The Syntax and Semantics of Punctuation and its Use in Interpretation

Ted Briscoe (ejb@cl.cam.ac.uk)
Computer Laboratory
Cambridge University
Pembroke St. Cambridge, CB2 3QG, UK

Abstract

In this paper, I argue for a declarative description of the syntax and semantics of punctuation marks (in English) couched in a feature/unification-based phrase structure formalism, describe how Nunberg's (1990) syntactic analysis of punctuation can be combined with Dale's (1991) suggested semantic analysis within this framework, and present experimental evidence that 1) the resulting text grammar should be interleaved with the lexical grammar to efficiently resolve ambiguity and 2) that the role of punctuation is not only to resolve syntactic and semantic ambiguity in the lexical grammar but also to encode and facilitate purely discourse relational links between text units in text sentences.

1 The Syntax of Punctuation

Nunberg (1990) argues that the punctuation system is, firstly, systematic and linguistic, in the sense that it obeys principles and constraints of a type familiar from work on other linguistic subsystems such as phonology or syntax, and secondly, it is a separate linguistic subsystem not reducible to principles of (prosodic) phonology or syntax. Nunberg develops a partial grammar of English text sentences which incorporates many constraints that (ultimately) restrict the syntactic and semantic/pragmatic interpretation of the text. For example, one such constraint is that textual adjunct clauses introduced by colons scope over following punctuation, as (1a) illustrates; another is that textual adjuncts introduced by dashes cannot intervene between a bracketed adjunct and the textual unit to which it attaches, as in (1b).

(1) a *He told them his reason: he would not renegotiate his contract, but he did not explain to the team owners. (vs. but would stay)

b *She left – who could blame her – (during the chainsaw scene) and went home.

Nunberg's analysis, therefore, incorporates constraints on grammatical sequences of punctuation marks and on the manner in which such punctuation serves to hierarchically structure text. Nunberg outlines a procedural, derivational account of 'text grammar'. I have developed a succinct, declarative grammar in a feature/unification-based phrase structure formalism (Briscoe, 1994). This grammar captures the bulk of the text-sentential constraints described by Nunberg with a grammar which compiles into 26 rules in a formalism which employs a syntactic variant of definite clause grammar rules incorporating (iterative) Kleene operators (Briscoe *et al.* , 1987).[1] The grammar treats all words uniformly and groups them iteratively into flat textual units according to surounding punctuation; for example, the text sentence in (2a) receives four analyses of which that in (2b) is correct.[2]

(2) a A Yale historian, writing a few years ago in the Yale review, said: "we in New England have long since segregated our children".

b
```
(T/txt-sc1/--
 (T/t_ta_t (T/w a_wd Yale_wd historian_wd)
  (Ta/comma _pco
   (T/w write_wd a_wd few_wd year_wd ago_wd in_wd the_wd
                                           Yale_wd review_wd)
   _pco)
  (T/t_ta-cl_t (T/w say_wd)
   (Ta/colon _pcl
    (T/txt-sc1/--
     (T/w we_wd in_wd new_wd England_wd have_wd
       long_wd since_wd segregate_wd our_wd child_wd))))))
```

The skeletal description shown in (2c) uses mnemonic rule names, based on Nunberg's description such as text unit (T), text adjunct (Ta) and text-sentence (T/txt-sc) which indicate that this text sentence consists of two text units (*a Yale historian*, *say: ...*) with an intervening text adjunct delimited by commas (*writing... ago*) and the final text unit contains a text adjunct introduced by a colon (*we... children*). Such analyses are a useful first step in interpretation as they clearly demarcate some of the syntactic boundaries in the text sentence and thus constrain the possible analyses that a syntactic parser can assign. In other cases, this step is indispensable because it also identifies the units for which a syntactic analysis should, in principle, be found; for example, in (3a), the absence of dashes would mislead a parser into seeking a syntactic relationship between *three* and the following names, whilst in fact there is only a discourse relation of elaboration between this text adjunct and pronominal *three*. The correct analysis in (3b) – one of six found by the system – makes this relationship clear.

[1] I do not deal with sentence-final periods or with quotation since these are paragraph-level punctuation markers. I also do not (yet) deal with some cases of 'promotion' of commas to semi-colons, since the phenomenon is rare in the textual corpora I have examined.

[2] These and the most following examples are drawn from the Susanne corpus (Sampson, 1994), the Spoken English Corpus (SEC, Taylor & Knowles, 1988), or the Lancaster Oslo-Bergen Corpus (LOB, Garside *et al.* , 1987).

(3) a The three – Miles J. Cooperman, Sheldon Teller, and Richard Austin – and eight other defendants were charged in six indictments with conspiracy to violate federal narcotic law.

b
```
(T/txt-sc1/--
  (T/t_ta-da+_t (T/w the_wd three_wd)
   (Ta/dash+ _pda
    (T/t_co_t (T/w Miles_wd J_wd Cooperman_wd) _pco
     (T/t_co_t (T/w Sheldon_wd Teller_wd) _pco
        (T/w and_wd Richard_wd Austin_wd)))
    _pda)
   (T/w and_wd eight_wd other_wd defendant_wd be_wd charge_wd
     in_wd six_wd indictment_wd with_wd conspiracy_wd to_wd
     violate_wd federal_wd narcotic_wd law_wd)))
```

The rules of the text grammar divide into three groups: those introducing text-sentences, those defining text adjunct introduction and those defining text adjuncts. An example of each type of rule is given below:

a) T/txt-sc1 : TxtS $\rightarrow$ (Tu[+sc])* Tu[-sc] (+pex|+pqu)

b) Ta/dash- : Tu[-sc] $\rightarrow$ T[-sc, -cl, -da] Ta[+da, -bal]

c) T/t_ta-bal_t : Ta[+da, -bal] $\rightarrow$ +pda Tu[-sc, -da]

These rules are phrase structure rewrite rule schemata employing standard operators, such as Kleene star, optionality and disjunction, preceded by a mnemonic name. Non-terminal categories are text sentences, units or adjuncts which carry features mostly representing the punctuation marks which occur as daughters in the rules (e.g. +sc represents presence of a semi-colon marker), whilst terminal punctuation categories are represented as +pxx (e.g. +pda represents a dash). For example, a) states that a text sentence can contain zero or more text units (with a semi-colon at their right boundary) followed by a text unit without the semi-colon, optionally followed by a question mark or exclamation mark. Features are unified between categories at parse time and serve to enforce constraints on presence/absence of marks and also enforce some scope constraints between them. For example, b) states that a text unit not containing a semi-colon can consist of a text unit or adjunct not containing dashes, colons or semi-colons followed by a text adjunct introduced by a dash. This type of 'unbalanced' text adjunct can only be expanded out by c) which states that it consists of a single opening dash followed by a text unit which doesn't itself contain dashes or semi-colons. The effect of the features on the first daughter of b) is to enforce dash adjuncts to have lower precedence and narrower scope than colons or semi-colons and to block interpretations of multiple dashes as sequences of 'unbalanced' adjuncts.

Nunberg (1990) invokes rules of (point) absorption which delete punctuation marks (inserted according to a simple context-free grammar) when adjacent to other 'stronger' punctuation marks. For instance, he treats all dash interpolated text adjuncts as underlyingly balanced, but allows a rule of point absorption to convert (4a) into (4b).

(4) a *Max fell – John had kicked him –.

b Max fell – John had kicked him.

The various rules of absorption introduce procedurality into the grammatical framework and require the positing of underlying forms which are not attested in text. For this reason, I make no use of such rules but rather capture their effects through propagation of featural constraints in parse trees. For instance, (4a) is blocked by including distinct rules for the introduction of balanced and unbalanced text adjuncts and only licensing the latter text sentence finally.

2 The Semantics of Punctuation

Dale (1991) proposes, expanding suggestions of Nunberg (1990:91f), that the semantics of punctuation be treated in terms of rhetorical or discourse relations (e.g. Hobbs, 1985) such as narrative/continuation, explanation, elaboration, parallel, contrast, and so forth (see Asher, 1993 for a detailed taxonomy). For example, the simple (artificial) examples in (5a,b,c,d), there is a discourse relation of explanation or elaboration, indicating a causative relationship between the second and first events described in the narrative. But in no case, are the two clauses describing the events related syntactically (by relations such as subject, adjunct or their semantic analogues).

(5) a Max fell. John had kicked him.

b Max fell; John (had) kicked him.

c Max fell – John (had) kicked him.

d Max fell – John (had) kicked him – and he died.

Dale points out that punctuation marks underdetermine the discourse relations obtaining between events denoted by text units. Nevertheless, it may be possible to identify broad constraints on such relations encoded by punctuation. For example, coordinating (e.g. narrative/continuation) and subordinating (e.g. explanation) discourse relations are often distinguished. Intuitively, in (5a) the period is compatible with either class and encodes no more than that some such relation holds, hence the obligatory pluperfect tense marker *had*. On the other hand, semi-colons and dashes appear to constrain the choice to the subordinating relations. Lee (1995) has explored this hypothesis by analysing examples from textual corpora. She argues that if we distinguish the use of semi-colons as a marker of conjuncts in syntactic coordination (where replacement by a comma is always possible if not always felicitous, see Nunberg, 1990:59f and elsewhere on promotion rules) then other uses of semi-colons do correlate with subordinating discourse relations. Similarly, Lee argues that colons, brackets, matched delimiting commas and dashes all correlate with subordinating relations. In addition, it seems that brackets are further restricted to 'digressive' relations, such as elaboration, and there are probably further such (absolute or probabilistic) constraints yet to be uncovered.

The analysis of the semantics and semantic effects of punctuation has only just begun and there are clearly other phenomena to address; for example, 'tone' indicators such as question and exclamation marks serve to alter or emphasise grammatical mood and information (topic, comment, focus) structure and can apply text sentence internally; anaphoric links between sentence-internal discourse subordinated textual adjuncts and other discourse 'segments' probably follow discourse structure (e.g. Grosz and Sidner, 1986) rather than obeying syntactic constraints like c-command; and so forth. The examples in (6a,b,c,d,e) illustrate some of these phenomena.

(6) a We may have grown accustomed to asking only – where is it this time? which service? what rank of officer? and have they taken over the radio station?

b Conditions for factory workers have improved?

c King James became associated with a Bible translation – the Authorized Version (which was never actually authorized!).

d *The woman who he_i really likes kissed Kim_i last night.

e ?Sandy – who he_i really likes – kissed Kim_i last night.

However, whatever the outcome of such investigations it is clear that the text grammar must allow for the incorporation of semantic rules and must integrate these with the semantic rules of the lexical grammar. Lee (1995) adds semantic rules to the purely syntactic text grammar described in Briscoe (1994) and integrates the result with the wide-coverage lexical grammar described by Grover *et al.* (1993), developed in the same formalism. The formalism supports rule-to-rule mapping from a syntactic to a semantic representation using beta reduction over formulae of a typed lambda calculus (in the style of Montague Grammar). Lee treats discourse relations as binary relations on propositions (identified and analysed by the lexical grammar) so that an example like (7a) receives the (simplified) analysis in (7b).

(7) a The host was blushing – Kim had apologized.

b

SubDR(The(x),Host(x),Blush(x)),(Apologize(kim))

The introduction of the SubDR relation is achieved simply by associating the syntactic rule which introduces the text adjunct with a semantic rule which applies this relation to the semantic interpretation of the immediate constituents:

Ta/dash- : Tu[-sc] $\rightarrow$ T Ta[+da, -bal] : λ(Q) SubDR(T$'$(Q), Ta$'$(Q))

In fact, this approach is not quite adequate because there are cases where the adjunct does not scope over the entire proposition expressed by the antecedent text unit: in *Kim made the discovery – Lee was the abbot* the text adjunct elaborates the discovery. This type of complication, though, can be dealt with in the same framework by a) ensuring that adjuncts can scope over phrasal and lexical constituents as well as clauses, and b) modifying the semantics so that SubDR relates individual variables denoting either events or other sorts of (discourse) referents in the semantic representation. Whether this type of extension can be reconciled with the semantics of subordinating discourse relations is a matter for further research.

3 Integration of Text and Lexical Grammar

Nunberg (1990) advocates loose coupling of textual and lexical grammars on the basis that text grammar does not reduce to lexical syntactic analysis. However, considerations of semantic analysis and also of efficient resolution of text grammatical ambiguity militate against this approach.

In Nunberg's and my analysis commas can function as syntactic separators in constructions such as coordination, or as textual delimiters of textual adjuncts. These uses cannot be resolved without access to (at least) the syntactic context of occurrence. The example in (8) contains eleven commas, a colon and an opening and closing bracket.

(8) Those three other great activities of the Persian[1], the bath[2], the teahouse[3], and the zur khaneh (the latter a kind of club in which a leader and a group of men in an octagonal pit move through a rite of calisthenics[4], dance[5], chant poetry[6], and music)[7], do not take place in buildings to which entrance tickets are sold[8], but some of them occupy splendid examples of Persian domestic architecture: long[9], domed[10], chalk-white rooms with dais of turquoise tiling[11], their end walls cut through to the orchard and the sky by open arches.

The bracketed textual adjunct and the colon adjunct provide some restriction on the relative scope of the commas. Nevertheless, this example has thousands of analyses in terms of punctuation alone. When we examine the syntactic context of these commas though, it is easy to see that [2-6] and [8] function as coordination separators with scope defined by the coordination construction, and [9-10] separate adjectival premodifiers. By contrast [1] until [7] and [11] until the period delimit textual adjuncts containing elaborative material. Whilst recognising these latter commas as delimiters requires, ultimately, recognition of the elaborative nature of the enclosed material, it is comparatively easy to determinately recognise the others as separators by integrating their recognition with that of the syntactic constructions which utilise commas in this way.

The integration of textual and lexical grammars is straightforward and remains modular: the text grammar is 'folded into' the lexical grammar, as text categories and syntactic categories use disjoint sets of features they can be merged and features can propogate according to independent principles (see Briscoe, 1994). The text grammar rules are represented as left- or right- branching rules of 'Chomsky-adjunction' to lexical or phrasal constituents. For example, the simplified rule for combining NP appositional or parenthetical text adjuncts is N2[+ta] $\rightarrow$ H2 Ta[+bal] which says that a NP containing a textual adjunct consists of a head NP followed by a textual adjunct with balanced delimiters (dashes, brackets or commas). If a textual adjunct intervenes between two constituents of the lexical grammar which nevertheless enter into a standard syntactic and semantic relationship expressed in the lexical grammar, as in (9a), then interleaving application of the syntactic and semantic rules of text and lexical grammar achieves exactly the desired interpretation (9b), where SubDR would plausibly be specialised to something like elaboration.

(9) a The rumour – the Prince had been unfaithful – appeared in a newspaper.

b The(x),Rumour(x),Appear(e,x),A(y),Newspaper(y),In(e,y), SubDR(x,e′),The(z) Prince(z),Be(e′,Unfaithful(z))

Given that the rule that introduces the dash-interpolated textual adjunct will Chomsky-adjoin it to the NP *the rumour*, the first argument of SubDR is appropriately identified, but more importantly the semantics proposed above ensures that the semantic type of the result remains that of a NP so that the semantic contribution of the text adjunct is 'invisible' to the standard semantic rule of the lexical grammar which combines subject NPs with VPs.

In addition to the core text grammatical rules which carry over essentially unchanged from the stand-alone grammar, some syntactic rules must include (often optional) comma separators (rules of pre- and post- posing, coordination, and so forth). Since the function of these commas seems to be to indicate syntactic boundaries and/or syntactic scope, they do not contribute to the semantics of the lexical grammar. Further details of the specific grammars are given in Briscoe (1994) and Lee (1995).

4 Coverage and Performance

4.1 The stand-alone text grammar

The text grammar has been tested on the Susanne corpus, a 138K word parsed subset of the Brown corpus (Sampson, 1994), and covers 99.8% of the text sentences extracted. The genuine counter examples found were mainly highly genre-specific forms of punctuation, such as citation punctuation from academic papers and variants of itemising punctuation, such as a colon followed by a sequence of dash-introduced items. It would be straightforward to extend the grammar to cover such cases, but this has not been undertaken since they occur rarely.

The number of analyses varies from one (71%) to the thousands (.1%). Just over 50% of Susanne sentences contain some punctuation, so this means that around 20% of the singleton parses are punctuated. The major source of ambiguity in the analysis of punctuation concerns the function of commas and their relative scope – a text sentence containing eight commas (and no other punctuation) has 3170 analyses.[3]

[3] An alternate iterative version of this grammar (where recursion was replaced with Kleene

4.2 The text grammar integrated with a PoS sequence grammar

The text grammar integrated with a part-of-speech tag sequence grammar has been used to parse punctuated sentences from the Susanne and SEC corpora (Briscoe and Carroll, 1995). To explore the role of punctuation in resolving syntactic ambiguity and extending coverage where punctuation cues discourse rather than syntactic relations between text units we took all in-coverage sentences from Susanne of length 8–40 words inclusive and containing internal punctuation; a total of 2449 sentences. The average parse base (APB), defined as the geometric mean over all sentences in the corpus of $\sqrt[n]{p}$, where n is the number of words in a sentence, and p, the number of parses for that sentence, for this set was 1.273, mean sentence length was 22.5 tokens, giving an expected number of analyses for an average sentence of 225. We then removed all sentence-internal punctuation from this set and re-parsed it. Around 8% of sentences now failed to receive an analysis. For those that did (mean length 20.7 words), the APB was now 1.320, so an average sentence would be assigned 310 analyses, 38% more than before. On closer inspection, the increase in ambiguity is due to two factors: a) a significant proportion of sentences that previously received less than 10 analyses now receive more, and b) there is a much more substantial tail in the distribution of sentence length vs. number of parses, due to some longer sentences being assigned many more parses. Manual examination of 100 depunctuated examples revealed that in around a third of cases, although the system returned global analyses, the correct one was not in this set.

Briscoe and Carroll (1995) also report experiments assessing parse selection accuracy for a probabilistic version of the integrated grammar. In order to assess the contribution of punctuation to the selection of the correct analysis, we applied the same trained version of the integrated grammar to the 106 sentences from our 250 sentence test set which contain internal punctuation, both with and without the punctuation marks in the input. A comparison of the GEIG (Harrison *et al.* , 1991) evaluation metrics for this set of sentences punctuated and depunctuated gives a measure of the contribution of punctuation to parse selection on this data. (The results for the depunctuated set were computed against a version of the Susanne treebank from which punctuation had also been removed.) As table 1 shows, recall declines by 10%, precision by 5% and there are an average of 1.27 more crossing brackets per sentence. These results indicate clearly that punctuation and text grammatical constraints can play an important role in parse selection. Further details of these experiments and examples of parse errors cause by depunctuation can be found in Briscoe and Carroll (1995).

operators) substantially reduced such ambiguity, but still resulted in hundreds of analyses for some examples containing multiple commas

	Cross.	Rec. (%)	Prec. (%)
With punctuation			
Top-ranked 3 analyses, weighted =	3.25	74.38	40.78
Punctuation removed			
Top-ranked 3 analyses, weighted =	4.52	65.54	35.95

Table 1: GEIG evaluation metrics for test set of 106 unseen punctuated sentences (mean length with punctuation 21.4 words; without, 19.6)

4.3 The text grammar integrated with a full lexical grammar

Manual evaluation of the parses for depunctuated versions of our Susanne test sentences indicated that in many cases the correct analysis was not in the set returned by the parser and that the parser was only able to find a globally coherent analysis because the the part-of-speech tag sequence grammar does not incorporate subcategorization constraints and therefore overgenerates considerably. Lee (1995) integrated the text grammar with the lexical grammar developed by Grover *et al.* (1993), which does incorporate subcategorization and is able to produce a semantic representation for sentences parsed (though it has less wide-coverage). Lee conducted a small experiment with 32 sentences chosen to be parallel to the examples used in Nunberg (1990) but with vocabulary drawn from the representative lexicon provided with the lexical grammar. Each of these sentences when punctuated is assigned a correct analysis (and possibly others) but when punctuation is removed over half of them do not receive any parse, a quarter receive the correct analysis because depunctuation preserves syntactic coherence (*the abbot appeared (in a scurry) and bellowed a message*), and the remainder are assigned incorrect analyses. This small experiment indicates that the true increase in coverage obtained by incorporating punctuation into text interpretation is likely to be much higher than the 8% found in the corpus experiment reported above. Further details of the experiment are provided in Lee (1995).

5 Conclusions

I have argued that a formal and declarative account of text grammar can be developed using a feature/unification-based phrase structure formalism utilising rule-to-rule construction of a (logical) semantic representation. The syntactic component of this grammar has been demonstrated to have wide-coverage on real data and to contribute significantly to resolution of syntactic ambiguity and increased parse coverage of actual text sentences when integrated with a lexical (PoS tag sequence) grammar. Although our understanding of the semantics of punctuation is not as well developed, I have argued that an adequate treatment can be devloped using a text grammar developed with a formalism of this type deployed in an interleaved fashion with a compatible lexical grammar. Finally, a preliminary experiment with the text grammar with semantics added, now integrated with a full lexical grammar with a compositional semantic component, further underlines the important role of punctuation in cueing the correct interpretation for many text sentences.

Acknowledgements

I would like to thank John Carroll, Christy Doran, Greg Grefenstette, Bernie Jones, Sherman Lee, Geoff Nunberg and Kiku Ribas for their numerous and varied contributions to the research reported here. Remaining errors are entirely my responsibility.

References

Asher, N. 1993. *Reference to Abstract Objects in Discourse.* Kluwer Academic, Dordrecht.

Briscoe, E.J. 1994. *Parsing (with) Punctuation etc..* Rank Xerox Research Laboratory, Grenoble, MLTT-TR-002.

Briscoe, E.J. and J. Carroll 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the ACL/SIGPARSE 4th Int. Workshop on Parsing Technologies (IWPT95)*, 48–58. Prague / Karlovy Vary, Czech Republic.

Briscoe, E., Grover, C., Boguraev, B. and Carroll, J. 1987. A formalism and environment for the development of a large grammar of English. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 703–708. Milan, Italy.

Dale, R. 1991. The role of punctuation in discourse structure. In *Proceedings of the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, 13–14. USA.

Garside, R., Leech, G. and Sampson, G. 1987. *Computational analysis of English.* Longman, London.

Grover, C., Carroll, J. and Briscoe, E. 1993. *The Alvey Natural Language Tools Grammar (4th Release).* Cambridge University Computer Laboratory, TR-284.

Grosz, B. and C. Sidner 1986. Attention, intention and the structure of discourse. *Computational Linguistics* 12.3: 175–204.

Harrison, P., Abney, S., Black, E., Flickenger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, B., Marcus, M., Santorini, B. and Strzalkowski, T. 1991. Evaluating syntax performance of parser/grammars of English. In *Proceedings of the Workshop on Evaluating Natural Language Processing Systems.* ACL.

Hobbs, J. 1985. *On the coherence and structure of discourse.* CSLP-TR.

Lee, S. 1995. *A syntax and semantics for text grammar.* MPhil. Dissertation, Engineering Dept., Cambridge University.

Nunberg, G. 1990. *The linguistics of punctuation.* CSLI Lecture Notes 18, Stanford, CA.

Sampson, G. 1994. Susanne: a Doomsday book of English grammar. In Oostdijk, N & de Haan, P. eds. *Corpus-based Research into Language.* Rodopi, Amsterdam: 169–188.

Taylor, L. and Knowles, G. 1988. *Manual of information to accompany the SEC corpus: the machine-readable corpus of spoken English.* University of Lancaster, UK, Ms..