The Domain Dependence of Parsing

Satoshi Sekine
New York University
Computer Science Department
715 Broadway, Room 709
New York, NY 10003, USA
sekine@cs.nyu.edu
http://cs.nyu.edu/cs/projects/proteus/sekine

Abstract

A major concern in corpus based approaches is that the applicability of the acquired knowledge may be limited by some feature of the corpus, in particular, the notion of text 'domain'. In order to examine the domain dependence of parsing, in this paper, we report 1) Comparison of structure distributions across domains; 2) Examples of domain specific structures; and 3) Parsing experiment using some domain dependent grammars. The observations using the Brown corpus demonstrate domain dependence and idiosyncrasy of syntactic structure. The parsing results show that the best accuracy is obtained using the grammar acquired from the same domain or the same class (fiction or non-fiction). We will also discuss the relationship between parsing accuracy and the size of training corpus.

1 Introduction

A major concern in corpus based approaches is that the applicability of the acquired knowledge may be limited by some feature of the corpus. In particular, the notion of text 'domain' has been seen as a major constraint on the applicability of the knowledge. This is a crucial issue for most application systems, since most systems operate within a specific domain and we are generally limited in the corpora available in that domain.

There has been considerable research in this area (Kittredge and Hirschman, 1983) (Grishman and Kittredge, 1986). For example, the domain dependence of lexical semantics is widely known. It is easy to observe that usage of the word 'bank' is different between the 'economic document' domain and the 'geographic' domain. Also, there are surveys of domain dependencies concerning syntax or syntax-related features (Slocum, 1986) (Biber, 1993) (Karlgren, 1994). It is intuitively conceivable that there are syntactic differences between 'telegraphic messages' and 'press report', or between 'weather forecast sentences' and 'romance and love story'. But, how about the difference between 'press report' and 'romance and love story'? Is there a general and simple method to compare domains? More importantly, shall we prepare different knowledge for these two domain sets?

In this paper, we describe two observations and an experiment which suggest an answer to the questions. Among the several types of linguistic knowledge, we are interested in parsing, the essential component of many NLP systems, and hence domain dependencies of syntactic knowledge. The observations and an experiment are the following:

- Comparison of structure distributions across domains
- Examples of domain specific structures
- Parsing experiment using some domain dependent grammars

2 Data and Tools

The definition of domain will dominate the performance of our experiments, so it is very important to choose a proper corpus. However, for practical reasons (availability and time constraint), we decided to use an existing multi-domain corpus which has naturally acceptable domain definition. In order to acquire grammar rules in our experiment, we need a syntactically tagged corpus consisting of different domains, and the tagging has to be uniform throughout the corpus. To meet these requirements, the Brown Corpus (Francis and Kucera, 1964) on the distribution of PennTreeBank version 1 (Marcus et.al., 1995) is used in our experiments. The corpus consists of 15

domains as shown in Appendix A; in the rest of the paper, we use the letters from the list to represent the domains. Each sample consists of about the same size of text in terms of the number of words (2000 words), although a part of the data is discarded because of erroneous data format.

For the parsing experiment, we use 'Apple Pie Parser' (Sekine, 1995) (Sekine, 1996). It is a probabilistic, bottom-up, best-first search, chart parser and its grammar can be obtained from a syntactically-tagged corpus. We acquire two-non-terminal grammars from corpus. Here, 'two-non-terminal grammar' means a grammar which uses only 'S' (sentence) and 'NP' (noun phrase) as actual non-terminals in the grammar and other grammatical nodes, like 'VP' or 'PP', are embedded into a rule. In other words, all rules can only have either 'S' or 'NP' as their left hand-side symbol. This strategy is useful to produce better accuracy compared to all non-terminal grammar. See (Sekine, 1995) for details.

In this experiment, grammars are acquired from the corpus of a single domain, or from some combination of domains. In order to avoid the unknown word problem, we used a general dictionary to supplement the dictionary acquired from corpus. Then, we apply each of the grammars to some texts of different domains. We use only 8 domains (A,B,E,J,K,L,N and P) for this experiment, because we want to fix the corpus size for each domain, and we want to have the same number of domains for the non-fiction and the fiction domains. The main objective is to observe the parsing performance based on the grammar acquired from the same domain compared with the performance based on grammars of different domains, or combined domains. Also, the issue of the size of training corpus will be discussed.

3 Domain Dependence of Structures

First, we investigate the syntactic structure of each domain of the Brown corpus and compare these for different domains. In order to represent the syntactic structure of each domain, the distribution of partial trees of syntactic structure is used. A partial tree is a part of syntactic tree with depth of one, and it corresponds to a production rule. Note that this partial tree definition is not the same as the structure definition used in the parsing experiments described later. We accumulate these partial trees for each domain and compute the distribution of partial trees based on their frequency divided by the total number of partial trees in the domain. For example, Figure 1 shows the five most frequent partial trees (in the format of production rule) in domain A (Press: Reportage) and domain P (Romance and love story).

domain A		domain P	
PP -> IN NP	8.40%	NP -> PRP	9.52%
NP -> NNPX	5.42%	PP -> IN NP	5.79%
S -> S	5.06%	S -> NP VP	5.77%
S -> NP VP	4.28%	S -> S	5.37%
NP -> DT NNX	3.81%	NP -> DT NNX	3.90%

Figure 1: Partial Trees

T\M	A	B	E	J	K	L	N	P
A	5.13	5.35	5.41	5.45	5.51	5.52	5.53	5.55
B	5.47	5.19	5.50	5.51	5.55	5.58	5.60	5.60
E	5.50	5.48	5.20	5.48	5.58	5.59	5.58	5.61
J	5.39	5.37	5.35	5.15	5.52	5.57	5.58	5.59
K	5.32	5.25	5.31	5.41	4.95	5.14	5.15	5.17
L	5.32	5.26	5.32	5.45	5.12	4.91	5.09	5.13
N	5.29	5.25	5.28	5.43	5.10	5.06	4.89	5.12
P	5.43	5.36	5.40	5.55	5.23	5.21	5.21	5.00

Figure 2: Cross Entropy of grammar across domains

For each domain, we compute the probabilities of partial trees like this. Then, for each pair of domains, cross entropy is computed using the probability data. Figure 2 shows a part of the cross entropy data. For example, 5.41 in column A, row E shows the cross entropy of modeling by domain E and testing on domain A. From the matrix, we can tell that some pairs of domains have lower cross entropy than others. It means that there are difference in similarity among domains. In particular, the differences among fiction domains are relatively small.

In order to make the observation easier, we clustered the domains based on the cross entropy data. The distance between two domains is calculated as the average of the two cross-entropies in both directions. We use non-overlapping and average-distance clustering. Figure 3 shows the clustering result based on grammar cross entropy data. From the results, we can clearly see that fiction domains, in particular domains K, L, and N are close which is intuitively understandable.

4 Domain Specific Structures

Secondly, in contrast to the global analysis reported in the previous section, we investigate the structural idiosyncrasies of each domain in the Brown corpus. For each domain, the list of partial trees which are relatively frequent in that domain is created. We select the partial trees which satisfy the following

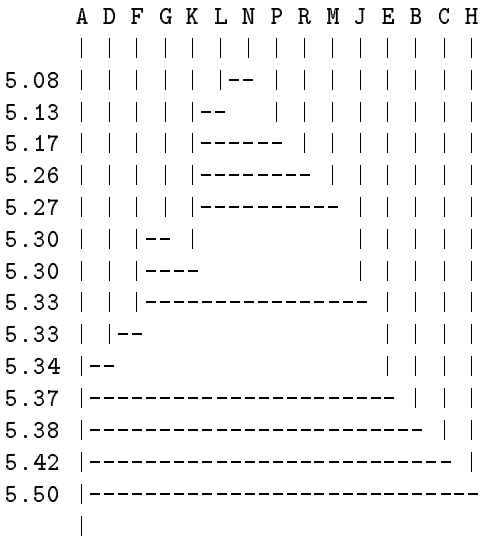

Figure 3: Clustering result

two conditions:

1. Frequency of the partial tree in a domain should be 5 times greater than that in the entire corpus
2. It occurs more than 5 times in the domain

The second condition is used to delete noise, because low frequency partial trees which satisfy the first condition have very low frequency in the entire corpus.

The list is too large to show in this paper; a part of the list is shown in Appendix B. It obviously demonstrates that each domain has many idiosyncratic structures. Many of them are interesting to see and can be easily explained by our linguistic intuition. (Some examples are listed under the corresponding partial tree) This supports the idea of domain dependent grammar, because these idiosyncratic structures are useful only in that domain.

5 Parsing Results

In this section, the parsing experiments are described. There are two subsections. The first is the individual experiment, where texts from 8 domains are parsed with 4 different types of grammars. These are grammars acquired from the same size corpus of the same domain, all domains, non-fiction domains and fiction domains.

The other parsing experiment is the intensive experiment, where we try to find the best suitable grammar for some particular domain of text and to see the relationship of the size of the training corpus. We use the domains of 'Press Reportage' and 'Romance and Love Story' in this intensive experiment.

Text	Same domain	All	non-fiction	fiction
A	**66.62/64.14**	64.39/61.45	65.57/62.40	62.23/59.32
B	**67.65**/62.55	64.67/61.78	65.73/**62.69**	63.03/60.36
E	64.05/60.79	65.25/61.51	**65.26/62.18**	62.87/59.04
J	**67.80/65.59**	65.87/63.90	65.57/64.58	63.04/60.77
K	70.99/68.54	71.00/68.04	70.04/66.64	**71.79/68.95**
L	67.59/65.02	68.08/66.22	67.32/64.31	**68.89/66.55**
N	73.09/71.38	72.97/70.27	70.51/67.90	**74.29/72.23**
P	**66.44/65.51**	64.52/63.95	62.37/61.55	64.69/64.50

Figure 4: Parsing accuracy for individual section

In order to measure the accuracy of parsing, recall and precision measures are used (Black et.al., 1991).

5.1 Individual Experiment

Figure 4 shows the parsing performance for domain A, B, E, J, K, L, N and P with four types of grammars. In the table, results are shown in the form of 'recall/precision'. Each grammar is acquired from roughly the same size (24 samples except L with 21 samples) of corpus. For example, the grammar of all domains is created using corpus of 3 samples each from the 8 domains. The grammar of non-fiction and fiction domains are created from corpus of 6 samples each from 4 domains. Then text of each domain is parsed by the four types of grammar. There is no overlap between training corpus and test corpus.

We can see that the result is always the best when the grammar acquired from either the same domain or the same class (fiction or non-fiction) is used. We will call the division into fiction and non-fiction as 'class'. It is interesting to see that the grammar acquired from all domains is not the best grammar in any tests. In other words, if the size of the training corpus is the same, using a training corpus drawn from a wide variety of domains does not help to achieve better parsing performance.

For non-fiction domain texts (A, B, E and J), the performance of the fiction grammar is notably worse than that of the same domain grammar or the same class grammar. In contrast, the performance on some fiction domain texts (K and L) with the non-fiction grammar is not so different from that of the same domain. Here, we can find a relationship between these results and the cross entropy observations. The cross entropies where any of the fiction domains are models and any of the non-fiction domains are test are the highest figures in the table. This means that the fiction domains are not suitable for modeling the syntactic structure of the non-fiction domains. On the other hand, the cross entropies where any of the non-fiction domains are

models and any of the non-fiction domains (except P) are test have some lower figures. Except for the case of N with the non-fiction grammar, these observations explains the result of parsing very nicely. The higher the cross entropy, the worse the parsing performance.

It is not easy to argue why, for some domains, the result is better with the grammar of the same class rather than the same domain. One rationale we can think of is based on the comparison observation described in section 3. For example, in the cross comparison experiment, we have seen that domains K, L and N are very close. So it may be plausible to say that the grammar of the fiction domains is mainly representing K, L and N and, because it covers wide syntactic structure, it gives better performance for each of these domains. This could be the explanation that the grammar of fiction domains are superior to the own grammar for the three domains. In other words, it is a small sampling problem, which can be seen in the next experiment, too. Because only 24 samples are used, a single domain grammar tends to covers relatively small part of the language phenomena. On the other hands, a corpus of similar domains could provide wider coverage for the grammar. The assumption that the fiction domain grammar represents domains of K, L and M may explain that the parsing result of domain P strongly favors the grammar of the same domain compared to that of the fiction class domains.

5.2 Intensive Experiments

In this section, the parsing experiments on texts of two domains are reported. The texts of the two domains are parsed with several grammars, e.g. grammars acquired from different domains or classes, and different sizes of the training corpus. The size of the training corpus is an interesting and important issue. We can easily imagine that the smaller the training corpus, the poorer the parsing performance. However, we don't know which of the following two types of grammar produce better performance: a grammar trained on a smaller corpus of the same domain, or a grammar trained on a larger corpus including different domains.

Figure 5 and Figure 6 shows recall and precision of the parsing result for the Press Reportage text. The same text is parsed with 5 different types of grammars of several variations of training corpus size. Because of corpus availability, we can not make single domain grammars of large size training corpus, as you can find it in the figures.

Figure 7 and Figure 8 shows recall and precision of the parsing result for the Romance and Love Story

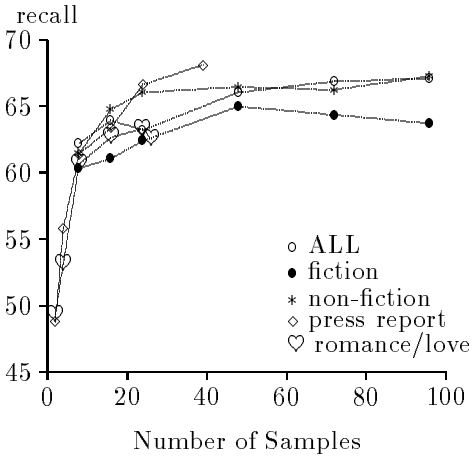

Figure 5: Size and Recall (Press Report)

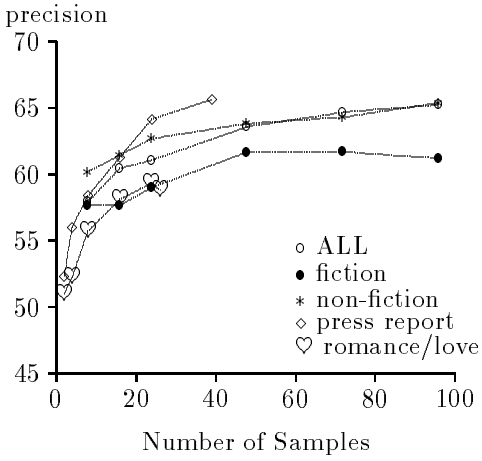

Figure 6: Size and Precision (Press Report)

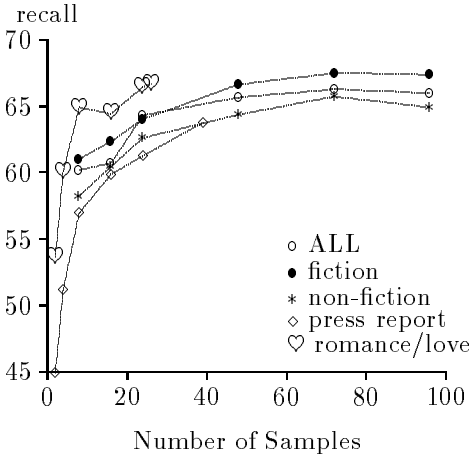

Figure 7: Size and Recall (Romance/Love)

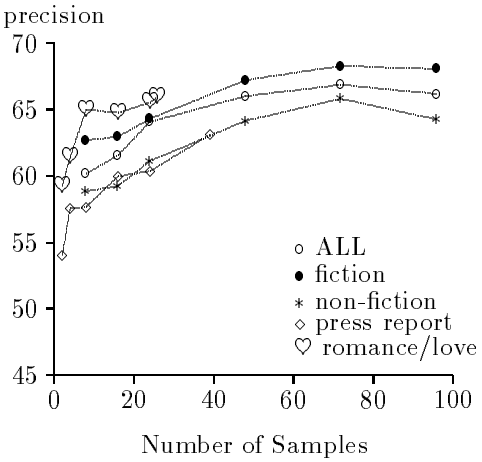

Figure 8: Size and Precision (Romance/Love)

text. This text is also parsed with 5 different types of grammars.

The graph between the size of training corpus and accuracy is generally an increasing curve with the slope gradually flattening as the size of the corpus increases. Note that the small declines of some graphs at large number of samples are mainly due to the memory limitation for parsing. Parsing is carried out with the same memory size, but when the training corpus grows and the grammar becomes large, some long sentences can't be parsed because of data area limitation. When the data area is exhausted during the parsing, a fitted parsing technique is used to build the most plausible parse tree from the partially parsed trees. These are generally worse than the trees completely parsed.

It is very interesting to see that the saturation point of any graph is about 10 to 30 samples. That is about 20,000 to 60,000 words, or about 1,000 to 3,000 sentences. In the romance and love story domain, the precision of the grammar acquired from 8 samples of the same domain is only about 2% lower than the precision of the grammar trained on 26 samples of the same domain. We believe that the reason why the performance in this domain saturates with such a small corpus is that there is relatively little variety in the syntactical structure of this domain.

The order of the performance is generally the following: the same domain (best), the same class, all domains, the other class and the other domain (worst). The performance of the last two grammars are very close in many cases. In the romance and love story domain, the grammar acquired from the same domain made the solo best performance. The difference of the accuracy of the grammars of the same domain and the other domain is quite large. The results for the press reportage is not so obvious, but the same tendencies can be observed.

In terms of the relationship between the size of training corpus and domain dependency, we will compare the performance of the grammar acquired from 24 samples of the same domain (we will call it 'baseline grammar'), and that of the other grammars. In the press reportage domain, one needs a three to four times bigger corpus of all domains or non-fiction domains to catch up to the performance of the baseline grammar. It should be noticed that a quarter of the non-fiction domain corpus and one eighth of the all domain corpus consists of the press report domain corpus. In other words, the fact that the performance of the baseline grammar is about the same as that of 92 samples of the non-fiction domains means that in the latter grammar, the rest of the corpus does not improve or is not harmful for the parsing performance. In the romance and love story domain, the wide variety grammar, in particular the fiction domain grammar quickly catch up to the performance of the baseline grammar. It needs only less than twice size of fiction domain corpus to achieve the performance of the baseline grammar.

These two results and the evidence that fiction domains are close in terms of structure indicate that if you have a corpus consisting of similar domains, it is worthwhile to include the corpus in grammar acquisition, otherwise not so useful. We need to further quantify these trade-offs in terms of the syntactic diversity of individual domains and the difference between domains.

We also find the small sampling problem in this experiment. In the press reportage experiment, the grammar acquired from the same domain does not make the best performance when the size of the training corpus is small. We observed the same phenomena in the previous experiment.

6 Discussion

One of our basic claims is the following. When we try to parse a text in a particular domain, we should prepare a grammar which suits that domain. This idea naturally contrasts to the idea of robust broad-coverage parsing (Carroll and Briscoe, 1996), in which a single grammar should be prepared for parsing of any kind of text. Obviously, the latter idea has a great advantage that you do not have to create a number of grammars for different domains and also do not need to consider which grammar should be used for a given text. On the other hand, it is plausible that a domain specific grammar can produce better results than a domain independent grammar. Practically, the increasing availability of

corpora provides the possibilities of creating domain dependent grammars. Also, it should be noted that we don't need a very large corpus to achieve a relatively good quality of parsing.

To summarize our observations and experiments:

- There are domain dependencies on syntactic structure distribution.
- Fiction domains in the Brown corpus are very similar in terms of syntactic structure.
- We found many idiosyncratic structures from each domain by a simple method.
- For 8 different domains, domain dependent grammar or the grammar of the same class provide the best performance, if the size of the training corpus is the same.
- The parsing performance is saturated at very small size of training corpus. This is the case, in particular, for the romance and love story domain.
- The order of the parsing performance is generally the following; the same domain (best), the same class, all domain, the other class and the other domain (worst).
- Sometime, training corpus in similar domains is useful for grammar acquisition.
- It may not be so useful to use different domain corpus even if the size of the corpus is relatively large.

Undoubtedly these conclusions depend on the parser, the corpus and the evaluation methods. Also our experiments don't cover all domains and possible combinations. However, the observations and the experiment suggest the significance of the notion of domain in parsing. The results would be useful for deciding what strategy should be taken in developing a grammar on a 'domain dependent' NLP application systems.

7 Acknowledgments

We would like to thank our colleagues, in particular Prof.Ralph Grishman for valuable discussions and suggestions.

References

Douglas Biber: 1993. Using Register-Diversified Corpora for General Language Studies. *Journal of Computer Linguistics Vol.19, Num 2*, pp219-241.

Ezra Black, et.al: 1991. A procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. *Proc. of Fourth DARPA Speech and Natural Language Workshop*

John Carroll and Ted Briscoe: 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. *Proceedings of Conference on Empirical Methods in Natural Language Processing.*

W. Nelson Francis and Henry Kucera: 1964/1979. Manual of information to accompany A Standard Corpus of Present-Day Edited American English. *Brown University, Department of Linguistics*

Ralph Grishman and Richard Kittredge: 1986. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. *Lawrence Erlbaum Associates, Publishers*

Jussi Karlgren and Douglass Cutting: 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *The 15th International Conference on Computational Linguistics*, pp1071-1075.

Richard Kittredge, Lynette Hirschman: 1983. Sublanguage: Studies of Language in Restricted Semantic domains. *Series of Foundations of Communications, Walter de Gruyter, Berlin*

Mitchell P. Marcus, Beatrice Santorini and Mary A Marcinkiewicz: 1993. Building a Large Annotated Corpus of English: The Penn TreeBank. *Computational Linguistics*, 19.1, pp313-330.

Satoshi Sekine: 1996. Apple Pie Parser homepage. *http://cs.nyu.edu/cs/projects/proteus/app*

Satoshi Sekine and Ralph Grishman: 1995. A Corpus-based Probabilistic Grammar with Only Two Non-terminals. *International Workshop on Parsing Technologies*, pp216-223.

Johathan Slocum: 1986. How One Might Automatically Identify and Adapt to a Sublanguage: An Initial Exploration. *Analyzing Language in Restricted Domains*, pp195-210.

APPENDIX

A Categories in Brown corpus

```
I. Informative Prose   (374 samples)
 A. Press: Reportage                (44)
 B. Press: Editorial                (27)
 C. Press: Reviews                  (17)
 D. Religion                        (17)
 E. Skills and Hobbies              (36)
 F. Popular Lore                    (48)
 G. Letters,Bibliography,Memories,  (75)
 H. Miscellaneous                   (30)
 J. Learned                         (80)
II. Imaginative Prose  (126 Samples)
 K. General Fiction                 (29)
 L. Mystery and Detective Fiction   (24)
 M. Science Fiction                 ( 6)
 N. Adventure and Western Fiction   (29)
 P. Romance and Love Story          (29)
 R. Humor                           ( 9)
```

B Sample of Relatively Frequent Partial Trees

```
SYM. DOMAIN (num.of type;total freq. of
                 qualified partial trees)

 ratio   frequency   rule (Example)
      (domain/corpus)

A. Press: Reportage  (30;507)
  9.40    11 / 14   NP -> NNPX NNX NP
  9.30     7 /  9   NP -> NP POS JJ NNPX
  8.70     8 / 11   S  -> NP VBX VP NP PP
  8.44    12 / 17   NP -> DT $ CD NNX
    `The $40,000,000 budget´
    `a 12,500 payment´
  8.30   77 /  111  NP -> NNPX NP
    `Vice President L.B. Johnson´
    `First Lady Jacqueline Kennedy´
B. Press: Editorial  (20;255)
 18.57    34 / 34   S  -> PP :
    `To the editor:´
    `To the editor of New York Times:´
 11.14    6 /  10   NP -> DT `` ADJP '' NNX
    `an ``autistic'' child´
    `a ``stair-step'' plan´
C. Press: Reviews   (19;267)
 26.27    8 /   9   WHADVP -> NNPX
 25.33   12 /  14   NP -> NP POS `` NNPX ''
D. Religion   (8;87)
 26.83   26 /  28   S  -> NP -RRB- S
 25.28   14 /  16   NP -> NNPX CD : CD
    `St. Peter 1:4´
    `St. John 3:8´
E. Skills and Hobbies   (17;219)
 10.58   22 /  22   NP -> CD NNX ''
 10.21   27 /  28   S  -> SBAR :
    `How to feed :´
    `What it does :´
F. Popular Lore   (12;86)
 10.58    8 /  8    NP -> DT NP POS NNPX
 10.58    6 /  6    NP -> NNX DT NNX PP
G. Letters,Bibliography,Memories,etc (12;125)
  6.59    8 /  8    WHPP -> TO SBAR
    `to what they mean by the concept´
    `to what may happen next´
  6.04   22 / 24    WHPP -> @OF SBAR
    `of what it is all about´
    `of what he had to show his country´
H. Miscellaneous   (69;2607)
 16.82   70 / 70    S  -> NP . S
 16.82   17 / 17    S  -> -LRB- VP . -RRB-
J. Learned   (22;295)
  6.51   28 / 28    NP -> CD : CD
  6.51   20 / 20    NP -> NNX :
  6.22   44 / 46    S  ->   S -LRB- NP -RRB-
     Sentence and name and year in bracket
     Sentence and figure name in bracket
K. General Fiction   (14;148)
 11.58    7 / 10    NP -> PRP S
 11.03    6 /  9    S  -> ADVP S : : S
 10.75   13 / 20    S  -> PP S , CC S
L. Mystery and Detective Fiction   (19;229)
 14.28    8 / 11    SQ -> S , SQ
     Tag questions
M. Science Fiction   (6;57)
 17.89    7 / 32    S  -> S , SINV
     "``Forgive me, Sandalphon´´, said Hal"
     "``sentence´´, remarked Helva"
 10.22    8 / 64    S  -> SBARQ '' .
N. Adventure and Western Fiction  (24;422)
 14.59   45 / 50    VP -> VBX RB
 12.97    8 / 10    VP -> VBX RB PP
P. Romance and Love Story   (31;556)
 15.99    7 /  7   S  -> CC SBARQ
 15.99    6 /  6   S  -> `` NP VP , NP ''
 12.23   13 / 17   S  -> SQ S
 11.99    6 /  8   S  -> `` VP , NP ''
R. Humor   (3;20)
  6.92    6 / 47   NP -> DT ADJP NP
  6.78    7 / 56   NP -> PRP @DLQ
  5.67    7 / 67   PP -> IN `` NP ''
     `as ``off-Broadway´´´
```