Towards Testing the Syntax of Punctuation

Bernard Jones*
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
United Kingdom
bernie@cogsci.ed.ac.uk

Abstract

Little work has been done in NLP on the subject of punctuation, owing mainly to a lack of a good theory on which computational treatments could be based. This paper described early work in progress to try to construct such a theory. Two approaches to finding the syntactic function of punctuation marks are discussed, and procedures are described by which the results from these approaches can be tested and evaluated both against each other as well as against other work. Suggestions are made for the use of these results, and for future work.

1 Background

The field of punctuation has been almost completely ignored within Natural Language Processing, with perhaps the exception of the sentence-final full-stop (period). This is because there is no coherent theory of punctuation on which a computational treatment could be based. As a result, most contemporary systems simply strip out punctuation in input text, and do not put any marks into generated texts.

Intuitively, this seems very wrong, since punctuation is such an integral part of many written languages. If text in the real world (a newspaper, for example) were to appear without any punctuation marks, it would appear very stilted, ambiguous or infantile. Therefore it is likely that any computational system that ignores these extra textual cues will suffer a degradation in performance, or at the very least a great restriction in the class of linguistic data it is able to process.

Several studies have already shown the potential for using punctuation within NLP. Dale (1991) has shown the benefits of using punctuation in the fields of discourse structure and semantics, and Jones (1994) has shown in the field of syntax that using a grammar that includes punctuation yields around two orders of magnitude fewer parses than one which does not. Further work has been carried out in this area, particularly by Briscoe and Carroll (1995), to show more accurately the contribution that usage of punctuation can make to the syntactic analysis of text.

The main problem with these studies is that there is little available in terms of a theory of punctuation on which computational treatments could be based, and so they have somewhat ad hoc, idiosyncratic treatments. The only account of punctuation available is that of Nunberg (1990), which although it provides a useful basis for a theory is a little too vague to be used as the basis of any implementation.

Therefore it seems necessary to develop a new theory of punctuation, that is suitable for computational implementation. Some work has already been carried out, showing the variety of punctuation marks and their orthographic interaction (Jones, 1995), but this paper describes the continuation of this research to determine the true syntactic function of punctuation marks in text.

There are two possible angles to the problem of the syntactic function of punctuation: an observational one, and a theoretical one. Both approaches were adopted, in order to be be able to evaluate the results of each approach against those of the other, and in the hope that the results of both approaches could be combined. Thus the approaches are described one after the other here.

2 Corpus-based Approach

The best data source for observation of grammatical punctuation usage is a large, parsed corpus. It ensures a wide range of real language is covered, and because of its size it should minimise the effect of any

* This work was carried out under an award from the (UK) ESRC. Thanks are also due to Lex Holt, Henry Thompson, Ted Briscoe and anonymous reviewers.

errors or idiosyncrasies on the part of editors, parsers and transcribers. Since these corpora are almost all hand-produced, some errors and idiosyncrasies are inevitable — one important part of the analysis is therefore to identify possible instances of these, and if they are clear, to remove them from the results.

The corpus chosen was the Dow Jones section of the Penn Treebank (size: 1.95 million words). The bracketings were analysed so that each node with a punctuation mark as its immediate daughter is reported, with its other daughters abbreviated to their categories, as in (1) – (3).

(1) [NP [NP the following] :] $\Longrightarrow$ [NP = NP :]

(2) [S [PP In Edinburgh] , [S ...]] $\Longrightarrow$ [S = PP , S]

(3) [NP [NP Bob] , [NP ...) ,] $\Longrightarrow$ [NP = NP , NP ,]

In this fashion each sentence was broken down into a set of such category-patterns, resulting in a set of different category-patterns for each punctuation symbol, which were then processed to extract the underlying rule patterns which represent all the ways that punctuation behaves in this corpus, and are good indicators of how the punctuation marks might behave in the rest of language.

There were 12,700 unique category-patterns extracted from the corpus for the five most common marks of point punctuation, ranging from 9,320 for the comma to 425 for the dash. These were then reduced to just 137 underlying rule-patterns for the colon, semicolon, dash, comma, full-stop.

Even some of these underlying rule-patterns, however, were questionable since their incidence is very low (maybe once in the whole corpus) or their form is so linguistically strange so as to call into doubt their correctness (possibly idiosyncratic misparses), as in (4).

(4) [ADVP = PP , NP]

Therefore all the patterns were checked against the original corpus to recover the original sentences. The sentences for patterns with low incidence and those whose correctness was questionable were carefully examined to determine whether there was any justification for a particular rule-pattern, given the content of the sentence.

For example, the NP=NP:VP rule-pattern was removed since all the verb phrases occurring in this pattern were imperative ones, which can legitimately act as sentences (5). Therefore instances of this rule application were covered by the NP=NP:S rule-pattern. A detailed account of the removal of idiosyncratic, incorrect and exceptional rule-patterns, with justifications, is reported in (Jones, 1996).

(5) [...] the show's distributor, Viacom Inc, is giving an ultimatum: either sign new long-term commitments to buy future episodes or risk losing "Cosby" to a competitor.

After this further pruning procedure, the number of rule-patterns was reduced to just 79, more than half of which related to the comma. It was now possible to postulate some generalisations about the use of the various punctuation marks from this reduced set of rule-patterns.

These generalised punctuation rules, described in more detail in (Jones, 1996), are given below for colons (6), semicolons (7), full-stops (8), dashes (9,10), commas (11), basic quotation(12) and stress-markers (13–15).

(6) $\mathcal{X} = \mathcal{X} : \{$ NP | S | ADJP $\}$ $\quad \mathcal{X}$:{NP, S}

(7) $\mathcal{S} = \mathcal{S} \; ; \mathcal{S}$ $\quad \mathcal{S}$:{NP, S, VP, PP}

(8) $\mathbf{T} = * \; .$

(9) $\mathcal{D} = \mathcal{D} - \mathcal{D} -$ $\quad \mathcal{D}$:{NP, S, VP, PP, ADJP}

(10) $\mathcal{E} = \mathcal{E} - \{$ NP | S | VP | PP $\} -$ $\quad \mathcal{E}$:{NP, S }

(11) $\mathcal{C} = \mathcal{C} \; , *$ $\quad \mathcal{C}$:{NP, S, VP, PP, ADJP, ADVP}
$\mathcal{C} = * \; , \mathcal{C}$

(12) $\mathcal{Q} =$ " $\mathcal{Q}$ " $\quad \mathcal{Q} : *$

(13) $\mathcal{Z} = \mathcal{Z} \; ?$ $\quad \mathcal{Z} : *$

(14) $\mathcal{Y} = \mathcal{Y} \; !$ $\quad \mathcal{Y} : *$

(15) $\mathcal{W} = \mathcal{W} \ldots$ $\quad \mathcal{W} : *$

3 A Theoretical Approach

The theoretical starting point is that punctuation seems to occur at a phrasal level, i.e. it comes immediately before or after a phrasal level lexical item (e.g. a noun phrase). However, this is a rather general definition, so we need to examine the problem more exactly.

Punctuation could occur adjacent to any complex structure. However, we want to prevent occurrences such as (16). Conversely, punctuation could only occur adjacent to maximal level phrases (e.g. NP, VP). However, this rules out correct cases like (17).

(16) The, new toy ...

(17) He does, surprisingly, like fish.

Clearly we need something stricter than the first approach, but more relaxed than the second. The notion of headedness seems to be involved, so we can postulate that only non-head structures can have punctuation attached. This system still does not rule out examples like (18) however, so

further refinement is necessary. The answer seems to be to look at the level of head daughter and mother categories under X-bar theory (Jackendoff, 1977). Attachment of punctuation to the non-head daughter only seems to be legal when mother and head-daughter are of the same bar level (and indeed more often than not they are identical categories), regardless of what that bar level is.

(18) the, big, man

From this theoretical approach it appears that punctuation could be described as being adjunctive (i.e. those phrases to which punctuation is attached serve an adjunctive function). Furthermore, conjunctive uses of punctuation (19,20), conventionally regarded as being distinct from other more grammatical uses (the adjunctive ones), can also be made to function via the theoretical principles formed here.

(19) dogs, cats, fish and mice

(20) most, or many, examples ...

4 Testing — Work in Progress

The next stage of this research is to test the results of both these approaches to see if they work, and also to compare their results. Since the results of the two studies do not seem incompatible, it should prove possible to combine them, and it will be interesting to see if the results from using the combined approaches differ at all from the results of using the approaches individually. It will also be useful to compare the results with those of studies that have a less formal basis for their treatments of punctuation, e.g. (Briscoe and Carroll, 1995).

For this reason the best way to test the results of these approaches to punctuation's role in syntax is to incorporate them into otherwise identical grammars and study the coverage of the grammars in parsing and the quality and accuracy of the parses. For ease of comparison with other studies, the best parsing framework to use will be the Alvey Tools' Grammar Development Environment (GDE) (Carroll et al., 1991), which allows for rapid prototyping and easy analysis of parses. The corpus of sentences to run the grammars over should ideally be large, and consist mainly of real text from external sources. To avoid dealing with idiosyncratic tagging of words, and over-complicated sentences, we shall follow Briscoe and Carroll (1995) rather than Jones (1994) and use 35,000 prepared sentences from the Susanne corpus rather than using the Spoken English Corpus.

5 Further Work

The theoretical approach not only seems to confirm the reality of the generalised punctuation rules derived observationally, since they all seem to have an adjunctive nature, but it also gives us a framework with which those generalised rules could be included in proper, linguistically-based, grammars. Results of testing will show whether either of the approaches are better on their own, and how they perform when they are combnined, and will, hopefully, show an improvement in performance over the ad-hoc methods used previously. The development of a theory of punctuation can then progress with investigations into the semantic function of punctuation marks, to ultimately form a theory that will be of great use to the NLP community.

References

Edward Briscoe and John Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the ACL/SIGPARSE 4th International Workshop on Parsing Technologies*, pages 48–58, Prague

John Carroll, Edward Briscoe and Claire Grover. 1991. A Development Environment for Large Natural Language Grammars. Technical Report 233, Cambridge University Computer Laboratory.

Robert Dale. 1991. Exploring the Role of Punctuation in the Signalling of Discourse Structure. In *Proceedings of the Workshop on Text Representation and Domain Modelling*, pages 110–120, Technical University Berlin.

Ray Jackendoff. 1977. X-bar Syntax: A Study of Phrase Structure. MIT Press, Cambridge, MA.

Bernard Jones. 1994. Exploring the Role of Punctuation in Parsing Real Text. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 421–425, Kyoto, Japan, August.

Bernard Jones. 1995. Exploring the Variety and Use of Punctuation. In *Proceedings of the 17th Annual Cognitive Science Conference*, pages 619–624, Pittsburgh, Pennsylvania, July.

Bernard Jones. 1996. Towards a Syntactic Account of Punctuation. To appear in *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, August.

Geoffrey Nunberg. 1990. The Linguistics of Punctuation. CSLI Lecture Notes **18**, Stanford, California.