An Analysis of English Punctuation: The Special Case of Comma

MURAT BAYRAKTAR, BILGE SAY AND VAROL AKMAN[1]
Bilkent University, Ankara, Turkey

Punctuation has usually been ignored by researchers in computational linguistics over the years. Recently, it has been realized that a true understanding of written language will be impossible if punctuation marks are not taken into account. This paper contains the details of a computer-aided exercise to investigate English punctuation practice for the special case of comma (the most significant punctuation mark) in a parsed corpus. The study classifies the various 'structural' uses of comma according to the syntax-patterns in which comma occurs. The corpus (Penn Treebank) consists of syntactically annotated sentences with no part-of-speech tag information about the individual words.

KEYWORDS: punctuation, structural punctuation marks, comma, the Penn Treebank

1. Introduction

Until recently, punctuation has been neglected by most researchers in theoretical and computational linguistics. This is due to the absence of a concise, formal background for the abstract problem. However, once we remember that punctuation is an orthographical component of written language, we see that research on punctuation makes reasonable sense. Accordingly, interest in the subject rose within the last decade because it has been realized that a fuller understanding and processing of written language is quite impossible without taking punctuation into account. Although punctuation was originally invented as a device for reflecting intonation in written text, it is now a linguistic "system on its own right'' (Nunberg 1990: 9).

This paper reports a study to analyze the English punctuation practice in a computer-aided exercise. The material used was a syntactically annotated (i.e., parsed) corpus, which was a part of the bracketed version of the Penn Treebank (Marcus et al. 1992). Due to its higher significance compared to other punctuation marks, only the comma was investigated. The purpose of the investigation was to classify various *structural*[2] uses of comma in the given corpus and observe the frequencies. The classification made by Ehrlich (Ehrlich 1992) was taken as a basis. The corpus consists of the parse trees of sentences with no part-of-speech tag information about the individual words. For the classification, abbreviated syntax-patterns containing the comma as an immediate daughter were extracted and intuitively assigned to appropriate classes by looking at sample sentences containing these patterns. Observing this classification, frequencies of the individual uses of comma in the analyzed corpus were reported.

The remainder of this paper is organized as follows. Section 2 is a brief appraisal of recent related work. Section 3 starts with a discussion on the significance of comma and ends with an enumeration of its uses. Information about the contents and structure of the Penn Treebank is offered in Section 4. Section 5 contains the details and results of the pattern extraction and classification process. The paper is concluded with Section 6, where a discussion and suggestions for further work can be found.

2. Background

If we look at the recent related works on punctuation, we detect (1) linguistic works treating punctuation, and (2) works within the framework of computational linguistics, which attempt to take punctuation marks into account in Natural Language Processing (NLP).

2.1 Related Works in Linguistics

Introductory, intermediate, and advanced composition handbooks and grammar books are pedagogical approaches to punctuation. These books usually contain lengthy passages on the individual punctuation marks. Discussions of punctuation addressing a general audience are found in style manuals, dictionaries and full-length books (Ehrlich 1992; Jarvie 1992; Paxson 1986). The common approach among these studies is that they employ a prescriptive treatment of punctuation: long lists of rules for correct punctuation are given, but the actual practice is not considered.

Meyer's work (Meyer 1987) is the first example of a wholly descriptive study of punctuation. Focusing on the American practice vis-à-vis structural punctuation marks and working with samples from the Brown Corpus (Francis and Kucera 1982), he classifies and illustrates the functions of punctuation and how these functions are realized. An important observation he makes is that the functions of marks and their realizations are distinct concepts; this is usually ignored within the prescriptive arena. According to Meyer, there may be three functions of punctuation: help the reader to understand the text easily, emphasize a concept, and vary the rhythm of the text. The realization of those functions, on the other hand, fall into two main categories: marks that separate and marks that enclose.

Nunberg's *The Linguistics of Punctuation* (Nunberg 1990) was the basic motivation for research on punctuation in the 90's. In this important study, he attacks the general opinion that punctuation is prescriptive and only a device for reflecting intonation, and claims that after the divergence of written and spoken languages, punctuation has become a linguistic system on its own right. He proposes to use two separate grammars to analyze texts. A *lexical grammar* accounts for the text-categories (text-clauses, text adjuncts, and text-phrases) occurring between the punctuation marks: A *text grammar* deals with the structure of punctuation, and the relation of punctuation marks to the text-categories they separate. Included within his text grammar are rules (for English orthography) that handle the interactions between various marks such as a point absorbing a comma when they are adjacent and a group of hierarchically-ordered rules for the presentation on paper of marks such as face- or font-alterations. Nunberg's characterization of punctuation has been the starting point for ensuing the computational research on the subject.

2.2 Related Works in Computational Linguistics

There are some NLP systems (Garside et al. 1987; Karlsson et al. 1994), parsers (Jones 1994; Briscoe and Carroll 1995), and generators (White 1995) taking punctuation into account. Briscoe and Carroll (Briscoe and Carroll 1995) report that treating punctuation marks within a grammar is useful for not only breaking the text into suitable units for parsing but also for resolving structural ambiguity. There are several studies such as (Dale 1991; Say and Akman 1996a; Say and Akman 1996b) on the semantic information carried by punctuation marks. A detailed survey can be found in (Say and Akman 1997).

The most closely related work to our study is (Jones 1996a).[3] Jones stresses the need for a new theory of punctuation which is suitable for computational implementation, and examines the true syntactic function of punctuation marks in the text. There may be two possible approaches to this problem: an observational one and a theoretical one. He tries to adopt both of these approaches, hoping to combine them suitably. For the observational part, he chooses the Dow

Jones section (approximately 2 million words) of the Penn Treebank (Marcus et al. 1992) and collects each node that has a punctuation mark as its immediate daughter in the parse tree, abbreviating its other daughters to their categories. This is shown in the following example:

[S [PP In Edinburgh] , [S ...]] $\Rightarrow$ [S $\rightarrow$ PP , S]

He groups different syntax-patterns into different sets for each punctuation mark and derives, using common properties among syntax-patterns within a set, rule-patterns representing the behavior of individual marks. As a result, he reduces the 12,700 unique syntax-patterns found in the corpus to just 137 rule-patterns for the colon, semicolon, dash, comma, and period. He reduces this number further to 79, employing a pruning procedure to remove idiosyncratic, incorrect, and exceptional rule-patterns. Using this reduced set of rule-patterns, he derives some generalized punctuation rules. He describes these in (Jones 1996b) in detail and suggests their integration into a grammar. He gives, for instance, the rule below (Jones 1996a: 364) for the potential syntax-patterns in which the comma may appear:

$C \rightarrow C,\ *$	C: {NP, S, VP, PP, ADJP, ADVP}
$C \rightarrow *\ ,\ C$	*: any category

In his theoretical approach, Jones starts with the following hypothesis, based on his observations (Jones 1996a: 364): punctuation seems to come "immediately before or after a phrasal level lexical item (e.g., a noun phrase).'' The combination of his theory and observations leads him to the conclusion that punctuation could be described as being either adjunctive or conjunctive. Adjunctive punctuation marks are the delimiting marks used in pairs and conjunctive ones are those used for separating. We investigate both functions of the comma in this sense.

3. The Comma

The comma has been described as "the most ubiquitous, elusive and discretionary of all stops'' (Jarvie 1992: 10). Meyer (Meyer 1987) observes that commas and periods each constitute around 45% of all marks in the Brown corpus (Francis and Kucera 1982). (The next most frequent mark has a frequency of 2%, a very sharp drop.) It may be argued that, other marks on the side, the period may be at least as important as the comma, since its frequency is almost the same. However, the comma beats the period with its versatility, which can best be illustrated by the interesting data obtained by Jones (Jones 1996a). As it was already mentioned in Section 2.2, Jones groups different syntax-patterns into different sets for each punctuation mark and observes 12,700 unique syntax-patterns in total for all punctuation marks. The cardinality of the set for the comma is 9,320, which makes about 73% of all patterns.

3.1 Classification of Uses of Comma

The number of classes mentioned for the uses of comma differ from two to 10 or 20 in different studies done on punctuation, depending on the potential audience of the study in question. Those works in the research camp (Meyer 1987, Nunberg 1990) prefer to be as general as possible, whereas those on the teaching side (e.g., style guides or punctuation usage books) try to illustrate all possible uses.

We need a detailed classification of potential uses of the comma in order to be able to group the syntax-patterns containing the comma later into these classes. At this point, it is more reasonable to refer to style guides or punctuation usage books. There are numerous such books (Ehrlich 1992; Jarvie 1992; Paxson 1986), each making a different classification. Since there is no consensus among these works, it would be wrong to say that one of them gives the 'correct' classification. Therefore, it is plausible to simply select one of them—preferably a popular one,

one which affects actual punctuation practice more widely—and complement its shortcomings with the others.

The following classification, which is used in our study, is mainly based on (Ehrlich 1992). Whenever this guide came short for the needs of the corpus, we referred to other books (Jarvie 1992; Paxson 1986). At some points, the classification is reorganized by making some classes subclasses of other classes. Furthermore, non-structural uses (such as the commas in numbers, dates, and addresses) are discarded, since they are outside the scope of this study. Every class is supported with examples to make its character more understandable. The examples are taken from the Penn Treebank (abbreviated as PT), whenever possible.

3.1.1 Elements in a Series

One of the frequent uses of comma is the separation of three or more elements listed in a series. The elements may be words, phrases, or clauses having the same syntactic type, cf. examples (1), (2), and (3). The last element is usually separated by a conjunction such as *and* or *or*, and sometimes by another comma.

(1) Elsewhere, share prices closed higher in *Amsterdam*, *Brussels*, *Milan* and *Paris*. (from PT)

(2) We innovated *telephone redemptions*, *daily dividends*, *total elimination of share certificates* and *the constant $1 pershare pricing*, all of which were painfully thought out and not the result of some inadvertence on the part of the SEC. (from PT)

(3) *John went shopping*, *Mary cooked the meal* and *David washed the dishes*.

In some cases, the conjunction may be preceded by a comma, in order to prevent misreading. Ehrlich names this as the *bacon-and-eggs* problem (Ehrlich 1992: 17):

(4) You may order anything you want at my diner as long as you order *sausage and eggs*, *ham and eggs*, or *bacon and eggs*.

(5) The chef said he needed *sausage*, *ham*, *bacon*, and *eggs*.

Independent clauses joined by a coordinating conjunction, such as *and*, *or*, *but*, etc., may be separated by a comma, if there is a risk of misreading:

(6) The Red Cross doesn't track contributions raised by the disaster ads, *but* it has amassed $46.6 million since it first launched its hurricane relief effort Sept. 23. (from PT)

Coordinate adjectives, which independently modify a noun, are separated by commas, if the meaning otherwise changes:

(7) And some US army analysts worry that the proposed Soviet redefinition is aimed at blocking the US from developing *lighter*, *more transportable*, *high technology* tanks. (from PT)

3.1.2 Sentence-initial Elements

A comma may delimit long phrases or clauses that appear sentence-initially as an introductory element, if there is a possibility of misleading the reader. This can be seen by looking at examples (8) and (9) for phrases and clauses respectively, and trying to read the sentences without the comma:

(8) *Under two new features,* participants will be able to transfer money from the new funds to other investment funds or, if their jobs are terminated, receive cash from the funds. (from PT)

(9) *Although the action removes one obstacle in the way of an overall settlement to the case,* it also means that Mr. Hunt could be stripped of virtually all of his assets if the Tax Court rules against him in a 1982 case heard earlier this year in Washington, D.C. (from PT)

Introductory modifiers, such as adjectives (10), adverbs (11), or participles (12), which usually consist of one word, are usually set off by a comma:

(10) *Victorious,* the army withdrew a thousand meters and encamped for the night. (Ehrlich 1992: 25)

(11) *Clearly,* the judge has had his share of accomplishments. (from PT)

(12) *Running,* he went up the stairs.

An absolute phrase may appear sentence-initially, in which case it is always delimited by a comma, since it modifies the entire sentence and has no grammatical connection to any other element in the sentence:

(13) *The party over,* the couple began to wash a sinkful of dishes. (Ehrlich 1992: 37)

It is noted that absolute phrases differ from other phrases in their capability of expressing a full idea, but unlike clauses, they only consist of a subject and a modifier.

3.1.3 Sentence-final Elements

Like sentence-initial introductory elements, sentence-final complementary elements are delimited by a comma, if there is a need for disambiguation. The element may be a phrase (14), a subordinate clause (15), or an absolute phrase (16):[4]

(14) A bomb exploded at a leftist union hall in San Salvador, *killing at least eight people and injuring about 30 others, including two Americans,* authorities said. (from PT)

(15) A face-to-face meeting with Mr. Gorbachev should damp such criticism, *though it will hardly eliminate it*. (from PT)

(16) She ran faster, *her breath coming in deep gasps*. (Paxson 1986: 31)

3.1.4 Nonrestrictive Phrases or Clauses

Postmodifiers of nouns, which may be phrases or clauses, are enclosed by commas if they are nonrestrictive. Restrictive modifiers identify, define, or limit the elements they modify, and thus are essential for the intended meaning. A nonrestrictive modifier, on the other hand, may be removed without changing the intended meaning since it only adds information concerning an element already identified, defined, or limited. Examples (17) and (18) illustrate restrictive phrases and clauses, respectively, whereas (19) and (20) show nonrestrictive ones:

(17) The man *at the left* is taller.

(18) He was the only student *who answered all the questions in the exam*.

versus

(19) A Western Union spokesman, *citing adverse developments in the market for high-yield "junk" bonds,* declined to say what alternatives are under consideration. (from PT)

(20) At one point, almost all of the shares in the 20-stock Major Market Index, *which mimics the industrial average,* were sharply higher. (from PT)

3.1.5 Appositives

Appositives, also known as noun repeaters, identify or point out to the nouns they succeed. Only nonrestrictive appositives are delimited by commas, as in the case of modifying phrases or clauses, mentioned in Section 3.1.4. Example (21) illustrates a restrictive appositive, whereas (22) shows a nonrestrictive one:

(21) Alexander *the Great* was a powerful emperor.

versus

(22) The new company, called Stardent Computer Inc., also said it named John William Poduska, *former chairman and chief executive of Stellar,* to the posts of president and chief executive. (from PT).

3.1.6 Interrupters

Commas are also used to delimit interrupters, which occur sentence-internally as a complementary or parenthetic element. This may be a single word (23), a phrase (24), or an entire clause (25), which breaks the expected logical flow of the sentence:

(23) The Brookings and Urban Institute authors caution, *however,* that most nursing home stays are of comparatively short duration, and reaching the Medicaid level is more likely with an unusually long stay or repeated stays. (from PT)

(24) The new bacteria recipients of the genes began producing pertussis toxin which, *because of the mutant virulence gene,* was no longer toxic. (from PT)

(25) Rebuilding that team, *Mr. Lee predicted,* will take another 10 years. (from PT)

3.1.7 Quotations

Direct quotations, indicating or repeating the exact words of the writer or the speaker, respectively, are set off by commas:

(26) "*The absurdity of the official rate should seem obvious to everyone*," the afternoon newspaper Izvestia wrote in a brief commentary on the devaluation. (from PT)

4. The Corpus

A suitable source for the observation of structural uses of the comma in real-life texts is a parsed corpus, since structural commas set off syntactical boundaries and depend on the grammatical structure of the sentence. Therefore, we have chosen the parsed version of the Penn Treebank (Marcus et al. 1992), which was produced using the skeleton parsing technique (Aarts, 1995).

The Penn Treebank, which is a 4.5 million word corpus of American English, was constructed by Marcus et al. (Marcus et al. 1992; Marcus et al. 1994) between 1989–1992. The part used in our study is a 307,089 word (14,823 sentences) portion of the parsed form (Version 0.75) of *Wall Street Journal* articles. It is available as part of the Penn Treebank Release 2 CDROM.[5] The sentences included in this particular piece of corpus are usually long and complex, which in turn means that they are also rich in punctuation.

The parsed version of the Penn Treebank consists of parsed sentences, which show the skeletal structure of the text. The appearance of a parsed sentence is in a bracketed, Lisp-like structure (equivalent to a syntax-tree diagram):

```
((S
      (NP (NP Mr. Smith)
          ,
          (NP 39)
          ,)
      (VP retains
          (NP the title
              (PP of
                  (NP (ADJP chief financial)
                      officer)))))
  .)
```

Bracketing groups words into phrases and/or clauses, and represents the hierarchical relationship which exists among these constructs. Left brackets are labeled with the type of construct they enclose. The types of constructs available in the syntactic tag-set of the Penn Treebank are listed in Table 2 in Appendix A.

Detailed guidelines for the bracketed (Treebank I style) version of the Penn Treebank are explained in (Santorini 1991), where a long list of problematic constructions and conventions (that were followed to represent them) are given.

5. The Experiment

Since the major function of the structural comma is setting off syntactic boundaries, the information contained in the parse trees should be enough to make the classification (of the uses of comma). The first step was the construction of a database of all syntax-patterns containing one or more commas. Then, the classification was made by assigning these syntax-patterns into appropriate classes.

5.1 Construction of Syntax-pattern Database

Construction of the database of all syntax-patterns containing one or more commas was done by a Prolog program, which analyzed all parse trees in the corpus and extracted each node with one or more commas as its immediate daughter(s), with the other daughters abbreviated to their syntactic category labels as in the following examples:

(NP (NP My uncle) , (NP 39 years old) ,)) ⇒ NP → NP , NP ,

(S (PP In London) , (S …)) ⇒ S → PP , S

(S Ultimately , (S …) , (S …)) ⇒ S → *** , S , S

The three consecutive asterisks denote any number of successive terminal words, not further labeled with any syntactic tag.

Each entry in the constructed database was recorded with the following fields:

- `Pattern` (primary key): The abbreviated syntax-pattern in question (cf. Appendix B).
- `Count`: Number of occurrences of this syntax-pattern in the whole corpus.
- `SampleSentence`: The first sentence, in which the syntax-pattern occurred, recorded in raw text format.

The outcome of this process was a database consisting of 1,978 unique (`Pattern, Count, SampleSentence`) triples.

5.2 Classification of Syntax-patterns

The aim of the construction of a database of syntax-patterns was to use it later in the classification of the uses of the comma in the corpus. This could be done manually, but classifying all 1,978 syntax-patterns would be a tremendous task. So, we decided to automatically classify only the most important patterns, such that, at the end, effectively 80% of all commas in the corpus would have been classified. This data would be sufficiently representative for the uses of the comma on the whole.

To determine the most important (frequent) syntax-patterns, the database was sorted according to the number of occurrences in the corpus. Starting from the top of this list, the number of occurrences were added cumulatively until the sum yielded 14,299, which is ~80% of the 17,883 commas in the corpus. The number of the syntax-patterns until this point was recorded as 211, which is only ~11% of the 1,978 unique syntax-patterns for the comma. In other words, it turned out that it was enough to classify the top 11% of the syntax-patterns in order to have effectively classified 80% of the commas in the whole corpus.

The last task to be accomplished was to assign each of the top 211 syntax-patterns to one of the classes listed in Section 3.1. These assignments were done via a simple user interface displaying each time the syntax-pattern and the recorded sample sentence. We had to read the sentence, find the comma, and then intuitively select (from a menu of classes) the class that the use of the comma in question matches. This class is the one to which the syntax-pattern has to be assigned. In this way, all of the 211 syntax-patterns were assigned to a class. Below is a list of all classes along with the topmost (i.e., most frequently occuring) syntax-pattern recorded for this class, its subclass, its total number of occurrences, and the sample sentence (with the underlined comma(s)):

1. Elements in a series:
   `Pattern: S → S , S`

Subclass: Coordinate Clauses in a Series [cf. Appendix B, (1.4)]
Count: 347
SampleSentence:
A SEC proposal to ease reporting requirements for some company executives would undermine the usefulness of information on insider trades as a stock-picking tool, individual investors and professional money managers contend.

2. Sentence-initial elements:
Pattern: S → PP , S
Subclass: Introductory Phrases [cf. Appendix B, (2.2)]
Count: 952
SampleSentence:
In an Oct. 19 review of "The Misanthrope" at Chicago's Goodman Theatre ("Revitalized Classics Take the Stage in Windy City," Leisure & Arts), the role of Celimene, played by Kim Cattrall, was mistakenly attributed to Christina Haag.

3. Sentence-final elements:
Pattern: S → NP *** VP , S
Subclass: Final Clauses [cf. Appendix B, (3.2)]
Count: 57
SampleSentence:
Jan Leemans, research director, said this gene was successfully introduced in oil-producing rapeseed plants, a major crop in Europe and Canada, using as a carrier a "promoter gene" developed by Robert Goldberg at the University of California in Los Angeles.

4. Nonrestrictive phrases or clauses:
Pattern: NP → NP , SBAR
Subclass: Nonrestrictive Clauses [cf. Appendix B, (4.2)]
Count: 570
SampleSentence:
The changes were proposed in an effort to streamline federal bureaucracy and boost compliance by the executives "who are really calling the shots," said Brian Lane, special counsel at the SEC's office of disclosure policy, which proposed the changes.

5. Appositives:
Pattern: NP → NP , NP ,
Subclass: none [cf. Appendix B, (5)]
Count: 1880
SampleSentence:
Howard Mosher, president and chief executive officer, said he anticipates growth for the luxury auto maker in Britain and Europe, and in Far Eastern markets.

6. Interrupters:
Pattern: S → NP , PP , VP
Subclass: none [cf. Appendix B, (6)]
Count: 82
SampleSentence:
The U.S., along with Britain and Singapore, left the agency when its anti-Western ideology, financial corruption and top leadership got out of hand.

7. Quotations:
Pattern: S → S --> " S , " SINV
Subclass: none [cf. Appendix B, (7)]
Count: 136

```
SampleSentence:
“The SEC has historically paid obeisance to the ideal of a level
playing field,” wrote Clyde S. McGregor of Winnetka, III., in
one of the 92 letters the agency has received since the changes
were proposed Aug. 17.
```

Each class and the range of patterns it covers are listed in descending order of pattern frequency (the number preceding the pattern) in Appendix B.

5.3 Results of the Classification

A summary is given in Table 1. The first column contains the general class and the second column, the more specific subclasses of this general class. The next two columns display the number of occurrences of the class (or subclass) and the percentage of this number to the whole number of effectively classified commas (14,299 ≅ 80% of 17,883), respectively. Column 5 shows the number of unique syntax-patterns assigned to the class (or subclass), and column 6 includes the percentage of this number to the whole number of classified patterns (211 ≅ 11% of 1,978). The last column contains the proportion of the number of commas to the number of patterns for the particular class or subclass, which we call the *stability* of that class or subclass:

$$stability = \frac{number\ of\ commas}{number\ of\ patterns}$$

Class	Subclass	#Commas	%Commas	#Patterns	%Patterns	Stability
Elements in a Series	Words in series	300	2.1%	5	2.4%	60
	Phrases in series	1105	7.7%	27	12.8%	41
	Clauses in series	235	1.7%	7	3.3%	34
	Coordinate clauses	1135	7.9%	13	6.1%	87
	Coordinate adjectives	121	0.9%	4	1.9%	30
	TOTAL	**2896**	**20.3%**	**56**	**26.5%**	**52**
Sentence-initial Elements	Introductory words	423	3.0%	4	1.9%	106
	Introductory phrases	1854	13.0%	17	8.1%	109
	Introductory clauses	602	4.2%	6	2.8%	100
	TOTAL	**2879**	**20.2%**	**27**	**12.8%**	**107**
Sentence-final Elements	Final phrases	387	2.7%	18	8.5%	22
	Final clauses	321	2.2%	13	6.2%	25
	Absolute phrases	14	0.1%	1	0.5%	14
	TOTAL	**722**	**5.0%**	**32**	**15.2%**	**23**
Nonrestrictive Phrases or Clauses	Nonr. phrases	975	6.8%	15	7.1%	65
	Nonr. clauses	1501	10.5%	14	6.6%	107
	TOTAL	**2476**	**17.3%**	**29**	**13.7%**	**85**
Appositives	**TOTAL**	**3738**	**26.1%**	**19**	**9.0%**	**197**
Interrupters	**TOTAL**	**946**	**6.6%**	**35**	**16.6%**	**27**
Quotations	**TOTAL**	**642**	**4.5%**	**13**	**6.2%**	**49**
	GRAND TOTAL	**14299**	**100%**	**211**	**100%**	**68**

Table 1: Results of the Classification[6]

According to Table 1, the most frequent use of comma in the corpus is the setting off of appositives, which is followed by elements in a series and sentence-initial elements, with sentence-final elements and quotations at the end. The most frequent elements listed in a series are

coordinate clauses followed by phrases. Phrases are also dominantly set off by commas as sentence-initial and sentence-final elements. Finally, nonrestrictive clauses delimited by commas were approximately 50% more than nonrestrictive phrases.

The stability measure of a class or subclass, introduced above, requires an explanation. This number shows the average number of commas per syntax-pattern assigned to a class or subclass, which is also a sign of the variety of these patterns for the class in question. The more the number of commas per pattern means the less variety of patterns; i.e., the more stable is the class in question (Bayraktar, 1996).

The most stable class of the use of the comma appeared to be the commas setting off appositives. This is followed by the commas delimiting sentence-initial elements, and nonrestrictive phrases or clauses. Conversely, the most versatile classes turned out to be the commas setting of interrupters and sentence-final elements, meaning the syntax-patterns occurring in these classes are less standardized. The stability of a class shows also the capability of its individual syntax-patterns to be reduced to more general rule-patterns. On the other hand, the calculated stability of the whole corpus, approximately 68, may be viewed as an indicator of the precision and consistency of the parsing and correction procedure applied on that particular corpus. Since this experiment was done with a single corpus, we cannot yet compare this parameter with the stabilities of other corpora.

6. Conclusion

The corpus contained material from a single type of origin: the *Wall Street Journal*, a respected business paper published in the strictest journalistic style. Therefore, this study could be extended by investigating other corpora, containing material from other types of journals or other domains of literacy such as fiction or learned writing, in which the punctuation practice might display variety. It is not difficult to guess that the frequencies given in Section 5.3 would then change and that new classes of uses could appear.

The uses of comma in the corpus were intuitively classified by means of the syntax-patterns in which they occur, each time looking at exactly one sample sentence for each pattern. In other words, it was assumed that all commas appearing in the same syntax-pattern fall into the same class of use, regardless of the sentence in which they occur. This assumption, however, needs to be verified for its degree of validity. For example, it is conceivable that two different uses of comma may have resulted in the same syntax-pattern during the parsing process.

Although the most significant punctuation mark is the comma, other structural marks, especially the colon, semicolon and dash, also deserve investigation. The experience obtained during our work could profitably guide such a study.

In this work, effectively only the 80% of all commas in the corpus were classified according to their uses. This number could be extended to 90%, or even 100%. In this case, the percentage of the abbreviated syntax-patterns that need to be investigated would rise from 11% to 36% and 100%, respectively. As a solution, the unique syntax-patterns could first be reduced to more general rule-patterns (Jones 1996a; Jones 1997). These rule-patterns could then be easily assigned to appropriate classes. Furthermore, the generality and coverage of such rule-patterns could be helpful in the determination of the class of a newly encountered syntax-pattern. A contribution of our work in furthering that of Jones' is establishing the link between syntax-patterns and semantic usage classes.

In fact, with the development of parsers with nearly full coverage in the near future, it may be possible to have punctuation checkers — along with grammar and spell checkers — which will ascertain the correctness or the consistency of the punctuation practice in a given text, according to a specific style.

Notes

[1]Please contact this author (akman@cs.bilkent.edu.tr) for future correspondence.
[2]Structural punctuation marks (Meyer 1987) are those which are conventionally considered as punctuation marks, and which do not set off constructions larger than the sentence or smaller than the syntactic constituents of the sentence (thus no paragraphs or hyphens, for example). Structural marks are a good working category to distinguish from text punctuation such as paragraphs or font changes.
[3]Our work was carried out independently from Jones's work.
[4]Sentence-final elements were omitted by Ehrlich (Ehrlich 1992), except for the case of subordinate clauses and absolute phrases, which he described as individual classes. In the corpus, we encountered sufficiently many examples involving sentence-final verbal phrases, so that it became mandatory to have this class.
[5]The corpus is available as part of the Penn Treebank Release 2 CDROM. (Refer to the web site http://www.ldc.upenn.edu/ for more information.)
[6]In order to arrive at a grand total of 100%, some percentages were slightly rounded off.

Bibliography

Aarts, J. 1995. "Corpus Analysis". In J. Vershueren, J. Ostman, and J. Blommaert (eds.): *Handbook of Pragmatics.* Amsterdam, the Netherlands: John Benjamins Publishing Company, 565–570.

Bayraktar, M. 1996. "Computer Aided Analysis of English Punctuation on a Parsed Corpus: The Special Case of Comma". Master's Thesis. Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey.

Briscoe, T. and J. Carroll. 1995. "Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels". In *Proceedings of International Workshop on Parsing Technologies.* Prague, Czech Republic, 48–58.

Dale, R. 1991. "Exploring the Role of Punctuation in the Signalling of Discourse Structure". In *Proceedings of a Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI,* Technical University of Berlin, Berlin, Germany, 110–120.

Ehrlich, E. 1992. *Theory and Problems of Punctuation, Capitalization, and Spelling.* Hong Kong: McGraw-Hill. 2nd edition.

Francis, W. N. and H. Kucera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston, Massachusetts: Houghton Mifflin.

Garside, R., G. Leech and G. Sampson (eds.). 1987. *The Computational Analysis of English.* London: Longman.

Jarvie, G. 1992. *Chambers Punctuation Guide.* Edinburgh, UK: Chambers.

Jones, B. 1994. "Exploring the Role of Punctuation in Parsing Natural Language". In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94).* Kyoto, Japan, 421–425.

Jones, B. 1996a. "Towards Testing the Syntax of Punctuation". In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL).* Santa Cruz, California, 363–365.

Jones, B. 1996b. "Towards a Syntactic Account of Punctuation". In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).* Copenhagen, Denmark, 604–609.

Jones, B. 1997. "What's the Point? A (Computational) Theory of Punctuation". PhD Thesis. Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

Karlsson, F., A. Voutilainen, J. Heikkila, and A. Antilla (eds.). 1994. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Berlin, Germany: Mouton de Gruyter.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz. 1992. “Building a Large Annotated Corpus of English: the Penn Treebank”. University of Pennsylvania, Philadelphia, Pennsylvania.

Marcus, M., G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz , and B. Schasberger. 1994. “The Penn Treebank: Annotating Predicate Argument Structure”. In *Proceedings of the Human Language Technology Workshop*. San Mateo, California: Morgan Kaufmann, 27–38.

Meyer, C. F. 1987. *A Linguistic Study of American Punctuation.* New York: Peter Lang Publishing Co.

Nunberg, G. 1990. *The Linguistics of Punctuation.* Stanford, California: CSLI Publications.

Paxson, W. C. 1986. *The Mentor Guide to Punctuation.* New York: Mentor Books.

Santorini, B. 1991. “Bracketing Guidelines for the Penn Treebank Project”. Draft. University of Pennsylvania, Philadelphia, Pennsylvania.

Say, B. and V. Akman. 1996a. “An Information-based Treatment of Punctuation in Discourse Representation Theory”. In *Second International Conference on Mathematical Linguistics.* Tarragona, Spain. An extended version to appear in a book to be published by John Benjamins Publishing Company.

Say, B. and V. Akman. 1996b. “Information-based Aspects of Punctuation”. In *Proceedings of the First International Workshop on Punctuation in Computational Linguistics.* Santa Cruz, California, 49–56. Available from Human Communication Research Centre, University of Edinburgh, UK. http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html

Say, B. and V. Akman. 1997. “Current Approaches to Punctuation in Computational Linguistics”. Accepted for publication in *Computers and the Humanities.*

White, M. 1995. “Presenting Punctuation”. In *Proceedings of the Fifth European Workshop on Natural Language Generation.* Leiden, the Netherlands, 107–125.

Appendix A Syntactic Tag-set of Penn Treebank

	Tag	Description
1.	ADJP	Adjective Phrase
2.	ADVP	Adverb Phrase
3.	NP	Noun Phrase
4.	PP	Prepositional Phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by subordinating conjunction
7.	SBARQ	Direct question introduced by *wh*-word or *wh*-phrase
8.	SINV	Declarative sentence with subject-aux inversion
9.	SQ	Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase
10.	VP	Verb phrase
11.	WHADVP	*Wh*-adverb phrase
12.	WHNP	*Wh*-noun phrase
13.	WHPP	*Wh*-prepositional phrase
14.	X	Constituent of unknown or uncertain category
	Null Elements	
1.	*	'Understood' subject of infinitive or imperative
2.	0	Zero variant of *that* in subordinate clauses
3.	T	Trace—marks position where moved *wh*-constituent is interpreted
4.	NIL	Marks position where preposition is interpreted in pied-piping contexts

Table 2: Syntactic Tag-Set of Penn Treebank (Marcus et al. 1992:10)

Appendix B Classified Syntax-patterns

```
(1) Elements in a Series
------------------------------------------------------------------------------
     (1.1) Words in a Series
          201    NP --> *** , ***
          28     NP --> *** , *** ,
          25     NP --> *** , *** and ***
          24     NP --> *** , *** , ***
          22     NP --> *** , *** , *** and ***
------------------------------------------------------------------------------
     (1.2) Phrases in a Series
          266    NP --> NP , NP and NP
          158    NP --> NP , NP , NP and NP
          88     NP --> NP , NP , and NP
          65     VP --> VP , and VP
          65     NP --> NP , and NP
          51     VP --> VP , *** VP
          45     NP --> NP , NP , NP , NP and NP
          36     NP --> NP , NP , NP , and NP
          36     NP --> NP , NP , NP , NP , NP and NP
          32     NP --> NP , NP , NP
          29     VP --> VP , VP and VP
          26     NP --> NP , and NP ,
          25     NP --> NP , NP , NP , NP , NP , NP and NP
          19     PP --> PP , and PP
```

```
          18     NP --> NP , NP , NP , NP , NP , NP , and NP
          18     NP --> NP , NP , NP , NP
          14     VP --> VP , VP , and VP
          14     VP --> VP , VP
          13     PP --> PP , PP
          13     NP --> NP , NP or NP
          12     NP --> NP , NP , and NP ,
          12     NP --> NP , NP , NP , NP , and NP
          12     NP --> NP , NP , NP , NP , NP , NP and NP ,
          11     X --> X , and X
          10     VP --> VP , VP , VP and VP
          9      X --> X , *** X
          8      VP --> VP , VP , VP , VP , and VP
-------------------------------------------------------------------------------
     (1.3) Clauses in a Series
          117    S --> NP VP , S
          49     S --> NP VP , SBAR
          21     S --> S , SINV .
          19     SBAR --> SBAR , and SBAR
          11     X --> VP , and X
          10     S --> *** S , S
          8      X --> X , *** VP
-------------------------------------------------------------------------------
     (1.4) Coordinate Clauses in a Series
          347    S --> S , S
          223    S --> S , and S
          166    S --> S , *** S
          124    S --> S , S .
          73     S --> S , *** S .
          54     S --> S , and S .
          48     S --> S , SINV
          30     S --> S , NP VP .
          20     S --> S , S , and S
          14     S --> `` S , and S , ''
          12     SBAR --> SBAR , *** SBAR
          12     S --> `` S , *** S , ''
          12     S --> S , S and S
-------------------------------------------------------------------------------
     (1.5) Coordinate Adjectives
          82     ADJP --> *** , ***
          18     NP --> *** , *** PP
          12     ADJP --> *** , *** , ***
          9      ADJP --> ADJP , ADJP
-------------------------------------------------------------------------------
(2) Sentence-initial Elements
-------------------------------------------------------------------------------
     (2.1) Introductory Words
          275    S --> *** , S
          129    S --> *** , S .
          10     S --> “ *** , S , ”
          9      S --> *** , NP VP
-------------------------------------------------------------------------------
     (2.2) Introductory Phrases
          952    S --> PP , S
          357    S --> PP , S .
          103    S --> NP , S
          89     S --> ADVP , S
          61     S --> VP , S
          56     S --> NP , S .
          53     S --> *** PP , S
          32     S --> VP , S .
          28     S --> PP NP , S
          21     S --> ADJP , S
          20     S --> PP , NP VP
          20     S --> ADVP , S .
          15     S --> *** PP , S .
          14     S --> NP , PP , S
          11     S --> “ PP , S
          11     S --> PP , NP *** VP
          11     S --> *** NP , S
-------------------------------------------------------------------------------
     (2.3) Introductory Clauses
          395    S --> SBAR , S
          123    S --> SBAR , S .
          35     S --> *** SBAR , S
          22     S --> “ SBAR , S , ”
          14     S --> “ SBAR , S
          13     SBAR --> *** S , S
-------------------------------------------------------------------------------
(3) Sentence-final Elements
-------------------------------------------------------------------------------
     (3.1) Final Phrases
          42     VP --> *** PP , ADVP
          35     S --> NP VP , VP
          34     S --> S , VP .
          33     S --> NP VP , PP
          31     VP --> *** PP , PP
          28     VP --> *** NP , VP
```

```
          27     VP --> *** NP PP , PP
          21     VP --> *** NP , ADVP
          18     S --> NP *** VP , PP
          17     VP --> *** NP PP , VP
          16     S --> NP *** VP , VP
          15     VP --> *** PP , VP
          14     VP --> *** NP PP , ADVP
          14     VP --> *** , PP
          12     VP --> *** PP , ADVP ,
          11     S --> NP VP , ADVP
          10     S --> S , PP
          9      VP --> *** NP PP , NP
-------------------------------------------------------------------------------
     (3.2) Final Clauses
          57     S --> NP *** VP , S
          45     VP --> *** , S
          44     S --> S , SBAR
          38     VP --> *** NP , S
          36     S --> NP *** VP , SBAR
          20     VP --> *** NP PP , S
          16     VP --> *** PP , SBAR
          16     VP --> *** NP , SBAR
          12     VP --> *** PP , S
          10     S --> S , SBAR .
          9      S --> NP VP ,
          9      S --> NP *** VP , S .
          9      S -->  *** VP , S
-------------------------------------------------------------------------------
     (3.3) Absolute Phrases
          14     S --> NP VP , NP
-------------------------------------------------------------------------------
(4) Nonrestrictive Phrases or Clauses
-------------------------------------------------------------------------------
     (4.1) Nonrestrictive Phrases
          348    NP --> NP , VP ,
          190    NP --> NP , VP
          104    NP --> NP , ADJP ,
          62     VP --> *** NP , PP
          55     NP --> NP , ADJP
          46     NP --> NP , PP ,
          26     S --> S , VP
          26     NP --> *** , PP
          25     NP --> *** PP , PP
          24     NP --> NP , PP
          19     NP --> NP ADJP , VP
          16     NP --> *** , ADJP ,
          14     NP --> *** PP , PP ,
          10     NP --> NP , VP , SBAR
          10     NP --> NP , ADVP ,
-------------------------------------------------------------------------------
     (4.2) Nonrestrictive Clauses
          570    NP --> NP , SBAR
          430    NP --> NP , SBAR ,
          170    NP --> NP , SBARQ ,
          76     NP --> NP , ADVP
          36     NP --> *** , SBAR
          35     NP --> “ *** , ”
          34     NP --> *** , SBAR ,
          33     NP --> NP , S
          26     NP --> NP , S ,
          26     NP --> *** , SBARQ ,
          20     S --> NP , *** VP
          19     NP --> NP , SBARQ
          14     NP --> *** “ *** , ”
          12     SINV --> VP NP , S
-------------------------------------------------------------------------------
(5) Appositives
-------------------------------------------------------------------------------
     1880   NP --> NP , NP ,
     1274   NP --> NP , NP
     144    NP --> NP , or NP ,
     139    NP --> NP , or NP
     47     NP --> NP , *** NP
     42     NP --> NP , *** NP ,
     40     NP --> NP , NP , PP
     33     NP --> NP , NP , NP ,
     28     NP --> NP , NP , SBAR
     18     NP --> *** NP , NP , ***
     13     NP --> NP NP ,
     13     NP --> NP , NP *** NP
     11     NP --> NP , ***
     10     NP --> NP ( NP , NP )
     10     NP --> *** , NP , ***
     9      NP --> NP , NP , SBARQ ,
     9      NP --> NP , NP , SBAR ,
     9      NP --> *** , NP
     9      NP --> ( NP , NP )
-------------------------------------------------------------------------------
```

(6) Interrupters

```
     82     S --> NP , PP , VP
     76     NP --> *** , PP ,
     66     S --> NP , S , VP
     58     S --> *** , S , S
     54     S --> NP , PP , *** VP
     48     S --> NP , *** , VP
     40     S --> PP , S , S
     40     S --> *** , PP , S
     32     S --> NP , S , *** VP
     30     S --> NP , PP , VP .
     26     S --> PP , PP , S
     26     S --> NP , *** , VP .
     25     SBAR --> *** , S
     22     S --> NP , VP
     22     S --> NP , PP , *** VP .
     22     S --> NP , *** , *** VP
     22     S --> *** , SBAR , S
     18     VP --> *** , *** , SBAR
     18     S --> PP , S , S .
     18     S --> , PP , S
     16     S --> SBAR , S , S .
     16     S --> SBAR , S , S
     16     S --> S , S , *** S
     16     S --> PP , *** , S .
     16     S --> *** , S , S .
     14     VP --> *** , PP , NP
     14     VP --> *** , NP
     14     S --> PP , SBAR , S
     14     S --> NP , *** , *** VP .
     13     VP --> *** , SBAR
     12     VP --> *** , PP , SBAR
     12     S --> NP , S , VP .
     10     VP --> *** , *** , NP
     10     S --> NP , S , *** VP .
     8      VP --> *** NP , PP , PP
```

(7) Quotations

```
     136    S --> “ S , ” SINV
     105    S --> “ S , ” S
     101    S --> “ NP VP , ”
     77     S --> “ NP *** VP , ”
     35     S --> “ S , ” SINV .
     34     S --> S , “ SINV
     31     S --> “ S , ” S .
     29     SINV --> “ S , ” VP NP .
     28     S --> S , “ S
     23     S --> “ S , ” NP VP .
     16     VP --> *** , “ S
     16     S --> “ PP , S , ”
     11     S --> NP “ *** VP , ”
```