | 38       |
| Unanalyzed                                 | 2,631       |          |
| Total                                      | 4,709       |          |

inductive relation is given from error 5 to the co-occurring error group. We can then finally obtain the inductive relations as shown at the bottom of Figure 10. This approach can trace the actual behavior of the parser precisely, and can therefore capture underlying dependencies which cannot be found only by observing error outputs.

4 Experiments

We applied our approaches to parsing errors given by the HPSG parser Enju, which was trained on the Penn Treebank (Marcus et al., 1994) section 2-21. We first examined each approach, and then explored the combination of the approaches.

4.1 Evaluation of descriptive approach

We examined our descriptive approach. We first parsed sentences in the Penn Treebank section 22 with Enju, and then observed the errors. Based on the observation, we next described the patterns as shown in Section 3. After that, we parsed section 0 and then applied the patterns to the errors.

Table 3 summarizes the extracted errors. As the table shows, with the 14 error patterns, we successfully matched 1,671 locations in error outputs and covered 2,078 of 4,709 errors, which comprised of more than 40% of the total errors. This was the first step of the application of our approach, and in the future work we would like to
add more patterns for capturing more phenomena.

When we focused on individual patterns, we could observe that the simple error phenomena such as the attachments were dominant. The first reason for this would be that such phenomena were among minimal linguistic events. This would make the phenomena components of other more complex ones. The second reason for the dominance would be that the patterns for these error phenomena were easy to implement only with argument inconsistencies, and only one or a few patterns could cover every probable error. Among these dominant error types, the number of prepositional attachments was outstanding. The error types which required matching with predicate types were fewer than the attachment errors since the limited patterns on the predicate types would narrow the possible linguistic behavior of the candidate words. When we focus on more structural errors, the table shows that the rates of the participant errors to matched locations were much larger than those for simpler pattern errors. Once our patterns matches, they could collect many errors at the same time.

4.2 Evaluation of empirical approach

Next, we applied our empirical approach in the same settings as in the previous section. We first parsed sentences in section 0 and then applied our approach to the obtained errors. In the experiments, some errors could not be forcibly corrected by our approach. The parser “cut off” less probable parse substructures before giving the predicate argument relation for reducing the cost of parsing. In this research, we ignored the errors which were subject to such “cut off” as “uncorrectable” ones, and focused only on the remaining “correctable” errors. In our future work, we would like to consider the “uncorrectable” errors.

Table 4 shows the summary of the analysis with our approach. Enju gave 4,709 errors for section 0. Among these errors, the correctable errors were 3,085, and from these errors, we successfully obtained 1,978 co-occurring error groups and 501 inductive relations. Figure 11 shows the frequency for each size of co-occurring groups. About a half of the groups contains only single errors, which would indicate that the errors could have only one-way inductive relations with other errors. The rest of this section explores examples of the obtained co-occurring error groups and inductive relations.

Figure 12 shows an example of the extracted co-occurring error groups. For the sentence shown at the top of the figure, Enju gave seven errors. By introducing our empirical approach, these errors were definitely classified into four co-occurring error groups (a) to (d), and there were no inductive relations detected among them. Group (a) contains two errors on the comma’s local behavior as apposition or coordination. Group (b) contains the errors on the words which gave almost the same attachment behaviors. Group (c) contains the errors on whether the verb “show” took “decades"
Table 5: Induction relations between errors for each linguistic phenomenon and other errors

| Error types | # of correctable errors | # of independent errors | Correction effect (errors) |
|-------------|-------------------------|-------------------------|---------------------------|
| **[Argument selection]** | | | |
| Prepositional attachment | 531 | 397 | 766 |
| Adjunction attachment | 196 | 111 | 352 |
| Conjunction attachment | 33 | 12 | 79 |
| Head selection for noun phrase | 22 | 0 | 84 |
| Coordination | 146 | 62 | 323 |
| **[Predicate type selection]** | | | |
| Preposition/Adjunction | 72 | 30 | 114 |
| Gerund acts as modifier or not | 39 | 18 | 62 |
| Coordination/conjunction | 36 | 16 | 61 |
| # of arguments for preposition | 24 | 23 | 26 |
| Adjunction/adjunctive noun | 8 | 6 | 10 |
| **[More structural errors]** | | | |
| To-infinitive for modifier/argument of verb | 75 | 27 | 87 |
| Subject for passive sentence or not | 8 | 3 | 9 |
| **[Others]** | | | |
| Comma | 372 | 147 | 723 |
| Relative clause attachment | 84 | 27 | 119 |
| **Total** | 1,646 | 979 | – |

Figure 13: Inductive relation between obtained co-occurring error groups

as its object or not. Group (d) contains an error on the attachment of the adverb “later”. Regardless of the overlap of the regions in the sentence for (c) and (d), our approach successfully classified the errors into the two independent groups. With our approach, it would be empirically shown that the errors in each group actually co-occurred and the group was independent. This would enable us to concentrate on each of the co-occurring error groups without paying attention to the influences from the errors in other groups.

Figure 13 shows another example of the analysis with our empirical approach. In this case, 8 errors for a sentence were classified into two co-occurring error groups (a) and (b), and our approach showed that correction in group (a) resulted in correcting group (b) together. The errors in group (a) were on whether “help” behaved as an auxiliary or pure verbal role. The errors in group (b) were on whether “save” took only one object “her teaching certificate,” or two objects “her” and “teaching certificate.” Between group (a) and (b), no “structural” conflict could be found when correcting only each of the groups. We could then guess that the inductive relation between these two groups was implicitly given by the disambiguation model of the parser. By dividing the errors into minimum units and clarifying the effects of correcting a target error, error analysis with our empirical approach could suggest some policy for parser improvements.

4.3 Combination of two approaches

On the basis of the experiments shown in the previous sections, we would like to explore possibilities for obtaining a more detailed analysis by combining the two approaches.

4.3.1 Interactions between a target linguistic phenomenon and other errors

Our descriptive approach could classify the parsing errors according to the linguistic phenomena they participated in. We then attempt to reveal how such classified errors interacted with other errors from the viewpoints of our empirical approach. In order to enable the analysis by our empirical approach, we focused only on the correctable errors.
Clark J. Vitulli was named senior vice president and general manager of this U.S. sales and marketing arm of Japanese auto Maker Mazda Motor Corp and is currently waiving management fees, which boosts its yield. Figure 14: Combination of results given by descriptive and empirical approaches (1)

Table 5 reports the degree to which the classified errors were related to other individual errors. The leftmost numbers show the numbers of correctable errors, which were the focused errors in the experiments. The central numbers show the numbers of independent errors, that is, the errors which could be corrected only by correcting themselves. The rightmost numbers show correction effects, that is, the number of errors which would consequently be corrected if all of the errors for the focused phenomena were forcibly corrected.

"Independent" errors are obtained by collecting error phenomena groups which consist of unions of co-occurring error groups and each error in which is not induced by other errors. Figure 14 shows an example of independent errors. For the sentence at the top of the figure, the parser had four errors on ARG1 of overseas, the comma, which and boosts. Our empirical approach then classified these errors into two co-occurring error groups (a) and (b), and there was no inductive relation between the groups. Our descriptive approach, on the other hand, matched all of the errors with the patterns for Adjunction attachment, Comma and Relative clause attachment. Since the error for the Adjunction attachment equals to a co-occurring group (a) and is not induced by other errors, the error is independent.

Table 5 shows that, for Prepositional attachment, Adjunction attachments, [# of arguments for preposition] and Adjunction/adjunctive noun, more than half of the errors for the focused phenomena are independent. Containing many independent errors would mean that the parser should handle these phenomena further more intensively as an independent event.

The correction effect for a focused linguistic phenomenon can be obtained by counting errors in the union of the correctable error set for the phenomenon and the error sets which were induced by the individual errors in the set. We would show an example of correction effect in Figure 15. In the figure, the parser had six errors for the sentence at the top: three false outputs for ARG1 of and, this and U.S., two false outputs for ARG2 of of and and, and missing output for ARG1 of sales. Our empirical approach classified these errors into two co-occurring error groups (a) and (b), and extracted an inductive relation from (a) to (b). Our descriptive approach, on the other hand, matched two errors on and with pattern Coordination and one error on this with Head selection for noun phrase. When we focus on the error for Head selection of noun phrase in co-occurring group (a), the correction of the error induced the rest of the errors in (a), and further induced the error in (b) according to the inductive relation from (a) to (b). Therefore, a correction effect for the error results in six errors.

Table 5 shows that, for Conjunction attachment, Head selection for noun phrase and Coordination, each correction effect results in more than twice the forcibly corrected errors. Improving the parser so that it can resolve such high-correction-effect erroneous phenomena may additionally improve the parsing performances to a great extent. On the other hand, Head selection for noun phrase contains no independent error, and therefore could not be handled independently of other erroneous phenomena at all. Consider-
ing the effects from outer events might make the
treatment of “Head selection for noun phrase” a
more complicated process than other phenomena,
regardless of its high “correction effect.”

Table 5 would thus suggest which phenomenon
we should resolve preferentially from the three
points of view: the number of errors, the number
of “independent” errors and its “correction effect.”
Considering these points, “Prepositional attach-
ment” seems most preferable for handling first.

4.3.2 Possibilities for further analysis
Since the errors for the phenomenon were system-
atically collected with our descriptive approach,
we can work on further focused error analyses
which would answer such questions as “Which
preposition causes most errors in attachments?”,
“Which pair of a correct answer and an erroneous
output for predicate argument relations can occur
most frequently?”, and so on. Our descriptive ap-
proach would enable us to thoroughly obtain such
analyses with more closely-defined patterns. In
addition, our empirical approach would clarify the
influences of the obtained error properties on the
parser’s behaviors. The results of the focused anal-
yses might reasonably lead us to the features that
can be captured as parameters for model training,
or policies for re-ranking the parse candidates.

The combination of our approaches would give
us interesting clues for planning effective strate-
gies for improving the parser. Our challenges for
combining the two approaches are now in the pre-
liminary stage and there would be many possibili-
ties for further detailed analysis.

5 Related work
Although there have been many researches which
analyzed errors on their own systems in the part of
the experiments, there have been few researches
which focused mainly on error analysis itself.

In the field of parsing, McDonald and Nivre
(2007) compared parsing errors between graph-
based and transition-based parsers. They observed
the accuracy transitions from various points of
view, and the obtained statistical data suggested
that error propagation seemed to occur in the
graph structures of parsing outputs. Our research
proceeded for one step in this point, and attempted
to reveal the way of the propagations. In exam-
ining the combination of the two types of pars-
ing, McDonald and Nivre (2007) utilized similar
approaches to our empirical analysis. They al-
lowed a parser to give only structures given by
the parsers. They implemented the ideas for eval-
uating the parser’s potentials whereas we imple-
mented the ideas for observing error propagations.

Dredze et al. (2007) showed the possibility
that many parsing errors in the domain adaptation
tasks came from inconsistencies between annotation
manners of training resources. Such findings
would further suggest that, comparing given errors
without considering the inconsistencies could lead
to the misunderstanding of what occurs in domain
transitions. The summarized error dependencies
given by our approaches would be useful clues for
extracting such domain-dependent error phenom-
ena.

Giménez and Màrquez (2008) proposed an au-
tomatic error analysis approach in machine trans-
lation (MT) technologies. They were developing
a metric set which could capture features in MT
outputs at different linguistic levels with different
levels of granularity. As we considered the parsing
systems, they explored the way to resolve costly
and non-rewarding error analysis in the MT field.
One of their objectives was to enable researchers
to easily access detailed linguistic reports on their
systems and to concentrate only on analyses for
the system improvements. From this point of view,
our research might provide an introduction into
such rewarding analysis in parsing.

6 Conclusions
We proposed empirical and descriptive approaches
to extracting dependencies among parsing errors.
In the experiments, with each of our approaches,
we successfully obtained relevant errors. More-
over, the possibility was shown that the combina-
tion of our approaches would give a more detailed
error analysis which would bring us useful clues
for parser improvements.

In our future work, we will improve the per-
formance of our approaches by adding more pat-
terns for the descriptive approach and by handling
uncorrectable errors for the empirical approach.
With the obtained robust information, we will ex-
plore rewarding ways for parser improvements.

Acknowledgments
This work was partially supported by Grant-in-Aid
for Specially Promoted Research (MEXT, Japan).
References

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João V. Graça, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In Proceedings of the CoNLL Shared Task Session of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1051–1055.

Jesús Giménez and Lluís Márquez. 2008. Towards heterogeneous automatic MT error analysis. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), pages 1894–1901.

Ronald M. Kaplan and Joan Bresnan. 1995. Lexical-functional grammar: A formal system for grammatical representation. Formal Issues in Lexical-Functional Grammar, pages 29–130.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In Proceedings of ARPA Human Language Technology Workshop.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 122–131.

Yusuke Miyao and Jun’ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL), pages 83–90.

Takashi Ninomiya, Takuya Matsuzaki, Yoshimasa Tsuruoka, Yusuke Miyao, and Jun’ichi Tsujii. 2006. Extremely lexicalized models for accurate and fast HPSG parsing. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 155–163.

Carl J. Pollard and Ivan A. Sag. 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press.

Mark Steedman. 2000. The Syntactic Process. THE MIT Press.