September 1998

Incorporating Punctuation Into the Sentence Grammar: A Lexicalized Tree Adjoining Grammar Perspective

Christine D. Doran
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-98-24.

Incorporating Punctuation Into the Sentence Grammar: A Lexicalized Tree Adjoining Grammar Perspective

Abstract
Punctuation helps us to structure, and thus to understand, texts. Many uses of punctuation straddle the line between syntax and discourse, because they serve to combine multiple propositions within a single orthographic sentence. They allow us to insert discourse-level relations at the level of a single sentence. Just as people make use of information from punctuation in processing what they read, computers can use information from punctuation in processing texts automatically. Most current natural language processing systems fail to take punctuation into account at all, losing a valuable source of information about the text. Those which do mostly do so in a superficial way, again failing to fully exploit the information conveyed by punctuation. To be able to make use of such information in a computational system, we must first characterize its uses and find a suitable representation for encoding them.

The work here focuses on extending a syntactic grammar to handle phenomena occurring within a single sentence which have punctuation as an integral component. Punctuation marks are treated as full-fledged lexical items in a Lexicalized Tree Adjoining Grammar, which is an extremely well-suited formalism for encoding punctuation in the sentence grammar. Each mark anchors its own elementary trees and imposes constraints on the surrounding lexical items. I have analyzed data representing a wide variety of constructions, and added treatments of them to the large English grammar which is part of the XTAG system. The advantages of using LTAG are that its elementary units are structured trees of a suitable size for stating the constraints we are interested in, and the derivation histories it produces contain information the discourse grammar will need about which elementary units have used and how they have been combined. I also consider in detail a few particularly interesting constructions where the sentence and discourse grammars meet-appositives, reported speech and uses of parentheses. My results confirm that punctuation can be used in analyzing sentences to increase the coverage of the grammar, reduce the ambiguity of certain word sequences and facilitate discourse-level processing of the texts.

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-98-24.

INCORPORATING PUNCTUATION INTO THE SENTENCE GRAMMAR: A LEXICALIZED TREE ADJOINING GRAMMAR PERSPECTIVE

Christine D. Doran

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1998

This research was partially supported by NSF Grant SBR8920230 and ARO Grant DAAH0404-94-G-0426.

Acknowledgements

It has seemed to me at various times during my grad school tenure that pursuing a PhD is a bit of an insane undertaking, but Penn has been great place to be insane in this particular way, and I have been lucky to have had a brilliant set of colleagues to share the experience. Thanks, all of you!

Special thanks should go to a number of people who have helped and supported me in many ways.

First, of course, to my advisor Aravind Joshi, who has been unwavering in his support and quiet prodding. He is always available, has read everything I've written, pushes me to send things to conferences, introduces me to people in the NL community—in short, everything one could hope for from an advisor. Also, my two advisors at Wellesley and at Sussex respectively, Andrea Levitt and Gerald Gazdar, who introduced me to linguistics and set me on the path that brought me to graduate school (and didn't even seem phased by the slightly indirect route).

The other NLP faculty at Penn have been terrifically helpful. Mark Steedman and Mitch Marcus gave me solid advice and direction during my first year at Penn, and they, along with Bonnie Webber and Martha Palmer, have continued to shape my research. Mark has also been a marvelous committee member, turning my drafts around in record time.

Geoff Nunberg wrote the book on punctuation that got me interested in this topic, and was always coming up with funny sentences to stump me with. Ted Briscoe has been incredibly generous with his time, talking with me about his own work on punctuation, looking through examples, giving comments on my work, and just being generally encouraging.

The members of the XTAG project have given me a team to play on and play with. I especially want to thank Beth Ann Hockey, B. Srinivas, Anoop Sarkar, Dania Egedi and Tilman Becker for being delightful to work with, in doing the research, and to travel with, in presenting it. My lovely officemates, Srini and Anoop, have been fabulous colleagues, collaborators and hackers-on-call.

Penn's Institute for Research in Cognitive Science has been an remarkable resource, providing both outstanding facilities and an intellectual community without peer. There have been so many students and visitors at IRCS in my years here that I will simply thank them all collectively, but Matthew Stone, Breck Baldwin, Mike White, Jeff Reynar, Laura Wagner, Laura Siegel, Al Kim and Mickey Chandrasekar

deserve individual mention. Jeff's graduate career at Penn has been precisely coextensive with my own, and I have shared with him many of the trials and tribulations of grad school. He is one of the most solidly grounded people I have ever known, and I can't think of a better seat-mate on the roller-coaster ride of dissertation writing. The administrative staff of IRCS make everything that happens here run more smoothly and enjoyably. Without Trisha Yannuzzi and Susan Deysher, in particular, IRCS would not have been such a pleasant place to work.

My first year cohorts in Linguistics and Computer Science made the work bearable and the play happen: Charles, Dave, Mike, Jeff N., Jeff R., Dean, Doug, Kyle, Srini and Lisa. My non-Penn friends have been exceptionally understanding, helped me to forge on in the lowest points of grad school, and reminded me that not everyone lives this strange grad school life: Guy Danner, Maria DePina, Stephanie Hornbeck, Kelly McGrath, Marya Postner, Kate Rice, Jonathan Risch and Mary Silveria. My roommate Teresa Halverson has been a great friend and house-mate, cheerfully taking up the household chores that I have neglected in these last months of dissertating and always ready to go off and do something fun when I found myself with a sliver of free time. And my cats Pele and Vinnie, and before them Emma, have made my apartment a real place to come home to.

I was fortunate to have two sets of grandparents who always took a great interest in my academic efforts: the Turvilles, both of whom preceded me in graduate work at Penn and had me playing language games as soon as I could talk (or maybe even before), and the Dorans, who always encouraged me and wanted to hear the details of my research, even when it was completely arcane.

Finally, my parents Clark and Karen Doran, instilled in me, by the example they set more than anything they said, a love of learning and a regard for higher (and higher and...) education. They have always stood behind me and believed I would succeed at my endeavors, even when success did not seem to me to be near at hand.

In line with the common practice of earlier times, the reader should feel free to insert into the dissertation additional punctuation marks to taste.

Abstract

INCORPORATING PUNCTUATION INTO THE
SENTENCE GRAMMAR: A LEXICALIZED TREE
ADJOINING GRAMMAR PERSPECTIVE
Christine D. Doran
Supervisor: Aravind K. Joshi

Punctuation helps us to structure, and thus to understand, texts. Many uses of punctuation straddle the line between syntax and discourse, because they serve to combine multiple propositions within a single orthographic sentence. They allow us to insert discourse-level relations at the level of a single sentence. Just as people make use of information from punctuation in processing what they read, computers can use information from punctuation in processing texts automatically. Most current natural language processing systems fail to take punctuation into account at all, losing a valuable source of information about the text. Those which do mostly do so in a superficial way, again failing to fully exploit the information conveyed by punctuation. To be able to make use of such information in a computational system, we must first characterize its uses and find a suitable representation for encoding them.

The work here focuses on extending a syntactic grammar to handle phenomena occurring within a single sentence which have punctuation as an integral component. Punctuation marks are treated as full-fledged lexical items in a Lexicalized Tree Adjoining Grammar, which is an extremely well-suited formalism for encoding punctuation in the sentence grammar. Each mark anchors its own elementary trees and imposes constraints on the surrounding lexical items. I have analyzed data representing a wide variety of constructions, and added treatments of them to the large English grammar which is part of the XTAG system. The advantages of using LTAG are that its elementary units are structured trees of a suitable size for stating the constraints we are interested in, and the derivation histories it produces contain information the discourse grammar will need about which elementary units have used and how they have been combined. I also consider in detail a few particularly interesting constructions where the sentence and discourse grammars meet—appositives, reported speech and uses of parentheses. My results confirm that punctuation can

be used in analyzing sentences to increase the coverage of the grammar, reduce the ambiguity of certain word sequences and facilitate discourse-level processing of the texts.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> "That there pass no mistakes of the punctuation. For...if the stops be omitted, or misplaced, it does...oftentimes quite spoil the sense." *Boyle, Style of Script., 1661*
>
> "The expectation of a settled Punctuation is in vain, since no rules of prevailing authority have been yet established." *Luckombe, Hist. Print., 1771*[1]

Despite the large number of style manuals published in the last 300 years, and the increase in uniformity of formal education, both of these quotes hold true today. The impact of misplaced punctuation can be quite severe, and yet there does not appear to be consistent usage in naturally occuring texts. Regardless of these difficulties, it has been intuitively apparent to linguists and computer scientists interested in the structure of texts that punctuation has much to contribute to language processing by both humans and computers. However, perhaps in part because of these difficulties, there has been surprisingly little research in this area.

1.1 What does punctuation do for us?

As Boyle noted over 300 years ago, the omission or insertion of punctuation often leads to confusion and misunderstanding. It also can have humorous results. Let us look at a few such cases as a way to illustrate how critical punctuation is to our ability to process texts.

What makes this cartoon funny is that without a comma between *John* and *Paul*, you interpret the sequence as describing one person, His Excellence John Paul, instead of two, John Lennon and Paul McCartney.

Having seen only variant (a) of example (1), you would be hard pressed to believe that the same sequence of words could, with "only" the punctuation changed, take on the exact opposite meaning. Example (2) is a similar but more compressed text.

[1] From the Oxford English Dictionary, 2nd edition.

(1) a. Dear John:
I want a man who knows what love is all about. You are generous, kind, thoughtful. People who are not like you admit to being useless and inferior. You have ruined me for other men. I yearn for you. I have no feelings whatsoever when we're apart. I can be forever happy–will you let me be yours?
Gloria

b. Dear John:
I want a man who knows what love is. All about you are generous, kind, thoughtful people, who are not like you. Admit to being useless and inferior. You have ruined me. For other men I yearn. For you I have no feelings whatsoever. When we're apart, I can be forever happy. Will you let me be?
Yours, Gloria

(2) a. Woman, without her man, is an animal.

b. Woman; without her, man is an animal.

Example (3) shows what can happen when you have a verb which can easily be understood either transitively or intransitively. Having the comma present ensures that only the intransitive reading is possible.

(3) a. Let's eat Grandfather before we go.

b. Let's eat, Grandfather, before we go.

The absence of punctuation around *suitable for lady* in (4) gives us a reading where the *with*-PP is more felicitously read as modifying *lady* rather than *desk*.

(4) For sale: an antique desk suitable for lady with thick legs and large drawers.

While entertaining, the forgoing examples[2] illustrate a serious point. Punctuation helps us to structure, and thus to understand, texts. When it is wrong, we are misled in sometimes unrecoverable ways. In (5), the lack of a comma between *Oklahoma* and *and* allows for an interpretation where the Ryder agency is a militant right-wing compound, since you cannot be quite sure whether the three phrases after *called* are a list or one noun phrase with a modifier. The comma before the *and* in a series is considered to be optional, but in cases like this it is extremely useful in steering the reader toward the intended meaning.

(5) That was also the day McVeigh called an Arizona Ryder agency, a militant right-wing compound in Oklahoma and an Arizona leader of the neo-Nazi National Alliance group that published the racist novel The Turner Diaries. [Rocky Mountain News, 1/18/98]

Sentences like (6) make it obvious that punctuation encodes both semantic relations and discourse structure—while (6) is orthographically one sentence, structurally it contains a veritable dialogue.

(6) The Usage Panel now has respected linguist Geoffrey Nunberg–not the fervent Edwin Newman of television fame–as its chair, and its composition, they proudly tell us, is closer to mainstream America: 112 men, 61 women, an average age of 61 (as opposed to 68 in previous editions).[3]

The problem we face is how to capture the information conveyed by punctuation in a systematic way which can be practicably incorporated into a computational system.

1.2 Punctuation in computational systems

Just as punctuation helps people to process texts they are reading, computers can use information from punctuation marks in trying to process texts automatically. Most current systems fail to take punctuation into account at all, losing a valuable source of information about the text. Those which do take it into account mostly do so in a superficial way, again failing to fully exploit the information conveyed by punctuation. To be able to make use of such information in a computational system, we must first characterize its uses and find a suitable representation for encoding them. Returning briefly to one of the examples above, let us consider (3) repeated below as (7). With the punctuation stripped out, most parsers would only get the transitive case, since they would not be able to recognize the vocative use of *Grandfather*. If the punctuation is left in, the system must then have a special

[2] All of the examples were collected from postings to the punct-l mailing list.

[3] *The Christian Century*, Book reviews, Vol 11 No 16 , 5/12/1993.

rule for handling non-argument noun phrases set off by commas. As it turns out, the grammar will need several rules for non-argument noun phrases. Certainly such rules could be added without reference to the punctuation marks, but then almost all NPs will be candidates for these rules. By taking punctuation into account, we significantly reduce the number of rules that can apply in constructions like this.

(7) a. Let's eat Grandfather before we go.

b. Let's eat, Grandfather, before we go.

My interest in punctuation grew out of my work on the XTAG English grammar, which is a wide-coverage computational grammar based on the Lexicalized Tree Adjoining Grammar (LTAG) formalism. The grammar was very large even then, but did not cover a number of critical multi-clausal constructions. One of the first major grammar development tasks I undertook was to expand the treatment of subordinate clauses from the wee number of subordinating conjunctions and clause types that were then part of the grammar. The grammar now recognizes 72 subordinating conjunctions, including multi-word constructions like *in order*, and has trees to handle subordinate clauses in four positions relative to the main clause (a few examples are shown in (8)-(10)). I also added an analysis for a class of constructions we call "bare adjuncts," which are clausal adjuncts without overt subordinating conjunctions, including infinitival purpose clauses (11).

(8) I put a lot more trust in my two legs than in the gun, *because the most important thing I had learned about war was that you could run away and survive to talk about it.* [ck09]

(9) *In order to accomplish the purposes of this Act*, the Secretary of the Interior shall... [ch09]

(10) *As he drove home through the thinning traffic*, Cady felt the unease growing. [cp27]

(11) Below, people line the steps, as though on bleachers, *to watch the sky and river.* [cg05]

It quickly became evident that multi-clausal constructions lay at the border (the "interface," dare I say) between what we standardly think of as syntax (the construction of single sentences) and discourse (the construction of larger extents of texts). Following Gardent's terminology [1997], I will refer to these two levels as the SENTENCE GRAMMAR and the DISCOURSE GRAMMAR. (Nunberg [1990] calls them the 'lexical grammar' and the 'text grammar.') The adjunct and subordinating clause constructions are two of the primary ways of combining into a single sentence information which could equally well be presented in multiple sentences. In addition,

the subordinating conjunctions represent a large subset of the class of **cue words** which are typically characterized as giving readers clues about the structure of the discourse.

It also became evident that punctuation is an important part of these complex constructions, sometimes appearing with subordinating conjunctions and sometimes functioning alone to combine clauses. In constructions where text-level elements are inserted into sentences, there is almost always some punctuational element required.[4] This is reflected in the grammar book characterization that things which are less closely 'connected' to the text are set off with punctuation. Punctuation is a system for demarcation of text constituents, and as such is a crucial point of contact between the sentence grammar and the discourse grammar.

1.3 The approach

As noted above, punctuation is useful in automatic text processing in many of the same ways that it is useful to human readers. The present work describes a computational model of punctuation, executed within the framework of Lexicalized Tree Adjoining Grammar. Punctuation marks will be treated as full-fledged lexical items, anchoring their own elementary trees and imposing constraints on the surrounding lexical items. Crucially, the analysis is developed and tested on data collected from naturally occurring texts. My goal in exploring punctuation within the framework of LTAG is to see to what extent the constructions involving punctuation which are realized at the level of the orthographic sentence, some of which introduce discourse-level relations, can be incorporated into an existing grammar. We do not want to turn to additional higher level processing mechanisms to handle the sentence-level phenomena, but do want their treatments to be compatible with the needs of the discourse grammar. Other treatments have looked at the syntactic and discourse level uses of punctuation independently, but have not sought to account for them in computational framework compatible with both levels of analysis.

To accomplish this, I have analyzed data representing a wide variety of constructions, and this work is discussed in Chapter 3; a few turned out to be of particular interest, and are discussed in more detail in the later chapters. That work explores a handful of constructions where the sentence and discourse grammars meet—appositives are ways to insert extra predicates, quoting clauses are text adjuncts that are closely related to embedded clausal complements, and parentheses can be used to either insert text which is syntactically completely unrelated to the surrounding text, or to set off some piece of text within the sentence grammar.

The LTAG syntactic account is assessed within the framework of the XTAG system, an existing system with a large English grammar. Prior to the current

[4] In fact, the word *comma* comes from the Greek 'to cut,' as in 'to cut off a piece' from the sentence.

work, this grammar did not attempt to handle any punctuation. There are three dimensions along which the punctuation analysis may be evaluated within the XTAG system:

1. Whether it improves the coverage of the existing grammar
2. Whether it constrains ambiguity in parsing, in particular where punctuation delimits constituent boundaries
3. Whether it improves the grammar's performance in particular applications

In this work I concentrate exclusively at sentence-level punctuation, where an orthographic sentence in English is taken to be a string of words beginning with a capital letter and ending with a period, exclamation point, question mark or ellipses, regardless of the syntactic structure of the string (i.e. it need not contain a verb). I do not consider morpheme-level punctuation (e.g. apostrophes, hyphens) or formatting punctuation (e.g. list elements preceded by dashes or bullets). The latter are better classed with other formatting information such as font changes and paragraph organization.

1.4 Underlying assumptions

1.4.1 Text and speech are different

As Parkes states in the start of the introduction to his book, *Pause and Effect: An Introduction to the History of Punctuation in the West*[5]:

> Punctuation is a phenomenon of written language, and its history is bound up with that of the written medium. In Antiquity the written word was regarded as a record if the spoken word, and texts were usually read aloud. But from the sixth century onwards attitudes to the written word changed: writing came to be regarded as conveying information directly to the mind through the eye....

There is undoubtedly a continuum between speech and text, with read speech closer to the text end and e-mail closer to spontaneous speech. The amount of editing done on texts, by oneself or others, varies across the continuum (newswire texts are heavily edited, email is edited lightly, if at all) and will affect the way punctuation and various types of formatting and layout information are used. When we study linguistic phenomena, we typically use texts to look at things like argument

[5]This is an absolutely fascinating book which I highly recommend to anyone interested in punctuation, with nearly 100 plates of manuscripts and discussion of the evolution of punctuation reflected in them.

structure and selectional restrictions, and speech for things that are more prescriptively marginal, like resumptive pronouns, or are unique to speech, like corrections. Since I am interested here in the more 'standard' uses of punctuation, in this work I concentrate on data from edited texts and stay away from the more speech-like end of the range.

1.4.2 Punctuation is not the written correlate of prosody

Again, the continuum from speech to text will reflect varying degrees of correlation between prosody and orthographic devices. In examining the relationship between punctuation and prosody, Schmidt [1995] uses the converse of read speech, "written conversation" (email and usenet news), precisely because it is more speech-like. With regard to read speech, people have naive intuitions that speakers make a conscious effort to reflect the written structure, including the punctuation marks, in their speech patterns, and this leads people to believe that they can "tell" what punctuation marks were used in the text.

Certain modern punctuation marks did originate as transcriptional devices, but they are no longer used this way (cf. discussion by [Parkes1993, passim] and [Nunberg1990, p. 12 ff.]). Other marks never indicated prosody. Quotation marks originated in the Middle Ages as angle brackets in the margin of the text, indicating quotation of passages from the bible [Parkes1993, p. 303]; they functioned more like footnote markers than markers of, say, pause length. As early as the mid-16th century, authors argued that the main role of punctuation was syntactic rather than prosodic. In 1566 Also Manuzio wrote *Orthographiae ratio*, which described a punctuation system quite like the modern Western one, using commas, colons, semi-colons, question marks and periods.

The "punctuation as a reflection of prosody" view, which dominated at the time of Manuzio's treatise, is still held in certain quarters; recent work continues to argue against that position. As Nunberg [1990] points out, the view that punctuation encodes prosody is seriously flawed. There are clear cases which illustrate the lack of correspondence between punctuation and prosody, in both directions. An example of a break down in the presumed mapping from punctuation to prosody is the use of the question mark. All English questions are written with a question mark but it is widely known that yes/no questions often have final rising prosody, but (non-echo) wh-questions do not. Going from prosody to punctuation, it is clear that at the very least that punctuation under-notates prosody. For instance, email correspondents have resorted to using *asterisk* notation to indicate prominence since there is no vehicle for doing this in standard written English. I take it as given that, while the functions of punctuation in writing and prosody in speech may overlap to a certain extent (one obvious correlation is between scare quotes in text and the rather unique rise-fall contour used to communicate similar information in speech), the primary function of punctuation is to structure texts.

Experimental work on the relation between punctuation and prosody

There has been little linguistic research on the connections between punctuation and prosody. Nunberg [1990] alludes to "informal experiments" in which speakers were unable to communicate differences in punctuation to hearers. While this is interesting, it is anecdotal and begs for follow up research. It was his discussion which inspired recent preliminary research which I have conducted with Beth Ann Hockey [Doran and Hockey1998], seeking to address two questions: In reading written texts, can people with any accuracy "encode" punctuation for their listeners? In listening to read texts, can people with any accuracy reconstruct the original punctuation? We had subjects listen to read versions of Wall Street Journal texts (from the LDC's ARPA/CSR corpus) and insert punctuation into printed copies; we then created punctuational variants of the texts based on areas where subjects differed in the punctuation they inserted, and had a second set of subjects read those variants aloud.

In analyzing the results the first part of the experiment, we found that subjects inserted quite widely varying punctuation marks on about half of the 28 sentences, even though they had all heard the identical production of each sentence. In the second part, the subject-read sentences were analyzed at the locations of punctuation marks, both for pitch range effects on the chunks delimited by punctuation (or the beginnings/ends of sentences) and for pauses at chunk boundaries. Pausing and pitch range effects frequently coincide with prosodic phrase boundaries [Liberman1975; Pierrehumbert1980], and these are the same prosodic effects have been argued to be represented by punctuation marks. Thus far, our analysis clearly indicates that particular types of prosody and punctuation do not always coincide, but there are places where they do to a certain extent. Parentheticals, for instance, do seem to consistently be marked in both systems, but their prosodic marking can vary (pauses vs. pitch range contraction), as can their punctuation (commas vs. dashes).

1.4.3 Punctuation is a rule-based system

Punctuation marks are used by authors to help structure the text both for themselves and for their readers. There are differences in how punctuation marks are used by different writers and across various genre, but readers clearly make generalizations about the uses of punctuation in much the same way that they make other types of grammatical generalizations. People have strong intuitions, for instance, about whether a particular pre-sentential modifier needs to be followed by a comma, or that a phrase is parenthetical and has to be set off with punctuation marks. These intuitions cannot be be easily dismissed as being the result of prescriptive brainwashing.

1.5 Overview of the dissertation

The first chapter of the dissertation has presented the motivations for finding a treatment of punctuation which is both syntactically and pragmatically well-founded, and discussed the basic approach that is to be taken. Section 1.4 laid out a few of the basic premises underlying this work.

Next, Chapter 2 surveys the other relevant work on punctuation, of which Nunberg [1990] offers the most comprehensive theoretical account and Briscoe and Carroll [1995; 1994] present the only sizable implemented analysis.

Chapter 3 presents a syntactic analysis of punctuation using Lexicalized Tree Adjoining Grammar (LTAG), with discussion of how the adequacy of this analysis can be evaluated using the XTAG system as a testbed. I argue that LTAG is an extremely well-suited formalism for encoding punctuation in the sentence grammar, because (1) its elementary units are structured trees of a suitable size for stating the constraints we are interested in, and (2) the derivation histories it produces contain information the discourse grammar will need about which elementary units have used and how they have been combined. A total of 55 trees handling punctuation were added to the existing XTAG English grammar.

Chapters 4, 5, and 6 present case studies of NP appositives, quoted speech and parentheses, respectively. These are all quite complex constructions, with interesting syntactic, semantic and pragmatic features, and they have superficially similar variants which appear to differ primarily in the presence or absence of punctuation.

Chapter 4 considers a class of complex NPs which look rather like appositives, and finds that they fall into two categories. Those which contain punctuation and are NP-level modifiers are non-restrictive, meaning that they add information about an entity without helping the hearer actually identify the entity. Those which do not contain punctuation and/or are attached lower are restrictive, helping to determine the reference of the NP.

Chapter 5 looks at reported speech, which is typically split into direct and indirect speech based on the presence or absence of quotation marks. A more useful distinction is found between argument quotes, which act like other clausal complements, and quotes where the verb of saying and its subject (the QUOTING CLAUSE) are attached to the the quote as TEXT-ADJUNCTS. The latter class has obligatory punctuation separating the quote from the quoting clause.

Chapter 6 examines the uses of parentheses in two corpora, a set of F16 repair instructions and a set of of academic papers. It then evaluates the uses identified with respect to Nunberg's [1990] binary classification of parentheticals into those which introduce ALTERNATIVES and those which RESTRICT the context of interpretation. Both corpora are found to have uses which do not fit either of these categories.

Chapter 2

Previous work on punctuation

Beyond the normative descriptions of punctuation found in style manuals and writers' guides, of which the Chicago Manual of Style [Chi1982] is a prime examplar, there has been little in the way of linguistic or computational work on punctuation until very recently. This chapter reviews the relevant research, classed into descriptive and linguistic approaches.

2.1 Descriptive studies

Quirk et al. [1985] is quite exceptional as descriptive grammars go, giving a very nice overview of the uses of punctuation marks in English. They do say that punctuation is the visual equivalent of prosody, but qualify this claim, saying that "the link is neither simple nor systematic, and traditional attempts to relate punctuation directly to (in particular) pauses are misguided".[1] They make the suggestive argument that punctuation and prosody differ quite distinctly in that the former has to be explicitly taught, while the latter is acquired. There is no simple argument to be made that punctuation is not acquired to a certain extent, given the lack of uniformity in the punctuation found in naturally occurring texts and the fact that peoples' judgements about the placement of punctuation are usually as strong as with other types of syntactic judgements. This is a very intriguing question, however.[2] They also give an interesting argument for why punctuation is/should be conventionalized, which is that the writer is often not present to interpret his/her material when it is being read. Despite this, they allow that there is a lot of variation in how people use punctuation. Regarding the actual uses of punctuation, they propose a hierarchy of

[1] Appendix III.1. NB: Nunberg [1990] makes a similar point to this and a number of others in Quirk et al.

[2] [de Beaugrande1984]:V.2.41 notes that in his experience, speech has a significant confounding effect on punctuation use with weaker writers. In particular, there is a tendency to overuse commas, placing them everywhere one would find a significant pause in speech. This sort of data might give useful clues as to how people "acquire" punctuation.

marks whereby a lower element such as a comma may be displaced when it co-occurs with a higher element such as a colon. They split punctuation into **specificational** (genitive *'s*) and **separating** marks (most other punctuation).

Sampson's description of how the SUSANNE annotation scheme [Sampson1995] handles punctuation is also quite detailed. Annotations are made to a number of formtags to indicate that a particular punctuation mark has been used in a particular way, for instance S indicates a clause and S! indicates an exclamative clause, typically ending with an exclamation mark. Sampson does not make any theoretical claims about punctuation, but there is a certain level of analysis implicit in the decisions about how to annotate constructions involving punctuation.

An interesting perspective on punctuation is presented by [de Beaugrande1984], whose central concern is the pedagogy of composition; he thinks writing teachers ought to focus on the motivations for punctuation rather than on rigid rules. He considers the various punctuation marks in light of his own general principles for **text linearization**. Some of the more interesting observations he makes are that: 'heavier' (length, content, focus) adjuncts are more likely to be separated from a main clause by punctuation; separation by punctuation gives modifiers wide-scope (although he describes this as "Looking Forward/Backward"); dashes and parentheses are unusual in allowing the writer to insert syntactically unrelated material without disrupting the syntax of the surrounding text; and ellipses, parentheses and questions marks indicate rhetorically 'lightness', while exclamation points and dashes indicate 'heaviness.'

2.1.1 Punctuation and prosody

Chafe [1988] thinks of punctuation as "the principal device" for encoding prosodic cues in written texts. In particular, punctuation is used by the writer to encode his or here "inner voice," and Chafe goes so far as to say that this is the main use of punctuation, with any other uses classified as "departures from its main functions." He conducts some experimental research on this point, the primary goal of which is to assess the correlation between **prosodic units** and **punctuation units** (the chunks of text between punctuation marks). In brief, his experiments find that (1) the prosodic units are about 40% shorter than the units delimited by punctuation in the same texts and (2) there is only about 50% correlation between the locations of punctuation marks and the locations of prosodic unit boundaries. I interpret these results as indicating that there is a considerable mismatch between the prosodic and punctuation units. Chafe, however, interprets them as supporting his thesis, saying "...the most broadly applicable finding of this study is that most writing most of the time does use punctuation in a way that respects the prosody of the language."

Schmidt [1995] looks for acoustic correlates to a small set of punctuation marks, some of which might be better classed as formatting information, e.g. all uppercase, as used in "Written Conversation" (email and usenet news postings). This is

a modality which shares many of the properties of both speech and written texts, and lies somewhere between them on the scale of "textiness." Some of the features Schmidt considers are unique to the genre (e.g. the use of emoticons/smiley faces). He primarily focuses on written markers of emphasis (capitalization, asterisks around text, etc.) and parenthetical statements (of the narrow sort—phrases enclosed in parentheses). His results are somewhat mixed, but he does not find that parentheticals are prosodically independent of their context to any significant extent, which he takes to suggest that their insertion does disrupt the surrounding context. Somewhat curiously, he uses read speech for his comparisons; it would have seemed more appropriate to compare spontaneous speech with such a spontaneous, unedited written medium.

Beeferman, et. all [1998] have built a trigram model, trained on the Treebank Wall Street Journal sentences, which inserts commas into text output from a speech recognizer, without any consideration of prosodic information from the input. Their aim is to develop a tool which would obviate the need for speakers to spell out punctuation marks when using an automatic dictation system. They achieve 54.0% per sentence accuracy on the set of 2317 sentences. It would be interesting if it turned out to be the case that the sentences they get right are the ones where there is no overlap between punctuation and prosody. This performance suggests that a model of punctuation may get some distance without taking into account prosodic information, but could benefit from prosodic information in those instances where the functions of the two systems overlap.

2.2 Linguistic studies

Meyer [1987] discusses the uses of punctuation in marking syntactic, semantic and prosodic boundaries from a more descriptive than formal point of view. He focuses on what he defines as *structural* punctuation, which includes everything above the lexical level (e.g. no hyphens) and up to the level of a single orthographic sentence. He also adopts the hierarchy view, with reference to Quirk, but divides punctuation into the categories *separating* (single marks) and *enclosing* (paired marks). One interesting claim is that punctuation functions as a "perceptual cue" in marking all of syntactic, semantic and prosodic boundaries, either functioning alone if there are no other indicators, or reinforcing other types of cues. This accords with psycholinguistic research on marking of syntactic and semantic constituents by Sevald and Trueswell, discussed briefly below.

Meyer gives some interesting statistics gathered over a 72,000 word subset of the Brown corpus. In particular, his Table 2.5, shown here as Table 2.1, gives the percentage of various types of sentences which contain punctuation.

[3]Where "A phrase...is a constituent consisting of one or more words centered around a head...." (p. 39)

Construction	Punctuated	Unpunctuated
Non-elliptical compound sentence	85%	15%
Elliptical compound sentence	17%	82%
Compound subordinate clause	23%	77%
Compound phrase[3]	13%	177%

Table 2.1: The Punctuation of Coordinated Constructions (Meyers' Table 2.5)

In addition, Meyer finds that the more complex the surrounding syntax, the higher the probability of a punctuation mark being used in a particular location. For instance, he finds that 90% of clausal adverbial elements are separated from a clause by punctuation, whereas single word adverbials are only punctuated 32% of the time. Semantically, he claims that elements that are less "semantically integrated" are more likely to be separated by punctuation. One such case is conjunctive vs. disjunctive coordination of clauses–only 13% of clauses coordinated with *and* did not have punctuation between the clauses, while 64% of clauses with *or* did. Meyer also reports on how the uses he finds in his corpus accord with usage guides (fairly well), and suggests some guiding principles for using punctuation based on his survey.

Nunberg [1990] offers the most comprehensive linguistic discussion of punctuation to date, with an extensive analysis of the interactions of different punctuation marks. He is primarily interested in characterizing punctuation as a formal system, independent from syntax. Nunberg proposes that punctuation be distinguished from the lexical grammar, and be part of a "text grammar" which controls how pieces of text are combined. Like Quirk and Meyer, he proposes that punctuation marks form a hierarchy, and when two marks are in competition at a given position, the higher one 'wins' and is used. Thus, all pairs of bracketing punctuation are produced, and then one is "absorbed" if it conflicts with another, higher ranked mark as is illustrated in example (1). Nunberg distinguishes **delimiting** punctuation marks (e.g. parentheticals) from **separating** punctuation marks (e.g commas between sequences of adjectives).

(1) John left, apparently,; Mary stayed. →
John left, apparently; Mary stayed. [Nunberg's 5.6]

However, as discussed in the review by Sampson [1992], Nunberg's account is based primarily on invented data, and not on analysis of naturally occuring texts. As a result, he makes some claims which are more prescriptive than descriptive. Nonetheless, Nunberg's book has been instrumental in stirring up interest in punctuation as a formal (rather than stylistic) object of study, and in arguing against the naive conception of punctuation as the translation of prosody in writing.

2.2.1 Punctuation and parsing

Briscoe [1994] presents an treatment of punctuation within the Alvey Natural Language Tools grammar. He and Carroll [1995] show that this analysis considerably reduces ambiguity in parsing the SUSANNE corpus (a subset of the Brown corpus) and work by Jones [1994] has shown similar results. Both add punctuation to grammars that work on strings of part-of-speech tags and ignore the actual lexical items, and find that punctuation adds more structure to their grammars and thus constrains ambiguity. It is crucial to note that the grammars they start with are highly unconstrained, given that they use only the POS tag, and usually produce a large number of parses for a given sentence. Briscoe and Carroll [1995] report a measure of ambiguity called "Average Parse Base" (APB), and find that on a 2449 sentence subset of the SUSANNE corpus, the unpunctuated grammar produced 38% more parses (310 vs. 225) for the average length sentence than the punctuated grammar. (Note that while the unpunctuated sentences are more ambiguous, we do not know for either case how many sentences received a correct parse. On a subset of 100 unpunctuated sentences, about 30% had no correct parse; we would expect that figure to be lower for the punctuated sentences.) They further find that the inclusion of punctuation improves their probabilistic LR parser's ability to select the correct parse for a given sentence, reporting an 8.84% improvement in recall and 4.83% in precision on the crossing-brackets metric, when evaluating 106 sentences against the hand-annotated SUSANNE bracketings. Briscoe and Carroll also find that adding punctuation improves the coverage of their grammar by 8% on the SUSANNE sentences (where coverage is measured as getting some parse for a sentence).

Briscoe [1994] also points out that if one takes a declarative approach, there is no real need for absorption rules, i.e. producing a punctuation mark and then removing it, as Nunberg suggests; rather one can think of certain uses of punctuation as occuring in pairs in the middle of sentences/clauses, but singly at boundaries delimited by other punctuation marks. Thus, his grammar contains sets of symmetric and asymmetric rules for bracketing punctuation marks like commas and dashes. Furthermore, in [Briscoe1996] he argues that by interleaving the text and lexical grammar, one can better take advantage of the combined power of both systems in disambiguating input, and that the semantics associated with the two sets of rules should not interfere with one another.

Jones tests his grammar on 50 sentences of manually punctuated text from the Spoken English Corpus, with an average length of 31 words. He reports that the average number of parses without punctuation but with the punctuation grammar (i.e. punctuation stripped, but no other changes) is $10^6 - 10^9$ (estimated), and with a "trimmed" grammar (rules relying on punctuation removed) is 1-2 orders of magnitude greater than the grammar with punctuation. Both Briscoe and Jones note that their punctuation grammars are not comprehensive and need further tuning.

In his dissertation, Jones [1996b] arrives at a number of generalizations about

punctuation, based primarily on corpus analysis (85 million words in the largest experiments). Over the nine data sources, and 85 million words of text, he finds that there are between 2 and 5 punctuation marks per sentence. He distills the syntax of punctuation first into 137 rules, and then further into a handful of ultra-simplified and underspecified schemas, along with a handful of hierarchies for handling absorption effects, relative scoping of different types of coordination and relative prominence of certain types of text-adjuncts. He argues that a theory of punctuation would be useful in both language understanding and generation, but concludes that the rules he arrives at are too permissive to be practical in a generation system.

It is not clear in the end whether Jones believes that the syntactic function of punctuation is crucial in parsing/understanding; he argues in several places that it is, both for humans and for machines (Ch 1, passim), but elsewhere claims that the syntax is already clear without the punctuation (p. 120 ff.). He does not seem to consider the fact that since sentences of any interesting length tend to be horribly ambiguous, any information providing additional constraints can be very useful. So even in cases where the punctuation marks simply reinforce boundaries already identifiable in principle in the syntax, they may provide useful information to a processing system. Jones finds that semantically, punctuation marks are for the most part too under-specified to provide any truly useful information, but that in the case of paired marks, they do indicate the relative importance of the delimited text (e.g. text in parentheses is less important to the overall content than the same text in commas). He criticizes Say and Akman [1996a; 1996b] for conflating syntactic and semantic information from punctuation, but in the absence of a more definitive semantic account, such a mixed approach would appear to be the best way to maximize the utility of punctuation marks in NL tasks.

Jones classifies punctuation into sub-lexical, inter-lexical and supra-lexical marks, and concentrates his attention primarily on the inter-lexical marks–commas, dashes, periods, etc. Inter-lexical marks are classified as either conjunctive or adjunctive. The conjunctive class roughly corresponds to Nunberg's **separating** marks, while the adjunctive class approximate the **delimiting** marks. Conjunctive marks are commas and semi-colons used in lists, and dashes used to conjoin two clauses. Adjunctive marks include most of the core punctuation phenomena, and they encode the nucleus-satellite relationship used in Rhetorical Structure Theory (RST). As with semantics, he finds that each punctuation mark can be used to encode a wide variety of RST relationships. Colons are the most constrained. He also divides punctuation marks into those which are "source-specific" and those core marks which are uniformly used across genre ("source-independent"). In the source-specific category are idiosyncratic symbols, like $ and @, as well as quotes and bracketing marks, which he finds are used/encoded very differently in the range of texts he examines.

Like Briscoe and Carroll, he decides to use a merged grammar rather than strictly adhering to Nunberg's proposal of distinct text and lexical grammars. In his final set of rules, he concludes, as Briscoe and Carroll did, that having balanced and

unbalanced variants of the rules is more practical than doing all of the absorption as a post-process, and that the merged grammar allows the parser to take advantage of the constraints imposed by both sets of rules.

2.2.2 Punctuation and discourse

Dale [1991] is the the first work which considers using punctuation to identify discourse structure, primarily in the context of natural language generation. He claims to be interested primarily in the semantics of punctuation, by which he means notions such as things linked by commas being more closely associated than things linked by semi-colons. This paper is quite preliminary, but it contains some interesting ideas. Some of the more intriguing points he raises are that:

1. Punctuation is a good indicator of discourse structure *within* text sentences. This is a very interesting point, since the minimal units in discourse are usually taken to be clauses. Ex: *John left. He was very unhappy with the situation.* vs. *John, who was very unhappy with the situation, left.*

2. Punctuation may mark the degree of semantic "importance" of various constituents, e.g. parentheticals are less important syntactically and semantically to the sentence as a whole. In addition, different punctuation marks may indicate different levels of "closeness" between elements, e.g. things separated by commas are more closely associated than things separated by semi-colons. This is also the underlying theme of rules in grammar books regarding punctuation use, but it is not clear that people actually think of punctuation in this way when they are using it.

3. Punctuation may reflect discourse relations (in particular, Rhetorical Structure Theory (RST) relations, such as **explanation** or **result**); like cue words, a given punctuation mark may correspond to more than one relation. Cue words, like *so*, *anyway* and *well*, can be used to identify the structure of the discourse, and the relations between sentences or whole pieces text. They often indicate the boundaries of discourse segments, and the relations between them. Ex: *But we need not mind too much, because Mr. Nagrin has expressed it through movement that is diverting and clever almost all the way.* vs. *But we need not mind too much – Mr. Nagrin has expressed it through movement that is diverting and clever almost all the way.*

Work by Say and Akman [1996a; 1996b; 1995] explores this third point, discussing how a number of constructions involving punctuation marks could be handled in Discourse Representation Theory (DRT). They represent the discourse structures and the roles of various punctuation marks in an extended form of DRT. One problem they encounter, which Dale also notes, is that each punctuation mark can encode

a variety of different rhetorical relations, i.e. they are each quite pragmatically ambiguous. In a case study of dashes [Say and Akman1995], Say identifies 12 major uses, the most common being **elaboration** and **apposition**. Elaboration is the most underspecified of the standard set of rhetorical relations, and amounts to saying the dash does not really provide any information. It is essentially the default relation between two adjacent text segments between which no more specific relation holds. In addition, it is clear from their examples that a great deal of world knowledge is needed to identify the appropriate use in a given context. Take her example given as (2) below–she identifies this relation as CONTRAST, which requires knowing that both the verbs embedded in relative clauses and the main predicates have contrastive meanings which are sufficient to make the entire clauses contrastive. If either of the pairs changes, different elements are contrasted, as seen in (3) and (4). So the semi-colon may indeed suggest a contrastive function, but we still need the sentence grammar in addition to other sources of information to help us identify which particular elements are contrasted.

(2) Those who lead must be considerate; those who follow must be responsive [Say's (27)]

(3) Those who lead must be considerate; those who lead must be responsive

(4) Those who lead must be energetic; those who follow must be energetic

Lee [1995] associates rhetorical relations with the punctuation rules in Briscoe's grammar, but only in a highly underspecified way. She associates a *subordinating* relationship between text adjuncts and the elements they modify, thus eliminating the class of *coordinating* relations, but not specifying what sort of subordinating relation holds.

2.2.3 Other linguistic work on punctuation

White [1995] is also concerned with generating punctuation, as part of a larger text generation system. To this end, he considers how Nunberg's account of punctuation might be best incorporated into a NLG system. His main concern is whether the "absorbed" punctuation marks need actually be present at any level of representation, and he concludes that they are in fact useful in capturing scoping effects.

There is also some related work in psycholinguistics. A number of experiments on language processing (e.g. [Clifton1995; Adams et al.1992]) have shown that disambiguating punctuation marks, usually commas, have a strong impact in on-line processing of otherwise ambiguous texts. Clifton notes that Adams, et al. find "the absence of a comma does not affect reading time when the comma is merely stylistically preferred and does not carry disambiguating grammatical information." Sevald

and Trueswell [1997] look at sentences with initial subordinate clauses with or without a comma separating the two, and find that, when reading the sentences aloud, "[s]peakers exaggerated prosodic cues, but only when other sources of information failed to disambiguate the sentence." [Hill and Murray1999] find quite conclusively that commas do help in disambiguation; for some constructions, they keep readers from re-reading sentences they have mis-analyzed, and for others they virtually eliminate the garden-path effects that would otherwise be expected.

2.3 How does my work differ?

The worked discussed in the remainder of the dissertation departs from that described in this chapter in a number of ways. It includes a analysis of the syntax of punctuation which has been integrated into a large English grammar that is being used on an everyday basis at the University of Pennsylvania. Jones' approach, while implemented to some extent, was never used in a working system. Briscoe's grammar is currently being used in two projects: SPARKLE, an EU effort using shallow parsing, and some work on simplifying texts for dyslexic readers. Thus far, there have been no reported results on the impact of punctuation in either of these applications. None of the other work described was implemented to any significant extent. In addition, the analysis differs considerably from those of Jones and Briscoe in treating punctuation within a framework which allows for more concise characterization of the non-local aspects of certain uses of punctuation. Furthermore, neither of their implementations cover the range of punctuated constructions my treatment does. The second part of the dissertation focuses on three sets of constructions where punctuation appears to play a crucial role: appositives, quoted speech and parenthesized text. Other work has looked closely at the use of one particular punctuation mark (e.g. dashes in [Say and Akman1995]), but not at specific sets of constructions.

Chapter 3

A TAG analysis of the syntax of punctuation

Many uses of punctuation straddle the line between syntax and discourse, because they serve to combine multiple propositions within a single orthographic sentence. They allows us to insert discourse-level relations at the level of a single sentence.[1] This taps into the debate about what the basic units of discourse are—sentences? clauses? How do we treat embedded clauses? Do we let the sentence grammar handle them, and feed its analysis to the discourse grammar or do we try to pull apart the relevant units in some other way? The sensible solution seems to be to let the sentence grammar do what it does best, i.e. assign structure to individual sentences, and then let the discourse grammar have access to the derivations produced. This, of course, requires a sentence grammar which produces structures compatible with information structure constituents, and not all traditional sentence grammar do this. I will argue that LTAG does.

For instance, relative clauses can themselves contain complex discourse relations, but do we really want to treat them syntactically like separate units? If we did that, we would need to have information about phenomena like agreement in both the sentence grammar and the discourse grammar. Instead, we can treat each sentence as one unit syntactically, and then pull them apart as needed for discourse-level processing. Granted that the orthographic sentence is really an artifact, i.e. we think of something as a sentence when it has a period, when things that end with colons, semi-colons and dashes can be syntactically identical, but it is nonetheless the basic unit that both human and automatic text segmenters most readily identify (cf. [Reynar and Ratnaparkhi1997] for one approach to automatic sentence detection). For this reason, it is most convenient to adhere to the convention of parsing/processing one orthographic sentence at a time.

To this end, the current effort focuses on extending the syntactic grammar to

[1] In fact, for generation Scott and de Souza [Scott and de Souza1990] argue that the best encoding of rhetorical relations is to have each relation encoded in a single sentence.

handle all of the phenomena occurring within a single sentence, in particular constructions containing punctuation, on the assumption that the resulting constituent analysis will be then passed on to the discourse grammar. The main job of the sentence grammar, then, is to produce a structure that makes the appropriate units easily accessible to the discourse grammar. Section 3.7 below gives an example of how this might work using a discourse grammar of the type proposed by Webber and Gardent [1998].

3.1 Why TAG?

Lexicalized Tree Adjoining Grammar (LTAG) is a linguistically attractive grammar formalism which has descended from Tree Adjunct Languages, introduced in [Joshi et al.1975], and non-lexicalized TAGs [Joshi1985]. LTAGs have been shown to have many linguistically useful properties, including an extended domain of locality – all of the arguments of an anchor, including both elements of a wh- dependency, are localized within a single elementary tree. Thus, both syntactic and semantic dependencies are expressed locally. For discussion of some linguistic issues and how these properties are advantageous in treating them, see [Kroch and Joshi1985; Frank1992]. This localization provides an elegant framework for handling clausal level information, since each simple clause (usually a verb and its arguments) is a single tree. One case where this property is obviously useful is with the verbs of saying used with direct speech, where the subject and verb may follow or be embedded in its complement clause. This is shown in sentence (1). This order absolutely requires punctuation, and the requirement is easily expressed in the LTAG framework. The quotation marks are handled by a single tree which has both sets of marks as anchors, ensuring that they appear in pairs no matter how large a constituent they enclose; the arguments of the verbs *love*, *audition* and *gush* all occur in the elementary structures associated with those trees.

(1) "I'd love to audition for you", she gushed. [cf09]

In addition, DERIVATION TREES are built for each parse that show which trees are used and how they are combined. From these derivation trees, we can extract the relations between the various components, and we can tell what the basic sentence type is (e.g. passive, question, imperative, etc.). In terms of identifying complex (multi-clausal) sentences, the derivation trees show precisely which clausal trees have been used and what their syntactic relationship is. TAG is quite unique in providing a structure which shows the "history" of the derivation in such a transparent way; in other formalisms, this information is difficult to track or may even be lost altogether.

The work discussed here has been incorporated into a large English LTAG which has been developed as part of the XTAG project. XTAG is a wide-coverage grammar which includes a morphological analyzer, a part-of-speech tagger, a large syntactic

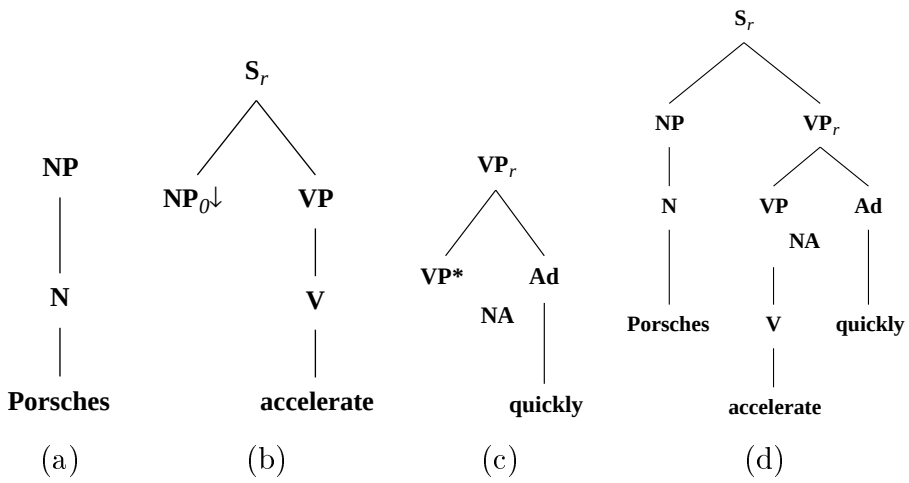
lexicon, and a parser. For more details on XTAG and the rest of the English grammar, see [XTAG-Group1995; Doran et al.1994].

One could certainly envision a similar analysis executed some other lexicalized formalism, like HPSG, LFG, or CCG. LTAG has two inherent features which make it more attractive in handling punctuation. First, its elementary units encode *constructions*, i.e. words in structured contexts, and these are the appropriate domains over which to state certain semantic and pragmatic constraints. For example, we have a tree for each verb subcategorization which allows some argument to be topicalized, and we can directly associate with that tree the requirement that the topicalized element be in a salient poset relation to other discourse entities (cf [Ward1985] for the particulars of topicalization). Second, LTAG produces derivational histories in the form of the derivation trees, which record the elementary trees used in a given derivation as well as the relationships between them. Many of the properties of punctuation relate to structural differences in how texts are built; for instance, where in the sentence a modifier is places can determine whether or not it needs to be enclosed by punctuation marks. An important benefit of LTAG is that it gives us access to that structure at both the individual tree level (in the grammar) and at the sentential level (via the derivation trees). Naturally, the other formalisms might turn out to have some advantages over LTAG lacks, for instance easy access to non-standard constituents in CCG.

3.1.1 LTAG in brief

The basic units of any TAG grammar are ELEMENTARY TREES, of which there are two types: INITIAL and AUXILIARY. Two combining operations are used: *substitution* and *adjunction*. Initial trees contain only argument positions, marked with $\downarrow$, where other initial trees must be substituted. Trees 3.1(a) and (b) are both initial trees, and (d) shows (a) substituted as the subject of (b). Auxiliary trees can also have argument positions, but they differ in having a distinguished leaf called the *foot* (marked with $*$) which has the same label as the root. These trees adjoin, or are spliced, into other trees. Tree 3.1(c) is an auxiliary adverb tree, and in (d) it has adjoined at the VP node.

With lexicalization [Schabes et al.1988], each elementary tree in the grammar is associated with at least one lexical ANCHOR (possibly more than one, for instance in handling idioms), and likewise every lexical item selects at least one tree in the grammar. The grammar used here is fully lexicalized, and uses feature structures [Vijay-Shanker and Joshi1991]. Figure 3.2 shows some LTAG trees and briefly illustrates the substitution and adjunction operations and how features are used.



(a) (b) (c) (d)

Figure 3.1: Basic LTAG trees: (a) initial NP tree, (b) initial S tree, (c) auxiliary adverb tree, and (d) S with NP substituted and adverb adjoined.

3.2 The current LTAG analysis of punctuation

Many parsers require that punctuation be stripped out of the input. Since punctuation is often optional, this sometimes has no effect. However, there are a number of constructions which must obligatorily contain punctuation and adding analyses of these to the grammar without the punctuation would lead to severe over-generation. An especially common example is noun appositives (example (2) has two appositives). Without access to punctuation, one would have to allow every combinatorial possibility of NPs in noun sequences, which is highly undesirable (especially since there is already unavoidable noun-noun compounding ambiguity). Aside from coverage issues, it is also preferable to take input "as is" and do as little editing as possible. With the addition of punctuation to the XTAG grammar, we need only do/assume the conversion of certain sequences of punctuation into the "British" order (this is discussed in more detail below in Section 3.5.2).

(2) But Tony Robinson, *the current sheriff of Nottingham* – **a job that really exists** – rejected the theory, saying that "as far as we are concerned, Robin Hood was a Nottinghamshire lad." [clari.living.celebrities]

The only grammar I know of with a systematic treatment of punctuation is a POS-tag sequence grammar developed by Briscoe and Carroll [1995] using the Alvey Natural Language Tools as a starting point, which includes Ted Briscoe's analysis of punctuation [Briscoe1994]. This grammar does not look at the particular lexical items in the input string, only the POS sequence. However, it does treat punctuation "lexically" to a certain extent, in that each punctuation mark occurs in a range of discourse grammar rules.

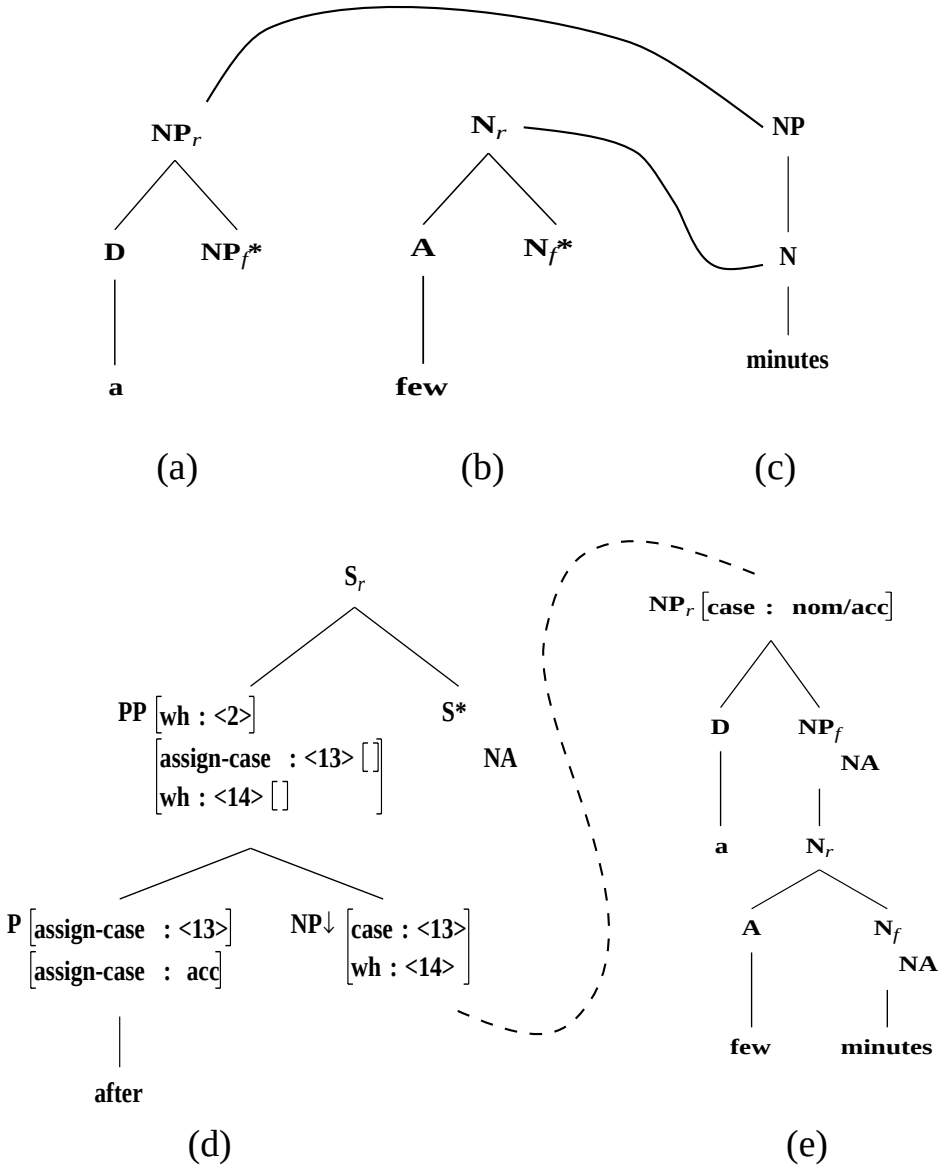


Figure 3.2: Sample LTAG trees: (a) and (b) are adjunction trees, which adjoin onto (c) as indicated by the solid lines. The resulting tree (e), then substitutes into the NP argument position of (d). (d) and (e) also show how features are used—the preposition which anchors (d) assigns accusative case, which will unify with the case feature at the root of (e). The NP has accusative/nominative as its case value, passed up from the head N, and received from the morphological analyzer. Figure 3.3 shows the resulting derived and derivation trees for this PP.

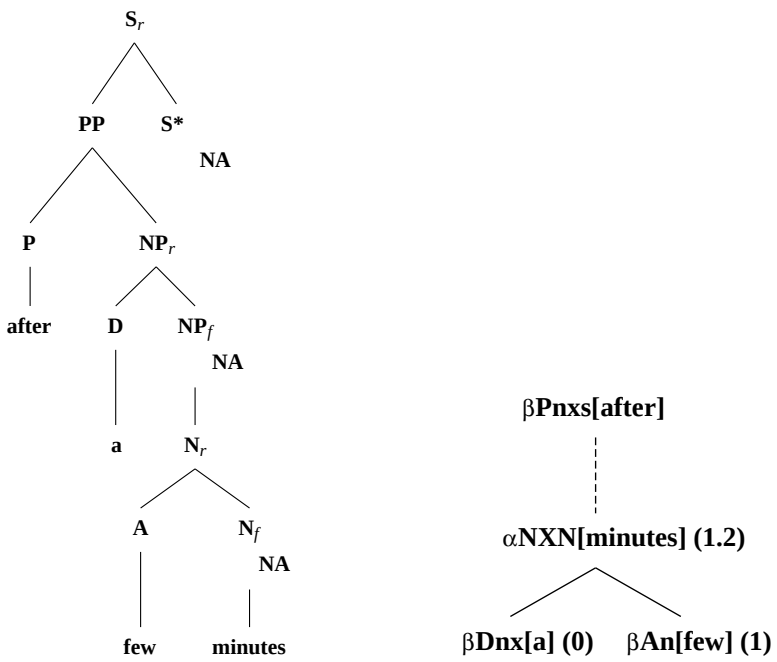


Figure 3.3: (a) Derived and (b) Derivation Trees for *after a few minutes* as a pre-sentential modifier. The derived tree shows the phrase structure which results from combining the elementary trees shown in Figure 3.2 and the derivation tree shows how those elements were combined. The solid lines indicate an adjunction operation and the dotted lines show substitution. The numbers in parentheses after the tree names give the Gorn-address at which the operation has taken place.

An analysis of punctuation has already been completed and integrated into the XTAG system. The analysis has been developed with some consideration of other work on the subject, but primarily through examination of naturally occurring data. The Brown Corpus has been the main source of data thus far, as it contains a variety of text genre. The new trees are of two types. The first have the punctuation marks as anchors, reflecting the fact that they do not specify the lexical content of the constructions they license/participate in. For example, any NP except a pronoun can be an appositive, and this is reflected in the analysis by having the NP position as a substitution site in the NP appositive tree (Figure 3.4). Figure 3.4 also illustrates how features can be used to constrain certain aspects of the argument positions (blocking the appositive NP from being a pronoun). This tree raises the interesting question of case assignment. There are no obvious case-assigners in most of these constructions (e.g. appositive, non-verbal parenthetical, vocative), and pronouns are blocked so it is hard to tell in English what the case is. One would have to either claim that they were "sharing" the case of the element they modify, or that they are

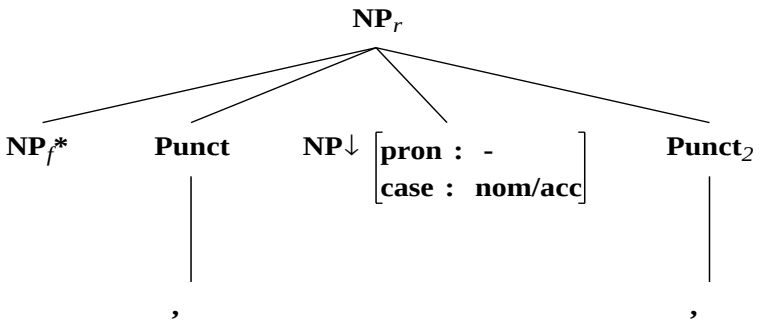


Figure 3.4: The non-peripheral NP appositive tree, showing relevant features.

receiving structural or default case; cross-linguistic investigation might shed some light on this issue. Some of the other punctuation trees are listed in Table 3.1, with examples of their use.

The XTAG part-of-speech tagger currently tags every punctuation mark as itself. These tags are all converted to the part-of-speech tag **Punct** before parsing.[3] This allows us to treat the punctuation marks as a single part-of-speech class. They then have features which distinguish amongst them. Wherever possible, we have the punctuation marks as anchors, to facilitate early filtering.

The full set of punctuation marks are separated into three classes: **balanced**, **structural**[4] and **terminal**. The balanced punctuation marks are quotes and parentheses, separating are commas, dashes, semi-colons and colons, and terminal are periods, exclamation points and question marks. Thus, the <**punct**> feature is complex (like the <**agr**> feature), yielding feature equations like <**Punct bal = paren**> or <**Punct term = excl**>. These three types of punctuation are essentially independent sub-systems, and a given constituent will typically have only one of each type. Separating and terminal punctuation marks do not occur adjacent to other members of the same class, but may occasionally occur adjacent to members of the other class, e.g. a question mark on a clause which is separated by a dash from a second clause. Balanced punctuation marks are sometimes adjacent to one another, e.g. quotes immediately inside of parentheses as in example (3). The <**punct**> feature allows us to control these local interactions.

(3) Each enjoys seeing the other hit home runs ("I hope Roger hits 80", Mantle says), and each enjoys even more seeing himself hit home runs ("and I hope I

[2]The names of the trees have the following key features: α denotes an initial tree and β denotes an auxiliary tree, the rest of the name encodes the frontier of the tree from left to right, and the capitalized elements are the anchor(s).

[3]Commas are assigned two POS tags, **Punct** and **Conj**. As **Conj**, the comma selects the coordination trees also anchored by lexical conjunctions.

[4]Meyers' term

Tree Name	**Anchor**	**Function**
αPU	Punct	Elementary tree for substituted punctuation marks
βPUs	Punct-comma	Adjoins after Pre-S modifiers *Last week , the market reported . . .*
βsPUs	Punct-dash,semi-colon	Combines clauses *Max left – he was very tired*
βPU$< x >$PU	Punct-parens, single quote, double quote	Used for optional parens or quotes around all categories *John "Bull Dog" Smith*
βnxPUnxpu	Punct- comma or dash	Non-peripheral Appositive NPs *John Smith, the leading cyclist...*
βpuARBpuvx	Adverb	Non-peripheral parenthetical adverb *Smith, however, has been found...*
βPUpxPUvx	Punct- comma or dash	Non-peripheral parenthetical PP *Smith, in recent months, began...*
βpunxVpuvx	Verbs of saying	Reported speech *Mary, John says, has vanished*
βsPUnx	Punct - comma	Vocative *You were there, Stanley*
βsPU	Punct - period, excl. point, question mark	Sentence-final punctuation

Table 3.1: Sample Punctuation Trees in Current XTAG Grammar.[2]

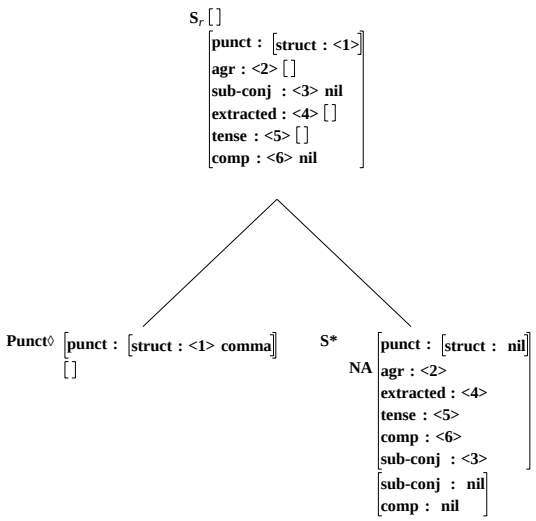


Figure 3.5: Tree for adjoining a comma after a Pre-S adjunct, showing **punct struct** feature (complete feature structure not shown for all features);e.g. *Along the way, he meets a solicitous Christian chauffeur*

hit 81"). [Brown:ca39]

The **structural** features are obviously the most interesting, and the most complicated. A sample tree is shown in Figure 3.5 with the relevant features displayed: the value of <**punct struct = comma**> is passed from the anchoring comma to the root—this allows us to block other structural punctuation from adjoining above the comma. The feature <**punct struct = nil**> on the foot blocks the comma from adjoining directly above another punctuation mark.

We also need to control non-local interaction of punctuation marks. Two cases of this are so-called quote alternation ([Quirk et al.1985],III.21), wherein embedded quotation marks must alternate between single and double, and the impossibility of embedding an item containing a colon inside of another item containing a colon. Thus, we have a fourth value for <**punct**>, <**contains colon/dquote/etc.** +/− >, which indicates whether or not a constituent contains a particular punctuation mark. This feature is percolated through all auxiliary trees. Things which may not embed are colons under colons, semi-colons, dashes or commas, and semi-colons under semi-colons or commas. Example (5) shows an impossible variation on the grammatical (4), with the *since* clause embedded under an earlier dash-separated adjunct. Although it is not extremely common, parentheses may appear inside of parentheses, say with an academic reference inside a parenthesized sentence.

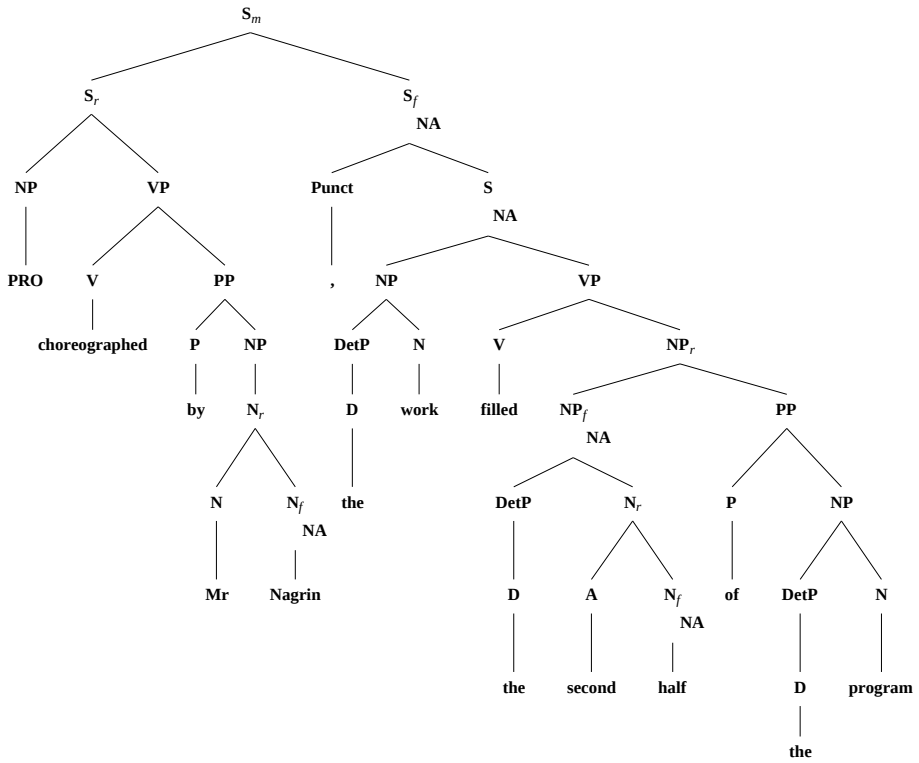


Figure 3.6: Sample tree containing punctuation; comma adjoins using the tree shown above.

(4) Now the basic question to be asked in this situation is what motivates the manipulators, that is, what are their values? – since, as Courtenay says, "Nobody should play with lives the way we do unless he's motivated by the highest ideals". [cg69]

(5) #Now the basic question to be asked in this situation is what motivates the manipulators– that is, what are their values? – since, as Courtenay says, "Nobody should play with lives the way we do unless he's motivated by the highest ideals".

One interesting question is how to control the interaction of various kinds of punctuation on the right periphery of the clause. For instance, an expression which would be bracketed by dashes sentence-medially only has a left dash when it is in sentence-final position (6). It is here that Nunberg's absorption takes place – being lower ranked the right dash is absorbed by the sentence-final punctuation. However, when an abbreviation ends a sentence which also has a period, the period does disappear (or get absorbed) (7), but if there is an exclamation or question mark, you get both marks (8). In the XTAG system, we handle this latter case with a tokenizer which identifies end-of-sentence boundaries and inserts spaces between the last word and the final punctuation mark, but not between abbreviations and their periods. The other cases are handled by having both symmetric and asymmetric rules for the relevant constructions (following [Briscoe1994]).

(6) The Kennedy plan alone would boost...the payroll tax to 6.5 per cent – 3.25 per cent each.

(7) ...as university alumni, as newspaper readers, etc.

(8) Are you indiscriminantly offering unnecessary medical services – flu shots, sun lamp treatments, etc.?

While in general I am trying to capture the uses of punctuation in American English as exemplified in corpora of American texts, there are culture/language-specific variations in how punctuation is used. One of the most extensively discussed is the placement of periods and other sentence-final punctuation marks relative quotation marks in American and British English. As noted above, this usage is not consistent in actual texts. My analysis assumes tokenization into the British format. In fact, marking of quotation is one of the areas where languages vary most; other punctuation marks are used rather more consistently, which ought to allow much of the current analysis to be transferred to languages other than English.

3.3 Strengths of the LTAG analysis

Nunberg wants to have the sentence grammar completely distinct from the discourse grammar. However, there are clear cases where the two must interact, such as when an extraposed argument must be separated from the main clause by a comma. Briscoe and Carroll [1995] argue that the two sets of rules need to be "interleaved" for efficient application, but that it is useful to access the discourse grammar rules independently of the rest of the grammar. One reason they give is so that the discourse grammar rules can be associated with individual semantic representations. They incorporate 26 "pure" text grammar rules as is into their existing sentence grammar, and then add information about punctuation into 44 syntactic rules as well. Jones [1994] treats punctuation marks as clitics that are realized as features on syntactic rules, which Briscoe and Carroll argue makes it virtually impossible to separate the two components.

The main distinction in the LTAG analysis is what items anchor the trees which introduce punctuation marks. The ones which are anchored by punctuation marks alone are purely text-grammatical. The ones which are anchored by other lexical items, such as verbs of saying, adverbs or prepositions, reflect places where the two grammars are intertwined. The need for these trees argues for a closer interaction between the sentence grammar and discourse grammar than Nunberg proposes, in line with Briscoe and Carroll's approach. It does mean that we need to make other trees in the grammar transparent to the punctuation features, but they still do not actually need to "know" anything about punctuation. Likewise, the punctuation trees need to be transparent to certain syntactic features, like agreement, but do not change any feature values. They do use features from the syntactic grammar to constrain the distribution of lexical material in text-adjuncts, for instance blocking pronominal NP appositives.

The TAG adjunction operation is advantageous in handling paired punctuation marks, because it allows us to keep both pieces of the complex object, e.g. a pair of parentheses or commas, in the same elementary tree, regardless of the size of the constituent they enclose. Another interesting fact about the auxiliary trees is that it seems to be the case that, in encoding rhetorical relations like those proposed in as theory like RST within an orthographic sentence, it is always the case that the satellite is realized syntactically as an adjunction structure.[5] This is certainly true in the discussion of encoding RST relations from [Scott and de Souza1990]. If this were found to be a correct generalization, it would further confirm the appropriateness of TAG as a representational framework for both the sentence grammar and discourse grammar.

[5]It could even be argued that clausal complement verbs, which anchor clausal auxiliary trees in LTAG, might encode a rhetorical relation like EVIDENCE or a modal operator.

3.4 Limitations of the LTAG analysis

There are a number of limitations of the current analysis, some due to the system itself, and others due to restrictions imposed by the LTAG formalism.

3.4.1 Restrictions resulting from the LTAG formalism

Because LTAG is tree-based and requires clausal trees to contain all arguments locally, the grammar can currently only handle balanced punctuation marks around constituents (but see [Sarkar and Joshi1996] for ideas on handling non-standard constituents in LTAG). There are rare instances of quotation marks and parentheses around non-constituents, and these are beyond the scope of the present LTAG implementation. For instance, example (9) has quotes which enclose the NP head and the VP of an infinitival relative clause, but not the object of the verb.

(9) Cartoonist Garry Trudeau is suing the Writers Guild of America East for $11 million, alleging it mounted a "campaign to harass and punish" him for crossing a screenwriters' picket line. [wsj0049]

A more significant concern in using LTAG is that the adjunction operation makes it very difficult to enforce certain right-periphery constraints, e.g. ensuring that the asymmetric variants of text-adjuncts only appear on the right-periphery of the clause, or that colon-expansions are also the rightmost elements. Ensuring that nothing adjoined to the right of these sorts of text-adjuncts would require an elaborate feature passing system. I believe it is better to impose these constraints in a later processing step, and simply disallow any derivations where the relevant text-adjuncts appear internally.

3.4.2 Restrictions resulting from the XTAG System

The grammar within which these rules are incorporated does not currently allow schematic trees. As a result, trees which could be quite naturally schematized, such as those allowing parentheses or quotation marks around any constituent (e.g. X $\rightarrow$ " X "), must be enumerated for every possible constituent. In principle, TAG could allow such trees and they would allow for a more streamlined and elegant treatment of these cross-categorical patterns.

Furthermore, the system is designed to take its input one sentence at a time, which prevents us from being able to handle, e.g. quotes around multiple sentences, or colon expansions containing multiple orthographic sentences. I have circumvented this in a rather crude way by allowing the terminal punctuation marks to select a tree which conjoins two clauses. Alternatively, one could handle the bracketing punctuation marks scoping over several sentences, or even paragraphs, outside the grammar altogether, for instance with some sort of stack-model integrated into the tokenization process.

3.5 Descriptions of the various trees

The following sections describe the tree TEMPLATES for punctuation which have been added to the XTAG English grammar. A template is a tree which has not yet been lexicalized, i.e. is unanchored. Some of them were already listed in Table 3.1, but are described in greater detail here. The template simply tells you the topology of the structure, with features assigned values only if those features are particular to the structure, rather than to the association between the structure and the lexical items which select it. The semantics of the tree cannot be abstracted to the template level, so the same tree with a different anchor will usually have a completely different meaning/function. The data these structures are based on was collected from primarily from the Brown Corpus, but also opportunistically from other sources (newsgroups, other on-line corpora, literary texts, etc.)

3.5.1 Appositives, parentheticals and vocatives

These trees handle constructions where additional lexical material is only licensed in conjunction with particular punctuation marks. Since the lexical material is unconstrained (virtually any noun can occur as an appositive), the punctuation marks are anchors and the other nodes are substitution sites. There are cases where the lexical material is restricted, as with parenthetical adverbs like *however*, and in those cases we have the adverb as the anchor and the punctuation marks as substitution sites.

When these constructions can appear inside of clauses (non-peripherally), they must be separated by punctuation marks on both sides. However, when they occur peripherally they have either a preceding or following punctuation mark. We handle this by having both peripheral and non-peripheral trees for the relevant constructions. The alternative is to insert the second (following) punctuation mark in the tokenization process (i.e. insert a comma before the period when an appositive appears on the last NP of a sentence). However, this is very difficult to do accurately.

β**nxPUnxPU**[6]

The symmetric (non-peripheral) tree for NP appositives, anchored by: comma, dash or parentheses. It is shown in Figure 3.7 anchored by parentheses.

(10) The music here , Russell Smith's "Tetrameron " , sounded good. [cc09]

[6]The XTAG tree naming convention is that initial trees start with α and auxiliary trees with β. The main part of the name consists of the frontier nodes traversing the tree from left to right, with the anchor or anchors capitalized. The node names should fairly intuitive, except that we use **x** for phrases because **p** is used for particles, so an NP, for example, is labeled **nx**. This tree's name tells us that it is an auxiliary tree with frontier nodes NP, Punct, NP and Punct, with the two punctuation marks as anchors.

(11) ...cost 2 million pounds (3 million dollars)

(12) ...some analysts believe the two recent natural disasters – Hurricane Hugo and the San Francisco earthquake – will carry economic ramifications.... [wsj]

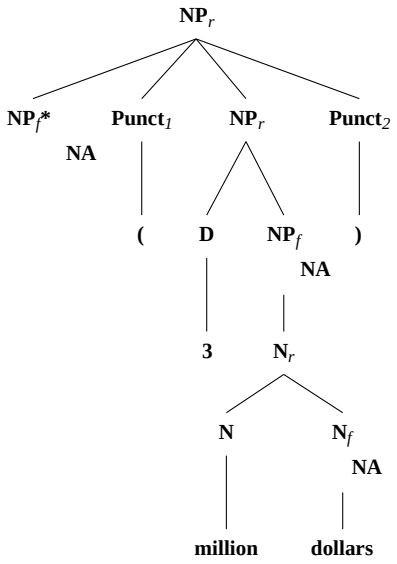


Figure 3.7: The βnxPUnxPU tree, anchored by parentheses

The punctuation marks are the anchors and the appositive NP is substituted. The appositive can be conjoined, but only with a lexical conjunction (not with a comma). Appositives with commas or dashes cannot be pronouns, although they may be conjuncts containing pronouns; likewise, they cannot modify pronouns. When used with parentheses this tree typically presents an alternative rather than an appositive, so a pronoun is possible. Finally, the appositive position is restricted to having nominative or accusative case to block PRO from appearing here.

Appositives can be embedded, as in (13), but do not seem to be able to stack on a single NP. In this they are more like restrictive relatives than appositive relatives, which typically can stack. This topic will be addressed in more detail in Chapter 4.

(13) ...noted Simon Briscoe, UK economist for Midland Montagu, a unit of Midland Bank PLC.

βnPUnxPU

The symmetric (non-peripheral) tree for N-level NP appositives is anchored by comma. The modifier is typically an address. Examples such as (14) show that these modifiers are attached at N, rather than NP. *Carrier* is not an appositive

on *Menlo Park*, as it would be if these were simply stacked appositives. Rather, *Calif.* modifies *Menlo Park*, and that entire complex is compounded with *carrier*, as shown in the correct derivation in Figure 3.8. Because this distinction is less clear when the modifier is peripheral (e.g. ends the sentence), and it would be difficult to distinguish between NP and N attachment, we do not currently allow a peripheral N-level attachment.

(14) An official at Consolidated Freightways Inc., a Menlo Park, Calif., less-than-truckload carrier , said...

(15) Rep. Ronnie Flippo (D., Ala.), of the delegation, says...

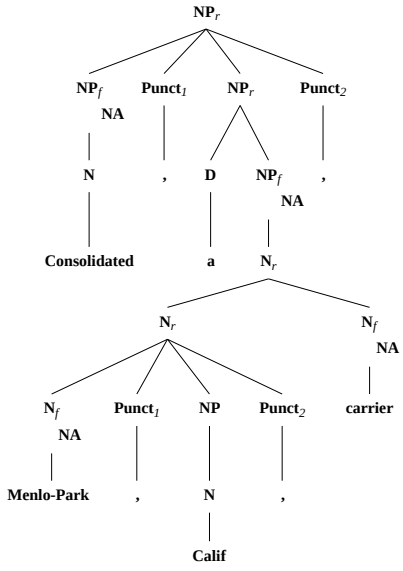


Figure 3.8: An N-level modifier, using the βnPUnx tree

βnxPUnx

This tree, which can be anchored by a comma, dash or colon, handles asymmetric (peripheral) NP appositives and NP colon expansions of NPs. Recall that the meaning of the tree derives only from its association with a particular anchor, so we can use the same structure for appositives and colon-expansions without making any claims that they have the same semantics. Figure 3.9 shows this tree anchored by a dash. Like the symmetric appositive tree, βnxPUnxpu, the asymmetric appositive cannot be a pronoun, while the colon expansion can. Thus, this constraint comes from the syntactic entry in both cases rather than being built into the tree.

(16) the bank's 90% shareholder – Petroliam Nasional Bhd. [brown]

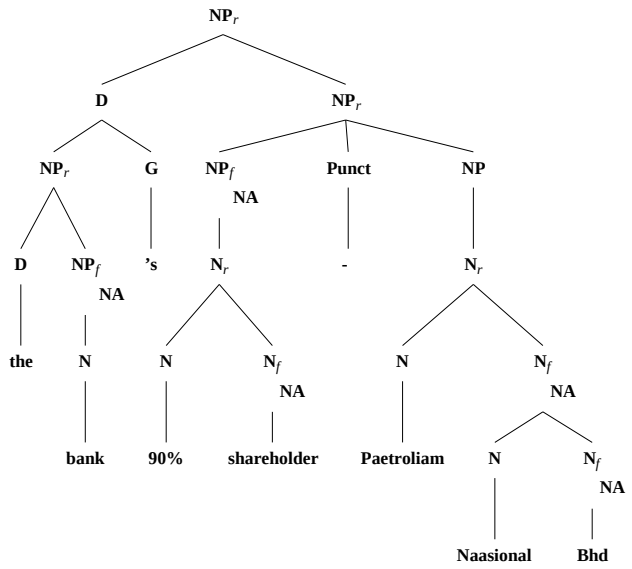


Figure 3.9: The derived tree for an NP with an peripheral, dash-separated appositive

(17) ...said Chris Dillow, senior U.K. economist at Nomura Research Institute. [wsj]

(18) ...qualities that are seldom found in one work: Scrupulous scholarship, a fund of personal experience,... [cc06]

(19) I had eyes for only one person: him.

The colon expansion cannot contain a second colon expansion, so the foot S has the feature NP.t:<punct contains colon> = $-$.

βPUpxPUvx

Tree for pre-VP parenthetical PP. It is anchored by commas or dashes rather than the preposition, because the class of prepositions that can appear here is unrestricted.

(20) John, in a fit of anger, broke the vase

(21) Mary, just within the last year, has totalled two cars

These are not interpretable as NP modifiers.

Figures 3.10 and 3.11 show this tree alone and as part of the parse for (20).

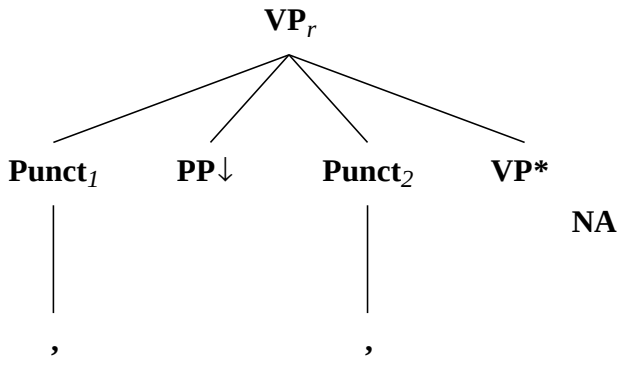


Figure 3.10: The βPUpxPUvx tree, anchored by commas

βpuARBpuvx

Parenthetical adverbs—*however*, *though*, etc. Since the class of adverbs is highly restricted, this tree is anchored by the adverb and the punctuation marks substitute (cf. previous tree where any PP can participate and punctuation marks are anchors). The punctuation marks may be either commas or dashes. Like the parenthetical PP above, these are not interpretable as NP modifiers.

(22) The new argument over the notification guideline, however, could sour any atmosphere of cooperation that existed. [WSJ]

βsPUnx

Sentence final vocative, anchored by comma:

(23) You were there, Stanley/my boy.

Also, when anchored by colon, NP expansion on S. These often appear to be extraposed modifiers of some internal NP. The NP must be quite heavy, and is usually a list:

(24) Of the major expansions in 1960, three were financed under the R. I. Industrial Building Authority's 100% guaranteed mortgage plan: Collyer Wire, Leesona Corporation, and American Tube & Controls.

A simplified version of this sentence is shown in figure 3.12. The NP cannot be a pronoun in the colon expansion case, which is compatible with it being an extraposed predicate. Both vocatives and colon expansions are restricted to appear on tensed clauses (indicative or imperative).

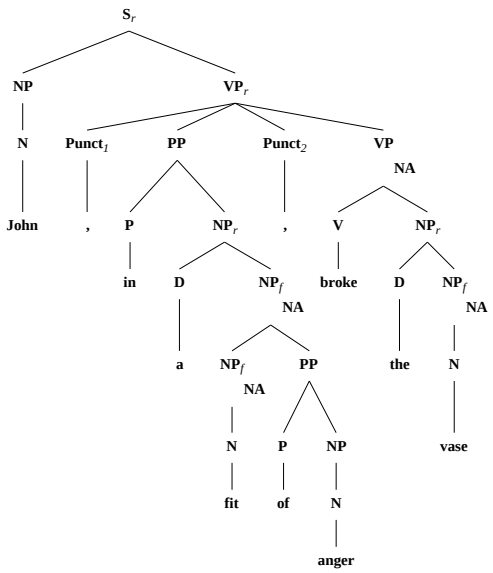


Figure 3.11: Tree illustrating the use of βPUpxPUvx

βnxPUs

Tree for sentence initial vocatives, anchored by a comma:

(25) Stanley/my boy, you were there

The noun phrase may be anything but a pronoun, although it is most commonly a proper noun. The clause adjoined to must be indicative or imperative.

3.5.2 Bracketing punctuation

Trees: βPUxPU, where x = any node label
These trees are selected by parentheses and quotes and can adjoin onto any node type, whether a head or a phrasal constituent. This handles things in parentheses or quotes which are syntactically integrated into the surrounding context. Figure 3.13 shows the βPUsPU tree anchored by parentheses, and this tree along with βPUnxPU in a derived tree.

(26) Dick Carroll and his accordion (which we now refer to as "Freida") held over at Bahia Cabana where "Sir" Judson Smith brings in his calypso capers Oct. 13 . [brown:ca31]

(27) ...noted that the term "teacher-employee" (as opposed to, e.g., "maintenance employee") was a not inapt description. [brown:ca35]

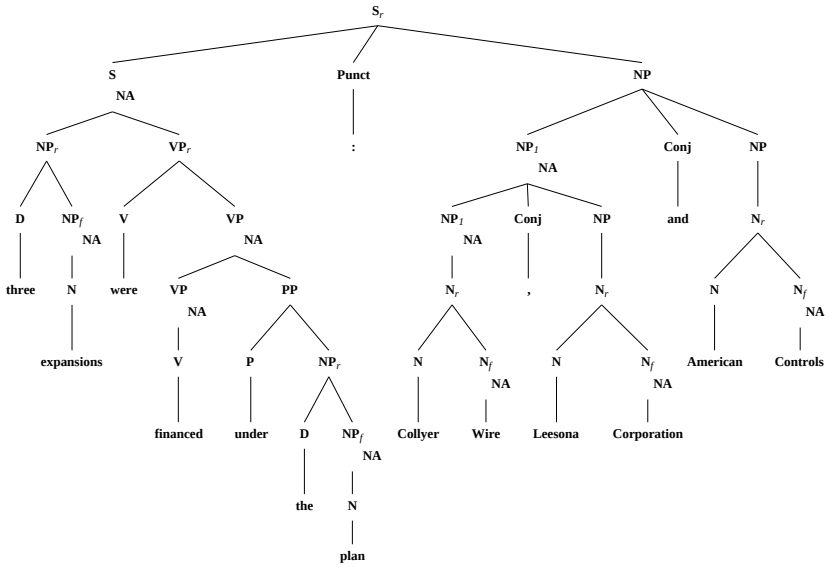


Figure 3.12: A tree illustrating the use of sPUnx for a colon expansion attached at S.

There is a convention in English that quotes embedded in quotes alternate between single and double; in American English the outermost are double quotes, while in British English they are single. The **contains** feature is used to control this alternation. The trees anchored by double quotation marks have the feature **punct contains dquote** $= -$ on the foot node and the feature **punct contains dquote** $= +$ on the root. All adjunction trees are transparent to the the **contains** feature, so if any tree below the double quote is itself enclosed in double quotes the derivation will fail. Likewise with the trees anchored by single quotes. The quote trees in effect "toggle" the **contains Xquote** feature. Immediate proximity is handled by the **punct balanced** feature, which allows quotes inside of parentheses, but not vice-versa.

In addition, American English typically places/moves periods (and commas) inside of quotation marks when they would logically occur outside, as in example (28). The comma in the first part of the quote is not part of the quote, but rather part of the parenthetical quoting clause. However, by convention it is shifted inside the quote, as is the final period. British English does not do this. We assume here that the input has already been tokenized into the British format.

(28) "You can't do this to us," Diane screamed. "We are Americans."

The βPUsPU can handle quotation marks around multiple sentences, since the sPUs tree allows us to join two sentences with a period, exclamation point or question mark. Currently, however, we cannot handle the style where only an open quote appears at the beginning of a paragraph when the quotation extends over multiple

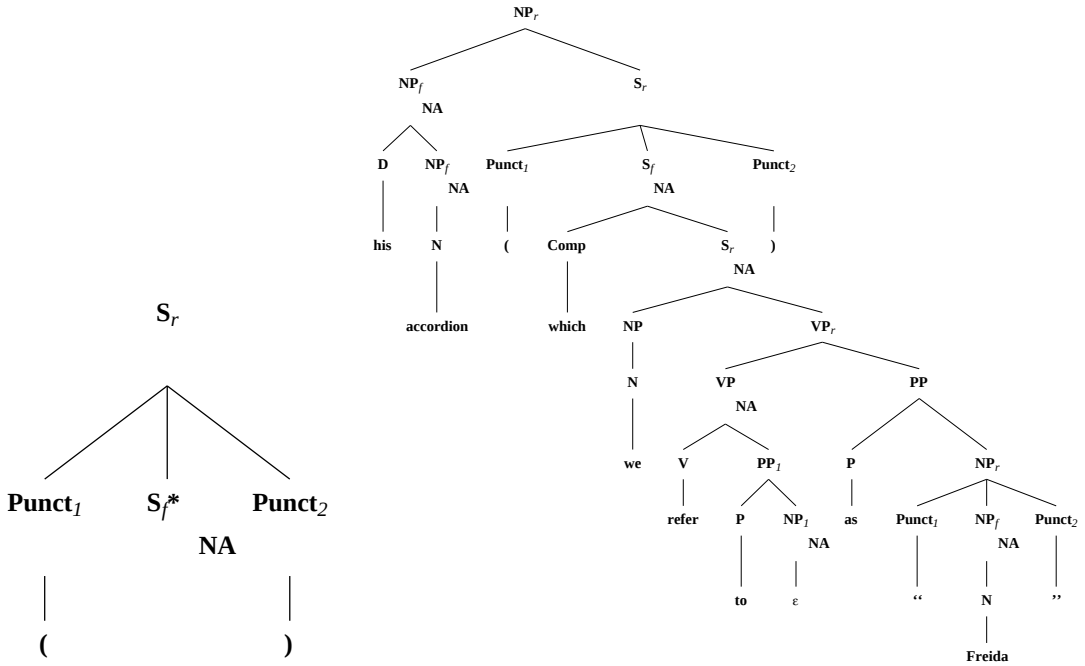


Figure 3.13: βPUsPU anchored by parentheses, and in a derivation, along with βPUnxPU

paragraphs. We could allow a lone open quote to select the βPUs tree, if this were deemed desirable for some application of the grammar.

Also, the βPUsPU is selected by a pair of commas to handle non-peripheral appositive relative clauses, such as (29). Restrictive and appositive relative clauses are not syntactically differentiated in the XTAG grammar (cf. [XTAG-Group1995]:Ch. 14).

(29) This news, announced by Jerome Toobin, the orchestra's administrative director, brought applause ... [brown:cc09]

The trees discussed in this section will only allow balanced punctuation marks to adjoin to standard constituents. We will not get them around non-standard constituents, as in (30).

(30) Mary asked him to leave (and he left)

3.5.3 Punctuation trees containing no lexical material

αPU

This is the elementary tree for substitution of punctuation marks. It is used in the adjunct reported speech trees (cf. Section 5.6.1), where including the punctuation

mark as an anchor along with the verb of saying would require a new entry for every tree selecting the relevant tree families. It is also used in the tree for parenthetical adverbs (βpuARBpuvx), and for S-adjoined PPs and adverbs (βspuARB and βspuPnx).

βPUs

Anchored by comma, it allows comma-separated clause initial adjuncts, (31)–(32).

(31) Here, as in "Journal", Mr. Louis has given himself the lion's share of the dancing... [cc09]

(32) Choreographed by Mr. Nagrin, the work filled the second half of a program

To keep this tree from appearing on root Ss (i.e. *, sentence*), we have a root constraint that <**punct struct = nil**> (similar to the requirement that root Ss be tensed, i.e. <**mode = ind/imp**>). The <**punct struct**> **= nil** feature on the foot blocks stacking of multiple punctuation marks. This feature is shown in the tree in Figure 3.14.

This tree can be also used by adjuncts on embedded clauses:

(33) One might expect that *in a poetic career of seventy-odd years,* some changes in style and method would have occurred, some development taken place. [cj65]

These adjuncts sometimes have commas on both sides of the adjunct, or, like (33), only have them at the end of the adjunct.

Finally, this tree is also used for peripheral appositive relative clauses.

(34) Interest may remain limited into tomorrow's U.K. trade figures, which the market will be watching closely to see if there is any improvement after disappointing numbers in the previous two months.

βsPUs

This tree handles clausal "coordination" with comma, dash, colon, semi-colon or any of the terminal punctuation marks. One clause must be tensed, either indicative or imperative. The second may also be infinitival or participial with the separating punctuation marks, but must be indicative or imperative with the terminal marks; with a comma, it may only be indicative. The two clauses need not share the same mode. NB: Allowing the terminal punctuation marks to anchor this tree allows us to parse sequences of multiple sentences. This is not the usual mode of parsing, but it is useful to development purposes.

(35) For critics, Hardy has had no poetic periods – one does not speak of early Hardy or late Hardy, or of the London or Max Gate period....

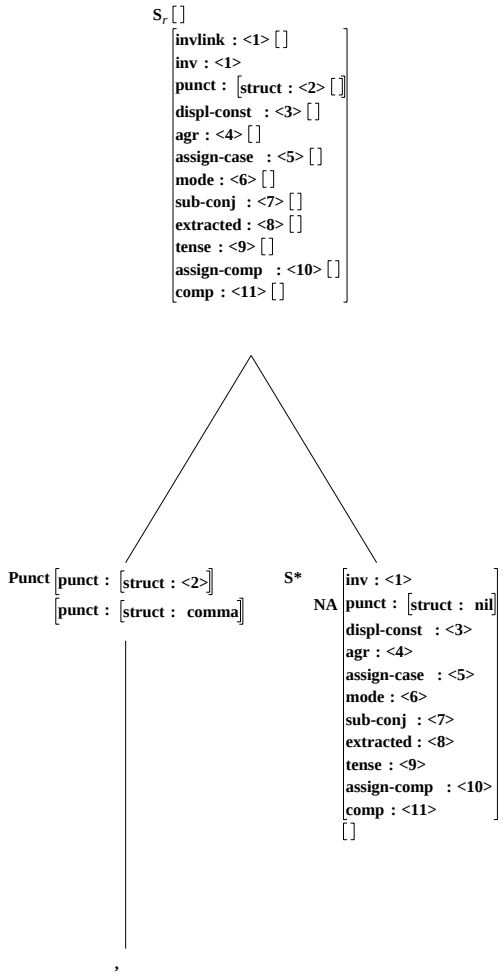


Figure 3.14: βPUs, with features displayed

(36) Then there was exercise, boating and hiking, which was not only good for you but also made you more virile: the thought of strenuous activity left him exhausted.

(37) Expressed differently: if the price for becoming a faithful follower... [cd02]

(38) Expressing it differently: if the price for becoming a faithful follower...

(39) To express it differently: if the price for becoming a faithful follower...

This construction is one of the few where two non-bracketing punctuation marks can be adjacent. It is possible (if rare) for the first clause to end with a question mark or exclamation point, when the two clauses are conjoined with a semi-colon,

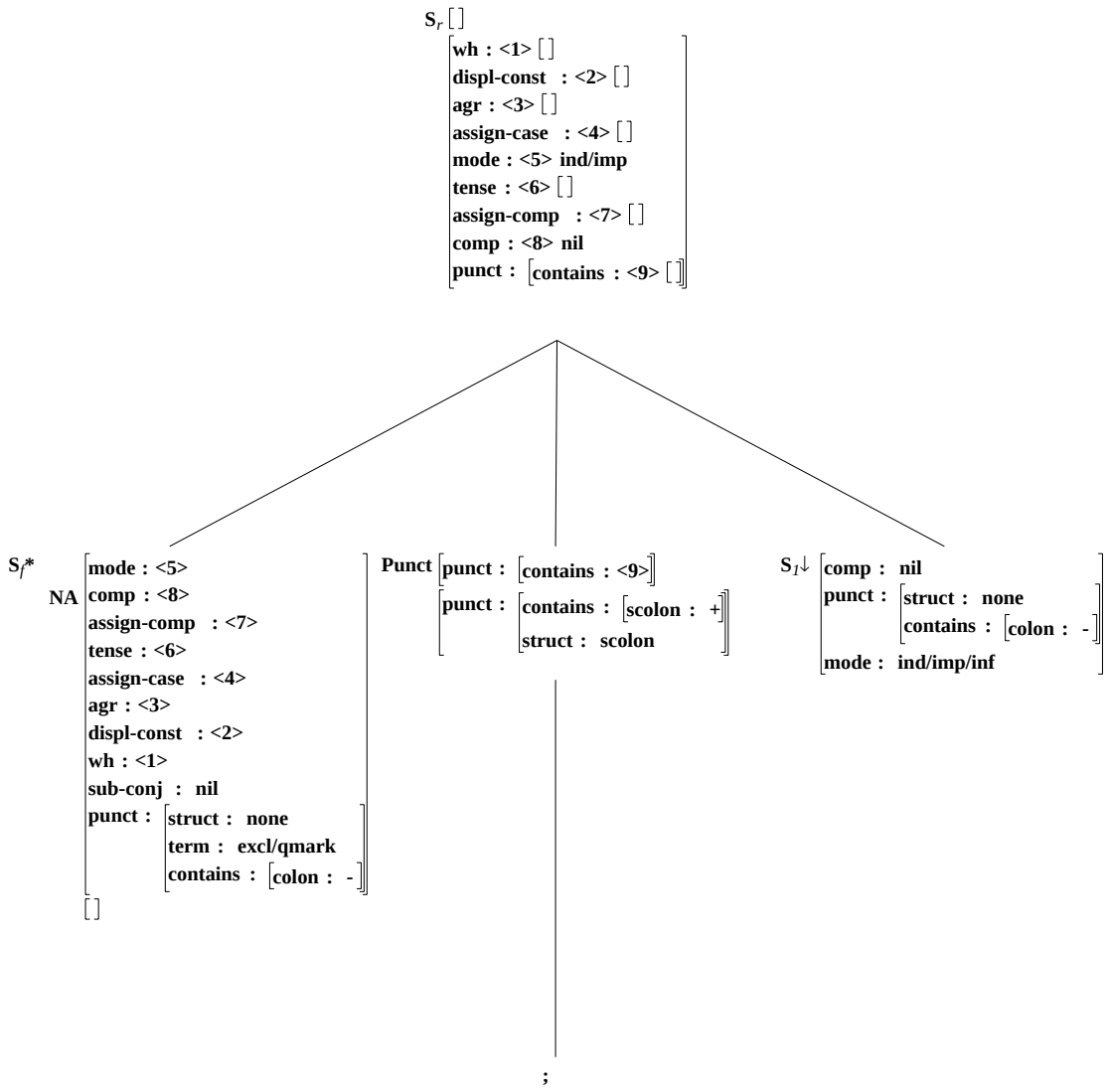


Figure 3.15: βsPUs, with features displayed

colon or dash. Features on the foot node, as shown in Figure 3.15, control this interaction.

Complementizers are not permitted on either conjunct. Subordinating conjunctions sometimes appear on the right conjunct, but seem to be impossible on the left:

(40) Killpath would just have to go out and drag Gun back by the heels once an hour; *because* he'd be damned if he was going to be a mid-watch pencil-pusher. [Brown, cl17]

(41) The best rule of thumb for detecting corked wine (provided the eye has not already spotted it) is to smell the wet end of the cork after pulling it: *if* it smells of wine, the bottle is probably all right; if it smells of cork, one has grounds for suspicion. [cf27]

β**sPU**

This tree handles the sentence final punctuation marks when selected by a question mark, exclamation point or period. One could also require a final punctuation mark for all clauses, but such an approach would not allow non-periods to occur internally, for instance before a semi-colon or dash as noted above in the description of βsPUs. This tree currently only adjoins to indicative or imperative (root) clauses.

(42) He left!

(43) Get lost.

(44) Get lost?

The feature **punct bal= nil** on the foot node ensures that this tree only adjoins inside of parentheses or quotation marks completely enclosing a sentence (45), but does not restrict it from adjoining to clause which ends with balanced punctuation if only the end of the clause is contained in the parentheses or quotes (46).

(45) (John then left.)

(46) (John then left).

(47) Mary asked him to leave (immediately).

β**vPU**

This tree is anchored by a colon or a dash, and occurs between a verb and its complement. These typically are lists.

(48) Printed material Available, on request, from U.S. Department of Agriculture, Washington 25, D.C., are: Cooperative Farm Credit Can Assist.... [Brown ch01]

βpPU

This tree is anchored by a colon or a dash, and occurs between a preposition and its complement. As with the tree above, this typically occurs with a conjoined complement.

(49) ...and utilization such as: (A) the protection of forage...

(50) ...can be represented as: Af.

3.5.4 Other trees

There are trees with punctuation substitution sites in both the object + sentential complement family (Tnx0Vnx1s2, e.g. *tell*) and the plain sentential complement family (Tnx0Vs1, e.g. *say*). These trees are anchored by the verbs of saying, and are discussed in Chapter 5. There are also subordinating conjunction trees which have punctuation sites. These are anchored by simple subordinating conjunctions (*until*) as well as complex ones (*in order, even if, as soon as*). Punctuation is required on both sides of the subordinate clause when it occurs between the subject and the verb, and before the clause when it follows the main clause.

βspuARB

In general, post-clausal modifiers are attached at the VP node, as you typically get scope ambiguity effects with negation (*John didn't leave today*—did he leave or not?). However, with post-sentential, comma-separated adverbs, there is no ambiguity—in *John didn't leave, today* he definitely did not leave. Since this tree is only selected by a subset of adverbs (namely, those which can appear pre-sententially), it is anchored by the adverb.

(51) The names of some of these products don't suggest the risk involved in buying them, either. [wsj]

βpuARBpuvx

This tree handles pre-verbal parenthetical adverbs. Like βspuARB, the tree is only selected by a subset of adverbs, e.g. *however*, *nonetheless*, so the adverb is the anchor.

(52) Most skilled industrial workers, nevertheless, still acquire their skills outside of formal training institutions. [Brown:cj38]

βspuPnx

Clause-final PP separated by a comma. Like the adverbs described above, these differ from VP adjoined PPs in taking widest scope.

(53) ...gold for current delivery settled at $367.30 an ounce, up 20 cents.

(54) It increases employee commitment to the company, with all that means for efficiency and quality control.

βnxPUa

Anchored by colon or dash, allows for post-modification of NPs by adjectives.

(55) Make no mistake, this Gorky Studio drama is a respectable import – aptly grave, carefully written, performed and directed.

3.6 Syntactic advantages of adding punctuation to a grammar

The XTAG grammar is intended to be very general, so as to offer the widest possible coverage of freely occurring English texts. On the other hand, one does not want to add constructions to the grammar which will cause rampant spurious ambiguity. This makes the system an ideal testbed for the syntactic aspects of the punctuation analysis. There are several ways in which punctuation should prove useful in large grammars, both with regard to the specific XTAG grammar and more generally: (1) improved coverage, (2) reduced ambiguity and (3) the ability to split long sentences into smaller, more manageable pieces. I will discuss each of these in turn. To a lesser extent, ambiguity reduction can also be used to evaluate the semantico-discourse component of the analysis, as will be discussed briefly in 3.6.2.

3.6.1 Improved grammar coverage

The first benefit of adding punctuation to the grammar is that it allows us to add some syntactically "exotic" constructions which we would have previously considered too unconstrained. Many such constructions occur with great frequency in naturally occurring texts. One case in point is the appositive construction, which allows an NP to modify another NP. Like appositive relative clauses, these NPs provide extra information (loosely speaking) about the nouns they modify, and they are separated from the surrounding syntax by commas – two if they are clause internal, and one if they are peripheral. Allowing these without punctuation would lead to severe ambiguity with strings of NPs. (See Chapter 4 for more detailed discussion of appositives.)

(56) This is in honor of John Ledyard, *class of 1773*, who scooped a canoe out of a handy tree and first set the course way back in his own student days. (Brown:cf30)

As is shown by example (56), these phrases must be "bracketed" by punctuation, which is precisely the sort of additional constraint we need. By adding a treatment of punctuation to the grammar, we should be able to recognize appositive constituents quite reliably.

Other reliably punctuated constructions which we will be able to add and use to evaluate improved coverage are:

- Parentheticals: *It may be,* **of course,** *that....*
- Reported speech: *But he,* **as I can now retort,** *was...*
  These are especially interesting, since they essentially have the subject and verb embedded within the complement of the verb. See Chapter 5 for more on these constructions.
- Compound sentences separated/conjoined by commas, semi-colons and dashes (alone, no lexical conjunct present): *For critics, Hardy has had no poetic periods – one does not speak of early Hardy or late Hardy...*
- Comma coordination: *...the detailed accents, phrasings and contours of the music...*
- Vocatives: *You were there,* **Stanley.**

None of these constructions were handled by the XTAG English grammar before it was extended to treat punctuation.

3.6.2 Reducing ambiguity in parsing

A second way to evaluate a punctuation grammar is by the additional constraints it provides for parsing. In developing a large grammar for any language, one of the fundamental concerns is the increase in ambiguity of derivations which invariably accompanies any increase in coverage of the language's constructions.

One source of disambiguating information, naturally, is pure statistics. The frequency of certain constructions can be calculated from a corpus of parsed sentences, and then these frequencies can be made use of by the parser. The XTAG system currently uses these sorts of general statistical information, which have proven extremely useful in improving the efficiency of the parser. Work by Joshi and Srinivas (1994) on "Supertagging" has taken advantage of the frequencies with which certain constructions (trees) are used with certain words or parts of speech. Frequency statistics collected from a large XTAG-parsed corpus are used to select the most

likely trees for each input word, which greatly increases the speed with which the parser produces analyses. In addition, the parser uses heuristics to rank the parses. Having selected the top trees for each word in the sentence, we add general preference weightings for certain trees and for certain lexical items, based on statistics collected from an LTAG parsed corpus. For instance, Topicalization is given a very low probability, as it is generally disprefered.

A second source of additional constraints is semantic knowledge. However, semantic knowledge is extremely difficult to assemble by hand, and is only practical within a highly constrained domain. For instance, in an airline reservation domain, one could use knowledge about the domain. Thus, a sentence like "Give me flights to Boston" would have the prepositional phrase unambiguously modifying "flights" (as opposed to modifying the verb phrase). Certain of this information can be inferred from statistics collected over corpora as described above, either from a single domain or mixed domains. The more restricted the domain, the more closely the statistics will reflect semantic "facts" about the domain (e.g. the probability of PPs contain proper names attaching to NPs rather than VPs). Statistical information from mixed corpora will reflect weaker generalizations, about the general grammatical preferences of the language (e.g. topicalized sentences are very rare, "of" PPs generally attach to NPs rather than VPs, etc.).

A third source of constraining information, and the one on which the present line of research will focus, comes from punctuation which can contribute much to the disambiguation process. Information from punctuation has only recently been taken into consideration in parsing and grammar development (see [Briscoe1994; Jones1996b]). Adding punctuation to the grammar will reduce the ambiguity of the current analyses by marking the boundaries of clauses and phrases. These may be either standard constituents, like clauses, or non-standard constituents, like a verb plus a determiner. Many "funny syntax" constructions, like topicalization or gapping, have punctuation as a critical element, and it can be used to identify these constructions when we encounter them. Furthermore, some discourse or text information can be used to further constrain ambiguity. Some particular resources include: how punctuation is used, the sequence of tenses in related clauses, the distribution of "cue" words (e.g. *because*, *in addition*), and the discourse status of nouns relative to the preceding discourse. As mentioned in Section 3.1, much of this information can be associated with the individual trees for which it is appropriate. Such information can be brought to bear even within single sentences, when they are composed of multiple clauses. Complex sentences are the natural place to start adding discourse information to the grammar, because parsing individual sentences is the primary function of the grammar. Looking at larger, multi-sentence pieces of texts, while clearly a more ambitious task, will then be able to contribute further information to the process. For instance, using a model of the preceding discourse we can supplement the statistical component with probabilistic "predictions" about the likelihood of certain marked syntactic constructions, like Topicalization or Inversion.

The next section discusses one group of constructions with regard to ambiguity reduction.

Complex sentences

Multi-clausal sentences occur very frequently in natural texts. The current XTAG grammar includes an extensive complementizer system which handles sentential complements (57) and sentential subjects (58). We have also have treatments of subordinating conjunction (59), adjunct clauses (60) and discourse conjunction (61).

(57) I have observed *that being upon a horse changes the whole character of a man. . .* (Brown:ck09)

(58) *That we are experiencing an upsurge of interest in the many formulations and preventive adaptations of brief treatment in social casework* is evident from even a small sampling of current literature. (Brown:j24)

(59) I put a lot more trust in my two legs than in the gun, *because the most important thing I had learned about war was that you could run away and survive to talk about it.* (Brown:ck09)

(60) *In order to accomplish the purposes of this Act*, the Secretary of the Interior shall – (A) conduct, encourage, and promote fundamental scientific research and basic studies. . . (Brown:ch09)

(61) *And* the automobiles that stream out of Hanover each weekend, toward Smith and Wellesley and Mount Holyoke, are no less rakish than those leaving Cambridge or West Philadelphia. (Brown:cf30)

To give a rough idea of the frequency of these phenomena, I looked at 100 sentences from 3 randomly selected passages of the Brown corpus. This sample contained 31 subordinate and adjunct clauses, most introduced by subordinating conjunctions and 26 having the adjunct or subordinate clause set off with punctuation (such as (59) above, with *because*). An analysis of punctuation is crucial in such sentences. In addition, analyzing these constructions is one step toward any discourse-level analysis of texts, as noted above. Adding analyses of such phenomena improves the coverage of the XTAG grammar by 6.6% [Doran1994], but also increases the ambiguity of the parses and provides further motivation for identifying other constraining factors in the text.

Adding punctuation to complex sentences

As can be seen in the examples, there is frequently a comma between a matrix clause and an adjunct clause; dashes, semi-colons and colons appear here less frequently (example (62)). Dashes, colons, semi-colons and (rarely) commas may all also serve to coordinate clauses when there is no lexical conjunction ("asyndetic" coordination). Example (63) has three clauses coordinated with semi-colons. Finally, commas and colons are also used between verbs and their sentential complements, as in example (64). I have found examples of all of these uses of punctuation in the Brown corpus.

(62) Expressed differently: if the price for becoming a faithful follower of Jesus Christ is some form of self-destruction, whether of the body or of the mind – sacrificium corporis, sacrificium intellectus – then there is no alternative but that the price remain unpaid [cd02]

(63) We know that we have hydrogen in water; water is Af [sic] and the H stands for hydrogen; there is also hydrogen in wood and hydrogen in our bodies. [cd13]

(64) In a long commentary which he has inserted in the published text of the first act of the play, he says at one point: "However, that experience never raised a doubt in his mind as to the reality of the underworld or the existence of Lucifer's many-faced lieutenants."[cd01]

With examples like (63) where the punctuation mark is the coordinator, a grammar with no treatment of punctuation cannot parse the sentence at all so this case also falls into the "Increasing Coverage" category of punctuation. However, the other cases would receive parses without punctuation, but far fewer with it. To give a concrete example, even a simplified version of sentence (62) gets 26 parses; with the colon, it gets 18 parses.

3.6.3 Chunking text using punctuation

A third area where punctuation can be put to good use is in chunking long input sentences into more manageable sized pieces before passing them to a parser, or to some other type of processing. Two critical assumptions are that (a) you can successfully identify text adjuncts and extract them without disrupting the syntactic coherence of the surrounding texts and (b) your domain and task are such that chunking is useful. One such task is discussed in work by Chandrasekar and colleagues [Chandrasekar et al.1996; Chandrasekar and Srinivas1997], which makes use of the information provided by punctuation to simplify texts.

(65) In order to accomplish the purposes of this Act, the Secretary of the Interior shall – (A) conduct, encourage, and promote fundamental scientific research

and basic studies to develop the best and most economical processes and methods for converting saline water into water suitable for beneficial consumptive purposes; (B) conduct engineering research and technical development work to determine, by laboratory and pilot plant testing, the results of the research and studies aforesaid in order to develop processes and plant designs to. . . [Brown:ch09]

Text chunking would be very useful in handling a sentence like (65) which is a marvel of punctuation. The "sentence" (admittedly, it is government legalese) goes on for a large paragraph, with five enumerated points separated by semicolons, one of which contains seven commas. Parsing the components of each clause separately and then reassembling them could be much more efficient than trying to parse the whole thing as a unit. Dashes, colons and semi-colons are obvious places to break texts. One would needly slightly more sophisticated heuristics to distinguish commas separating clauses, either from each other or within sentences.[7]

3.7 How would this analysis combine with a discourse grammar?

Work by Gardent, Joshi and Webber [Gardent1997; Gardent and Webber1998; Webber and Joshi1998] uses TAG trees for discourse grammars. They represent the propositional content of a single clause as a leaf node, and encode rhetorical relations between the clauses, along with their combined semantic representations, at the root nodes. Thus, a complex sentence like (66) (Gardent's (7)) would yield the discourse structure in Figure 3.16.

(66) a. Dick did not come to work

b. because the trains aren't running.

c. Buses aren't either.

Webber and Joshi lexicalize their discourse trees on subordinating conjunctions and cue words. The anchors have rich feature structures associated with them, and may be empty when there is no explicit cue word, but nonetheless encode the discourse relation in the features. Example (66) would be represented as in Figure 3.17 in their framework, with an empty anchor (just a set of features) realizing the relation between sentences (b) and (c). The structure introduced by the *because* relation is shown in bold lines, and that introduced by the period in dashed lines.

[7] Because commas are used for so many things, they will always be trickier to manage – it would be very interesting to see if one could train part-of-speech tagger or rule-based system to distinguish commas as conjunctions vs. separating punctuation marks.

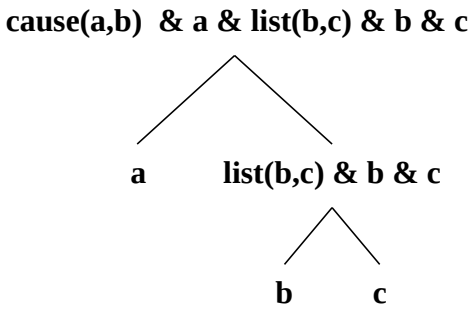


Figure 3.16: Discourse tree for Gardent's example (7), repeated above as (66)

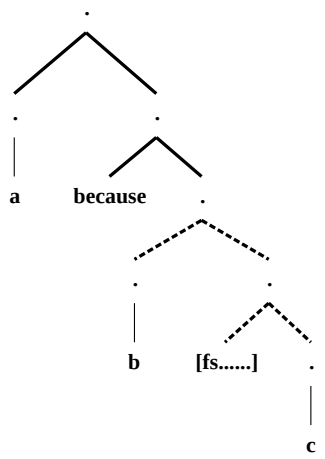


Figure 3.17: Discourse tree for Gardent's example (7), a la Webber and Joshi [1998]

Both sets of work assume that a text has already been split into the appropriate minimal clauses. One obvious way of doing that is with a sentence grammar like the LTAG discussed here for English. The LTAG derivation tree records the history of the derivation and is likely to be the primary object of interest in deriving a semantic description of the sentence. The following example illustrates one such case.

(67) a. Mary was awake, because she had napped earlier.

b. Because she had napped earlier, Mary was awake.

c. Mary, because she had napped earlier, was awake.

d. Mary was awake–she had napped earlier.

Figure 3.18 shows derivation trees for sentences (67)a-d. The details of the derivations are not crucial. The important thing to notice is that each derivation has a

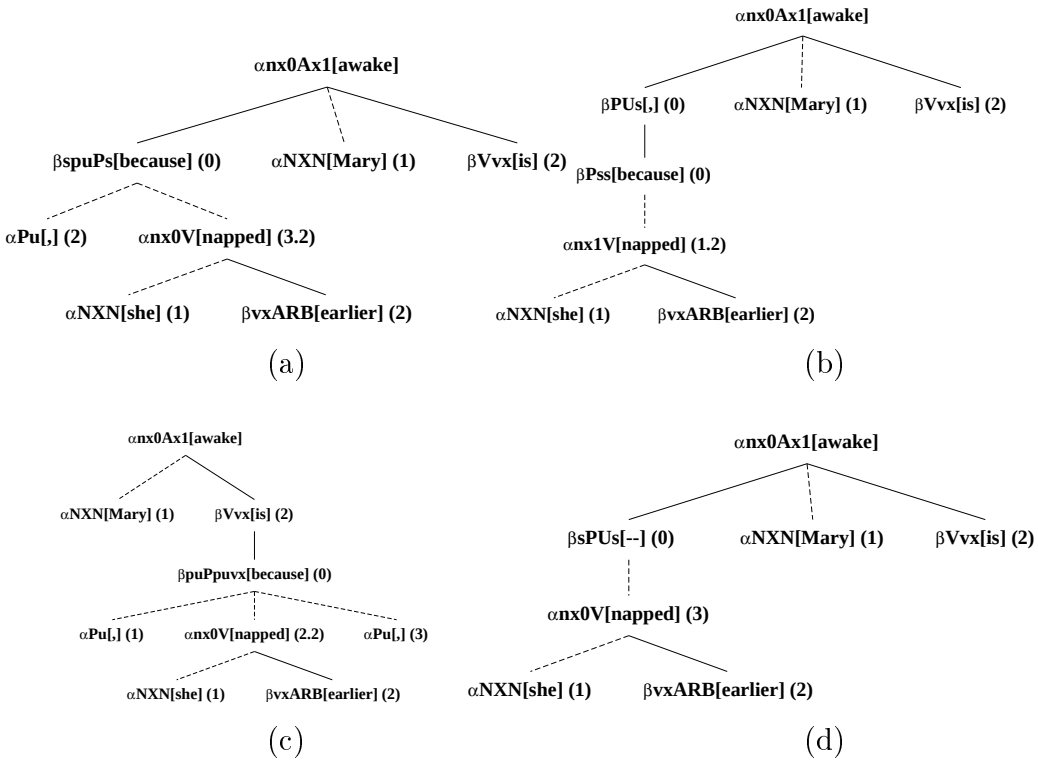


Figure 3.18: Derivation trees for the sentences in Example (67)

clearly defined and easily identified sub-tree rooted at *because* for the causal clause, regardless of its syntactic position. Likewise, the clause set off by dashes in (67)d has a distinct sub-derivation. This is the information one would need to create a discourse relation, in this case as simple causal relation such as:

where b represents the propositional content of the *because* clause and a of the *awake* clause.

The treatment of punctuation in LTAG described here provides a syntactic analysis which makes available to the discourse grammar the appropriate constituents, via the derivation structures produced. The approach seems to be completely compatible with the work being done on TAGs for discourse by Joshi, Webber and Gardent.

Token	Supertag	Description of supertag
Rockwell	B_Nn	Noun that modifies a noun on its right (1)
International	B_Nn	(1)
Corp.	A_NXN	Simple argument NP
won	A_nx0Vnx1	Transitive verb, one NP arg. to left, one to right
a	B_Dnx	Determiner looking for NP on right to adjoin to
contract	A_NXN	Simple argument NP
for	B_nxPnx	PP modifying NP on left, with NP arg. on right
gunship	B_Nn	(1)
replacement	B_Nn	(1)
aircraft	A_NXN	Simple argument NP
.	B_sPU	Punctuation adjoining to S on left

Table 3.2: A sample sentence with a unique supertag assignment to each token.

3.8 Evaluating the syntactic account

Ideally, we would evaluate the punctuation rules using the full XTAG parser—take a corpus of sufficiently complex sentences, parse it both with and without the punctuation marks, and measure the improvements in coverage and accuracy when the punctuation is taken into consideration. However, such an experiment is impossible for practical reasons because our parser would definitely run out of memory on sentences of any interesting length with their punctuation stripped, and might also on highly ambiguous sentences containing punctuation.[8]

Another way to measure the improvement in the grammar is to use the supertagging technique developed by Srinivas [1997] as an alternative to full parsing. Supertagging takes the trees of the LTAG English grammar, and uses them as complex part-of-speech tags. Each supertag encodes not only the part-of-speech of the lexical item, but also the number, type and direction of its arguments or the element it modifies. Using a trigram model like those used in standard part-of-speech tagging, supertags can be assigned to a new text based on probabilities derived from pre-tagged training data. Once a sequence of supertags is assigned to a sentence, the structure of that sentence can be quite easily determined (but for attachment ambiguities). Table 3.2 shows the sample sentence *Rockwell International Corp. won a contract for gunship replacement aircraft.* with supertags assigned to each word; Figure 3.19 shows the derivation that would result from combining the supertags.

To evaluate the LTAG punctuation analysis, I used a supertagger trained on just over 1 million words of Wall Street Journal data whose supertags were derived by

[8]Jones [1996a] encounters the same problem in attempting to evaluate his grammar. His chart-based parser cannot enumerate the number of parses possible for many of the unpunctuated sentences in his test set, and he has to turn to a special estimation process which interrupts the parser before it actually builds any parses.

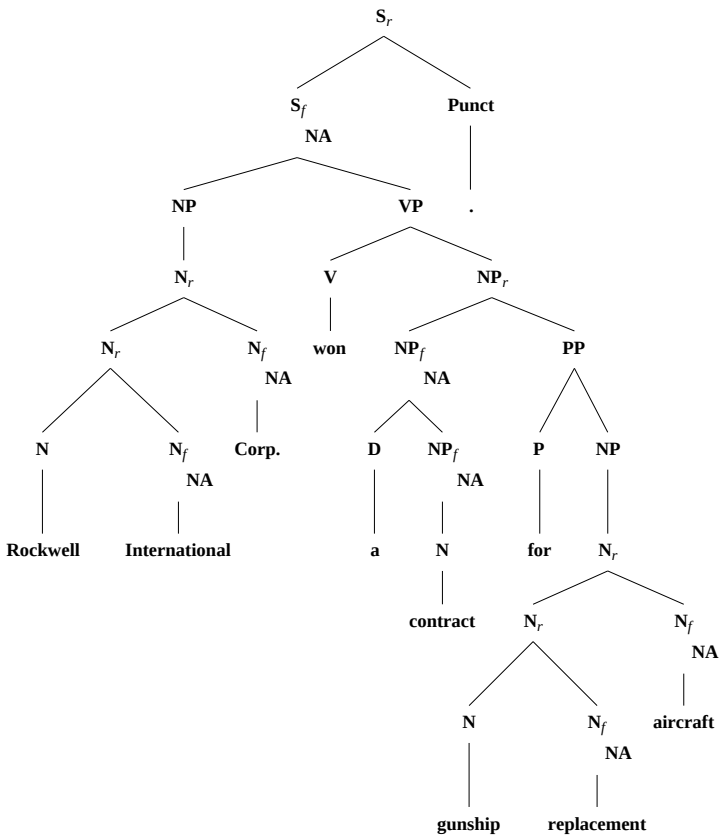


Figure 3.19: The derivation resulting from combining the trees assigned in Table 3.2 (other structures for the noun-noun compounds can be derived with the same supertags).

conversion from the (hand-corrected) Treebank parses. I first trained the tagger on the data with all punctuation stripped, and tested it on 2012 held-out sentences, also with punctuation stripped. I then retrained the tagger on the full million words, and tested it on the same test data with punctuation retained. The performance is shown in Table 2.1. The most important line is the middle one, showing performance of both sets of training data on exclusively non-punctuation tokens. The improvement in performance of 1.4%, while small, indicates that the presence of punctuation does indeed improve the accuracy of analysis of the surrounding texts. My result reflects an increase in the number of non-punctuation tokens to which the correct structural tag was assigned only when punctuation was present. This figure is not directly comparable to the coverage improvement obtained by Briscoe and Carroll [1995] of 8% (cf. Section 2.2.1), which reflects an increase in the number of sentences for which some parse (not necessarily correct) was obtained. Nor can it be compared with their improved crossing brackets performance on SUSANNE sentences, which looks at the number of correct constituents. Supertagging accuracy is measured on

a per word basis, and always assigns a tag to every word, so there is no notion of complete failure on a sentence. In that sense, supertagging does assign a structure to every sentence, but without assembling the supertag sequence assigned, you do not know what the hypothesized constituents are. The most appropriate comparison is with the evaluation presented in [Briscoe1994], where he finds a 2% improvement in "rule application" on SUSANNE sentences (i.e. the correct derivational step applied at a given point) since we can think of each LTAG tree as a rule (or possible several rules) to be applied.

One important thing to remember is that the supertagger has only a three-token window in assigning tags, and constructions involving punctuation often span a fairly large number of tokens (e.g. the comma around a relative clause, parentheses around sentences). This suggests that performance might be much more dramatically improved if we were able to use the full parser. The baseline performance for supertagging punctuation marks (i.e. assigning simply the most likely tag to each mark) is 65.9%. This is considerably lower than regular part-of-speech tagging at around 90% and supertagging overall at 77.2% for this corpus. The baseline for punctuation is lower because the average number of supertags is higher: 6.5 supertags per punctuation mark compared with 1.5 parts-of-speech per word in standard part-of-speech tagging.

	Trained and tested on text without punctuation	with punctuation
% Correct Overall	87.1%	88.0%
On non-punct tokens	**87.1%**	**88.5%**
On punct tokens	—	83.7%

Table 3.3: Accuracy of supertagging with and without punctuation

This difference in performance can be seen on the following examples. One very common error in the un-punctuated text is for NPs preceding appositives to be tagged as noun modifiers. This is shown in example (68), with the tag assigned by the no-punctuation trained supertagger on top and the punctuation trained one on the bottom. This is exactly what we would predict—if there is no punctuation to demarcate the NP boundaries, the only possible analysis of appositives and the NP they modify is as one giant NP. Other types of commas occuring between two NPs cause similar mistakes, as in examples (69) and (70), both of which have commas separating a modifier ending with an NP from the matrix clause. Example (71) is rather different. When the comma preceding the lexical conjunction *and* is removed, the supertagger incorrectly assigns a relative clause tag to the verb *gave*. With the comma present, the verb correctly gets a main verb tag.

(68) shares of $\begin{matrix}\text{UAL\_Nn}\\ \text{UAL\_NXN}\end{matrix}$, United 's parent company , dived .

(69) Under the existing $\begin{matrix}\text{contract\_Nn} \\ \text{contract\_NXN}\end{matrix}$, Rockwell....

(70) On a day some United Airlines employees wanted Mr. Wolf fired and takeover stock speculators wanted his $\begin{matrix}\text{scalp\_Nn} \\ \text{scalp\_NXN}\end{matrix}$, Messrs. Wolf and Pope....

(71) He left his last two jobs at Republic Airlines and Flying Tiger with combined stock-option gains of about $22 million $\begin{matrix}\text{and UAL gave\_N0nx0Vnx1nx2} \\ \text{, and UAL gave\_nx0Vnx1nx2}\end{matrix}$ him a $15 million bonus when he was hired .

Obviously performance with this method will not be quite as good as if we were able to hand-correct the supertags on the training data, but the hope is that the volume of training data will compensate for some of the errors generated in translating from Treebank to supertags. Unfortunately, there are a few systematic difficulties in translation which have not yet been addressed. One known problem with this method is that it is very difficult to accurately distinguish commas used to conjoin sequences of NPs from those used in appositives in building the training files. The Treebank annotation of the two is identical, and the presence the lexical conjunction at the very end of a conjoined sequence is the only distinguishing feature. Improvements in the translation would likely improve the performance of the supertagger on both punctuation and non-punctuation tokens.

Chapter 4

NP Appositives

The contrast which drove me to scrutinize appositives in some detail is the seemingly minimal difference between the constructions shown in examples (1) and (2). The two expressions appear to be synonymous, and are typically interchangeable in a given context. I will call the first type a *classic appositive* (the LTAG treatment of which is described in Section 3.5.1) and the second, a *pseudo-title* (following Meyer–McCawley calls it the "journalese construction"). The latter construction is sometimes given as an example of a "restrictive appositive." Superficially, all that differs is the order of the components, and the presence or absence of the comma.[1] In (2) there is no comma between the post description and the name of the person, while in classic appositives like (1) the comma is obligatory.

(1) George Scandalios, Chase Senior Vice President

(2) Chase Senior Vice President George Scandalios [wsj1630]

4.1 Possible analyses

A number of attributes must be considered in looking at appositive-like constructions. One standard claim in the literature is that appositives are reduced clauses, so the first thing we need to assess is whether both elements are really nominal or the predicative element has a reduced clausal structure. In principle, they could also be reduced prepositions of the sort Larson [1985] describes, although I have never seen such a claim made.

Second, there is the question of whether the elements are in a restrictive or non-restrictive relationship. Researchers have, if glancingly, given pseudo-titles as

[1] Appositives can also be separated by dashes, but this distinction is not relevant for the discussion here. All conclusions about commas apply equally to dashes.

(i) But Tony Robinson, the current sheriff of Nottingham – a job that really exists – rejected the theory, saying that "as far as we are concerned, Robin Hood was a Nottinghamshire lad." [Clari UK news]

examples of restrictive appositives. In relative clauses, the presence of the comma is correlated with a non-restrictive interpretation. If appositives prove to be reduced relative clauses, the commas may turn out to be similarly significant.

Finally, we need to ascertain what the relationship between the nominal elements is: Spec-Head, Head-Argument or Head-Adjunct. Under standard $\bar{X}$ assumptions for nominals these will be realized as sisters of the NP, sisters of N' or sisters of N, respectively. In the XTAG English grammar, we only use N and NP labels, so I will talk about modifiers attaching at only those two levels. For the present purposes, I am interested in making a distinction between things which act more like specifiers, arguments or adjuncts rather than in ascertaining specific attachment points.

Upon closer examination of the data, we find still more constructions beyond the two shown above which might be appropriately classified as kinds of NP appositives. Section 4.2 gives an overview of this collection of constructions. Sections 4.3-4.5 address three main issues to be considered in looking at the class of appositive-like constructions: what their internal structure is, whether the components of the construction are in a restrictive or non-restrictive relationship and what the syntactic relationship between the elements is. Lastly, Section 4.7 presents some proposed semantic representations, and looks at how they fit the range of data discussed here.

4.2 Defining apposition

Apposition is an extremely complex class of phenomena, potentially encompassing a wide range of constructions. Hollenbach [1983] suggests that the appositive construction and parentheticals might be "two ends of a single worm," with a whole range of constructions falling in between (in the body?), and Meyer [1987] includes everything in the list below and more in his book on apposition.

The classic appositive construction has an NP modifying another NP, as in *James B. Lee, head of syndications*; *her passion in life, acting*; or *vexilloids–objects that function as flags.* As with relative clauses, there are what have been argued to be restrictive appositives: *Actor Lionel Barrymore*, *the number six.* Looking at syntax alone, appositives with punctuation removed will be highly ambiguous. In (3) is the complement a small clause, or a NP with an appositive? We must clearly take punctuation into account with appositives, but is the punctuation a defining characteristic? The Chicago Manual of Style [1982, Section 5.44] claims it is—they say that "[i]f the appositive has a restrictive function, it is not set off by commas." Likewise, the SUSANNE annotation scheme [Sampson1995] takes the presence of punctuation to be a characteristic, though not mandatory, feature of appositives.

(3) The late Secretary of State John Foster Dulles considered *the 1954 Geneva agreement a specimen of appeasement...*

Some of the range of candidates for the appositive family, not all of which require punctuation between the two elements, include:

The "reduced namely" construction [McCawley1982]: *the president, (namely) Bill Clinton*
Non-restrictive. Unlike classic appositives, these are not paraphrasable with relative clauses.

Reduced partitives [Lasersohn1986]: *The two professors, each (of them) an artichoke-expert, debated the issue for hours.*
Non-restrictive. These may actually be sequences of three NPs, since the only determiners that can appear on the second piece are those which also occur as pronouns.

Titles and Pseudo-titles: *Mr. Smith; President of Wellesley College Diana Chapman Walsh*
Where titles are simply honorifics, they are not referential in any sense. Some "titles" do refer independently, and thus can appear alone. Are these modifiers or are they closer to classic appositives, with the second part behaving more like a modifier?[2] It is difficult to know whether to classify these as restrictive or non-restrictive.

Pseudo-appositives [Lasersohn1986]: *my cousin Janet; Lawrence the novelist*
Restrictive. Unlike pseudo-titles, these always have determiners on the common noun part.

Some colon expansions: *Three people left: Maude, Claude and Rimbaud.*
Non-restrictive.

The N-E construction (Jackendoff): *the word artichoke*
Restrictive. The underlined portion here can be of any category–phrase, word, morpheme, sound, even a gesture.

Addresses: *An official at Consolidated Freightways Inc., a Menlo Park, Calif., less-than-truckload carrier , said...*
Restrictive. Address or location modifiers like *Calif.* must be attached at N, rather than NP, because they can occur in the middle of compound nouns. *Carrier* is not an appositive on either *Menlo Park* or *Calif.*, as it would be if these were simply stacked appositives. Rather, *Calif.* modifies *Menlo Park*, and that entire complex is compounded with *carrier*.

In this work, I will be concentrating primarily on classic appositives and pseudo-titles, but will also attempt to make some generalizations about the other constructions listed here. Let us look at classic appositives and pseudo-titles in a bit more

[2] According to [Sabin1996, Section 312], "Occupational titles can be distinguished from official titles in that only official titles can be used with a last name alone. Since one would not address a person as 'Author Mailer' or 'Publisher Johnson,' these are not official titles...."

detail, to see whether either or both behave as if they contain a separate predicate, or whether we are happy to treat them as single, albeit complex, NPs.

4.3 Reduced clauses or noun phrases?

Historically, a number of linguists (e.g. [Smith1969; McCawley1995]) have argued that appositive NPs are reduced relative clauses—the famous phenomena of "whiz drop" (dropping the wh- element and the copula). One piece of data which supports the claim that the basic appositives are reduced relative clauses is that they take both adverbial and adjectival modifiers, as shown in (4) and (5). This is compatible with a reduced clausal structure that has an NP predicate. The two alternative structures are shown in Figure 4.1. You can see in (a) that there are attachment sites for both adjectives, at NP_1, and adverbs, at VP, whereas (b) has only nominal attachment sites. Sampson [1995], in tagging the SUSANNE corpus, takes the presence of an adverbial as a clear indicator that the appositive is a reduced clause and gives it a clausal rather than nominal tag. Pseudo-titles, however, only allow adjectival modifiers, suggesting that they are underlyingly nominal (examples (6) and (7)).

(4) Alberto M. Paracchini, **currently** chairman of BanPonce, will serve.... [wsj]

(5) Alberto M. Paracchini, **current** chairman of BanPonce, will serve...

(6) **Former** Democratic fund-raiser Thomas M. Gaubert, whose saving and loan.... [wsj]

(7) ***Formerly** Democratic fund-raiser Thomas M. Gaubert, whose saving and loan....[3]

There are instances of constructions which look like pseudo-titles with commas and adverbial modifiers, like that in (8). Note, however, that there is no comma following *Mr. Lang*. This is also unlike the pseudo-title construction in that a determiner is possible on the first NP (9), and only an adverbial modifier is possible either with or without the determiner. This indicates that such modifiers must be true clausal adjuncts, and interpreted more like subordinate clauses, e.g. *While he was formerly the president and treasurer....*

(8) **Formerly/*former** President and Treasurer, Mr. Lang remains Chief Executive Officer. [wsj]

(9) Formerly/*former **both** President and Treasurer, Mr. Lang remains Chief Executive Officer.

[3] On the relevant reading, where he has not changed his party allegiance.

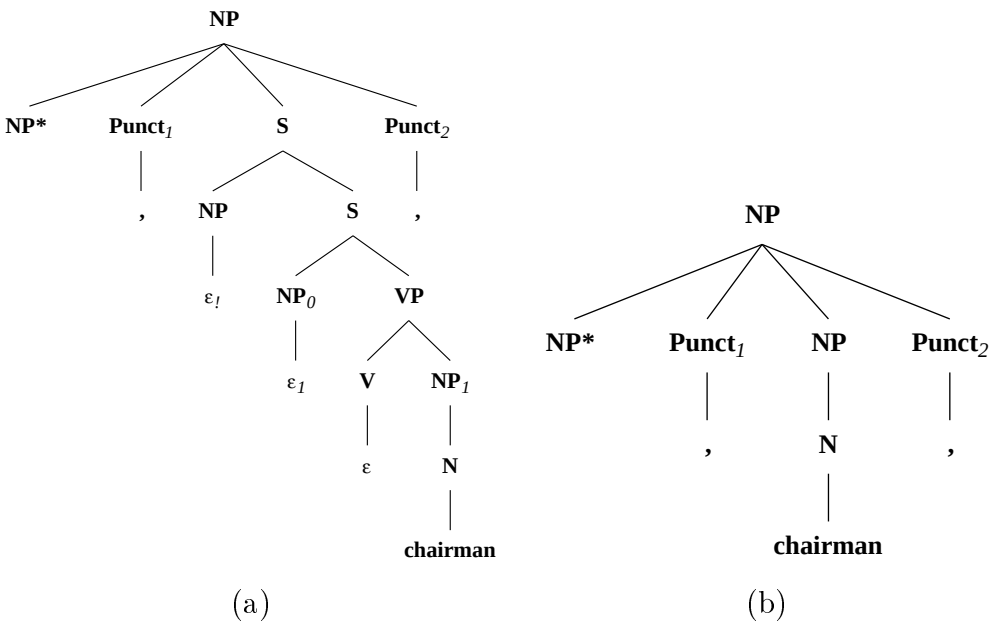


Figure 4.1: Clausal and Nominal Structures for Appositives

Evidence from case-assignment might also indicate what the internal structure of the appositive clause is. If it is always accusative, even when the appositive is on a subject, it would suggest that there is reduced clausal structure and the appositive is the object/predicate. If the case is always the same as the NP it modifies, this would suggest that case is being shared by the two NPs, as in a coordinate structure. Unfortunately, it is not easy to evaluate this situation in English, as pronouns make quite bad appositives (or predicates of any type). However, one possible piece of evidence is from examples like (10), which are not possible in writing, but seem possible in speech.

(10) President Rodin, her/*she [pointing], runs from one meeting to another.

The deictic pronoun must have accusative case here (which is a bit funny, given that one might alternatively analyze this as a correction or restatement rather than an appositive). Cross-linguistic inquiry is needed to pursue this point any further. There is also the *we graduate students/us graduate students* construction, which may or may not be a type of appositive (depending whether you think the pronoun has a determiner/specifier role here). Both the nominative and accusative pronouns are grammatical here; strangely, while Delorme and Dougherty [1972] discuss these forms quite a lot, they do not mention anything about possible mechanisms for case-assignment.

Thus far, the evidence argues for treating classic appositives as reduced clauses and pseudo-titles as entirely nominal.

4.3.1 Other shared properties of appositives and full relative clauses

Classic appositives appear to share other properties with non-restrictive (aka "appositive") relative clauses which support the claim that classic appositives are reduced clauses. Appositives are generally paraphrasable by non-restrictive relative clauses, as in the pair (11) and (12). Often these are copular relative clauses, and for most of this type of relative clause, the reverse holds and the relative clauses can be paraphrased with appositives, as in (13) and (14). This is not generally true of non-copular non-restrictive relatives, as shown in (15) and (16).

(11) ...a concerto he has recently recorded, " The Emperor ".... [Brown]

(12) ...a concerto he has recently recorded, (which is) called "The Emperor"

(13) ...during May, which is National Salvation Army Week.... [Brown,sic]

(14) ...during May, National Salvation Army Week...

(15) Rudy Vallee, who shares star billing with Mr. Morse... [Brown].

(16) *Rudy Vallee, star billing with Mr. Morse....

Pseudo-titles cannot be paraphrased with relative clauses. Even with a comma inserted, a pseudo-title yields an instance of the "reduced namely construction" (17).

(17) a. Chase Senior Vice President, (namely) George Scandalios,..

b. *Chase Senior Vice President who is George Scandalios,..

One question is how far the parallel between non-restrictive relative clauses and classic appositives extends, and whether the appositives consistently pattern with non-restrictive rather than restrictive relatives. Some of the differences between the two types of relative clauses are listed below, along with discussion of whether classic appositives and/or pseudo-titles appear to behave like one or the other type of relative clause. A number of these points are raised by Emonds [1979], in an excellent article surveying the differences between restrictive and non-restrictive relative clauses.

- There may be more than one restrictive relative per head:

  (18) The very places which he discusses in his book where language is at its most conventional.

  Appositives, like non-restrictive relatives, cannot typically stack on a single NP:

(19) *Max, who is very upset, who I saw at the party, recently lost his job.

(20) *Leroy Barnes, an avid golfer, current VP for marketing, recently lost his job.

Meyer gives one example of stacked appositives (reproduced as (21)), and I found one other in the Brown corpus ((22) below). Neither of these sounds horribly ungrammatical, but to me they have the feel of asyndetic coordination, which the stacked restrictive example in (18) does not.

(21) After all, a figure had to be arrived at, a definite sum in pounds, shillings and pence, a last chord in which all the conflicts and problems of the return would be resolved. [Meyer's (120)]

(22) The monthly cost of ADC to more than 100,000 recipients in the county is 4.4 million dollars, said C. Virgil Martin, president of Carson Pirie Scott & Co, committee chairman. [Brown]

Pseudo-titles cannot stack:

(23) *Former President current CEO Leroy Barnes

(24) Former President and current CEO Leroy Barnes

- Non-restrictive relatives can modify categories other than NPs:

(25) Arthur acted like a complete cad at Sue's party, which I found appalling/where I had never expected to see him.

(26) John drives too quickly, which is a very bad habit.

Appositives are unable to modify categories other than NP. You can have what look to be PP appositives, but only on other PPs (28).

(27) ?*John drives too quickly, a very bad habit.

(28) Mary put the book right here, on the corner of the table.

All of the elements of pseudo-titles are obligatorily nominal (it is difficult imagine what it would mean for either a proper name or a title to be anything else.)

- Restrictive relatives, and not non-restrictives, may be postposed:

(29) A cry was first used 500 years ago which white crane boxers still imitate today.

(30) *Max was very upset, who I saw at the party last night.

Appositives are quite hard to extrapose.[4] One example that Meyer gives as extraposition is shown in (33), but, as with his other example discussed above, I think it looks more like a conjoined predicate (e.g. *difficult to work with and an unsurly....*).

(31) The people were all really nice, who I met at Bill's party the other day.

(32) *The music here sounded good, Russell Smith's stunning new composition "Tetrameron."

(33) The man is difficult to work with, an unsurly [sic] individual who scowls at just about everyone he encounters. [Meyer, 1.20a]

Pseudo-titles act more like single NPs, and cannot be separated at all:

(34) *Former Democratic fund-raiser Thomas M. Gaubert*, whose savings and loan was wrested from his control by federal thrift regulators, has been granted court permission to sue the regulators. [wsj]

(35) **Former Democratic fund-raiser*, whose savings and loan was wrested from his control by federal thrift regulators, *Thomas M. Gaubert* has been granted court permission to sue the regulators.

(36) **Thomas M. Gaubert*, whose savings and loan was wrested from his control by federal thrift regulators, *Former Democratic fund-raiser* has been granted court permission to sue the regulators. [wsj]

- Restrictive relatives may modify a quantificational element (37) or contain a pronoun with a bound variable (39); appositive relatives cannot be used to modify a a quantificational head (38) and never have a bound variable reading with a quantificational element (40). (The judgement in (38) is a bit subtle, because one tends to force a restrictive reading onto it. With a non-restrictive interpretation it can only mean something like 'A person, named Every Person, who reads the daily newspaper....')

(37) Every person who reads the daily newspaper is well informed.

(38) *Every person, who reads the daily newspaper, is well informed.

[4] This is not extraposition per se, but Jesperson has a great example from Shakespeare of an appositive on a genitive noun: "This same skull, sir, was Yoricks skull, the kings jester...." [Jesperson1966, Section 9.6]

(39) Every person reads the daily newspaper that he finds on the front step.

(40) a. *Every person reads a newspaper, that he finds on the front step.

b. John reads a daily newspaper, that he finds on the front step.

Appositives can modify some quantificational elements:

(41) *Every person/both men/all men, prolific authors

(42) Five/some men, prolific authors

Lasersohn [1986] discusses which quantifiers are possible, concluding that ones which only allow distributive (as opposed to collective) readings are grammatical.

Pseudo-titles cannot have determiners of any sort.

Additionally, parasitic gaps ((43) and (44)) and weak-crossover effects ((45) and (46)) occur in restrictive but not appositive relatives. However, this distinction is not relevant to appositives because they never contain overt verbs and cannot have object traces; as reduced copular clauses, they will always have subject traces.

(43) Clinton is a person who$_i$ everyone who knows e$_i$ admires t$_i$ greatly

(44) *Clinton is a person who$_i$ Smith, who knows e$_i$, admires t$_i$ greatly (cf. who knows him$_i$)

(45) *The cat$_i$ who her$_i$ family loves t$_i$ is very happy.

(46) OC$_i$, who her$_i$ family loves t$_i$, is very happy.

Overall then, classic appositives continue to behave more like relative clauses, i.e. a head and a modifier, and pseudo-titles act like single NPs. In particular, appositives pattern with non-restrictive relative clauses, but we need to look a bit more carefully before deciding whether they themselves are non-restrictive.

4.4 Restrictive vs. non-restrictive

Restrictive and non-restrictive relatives have a number well-known syntax, semantics and pragmatics differences. Non-restrictive relatives are marked by commas on either side in writing and an intonation break in speech (anecdotally, at any rate), while restrictives cannot be thus marked. Like many other things set off by punctuation, non-restrictive relatives act like they are syntactically independent of the rest

of the clause, or at least not completely integrated into the syntax of the matrix sentence. In general, there are no dependencies between elements outside and elements inside of an a non-restrictive relative, i.e. the relative clause is opaque. Roughly speaking, non-restrictive relatives convey independent propositions which give you more information about the NPs they modify, and they can be paraphrased with a separate matrix clause. Restrictive relatives typically do just that—they restrict the reference of their heads to a smaller set of discourse entities. In file-change semantics terms, restrictive relatives help the hearer decide which card to choose (i.e. the reference of the head is affected by the content of the relative clause), while non-restrictive relatives add information to an existing card (the reference of the head is determined independently of the relative clause).

Like relative clauses, appositives also have been argued to have restrictive and non-restrictive realizations, as illustrated in examples (47) and (48). As noted above, Lasersohn calls constructions like (47) "pseudo-appositives." He says they are restrictive, and that both parts must be definite because of a pragmatic restriction—the NP picks out a singleton set and the indefinite is under-informative (Griceanly speaking).

(47) my brother Bill → I might have more than one brother

(48) my brother, Bill → Bill is my only brother

It is clear that the classic appositive in (48) is not restricting the reference of *my brother*, while the superficially similar example (47) is. In his corpus of 2800 examples, Meyer claims to have found about the same proportion of restrictive to non-restrictive appositives as has been found for relative clauses–about 60% restrictive and 40% non-restrictive.

Furthermore, like non-restrictive relatives, many of the constructions one might want to classify as appositives have a second part which provides more information about the head without necessarily restricting its reference. The listing in Section 4.2 categorizes each appositive-like construction is as restrictive or non-restrictive. Based on the data in this list, there is a correlation between commas and non-restrictiveness. The only exception is the N-modifying addresses, which are clearly restrictive (*Oxford (= England)* vs. *Oxford, Ohio*), and the pseudo-titles, which are hard to classify on this dimension. So let us summarize are results so far as showing that, of the complex nominal constructions we are considering, those not separated by commas along with the address construction are restrictive.

4.5 Syntactic relationships

We have shown that classic appositives are reduced clausal predicates, which are in a non-restrictive relationship with the noun phrase they modify. There is plenty

of debate in the literature about the attachment site of relative clauses, but let us follow the XTAG grammar's convention here and say that they are NP adjuncts, in Head-Adjunct relation.

Up to this point, I have assumed that the structures of pseudo-titles and classic appositives was parallel, i.e. the second part was modifying the first, but this is not necessarily the case. Perhaps it is the other way around. This is, in fact, what Meyer thinks. He finds that overall, pseudo-titles are lighter (have fewer modifiers) than classic appositives, which he attributes to the preference in English for light pre-modification.

(49) a. Governor of New Jersey Christie Todd Whitman

b. Governor Christie Todd Whitman

c. Governor Whitman

d. *Governor of New Jersey Whitman (cf. New Jersey's Governor Whitman)

Where would one draw the line between titles and pseudo-titles? From the examples in (49), (d) might suggest that only when the bare title appears with a last name do we have a genuine title, as it cannot take any further modification. However, this would force us to claim that (a) and (b) are significantly different. It is certainly more attractive to treat them all alike. If we decide to group titles and pseudo-titles together, we have little choice but to say that title is a pre-modifier. It would be extremely odd to say that *Whitman* is modifying *Governor* in the expression *Governor Whitman*. This would also account for the lack of the comma, since pre-nominal modifiers are not separated from their heads unless they occur in a particular kind of list (*a long, hot summer* vs. *a typical hot summer*). If the classic appositive is a separate piece of discourse or, in Nunberg's terms, a text-adjunct like a non-restrictive relative, we would expect it to need to be separated from the head by punctuation.

Also, as can be seen in examples (50b) and (51b), inserting a determiner before the post in the pseudo-title is impossible, but it is acceptable in the classic appositive. (Since the other part is always a proper name, we would not expect to find a determiner before it in either construction.)

(50) a. George Scandalios, Chase Senior Vice President

b. George Scandalios, a Chase Senior Vice President

(51) a. Chase Senior Vice President George Scandalios [wsj1630]

b. *a/*the Senior Vice President at Chase George Scandalios

It the title or pseudo-title is acting a specifier, we would not expect to get a determiner as well. All evidence supports the finding that the relationship between the title or pseudo-title and the proper name is Spec-Head.

4.6 Summary of findings

The previous sections have discussed a range of NP appositive or appositive-like constructions, and attempted to answer the question of whether there is any commonality amongst those constructions which have punctuation as an integral component. In particular, we have looked at pseudo-titles and classic appositives. The discussion here has found that in pseudo-titles, the name is the head and the title is the specifier. In classic appositives, the name is the head and the appositive is an adjunct, predicated of the head. Where there is a comma, we have a non-restrictive relationship between the two constituents. Where there is no comma and we have modification at the NP level, we have a restrictive relationship (recall the restrictive, N-attached addresses). The pseudo-title thus acts like a single NP, despite the fact that the title itself may be a complex NP. Classic appositives are clearly composed of two distinct constituents, which may even be separated.

4.7 The semantics of appositives

There are two basic takes on the semantics of appositives, roughly correlated with whether one takes the underlying structure of the second constituent to be a reduced clause or an full NP. If the former, the relationship is a more predicative one (a property is predicated of the first NP), assuming the clause is a reduced copular relative clause. If the latter, the semantic relationship is argued to be more like equation of the two NPs.

As we have seen, apposition is far from being a unitary phenomenon, so we can hardly expect a uniform semantic analysis. The NP account fails to account for appositives with explicit "markers of apposition" [Meyer1992]. These are patently asymmetric, with the second part behaving as a modifier. Meyer identifies a number of such explicit markers, including *particularly, namely, primarily, i.e., or* and *like*. At first glance these looked like prepositional phrases, but many of the markers are adverbs, which we expect from the discussion in Section 4.3. He claims that the unmarked appositives have a default coreference-like relationship (the NP-NP angle), and that explicit markers are needed to specify other relations like PART-WHOLE and INSTANCE-OF. *Or* is particularly interesting in this usage, which is distinctly different from its disjunctive use. In example (52), *6.4% of the GNP* is alternative (and equivalent) to the figure of *$150 billion*, rather than a second amount altogether as you would have in a truly conjunctive NP like (53), where there are two possible deadlines which are not equivalent. The non-disjunctive use requires separating commas.

(52) Shippers cut...truck and rail costs, to about $150 billion, or about 6.4% of gross national product....[wsj]

(53) ...scheduled for this fall or early next year.[wsj]

A strict predication account fails on appositive/appositive-like constructions which do not appear to be reduced clauses, i.e. everything in the class but classic appositives. In particular, pseudo-titles are well-handled by a coreference account.

Chapter 5

Quoted Speech

This chapter looks at the quoted speech construction to see which of the punctuation marks typically involved is really critical to its structure. Based on information from punctuation and other distributional clues, I conclude that there are two types of reported speech: in one the quote is an argument of the verb of saying, and in the other the verb of saying and its subject adjoin into the quote. This second class patterns in many ways like other types of parenthetical modification. Examples of the two cases are shown in (1) and (2).

(1) When *she said that* she didn't have the money, *he said that* she could come in for treatment with his office model until she was ready to buy one. [cf10]

(2) "The primary objective of non-violence", *writes the outstanding Mennonite ethicist*, "is not peace, or obedience to the divine will, but rather certain desired social changes, for personal, or class, or national advantage". [cf48]

5.1 Motivation

In looking for constructions where punctuation plays an integral role, reported speech is an obvious candidate. Not only do we find commas, dashes or colons separating the quote from the speaker, we also expect to find quotation marks around the reported content. We would expect that the quotation marks might be useful in distinguishing direct and indirect speech[1], which appear to have radically different forms and functions, and that this distinction would facilitate text processing of such constructions, whether via full syntactic parsing or some more superficial analysis, such as regular expression matching. Unfortunately, the situation is not quite so

[1] While other familiar punctuation marks are used in much the same way throughout the Americas, Europe and Russia, one area where there is more than average variation is in marking reported speech. The quoted material can be marked by *guillemets*, dashes, or double or single apostrophes, either both raised or one set raised and one unraised.

tidy. In particular, the distinction between direct and indirect quoted speech is very blurry.

To start with, let us consider how to distinguish quoted material from material in quotation marks. The latter are a subset of the former–text in quotation marks is always quoted, but not all quoted material is enclosed in quotation marks. It might appear that the quotation marks themselves would be extremely useful in identifying these structures in texts. I will argue that quotation marks are not adequate for either identifying or constraining the syntax of quoted speech. More useful information comes from the presence of a quoting verb, which is either a verb of saying or a punctual verb, and the presence of other punctuation marks, usually commas. Using a lexicalized grammar, we can license most quoting clauses as text adjuncts. A distinction will be made not between direct and indirect quoted speech, but rather between adjunct and non-adjunct quoting clauses.

The framework within which the present work is couched is Lexicalized Tree Adjoining Grammar; the treatment of punctuation of which this construction is a part is discussed in more detail in Chapter 3. I will argue in this chapter that the direct/indirect split is not the correct one, and that the choice of the verb and the other punctuation marks involved are more informative than the quotation marks.

5.2 What do the quotation marks tell us?

The first problem is to identify a class of constructions identifiable as Quoted Speech. Punctuation-wise, we canonically expect a comma after the quoting verb and quotation marks around the speech for direct speech, and neither of these for indirect speech.[2] And indeed, there are clear cases of direct speech, like (3), and indirect speech (4).

(3) A Lorillard spokeswoman said, "This is an old story. We're talking about years ago before anyone heard of asbestos having any questionable properties. There is no asbestos in our products now." [wsj0003]

(4) However, Mr. Dillow said he believes that a reduction in raw material stock-building by industry could lead to a sharp drop in imports. [wsj1500]

However, there are also cases which blur the distinction, such as (5) and (6):

(5) Some bulk shipping rates have increased "3% to 4% in the past few months," said Salomon's Mr. Lloyd. [wsj1500]

[2] I am leaving aside a possible third category, "Free Indirect Speech," which is argued to be an intermediary type, reflecting the sequence of tense effects of indirect speech and the deictic use of direct speech. For the features I am considering, it appears to pattern with direct speech.

(6) And, they warn, any further drop in the government's popularity could swiftly make this promise sound hollow. [wsj1500]

(7) Republican Sen. William Cohen of Maine, the panel's vice chairman, said of the disclosure that "a text torn out of context is a pretext, and it is unfair for those in the White House who are leaking to present the evidence in a selective fashion." [wsj1500]

Example (5) is partly a direct quote (the object of the verb) and partly indirect. Example (6) has the usual subject-verb-complement order (SVO), but has the subject and the verb of saying separated from the speech by a comma. Example (7) has the syntax of an indirect quote (i.e. a complementizer and no comma), but uses quotation marks.

Furthermore, how are we to distinguish quoted material in the running text from quoted speech proper? Examples (8)-(11) show several such variants. Text in scare quotes, terminology and other quoted material included in running text are often only identifiable by their enclosure in quotation marks, and they are not distinguished syntactically from the surrounding material.

(8) ...noted that the term "teacher-employee" (as opposed to, e.g., "maintenance employee") was a not inapt description. [wsj1500]

(9) Unable to persuade the manager to change his decision, he went to a "company court" for a hearing. [wsj1500]

(10) Mr. Nagrin has described four "places", each with its scenery and people, added two "diversions"... [Brown:cc09]

(11) Types of loans SBA business loans are of two types: "participation" and "direct" [Brown:ch01]

Based on data such as this, we find that the quotation marks are not a useful indicator of any particular construction. While text in quotation marks is always a quotation of some sort, not all quotations are enclosed in quotation marks. Direct speech is simply a subset of the more general class of verbatim text (or at least text which is presented as if it were verbatim). Quotation marks typically have the same approximate interpretation: they mark what someone else, possible the author him/herself in different circumstances,[3] says/said/thinks/thought. As with scare quotes, the Other need not be identified explicitly. However, the quotation marks themselves are not an indicator of the larger syntactic context. Syntactically, we simply need a tree or a rule like those in Figure 5.1 to handle quotation marks.

[3]Nunberg [1990] describes quotes as "mark[ing] a text-expression that is to be construed as having been produced in circumstances that differ from those of the surrounding text...."

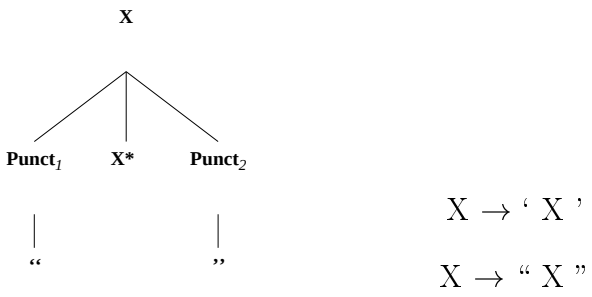


Figure 5.1: The schematic tree and phrase-structure rule for handling quotation marks, where X can be any node label. The tree is lexicalized on both the opening and closing quotation marks, so we are guaranteed to always get matching pairs of quotes.

But all is not lost — I will show that the comma (or, less commonly, the dash or colon) which is used in direct speech is actually the important cue, along with the particular verb used in the quoting clause. In the remainder of the paper, I will argue that the relevant distinction is between quoted speech in the normal SVO order and all other quoted speech, rather than between direct and indirect speech.

5.3 Characterizing reported speech

Having concluded in the previous section that quotation marks are not useful in characterizing the various types of reported speech, let us look in this section at some features which may be more useful, including the choice of verb, presence of absence of complementizer and punctuation marks, and the order in which the constituents appear.

Typically, indirect speech is shown as the complement to a verb of propositional attitude, like *say* or *believe*, as in (12). Direct speech may also use the same syntax, as shown in (13). Although it is typically restricted to occuring with verbs of saying (14), this appears to be a pragmatic rather than a syntactic/lexical constraint. In a context where it is possible to know what the speaker is thinking, in particular in text with an omniscient narrator, this construction is fine (15). There are also differences in the point of view (i.e. choice of first or third person pronouns, other deictics) and in sequence of tense effects.

(12) After a few minutes he said (that) he couldn't use her if she danced like that.

(13) After a few minutes he said, "I can't use you if you dance like that." [Brown:cf09]

(14) #After a few minutes he believed/thought, "I can't use you if you dance like that."

(15) Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice, 'without pictures or conversation?' [First line of *Alice's Adventures in Wonderland*]

However, direct speech has further options unavailable to indirect speech. Direct speech may be introduced by punctual verbs like *begin* and *continue*, as in (16); these typically take infinitival complements, so they cannot be used with indirect speech (17).

(16) A Birmingham newspaper printed in a column for children an article entitled "The Story of Guy Fawkes", which began:"When you pile your "guy" on the bonfire tomorrow night... [Brown:cd03]

(17) *A Birmingham newspaper printed an article which began that when you pile your "guy" on the bonfire tomorrow night...

Corpus analysis shows that direct speech is far less likely to occur with a complementizer (although it can), and is more likely to have a comma (dash, colon) before the complement clause. Direct speech often has quotation marks around the speech, but as noted above in Section 5.2, they are not required and sometimes are dropped altogether in cases such as dialogues in works of fiction.

In addition, both types of speech can occur with intransitive or transitive clausal complement verbs, as in examples (18) and (19):

(18) Because of deteriorating hearing, she told colleagues she feared she might not be able to teach much longer. [wsj0044]

(19) Richard Driscoll, vice chairman of Bank of New England, told the Dow Jones Professional Investor Report, "Certainly, there are those outside the region who think of us prospectively as a good partner." [wsj0067]

A preposition/subordinating conjunction is possible before a quoting clause, as in (20). *As* is the most commonly used.

(20) But he, as I can now retort, was the man who could see so short a distance ahead... [Brown:cg70]

Finally, both indirect and direct speech allow for multiple locations of the QUOTING CLAUSE (the subject and the quoting verb) relative to the quoted material: sentence initially, sentence finally and sentence internally. In all of these positions the verb of saying and its subject may be inverted. The next two sections will address these issues in more detail, as they are both unusual behaviors for a matrix verb in English.

5.4 Inversion in the quoting clause

Only with the intransitive clausal complement verbs, but in all three positions where the quoting clauses can appear, the subject and quoting verb may be inverted, as in (21). One rarely finds inversion in the sentence initial position in modern texts, but it is quite common when the quoting clause is either embedded or quote-final. Inversion of pronouns is also rare in modern texts — example (22) is from Jane Austen's *Persuasion*. No complementizers are permitted with the embedded and sentence-final orders.

(21) "The morbidity rate is a striking finding among those of us who study asbestos-related diseases," said Dr. Talcott. [wsj0003]

(22) "That is the woman I want", said he. "Something a little inferior I shall of course put up with, but it must not be much. If I am a fool, I shall be a fool indeed, for I have thought on the subject more than most men."

The inversion is unusual in that it involves a main verb, and English does not generally allow main verbs to invert. The syntactic details are not crucial for the current purposes, but for a detailed Minimalist account of quotative inversion, see Collins and Branigan [1996]. Their basic argument is that there is a null operator in Spec/CP. The operator raises from the complement position of the verb, where it leaves a co-indexed trace. (They claim that it can occasionally be lexicalized as *so* — "So Mary said.") The operator is bound by the quoted clause at a discourse level (similar to PRO_{arb}).

Given the syntactic free choice between inverted and non-inverted quoting verbs, there is clearly more to say about why one form or the other is used. Birner [1992] finds that quotative inversion does not pattern with the other types inversion she considers. If you think of the quoted material preceding the quoting clause as "preposed," you might expect it to pattern with other preposed elements. In true inverted constructions, where the subject is postposed and some other element is preposed, Birner finds that the preposed constituents are always discourse older than the subjects. However, with quotative inversion, she finds, contra other claims e.g. [Penhallurick1984], a number of examples where the preposed element is brand new. Also, the various positions in which the quoting clause occur, the fact that it occurs with transitive main verbs, and the the fact that it would be the only type of inversion to allow preposing of full clauses all argue against grouping quotative inversion with other types of inversion.

5.5 Positions available to the quoting clause

In addition to preceding the speech, the quoting clause verb may follow (23) or be embedded in the speech (24). If a verb can occur with reported speech, it can occur

in any of these three positions. Such behavior would be very surprising if the speech were always the complement of the verb, since subject-verb units cannot usually float around the sentence the way adverbs can.

(23) "You can't do this to us", Diane screamed. "We are Americans". [Brown:cf09]

(24) "Today 's action," Transportation Secretary Samuel Skinner said, "represents another milestone in the ongoing program to promote vehicle occupant safety in light trucks and minivans through its extension of passenger car standards." [wsj0064]

This positional variation raises interesting syntactic questions: are the various orders derived from the sentence-initial order, with the quoted clause always being an argument of the quoting verb? or are the quoting clauses text adjuncts, like parentheticals, adjoining into clauses at will? If the latter, do all of the orders behave alike? Emonds [1976; 1973] argues that both the sentence-initial and sentence-final orders are basic, and that the sentence-medial order is derived from the latter. In that case, do the sentence-initial orders of both direct and indirect speech have the same syntax, or do they diverge? Let us consider each of the positions for the quoting clause in turn, and see what they have to tell us about the larger syntactic picture.

5.5.1 Sentence-internal order

Our first case is where the quoting clause is embedded in the quote itself. A movement analysis where the speech starts out as the complement of the quoting verb and moves would be very surprising. It would require us to suppose that either the quoting clause moved to the left, into the quoting clause (some sort of "intraposition"), or the quote moved and wrapped itself around the quoting clause. While examples such as (25) suggest the possibility of a movement analysis which treats the subject of the quoted clause as topicalized (syntactically, not pragmatically), the portion preceding the quoting clause is frequently not just its subject. Examples (26) and (27) show a quoting clause coming between the verb and complement of the quoted material. This is not typical of a (syntactic) topicalization structure. Also, in topicalization you only get a comma after the topicalized element (and sometimes not even there), not in the site from which the element was moved.

(25) "Today 's action," Transportation Secretary Samuel Skinner said, "represents another milestone in the ongoing program to promote vehicle occupant safety in light trucks and minivans through its extension of passenger car standards." [wsj0064]

(26) "I rather resent", she said, "you speaking to those groups in Portland as though just the move accomplished this." [Brown:ca23]

(27) The appestat, which adjusts the appetite to keep weight constant, is located, says Jolliffe, in the hypothalamus–near the body's temperature, sleep and water-balance controls. [Brown:cc17]

One way of simulating "wrapping" without movement is to allow the quoting clauses to be independent text adjuncts, which can then be adjoined at any of a number of places in the quoted clause. LTAG is very well-suited to such an analysis, because, as noted in Chapter 3, the clause into which the quoting clause adjoins is in itself a complete matrix sentence. Thus, there are no concerns about passing agreement or other clausally local information "across" the parenthetical quoting clause. Sample LTAG trees for pre-VP and post-V quoting clauses are shown in Figure 5.2.

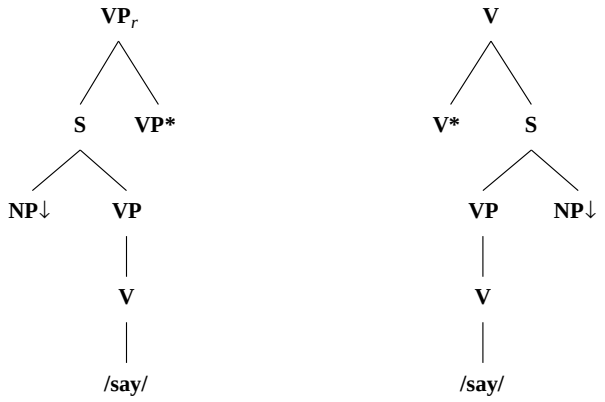


Figure 5.2: The trees used for a non-inverted quoting clause: (a) pre-VP e.g. *"Today's action," Transportation Secretary Samuel Skinner said, "represents another...* and (b) post-V, e.g. *"I rather resent", she said, "you speaking..."*

Because the grammar is lexicalized, we can elegantly capture the generalization that only verbs taking clausal complements can select this structure. The LTAG lexicon groups clausal trees into *Tree Families*, which contain all of the constructions allowed for a single subcategorization frame (active, passive, wh- question, relative clauses, etc.). These adjunct trees would simply be members of the clausal complement tree families.[4] Furthermore, as mentioned earlier, the LTAG trees have a larger domain of locality than context-free grammars. In the tree shown in 5.2, the relationship between the quoting verb and the clause it adjoins into is expressed in a single rule, allowing us to directly state constraints imposed by the quoting verb on the quoting clause.

[4]It is important to note that each tree in a family can be associated with distinct semantic and pragmatic content.

On this analysis, the quoted clause is not overtly a complement of the quoting verb. However, each tree in a tree family is associated with a type, which gives the number and type of arguments the verb requires. The transitive family would have the type NP $\times$ NP and the intransitive clausal complement family, the type NP $\times$ S. When an argument is not overtly realized, as in an agentless passive, this information is available to the semantic and discourse modules. For the passive, we can look for the agent in the discourse context, while for the quoting clause, the semantic component could associate the matrix clause with the missing complement. If one wanted a more explicit connection, a null operator as in [Collins and Branigan1996] could be built into the adjunct quoting clause tree.

There are a number of reasons to believe that the adjunct clauses analysis is correct analysis for quoting clauses separated by punctuation. In this construction, verbs lose many of their selectional restrictions. As noted above, punctual verbs usually select for infinitival complements, yet they can be embedded in tensed quoted clauses. Verbs also lose their selectional restrictions as to *wh-* features, with verbs like *insist* embedding in questions as in (28). Ross [1973] claims that verbs do retain some selectional restrictions in this construction, which he classifies as parenthetical. He notes that the parenthetical and canonical SVO orders share sequence of tense restrictions, factivity effects, and a few other restrictions, most of which seem to be of a more semantic nature. The structure he suggests for this type of parenthetical ends up looking just like the adjunct structure proposed here, although it is derived from the SVO order via transformations. (The first transformation yields the sentence-final order, and then further transformations apply to move the parenthetical into other positions in the clause.) Many of the selectional restrictions he discusses would be handled at the tree family level as presented above, and do not appear to be incompatible with the current analysis.

(28) Who, Mary insisted, has ever seen a purple elephant?

Furthermore, the embedded quoting clauses are frequently interchangeable with other kinds of parentheticals: *John, I presume/presumably/it seems, bought a new car*. Like other parentheticals, quoting clauses are also argued to be set off with "comma intonation" in speech. [Schmidt1995] finds that there is a significant pitch range restriction across parenthetical types, but he does not give any examples of direct quotation. While we should be cautious about drawing analogies between prosody and punctuation for the reasons noted in Section 1.4.2, if quoted clauses were shown to have a similarly restricted pitch range, this would be further evidence for the similarity of the constructions.

In his discussion of parentheticals as discontinuous constituents, McCawley [1982] argues that the parenthetical does not behave as part of the constituent that contains it. The ellipsis tests he use to support his argument suggest that the quoting clauses behave similarly.

(29) John, Mary said, bought a house, and Sue did too = Sue bought a house OR Sue said John bought a house ≠ Mary said Sue bought a house

In (29), the antecedent for the ellipsis is *said* or *bought*, but not *said..bought*. This is what would be predicted if the complete sentence is not a constituent available as an antecedent.

Abeillé (p.c) has pointed out that there are some sentential complement verbs which cannot occur parenthetically, or can only occur with particular subjects. However, this seems again to be a pragmatic rather than syntactic restriction.

(30) Mary recently quit her job, I/*you/Bill heard yesterday.

In (30), the second person subject is nonsensical, as the speaker would be telling the hearer what the hearer had heard. However, the first and third person subjects are acceptable. Likewise, negative verbs (*doubt, deny*) cannot generally be used parenthetically; Ross claims that these can surface as negated parentheticals.

(31) I doubt I will go to the party tonight.

(32) *I will go to the party tonight, I doubt.

We could give either a syntactic or pragmatic explanation for this contrast. If we take the empty operator seriously, we could argue that it does not allow negative features which others have argued are introduced by negative verbs [Mugarza1992]. Alternatively, we could argue that this construction asserts the quoted clause, and we cannot then retract it with the quoting clause. We will discuss this latter claim in section 5.7.

5.5.2 Sentence-final position

In this order, it is certainly more plausible that the quoted clause is a fronted complement. However, Emonds gives some compelling examples against a derivational relation.

(33) John hasn't completed his book, I don't think. [Emonds' II.91]

(34) John hasn't completed his book, I think.

(35) I don't think John hasn't completed his book.

(36) I think John hasn't completed his book.

Sentences (33) and (34) are synonymous for speakers who accept both variants, i.e. the negation in the quoting clause has no effect. However, in the sentences these would have to be derived from, (35) or (36), the presence or absence of the matrix or negation does change the meaning of the sentence.

Additionally the sentence-final order for quoting clauses shares the features of embedded quoting clauses discussed in the previous section, i.e. the loss of some selection restrictions, the obligatory presence of punctuation, synonymity with other parentheticals, and we again are led to decide against a movement analysis and for an adjunct analysis. Figure 5.3 shows the relevant tree.

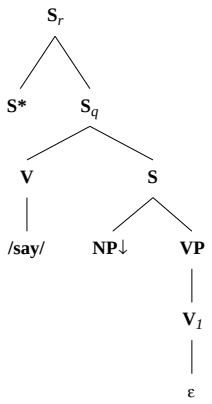


Figure 5.3: The tree used for an inverted, post-S quoting clause, e.g. *'Come, let's try the first figure!' said the Mock Turtle to the Gryphon.* [Carroll:AAIW]

5.5.3 Sentence-initial position

Finally, we come to the most difficult case — the quoting clause in sentence initial position. As in the other cases, direct and indirect speech are identical in the left to right order of constituents. In the previous two sections, direct and indirect speech patterned together, as parenthetical clauses. However, the question here is whether they will continue to pattern together. The LTAG analysis for normal clausal complement structures is shown in Figure 5.4. The tree adjoins at the root of the complement clause tree for indicative clausal complements and below the extracted element in extracted clause. This analysis gives an elegant treatment of long-distance extraction (see [Kroch and Joshi1985]).

We could simply allow the additional punctuation to adjoin to this tree for sentences like (37).

(37) Alice replied very readily: 'but that's because it stays the same year for such a long time together.' [Carroll:AAIW]

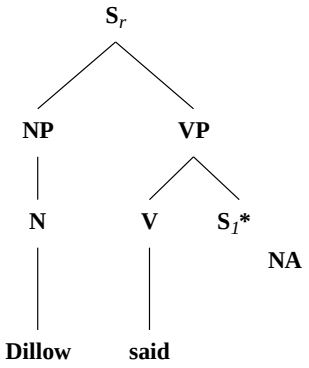


Figure 5.4: The basic LTAG tree for clausal complements.

However, if we look more closely at the two kinds of speech, we find several differences. For one, direct speech requires that questions be inverted, (38) and (39), while normal clausal complements cannot be inverted, (40) and (41).

(38) Alice asked, 'Has anyone seen the Cheshire Cat?'

(39) *Alice asked, 'Anyone has/had seen the Cheshire Cat?'

(40) *Alice asked whether had anyone seen the Cheshire Cat.

(41) Alice asked whether anyone had seen the Cheshire Cat.

Secondly, you cannot get embedding in the quoting clause of direct speech, whereas you can have (in principle) unbounded embedding in clausal complements:

(42) *The queen said the White Rabbit whispered Alice asked, 'Has anyone seen the Cheshire Cat?'

(43) The queen said the White Rabbit whispered...that Alice asked whether anyone had seen the Cheshire Cat.

These differences suggest that Emonds was correct in concluding that the quoted clause in the parenthetical type of quoted speech is a matrix clause, rather than an embedded one. This leaves us with two kinds of possible derivations for sentence initial quoting clauses. If there is no punctuation other than quotation marks after the quoting verb, we use the tree in Figure 5.4. This will mean giving sentences like (44) the same analysis as indirect speech, i.e. the non-parenthetical analysis. If there is punctuation, we would use the LTAG tree would that shown in Figure 5.5.

(44) Gemina said in a statement that "it reserves the right to take any action to protect its rights as a member of the syndicate." [wsj1371]

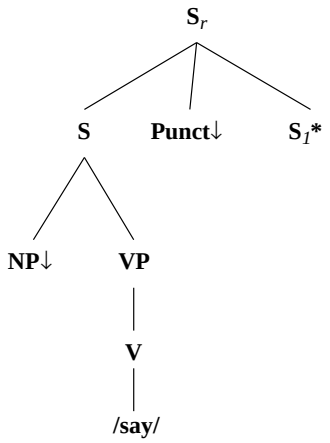


Figure 5.5: The LTAG tree for sentence-initial adjunct quoting clauses.

5.5.4 Conclusions about handling quoting clauses

Based on independent consideration in the previous three sections, I have argued that quoting clauses in all three positions relative to the quoted material ought to be treated as text adjuncts. Those which appear internally and quote-finally are exclusively text adjuncts, while those which precede the quote can be either text adjuncts or clausal complement structures. In the remaining sections, I discuss how the punctuation required by the adjunct structures should be handled, as well as the semantic implications of the analysis. Section 5.9 describes the performance of the analysis in an information extraction task.

5.6 Punctuation in reported speech

As the alert reader will have noticed, there has been little discussion about handling the punctuation marks present in the quoting clauses. The quotation marks would be handled as shown in Figure 5.1 above and repeated below as Figure 5.6, simply adjoining onto the quoted constituent.

Having concluded that the two main classes of quoting clauses are parenthetical and non-parenthetical, there is obviously more to say about the comma, dash or colon separating the quoting clause from the quote.

5.6.1 Quote transposition

Since we are treating the quoting clause like a parenthetical, the commas around it are Nunberg's "delimiting" punctuation marks, and absorption applies to the second mark. Because we are pre-compiling the absorption effects, as it were, only the sentence-internal adjuncts (adjoined to V or VP) will have both commas. The sentence-initial and sentence-final adjunct are inherently unbalanced, i.e. one mark will always be absorbed at the beginning or end of the clause. This is easily captured

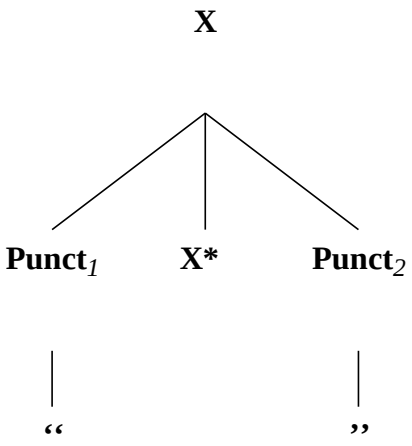


Figure 5.6: Schematic tree for quotation marks

in the LTAG treatment, as each position for the quoting clause has its own tree. This also allows us to license a colon in the sentence initial order, but not in either of the other orders, but let us leave aside the colon until the next section. The tree for post-subject quoting clauses is shown in Figure 5.7; the trees for pre-S and post-S clauses are similar, but would have only one PUNCT node. The punctuation nodes are substitution sites, built into each tree, and are thus required to be instantiated for the tree to be licensed. This declarative formulation is useful in parsing with a lexicalized grammar, where syntactic structures are only licensed by lexical items in the input string.

As Nunberg [1990] discusses at some length, American English and British English differ in how they treat certain punctuation marks when they occur adjacent to a closing quote. In American English, commas and all terminal punctuation marks (periods, question marks and exclamation points) are transposed with closing quotation marks (e.g. .”), whether they are logically associated with the entire sentence or only with the quoted portion. In British English, the comma or terminal mark remains outside of the quote (e.g. ”.), unless it is logically a part of the quoted material.[5] An examination of the Brown corpus (exclusively American texts) shows this distinction to be unhelpful in processing corpus data: there are only 39 commas and 28 periods inside of quotation marks (both single and double), but 1823 commas and 1023 periods outside. The so-called British system is massively predominant. This may be the result of post-processing on the corpus, since analysis of 2.5 million words of Wall Street Journal data turns up only 15 commas and 15 periods in

[5]No one seems to have a good account of why transposition occurs. There are some claims that it was a move made by type-setters for either practical or aesthetic reasons, but then one has to wonder why only American type-setters took up the practice. Jones [1996b] states quite definitively that it is an aesthetic move because “the white-space underneath the final quotation marks is seen as disrupting the natural reading movement of the eye,” but then why are dashes not transposed?

the British system. Sampson [1992] also cites an example from the LOB corpus of British English which uses the American system. Jones [1996b] finds both variants throughout the nine sources he used for his corpus analyses. In any event, if one is dealing with naturally occurring data, one is likely to encounter both systems in varying proportions.

So, how is one to (a) require the separating punctuation mark to be present on the right of the quoting clause and (b) allow it to occur in either of two locations (inside or outside the closing quotes)?

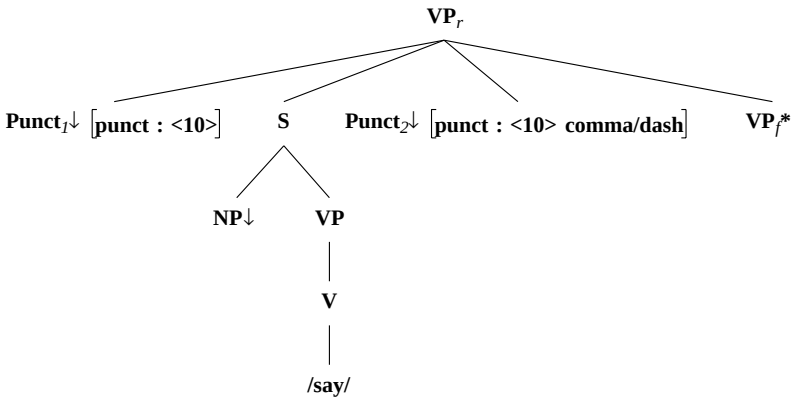


Figure 5.7: Tree for embedded quoting clause, with punctuation argument positions.

Using tree 5.7 on a simplified version of (45), the first pair of quotation marks is around the subject NP and the second is around the VP. With this tree, we can only derive the British order, Figure (5.8). The simplest solution is to do some "normalization" in tokenizing the data, in this case into the British form. This is the option which we are currently pursuing with the XTAG English grammar.

(45) "Today 's action," Transportation Secretary Samuel Skinner said, "represents another milestone in the ongoing program to promote vehicle occupant safety in light trucks and minivans through its extension of passenger car standards." [wsj0064]

An alternative would be treat quote inversion as something like clitic-climbing by the punctuation mark. This would allow us to use the same tree for both orders, but the American order would use a multi-component tree set. Briefly stated, a multi-component set allows one to force a set of trees to act as single tree — if one tree in the set is used in a derivation, all of the trees must be used. The two components of this set would be a tree anchored by the trace, which would substitute into the argument position, and a tree anchored by the comma, which would adjoin to the closing quote. The same multi-component set would be selected by both the comma

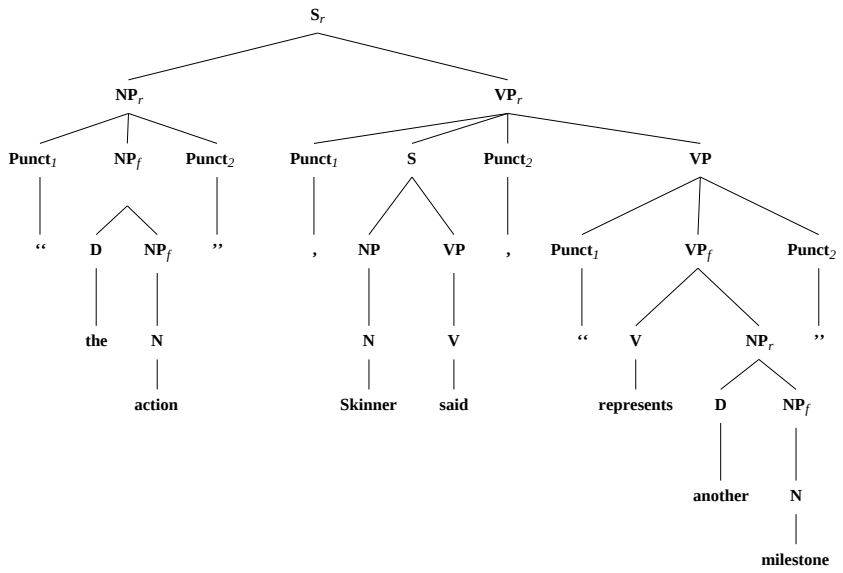


Figure 5.8: Parsed sentence with embedded quoting clause and quotation marks, British order.

and the terminal, but not by the dash, semi-colon or colon as they do not undergo quote inversion.[6]

5.6.2 How to treat the colon

With sentence-initial quoting clauses, a colon can sometimes follow the quoting verb. Given that the colon is not typically a delimiting punctuation mark, and that it does not participate in quote transposition, we might well want to group constructions like (46) with non-parenthetical quoted speech, allowing a colon to adjoin into Tree 5.4.

Recall that the colon is possible only in the sentence-initial order. Also, complementizers are more freely permitted here. It remains to be seen whether this construction takes inverted complements — if it does not, then it surely ought to be classed with the non-parenthetical quoting clauses.

(46) Indicating the way in which he has turned his back on his 1910 philosophy , Mr. Reama said: “A Socialist is a person who believes in dividing everything he does not own”. [Brown:ca05]

[6] In fact, dashes quite rarely set off quotative clauses and typically only do so in the clause internal position. It would be straightforward to capture this with the features in the LTAG quoting clause trees.

5.6.3 Quote alternation

American English requires that nested quotation marks alternate between single and double marks, with double-quotes on the outermost pair. In British English, the outermost quotes are single but, again, alternation is required. This is handled in the LTAG account by a CONTAINS feature (introduced in Section 3.2), which is also used to block self-embedding of other text-adjuncts (cf. discussion in Section 3.2). The feature has the value DQUOTE+ at the root of the tree anchored by double quotation marks (shown in Fig. 5.9), to indicate that the subtree contains double quotes, and the value DQUOTE– on the foot node, to block the tree from adjoining to any subtree which already contains double quotes. The same feature is used with the value SQUOTE+/– for single quotes. Note that since the grammar is lexicalized (here, on the punctuation marks themselves) the features come from different instantiations of a single tree (i.e. we do not need separate trees for each type of quotation mark). Other trees in the grammar are simply transparent to the CONTAINS feature, passing up its value in the relevant contexts. The quote trees themselves are opaque to all other values of CONTAINS, so that for instance, while colon-expansions cannot usually be embedded, they can be embedded if the inner expansion is inside of quotation marks. This treatment handles both the English and American styles, as the features merely require alternation of single and double quotes without specifying what type the outermost quotes should be.

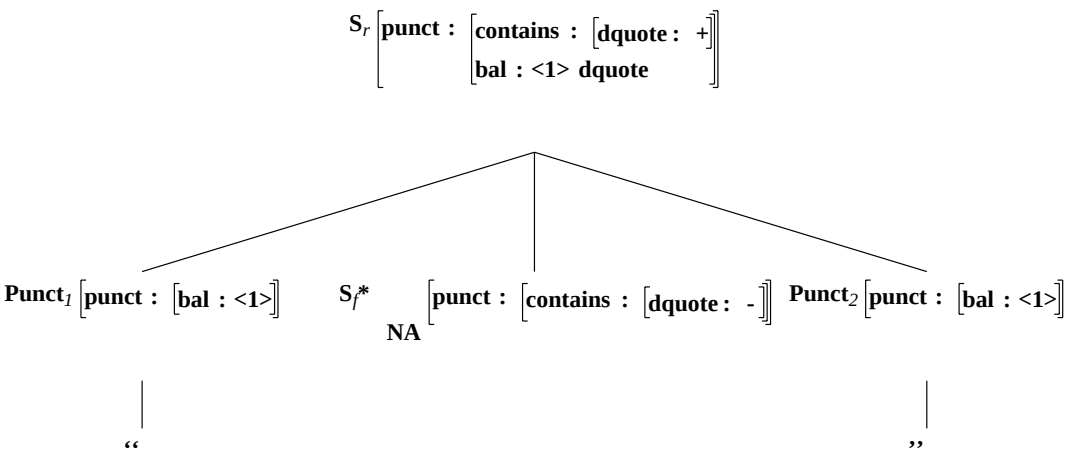


Figure 5.9: The LTAG tree for quotes around a clause, with the punctuation features shown.

5.7 Interpretive issues for this analysis

5.7.1 Traditional semantic accounts

The semantics of embedded clauses of all types has been the subject of centuries of consideration especially with regard to the issue of referential opacity. I will not attempt to synthesize the relevant literature here, but will simply mention a few points which are especially relevant to the parenthetical/non-parenthetical quoting clause distinction. Adding semantic information to the proposed syntactic account ought to be straightforward, if one assumes a compositional semantics where a meaning is associated with each LTAG tree as in [Shieber and Schabes1990; Stone and Doran1996; Stone and Doran1997]. The fact that all of the parenthetical trees adjoin to the VP spine, and all projections of V are S trees guarantees that both the parenthetical and non-parenthetical quoting verbs will adjoin to complete clauses and therefore will be semantically composed with complete propositions.

Much of the relevant discussion concerns "parenthetical constructions" defined to exclude reported speech, but most of the generalizations hold for reported speech as well. Urmson [1963] is the first of a number of authors to suggest that the main informational contribution comes from the "complement" and the parenthetical clause is more like a qualifier when it is in adjunct position. Urmson says "They function rather like a certain class of adverbs to orient the hearer aright towards the statements with which they are associated....They help the understanding and assessment of what has been said rather than being a part of what is said." Li [1986] describes these clauses "epistemic quantifiers"; Bolinger [1972] calls them "adverbialized message verbs" and talks about the "autonomy of the message"; Hand [1993] says that the illocutionary force of the expression is carried by the "complement" and not by the "matrix"; and Thompson and Mulac [1991a] call them "epistemic parentheticals" acting like adverbs, saying that "...when there is no *that*, the main clause subject and verb function as an epistemic phrase, not as a main clause introducing a complement."

This view in in contrast to the paratactic account put forth by Davidson [1984] and extended by Lepore and Loewer [1989], which argues that both clauses are semantically main clauses, with *that* serving as a deictic pointer to the complement clause. Counter-arguments to this account presented in the abovementioned works are quite compelling, and include:

- If the "complement" is an independent clause, why does it not necessarily have to be a complete clause? Also, why can there be semantic links, e.g. bound pronouns, negative polarity items licensed by the "matrix"?

- *That* in this context behaves like a complementizer, not a pronoun - it undergoes phonological reduction, is deletable and cannot refer to non-S constituents [Segal and Speas1986]

Less attention has been paid to the pragmatics of these constructions, but Bolinger's [1972] discussion of the use of the complementizer *that* is quite relevant. Bolinger finds that *that* is more likely to appear in ambiguous contexts, in particular in cases where the subject of the complement clause could be interpreted as the object of the higher verb, and in those contexts more likely with verbs which can also occur transitively. These findings are verified by psycholinguistic experiments (cf. work by Trueswell and colleagues, e.g. [Trueswell and Tanenhaus1994; Trueswell1993]), which confirm that verbs which occur less frequently as transitives are correspondingly less likely to be misinterpreted as transitive when *that* is dropped in ambiguous contexts. Bolinger also observes that the complementizer is much more likely to be dropped when the proposition conveyed by the message clause is new information in the discourse. He discusses the parenthetical message construction itself briefly, calling the reporting clauses "adverbialized message verbs" All of these characterizations fit with the parenthetical quoting constructions. The message typically is usually new information, making *that* less likely. Additionally, *that* is not needed to disambiguate the constituent structure as there is a comma (or vice-versa, the comma is required because there is no complementizer).

5.7.2 An alternative approach

One interesting possibility is that, when the quoting clauses are text-adjuncts, they are related at the level of the discourse grammar rather than the sentence grammar.[7] Along the lines of the approach described in Section 3.7, we might look at the quoting clause and the quoted material as being connected by a discourse relation such as EVIDENCE. How such a treatment would work for the sentence-initial and sentence-final cases is clear enough, as they behave just like subordinate clauses. There is even a subordinating conjunction which can optionally surface on quoting clauses, *as*. However, the sentence-internal quoting clauses look as if they might need a different treatment.

As work by Prince and her students ([Birner1992; Ward and Prince1991; Prince1986; Ward1985] et al.) has shown, all syntactic paraphrases exist for a reason, and the choice of one variant over another is driven by pragmatic concerns. Thus, it must surely be the case that writers (for this is primarily a construction of written genre) have some reason for choosing to insert a quoting clause, or any other text adjunct, into the quote/matrix clause. Webber and Joshi [1998] briefly consider this issue, but only for simple (non-clausal) cue-phrases. They conclude that adjoining them internally is a way to split the matrix into theme and rheme, with the relation triggered by the cue-phrase holding between one of these parts and the preceding

[7] Espinal [1991] makes a similar such claim about a class of "disjunct constituents," which from her examples would appear to include this sort of quoting clause. Her proposal is to treat them as completely separate syntactic constituents within a multi-dimensional syntactic structure, which are linked only when the entire utterance is interpreted.

discourse, instead of between the entire proposition and the preceding discourse.

(47) a. Although the episodic construction of the book often makes it difficult to follow,

b. it **nevertheless** makes devastating reading.

b'. #**nevertheless** it makes devastating reading. [Webber & Joshi's (9)]

Example (47) illustrates a case where the placement of the cue phrase is significant. Could the same be true of quoting clauses? (Subordinating clauses, which can also appear between the subject and main verb, ought to behave similarly.) Example (48) is extracted from the Wall Street Journal.

(48) a. The Transportation Department, responding to pressure from safety advocates, took further steps to impose on light trucks and vans the safety requirements used for automobiles.

b. The department proposed requiring stronger roofs for light trucks and minivans, beginning with 1992 models.

c. It also issued a final rule requiring auto makers to equip light trucks and minivans with lap-shoulder belts for rear seats beginning in the 1992 model year. Such belts already are required for the vehicles' front seats.

d. *"Today's action,"* **Transportation Secretary Samuel Skinner said,** "represents another milestone in the ongoing program to promote vehicle occupant safety in light trucks and minivans through its extension of passenger car standards."

The topic (using the term in its most casual sense) of sentences (a)-(c) is the transportation department, with the subject NPs shrinking in a most pleasing fashion from the full description of the department in (a) to *it* in (c). Suddenly in (d), the topic is the action taken by the department. This example certainly is suggestive that inserting the quoting clause into the quote has some sort of topic-marking or focusing function, but more work remains to be done here. In particular, we need to look at the cases where the quoting clause comes between the verb and its object. There is no a priori reason to believe that, like movement-driven types of topic marking, the topic here has to be an NP or even a standard syntactic constituent; in this, the construction may well pattern with certain types of prosodic topic-marking [Prevost and Steedman1993].

5.8 Cross-linguistic generalizations about reported speech

Coulmas' [1986] volume on reported speech brings to light a number of interesting cross-linguistic generalizations. Coulmas, in her introduction, notes that all speech is processed by the reporter before they report it. Tannen [1986] takes an even stronger position, saying that there is no reported speech as such, only "constructed dialogue." Her bottom line is that things are never truly reported verbatim, no matter how they are couched.

Nonetheless, reporting what others have said with some degree of accuracy is a basic function of language. Coulmas claims that all languages have some way to report speech. For those languages which attempt to differentiate direct and indirect speech, the line between the two is extremely blurry. In fact a number of articles in the volume note that the deictic center of the sentence may be switched mid-sentence, indicating that the speaker has mixed direct and indirect reporting.

The reported clause may be more or less tightly syntactically and semantically "integrated" or "fused" with the reporting clause. This integration is shown by: shifting deictics—pronouns, tense, agreement markers; sequence of tense/mood effects; presence or absence of complementizer on reported clause; different word orders; or the use of particles. Many of these effects are very similar to what is discussed above for English. Like English, Yoruba [Bamgboṣe1986] allows a complementizer on both direct and indirect reports, but it is more likely with indirect. Slave (an Athabaskan dialect) [Rice1986] and Hungarian [Fónagy1986] only allow a complementizer on indirect speech. Danish data [Haberland1986] suggests that indirect speech uses subordinate clause word order (V2), while direct speech uses main clause word order; this parallels the English adjunct/main-clause distinction. Likewise, in English, inverted clauses are possible in what I am classifying as parenthetical reported speech, but not in non-parenthetical/complement clauses. Fónagy says that a number of Indo-European languages allow inversion of reporting clauses in the clause-final position. Other properties correlate with the English findings that some quoting clauses are text adjuncts, but are not directly comparable; for instance, in Hungarian, Fónagy finds that object marking on a verb of saying indicates tighter integration of quote, even when the quoting clause follows the quote.

As in English, quotation marks are found not to be a strong indicator of direct speech in Japanese [Maynare1986], Danish [Haberland1986] or French (Anne Abeillé, p.c.). Although it is not made explicit in most of the articles, it appears from the data as if a number of widely differing languages use punctuation in a manner similar to English: Yoruba [Bamgboṣe1986], Swahili [Massamba1986] and Danish [Haberland1986] use quotation marks, and separate the quoting clause with a comma or even a colon or dash (in Danish); Georgian [Hewitt and Crisp1986] and Hungarian [Kiefer1986] separate the quoting clause with a comma. At least French and Hungarian [Fónagy1986] also allow multiple positions for quoting clause; the

data given for other languages is not extensive enough to tell whether they do as well.

5.9 Evaluating the analysis

This analysis of reported speech was used in a template filling task, conducted at the University of Pennsylvania, and sponsored by Lexis-Nexis. The effort focused on employing existing technologies, such as tokenizers, parsers and part-of-speech taggers, to do information extraction from news data. The task was similar to the MUC-6 template filling task, but the actual fields to be filled were rather different. See [Doran et al.1997] and [Baldwin et al.1997] for descriptions of the project. The LTAG grammar was used to do parsing using the Supertagging technique developed by [Srinivas1997] and introduced in Section 3.8 above. Recall that Supertagging uses the LTAG trees as complex part-of-speech tags and uses standard statistical tagging techniques to assign the correct tree to each word. It is then quite straightforward to connect the trees and construct a derivation for each sentence.

Two of the template fields to be filled were COMMENT and COMMENTER; these were filled with direct quotes relevant to the topic of the template. COMMENT and COMMENTER pairs were identified in the news texts using patterns written in a regular expression language called MOP [Doran et al.1997]. A set of patterns were written using the LTAG trees described above for verbs of saying. Supertags allowed for just the right level of generalization in these rules, as listing the verbs one by one is not practical for a class of this size, and yet allowing any verb at all would be too permissive. The performance of the comment patterns was excellent: in our evaluation of the entire system, the COMMENT field was our most accurately filled; of 106 evaluation templates, there was only 1 for which our patterns failed to find a quote.

Even more interesting than the patterns themselves is the accuracy of the Supertagged text they rely upon. If the verbs of saying are not assigned the correct LTAG tree, any later processing will invariably fail. To determine the correctness of the Supertag assignment, just over 2000 Wall Street Journal sentences were tagged using a Supertagger trained on 200,000 hand-corrected words of WSJ data, and evaluated. Of the 192 instances of verbs of saying used in parenthetical reported speech, 162 (84%) were assigned the correct parenthetical Supertag, 5 were tagged with an incorrect parenthetical tree, and 25 were tagged as non-parenthetical verbs of saying. There were also 25 instances of non-parenthetical reported speech (out of 602 total, 4%) which were incorrectly tagged as parenthetical, and 4 instances of non-reported speech which were likewise incorrectly tagged. Overall, the correct subcategorization frame (sentential complement) was assigned 99.5% of the time. This suggests that the combination of lexical probabilities for the verbs taking clausal complements and the presence of the separating punctuation mark in the parenthetical constructions allows a simple trigram model to correctly identify the appropriate reported speech

construction.

5.10 Summary

In this chapter, I have shown that the presence of punctuation is correlated with the parenthetical nature of some reported speech. Quotation marks do not provide additional constraints on the construction, nor is the traditional distinction between direct and indirect speech found to be a useful one. The parenthetical properties of those quoting clauses set off with punctuation also correspond to a lack of syntactic integration between the quoted clause and the quoting clause, which is reflected in many languages other than English.

Using a lexicalized grammar, we can license the parenthetical quoting clauses as text adjuncts, anchored by the appropriate subset of verbs, and selecting the relevant punctuation marks as arguments. Lexicalization and features as utilized by LTAG allow us to elegantly capture the distribution of both the verbs and the punctuation marks in the relevant constructions. Quotation marks may optionally adjoin, in a separate step. Quoting clauses which are sentence-initial and are not separated from the quote by a comma or dash are treated as normal clausal complement verbs, with the quoted material as the internal argument of the verb.

I have also argued that the account presented here is compatible with either of two classes semantic treatments, and observed that many of the properties of reported speech in English are likewise found in other, very dissimilar languages.

Chapter 6

Parentheses

6.1 Background on parentheticals[1]

Parentheses have a number of functions, which are typically characterized in grammar books and casual analyses as providing background information. In cases where they are used interchangeably with another punctuation mark, for instances dashes or commas, the material they enclose is standardly described as less closely integrated into the rest of the text than if one of the other marks is chosen. The Chicago Manual of Style [1982, 5.97] says that parentheses, "...like commas and dashes, may be used to set off amplifying, explaining, or digressive elements" and Quirk et al. [1985, III.3 and III.20] describe them as marking an "obtrusive" or "sharp interruption in the structure within which they are inserted."

Structurally, parenthesized material may be of any grammatical category, from multiple sentences down to a letter, and may occur anywhere in a sentence except at the left-most edge of a clause. Parentheses are obligatorily paired, but unlike paired dashes and commas, do not undergo absorption. They do absorb non-terminal punctuation marks which would occur inside the closing parenthesis, but are not themselves absorbed by anything else. Unlike quotation marks, they do not undergo transposition, i.e. if you have a sentence ending with parenthesized material, you do not move the sentence ending punctuation mark inside the parentheses. This is exemplified below in examples (1) and (2).

(1) Innumerable motels from Tucson to New York boast swimming pools ("swim at your own risk" is the hospitable sign poised at the brink of most pools). [Brown:ca17]

(2) Each enjoys seeing the other hit home runs ("I hope Roger hits 80", Mantle says), and each enjoys even more seeing himself hit home runs ("and I hope I hit 81 "). [Brown:ca39]

[1] In this chapter, I will use the term *parenthetical* to mean anything enclosed in parentheses; elsewhere in the document, the term is used in its more general sense.

Nunberg points out that parenthetical material is not available for later reference [1990, Ch. 6]. This can be seen in the continuations which are and are not possible for example (3).

(3) a. Steps 1 through 4 may be omitted if reducer and elbow were not removed from *(or have already been installed on)* pressure switch. [F16]

b. In that case/if they were not removed, skip directly to step 5.

b'. #If they were not removed, skip directly to step 5a; if they were already installed, go to 5b.

He attributes this referential isolation to the "semantic function of parentheticals, which ensures that their content is not actually incorporated into the text proper, and so is unavailable for any external reference." [Op. Cit, p. 105] If parentheticals are neither syntactically nor semantically connected to the sentence, what then is their role? Nunberg describes it thus: "if the content of the parentheticals is to figure in interpretation, it must be relative to some other circumstances of interpretation, which are distinct from the context associated with the primary text." [P. 106]

What this means is that there is a situation in which the sentence containing the parenthetical is expected be interpreted, but that sometimes the author wants or needs to explicitly acknowledge that the expected situation may not hold along some dimension of contextualization. In this, text is different from many spoken genre in not having a particular time of utterance, addressee, indexical context, etc. Texts are typically addressed to a wide-range of people who will be reading the text at some unknown time in the future. As a result, most texts are addressed to what Nunberg calls the "presumptive reader," and parentheticals allow the writer to "accommodate circumstances in which the values of relevant contextual parameters depart from those of the presumptive context." [P. 110]

6.1.1 Kinds of parentheticals

Nunberg identifies two main classes of parentheticals: ones that introduce alternative texts and ones that restrict the context of interpretation. He also distinguishes lexical parentheticals from text-level ones, with all lexical parentheticals falling into the alternative category and text-level ones being of both types. Lexical parentheticals are characterized as not disrupting the surrounding syntax, i.e. you could simply remove the parentheses and have a syntactically well-formed sentence. Textual parentheticals do not disrupt the syntax of the sentence containing them *per se*; rather, it is that they have no syntactically licensed way of attaching to the sentence, and are simply spliced in. As noted in Chapter 3, this distinction is clear in the LTAG analysis of punctuation, as text adjunct trees will introduce both punctuation marks and additional lexical material, while lexical adjunct trees will contain only the punctuation marks themselves. Figure 6.1(a) shows a lexical parenthetical tree (cf. also

Section 3.5.2), which would simply insert parentheses around an adjective in normal pre-modifying position, while (b) shows the tree for a parenthetical NP appositive (cf. Section 3.5.1).

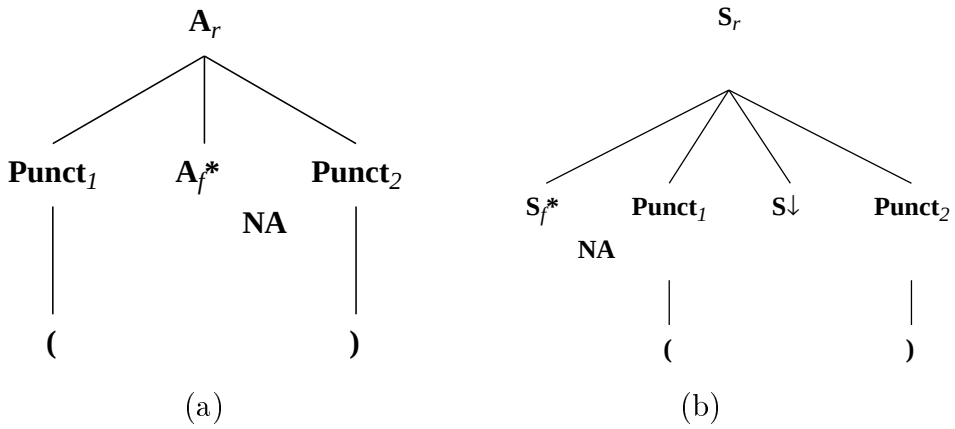


Figure 6.1: Two trees for introducing parentheses: (a) for a lexical parenthetical adjective (e.g. *the (usually agent-less) passive*), and (b) for a text-level parenthetical NP appositive (e.g. *100,000 francs (about $300)*)

The following examples illustrate some of the realizations of parentheticals.

- Alternative texts as lexical adjuncts:

  (4) Obviously hydrophobic *(oleophilic)* substances such as greases, oils, or particles having a greasy or oily surface.... [Brown:cj05]

  (5) Fearless Freddy Bryan could take credit, if he cared to *(and he did)*, for the second time. [Brown:cp29]

- Alternative texts as text adjuncts:

  (6) The Greek evidently fell for her, "Monsieur X" recounted, and to clinch what he thought was an affair in the making he gave her 100,000 francs *(about $300)* and led her to the roulette tables. [Brown:cf09]

  (7) Printed material Available, on request, from U.S. Department of Agriculture, Washington 25, D.C., are: Cooperative Farm Credit Can Assist In Rural Development *(Circular No. 44)*, and The Cooperative Farm Credit System *(Circular No. 36-A)*. [Brown:ch01]

- A sub-class of the alternative texts, which Nunberg calls "in case you're interested" parentheticals. (8) and (9) are lexical, and (10) is a text adjunct:

(8) There's a memorable passage in which Mr. Shields, having finally learned of the practice, expresses his outrage to Bill Clements, then a university governor *(and now the governor of Texas!)* and oil man Edwin Cox, chairman of the board of trustees. [wsj0966]

(9) Dick Carroll and his accordion *(which we now refer to as "Freida")* held over at Bahia Cabana where "Sir" Judson Smith brings in his calypso capers Oct. 13 . [Brown:ca31]

(10) Innumerable motels from Tucson to New York boast swimming pools *("swim at your own risk" is the hospitable sign poised at the brink of most pools).* [Brown:ca17]

- Context restricting parentheticals, text adjunct:

(11) The best rule of thumb for detecting corked wine *(provided the eye has not already spotted it)* is to smell the wet end of the cork after pulling it.... [Brown:cf27]

There are a number of examples which do not fit in any direct way into either of Nunberg's classifications. These present what I would consider parallel texts in parentheses. In (12), the text is interspersed with commentary from the person being discussed.

(12) Each enjoys seeing the other hit home runs *("I hope Roger hits 80", Mantle says)*, and each enjoys even more seeing himself hit home runs *("and I hope I hit 81").* [Brown:ca39]

Similarly, in (13) and (14),

(13) The man most firmly at grips with the problem is the University of Minnesota's Physiologist Ancel Keys, 57, inventor of the wartime K (for Keys) ration and author of last year's bestselling Eat Well And Stay Well. From his birch-paneled office in the Laboratory of Physiological Hygiene, under the university's football stadium in Minneapolis *("We get a rumble on every touchdown")*...Despite his personal distaste for obesity *("disgusting")*, Dr. Keys has only an incidental interest in how much Americans eat. [Brown:cc17]

(14) But Wisman, too, does not know the go code. He must take it from "the red box"....The box is internally wired so the door can never be opened without setting off a screeching klaxon *("It's real obnoxious").* [Brown:cg03]

The remainder of this chapter will look at the uses of parentheses in a corpus of military instructional data and a set of academic papers, and will consider how well they fit with Nunberg's classification.

6.2 Parentheses in the F16 Technical Orders

6.2.1 Structure of the F16 Technical Orders

The F16 Technical Orders (maintenance manuals, [T.O. 1F-16CG-2-28JG-20]) are divided into five main sections: the first specifies the pre-conditions for performing the procedure; the second lists the participants in the repair; the third lists equipment that will be needed; the fourth enumerates the repair procedure step-by-step; and the fifth lists follow-up procedures, if any are required. All of the tasks and preparatory activities, including the overarching Technical Order for the repair, are assigned code numbers and are cross-referenced using the codes. Parentheses are used in a number of different ways in the manual as a whole, but what is especially interesting is that most usages are associated with particular section types. Thus, to accurately interpret (or correctly produce) the material in parentheses, one must know which section one is in.

6.2.2 Labeling parentheticals

T.O. Section: Required Conditions

In the "Required Conditions" section, each condition which must hold before executing the primary maintenance task is named, and then has its procedure code (a number or "General Maintenance") following it in parentheses, as shown in examples (15) and (16). This allows the technician to easily find the relevant description of the repair manual without having to use the index, since the pages are labeled with the procedure codes. These procedure codes are sometimes cross-referenced in the repair steps as well, cf. example (21) below. This type of parenthetical could be classified as alternative-text type, with the code number as an alternative (and more precise way of referring to a particular set of maintenance instructions).

(15) Aircraft safe for maintenance *(JG10-30-01)*

(16) Access panel 3416 removed *(General Maintenance)*

T.O. Section: Personnel Recommended

The "Personnel Recommended" section uses parenthesized labels in two different ways. Each technician is listed along with a brief description of his/her role. If more than one technician is required for the task, the location of each participant is specified in parentheses after this description. (If there is only one technician, his/her location is obvious from the task.) They can be thought of as context-restricting in the sense that some technicians may already know where they needed to be to perform their assigned task, in which case the parenthetical information can be ignored.

(17) Technician A performs removal and installation *(access panel 3428)*.

(18) Technician C assists in checkout *(forward cockpit)*.

(19) Technician D acts as refueling supervisor *(between aircraft and fuel truck in full view of refueling operation)*.

In addition, the "Personnel" section associates a label with each technician, starting with "A" and assigning letters in order as needed ("A", "C" and "D" are shown in the examples above). For the rest of the repair specification, each step uses these labels to designate who should perform that step. In the examples below, Technician A performs step 2.1, and then Tech. B performs step 3.

(20) 2.1. *(A)* Install leak check panel on access area 3430 using 28 washers and 28 bolts.

(21) 3. *(B)* Connect hydraulic test stand to system A. (General Maintenance)

T.O. Section: Results

If a "Result" is specified for some group of sub-steps, a status code is sometimes shown, as in (22) and (23).

(22) RESULT: No leakage allowed. *(28-23-FD)*

(23) RESULT: (A) Ground test panel FUEL PUMP NO. 1 to 5 and FFP advisory lights come on (access door 3308). *(28-23-DD, 28-23-DE, 28-23-DF)*

It is not entirely clear to me whether these codes refer to information in other documents about these states, or whether they are labels assigned in this T.O. for reference by other documents. If the later, this is an especially interesting use of parentheses, since it creates a label rather than simply referring to one (as in the Required Conditions section).

6.2.3 T.O. Section: Maintenance Enumeration

In the maintenance descriptions themselves, parentheticals are used as already noted, to specify which technician performs each step, but also to give part numbers, part names and alternative descriptions of states. For the most part, these uses can be categorized into Nunberg's categories of alternatives, but there are also cases which are better characterized as elaborations of descriptions than complete alternatives.

Alternatives

Parentheses are frequently used for alternative descriptions of entities or situations. Sometimes alternative descriptions of indicator positions are given, as in (24), (25) and (26). In task where *in/outboard* or *open/closed* are used, either all uses are parenthetically or none are.

(24) 5. Position FFP control valve handle in down *(closed)* position.

(25) 13. Position FFP control valve handle in up *(open)* position.

(26) RESULT: 6.c.(B) Engine fuel shutoff valve actuator indicator does not move off full CLOSED *(inboard)* position once in full CLOSED *(inboard)* position. (28-23-FH)

Instruction (27) is slightly different, in that it is an alternate description of an event rather than an entity, i.e. the *drain* event. It is also unusual for this corpus in having a full matrix clause as a parenthetical.

(27) 3.(A) Position waste fluid container under receptacle.
4.(A) Connect nozzle to receptacle and drain residual fuel. *(Approximately 2 gallons will drain.)*

Context restricting

Optionality in certain aircraft configurations is sometimes specified with context-restricting parentheticals, as in (28)-(30). Example (29) is taken from the very beginning of a task description. In (30) the *four washers* have just been mentioned in the preceding step, but no mention was made of washers in excess of those four (which, presumably, are the minimum required for the task).

(28) 5. Position one fire extinguisher near aircraft servicing connection point and one fire extinguisher upwind and near generator set *(if operating)*.

(29) NOTE ... Steps 1 through 4 may be omitted if reducer and elbow were not removed from *(or have already been installed on)* pressure switch.

(30) NOTE: Stop bolts shall be adjusted by distributing a total of four washers under bolthead and nut. Up to four washers may be used under head of stop bolt. Remaining washers *(if any)* shall be placed under nut....

In (31), the first two instructions are conditional on the presence of the *centerline tank*, and instruction 30 is likewise only relevant under the same circumstances. Strangely, step 29 which instructs the technician to remove the *shortening plug* which was also conditionally installed (in step 2) does not have the conditional clause in parentheses. This lack of parallelism is striking, given the consistency of the other uses of parentheticals.

(31) NOTE: If centerline tank is installed, omit steps 1 and 2.
1. (B) Remove protective cap or pylon connector from receptacle (J236).
2. (B) Install shorting plug on receptacle (J236).
⋮
29. (B) Remove shorting plug from receptacle (J236), if installed.
30. (B) Install protective cap or pylon connector *(if removed)*.

Elaborating descriptions

Parentheses are also used to elaborate descriptions. In example (32) the parenthetical information tells us where in the tables to find the relevant information, while (33) and (34) specify the positions equipment should be in at the end of the action. In (35), there are an unspecified number of washers to be removed (although it is hard to know why they don't just say *all washers*); the second parenthetical bit is ambiguous to me as a naive reader–it may mean that there are two places where things need to be removed, or that the bolts may not need to be removed at all.

(32) 8. Operate hydraulic test stand fill pump until quantity gage indicates in accordance with tables 6 and/or 7 *(depressurized column)*.

(33) NOTE: Coupling remover shall be installed in open position *(lever all the way forward)*.
12. (A) Install coupling remover on external vent and pressurization valve and external tank and pressure tube.

(34) 35. (A) Connect drain tube to FFP and drain fitting. Torque to 72-78 inch-pounds *(two places)*.

(35) 18. (A) Remove two nuts, washers *(as required)*, and two bolts from aft slipway wall support *(two places if required)*.

Often the part number for a piece of equipment is listed parenthetically, as in (36) and (37). They are present when the part is first mentioned, and are used with parts which are likely come in a variety of hard-to-distinguish sizes, e.g. washers and bolts. The part number could be looked at as an alternative way of describing the washer, but could also be seen as adding information about the washer in a way that will help the technician find the right one.

(36) 6. (A) Install lower inboard bolt, washer *(AN960PD416)* under bolthead, sealing washer, washer *(AN960PD416)* under nut, and nut. Do not torque.

(37) (A) Lubricate packing *(M25988/1-904)* and install on union.

Parallel instruction sets

In the title of the procedure, parentheticals are only used when the procedure applies to pairs of components, e.g. *Flow Divider Outlet Check Valve, 2823FV17 (Right) or 2823FV5 (Left), Removal and Installation.* These are a bit like to the interleaved dialogue in examples (12)-(14) above, but can more easily be assimilated to the class of context-restricting parentheticals. What is different is that a particular alternative context (*left*, say) is salient throughout a large span of text. This results in interestingly parallel texts with repairs applying to a set of components (e.g. right, left, front or aft), and the T.O. written to handle all cases. In cases where *left* or *right* is specified parenthetically after access panel numbers or part numbers/names, it is because the procedure is the same for both sides, modulo these specifics. The sentence in (38) is typical of how this is notated at the start of the instruction, and (39) shows show the repair is specified.

(38) NOTE: Removal procedures for left, right, and aft scavenge pumps are similar; therefore, only right procedures are given, except where noted.

(39) 1. (A) Remove access cover 5419 *(left)* or 6420 *(right)*. (General Maintenance)

6.2.4 Non-genre-specific uses

There are a few other uses of parentheses that are not specific to this genre:

- Cross references to other sections of the document, or to tables or figures

  (40) 5. Illustrations in this job guide include a location view of the equipment on which the task is being performed and key numbers that are numerically identical to the task steps *(figure 1)*. When a part/component within an illustration is referenced in more than one step in the procedure, the illustration will be keyed to each of the steps *(figure 1, key numbers 4 and 5)*.

- Abbreviations and acronyms

  (41) Forms 22 will be forwarded to the F-16 Central Technical Order Control Unit *(CTOCU)* for processing. Address is as follows:

- Citations

  (42) WARNING: This document contains technical data whose export is restricted by the Arms Export Control Act *(Title 22, U.S.C. Sec. 2751 et seq)* or the Export Administrative Act of 1979 as amended *(Title 50, U.S.C. app. 2401 et seq)*. Violations of these export laws are subject to severe criminal penalties. Disseminate in accordance with provisions of AFR 80-34.

6.3 Parentheses in academic papers

As with the technical orders, academic papers contain some uses of parentheses which are fairly generic and some which are idiosyncratic. In looking at four computational linguistics research papers, we again find both of Nunberg's classes, alternative text parentheticals and context restricting ones.

6.3.1 Alternative texts

Some parentheticals present simple alternatives:

(43) ...and a part-of-speech tagger *(also referred to as simply 'tagger')*....

(44) As it turned out, the 78 categories generated in this stage of the translation plus a few later additions cover *(account for the syntactic phenomena handled by)* just over 400 of the 566 XTAG trees.

There are also identifiable sub-classes of alternate texts. Some give examples or instantiations of the antecedent:

(45) It consists of approximately 317,000 inflected items, along with their root forms and inflectional information *(such as case, number, tense).*

(46) Each entry in the lexicon is restricted via the FS field to only a certain form of the auxiliary verb *(present, past, ppart, etc)*....

Others give more specific enumerations:

(47) On translation, this set collapses into an active, two passive *(with and without by-phrase)* and one gerund category.

(48) There are 19 different frames that the verbs can select, including transitive, intransitive, sentential complement, sentential subject, verb particle constructions *(transitive and intransitive)*...

There are also "in-case-you're-interested" parentheticals:

(49) The Proteus Project at New York University is developing the Comlex Syntactic Dictionary from scratch for release as one of the lexical resources in COMLEX *(available through the Linguistic Data Consortium).*

(50) Each lexical entry contains the root form ("INDEX"), all the categories the root form selects ("CAT"), and, optionally, features associated with that lexical item. *(We also allow idioms to be entered in the lexicon as single units.)*

(51) As a result, the CCG syntactic lexicon contains multiple categories for each wh− word, but only one for each extraction position irrespective of the valency of the verb, as opposed to LTAG which has a tree for each extraction possibility for each verb. *(There are approximately 35 extraction categories in the CCG as compared with 211 extraction trees in the LTAG.)*

6.3.2 Context restricting

There is a particularly interesting type of context restricting parenthetical which I expect is common in other types of persuasive writing as well. I call them "precautionary" parentheticals, because the context they are addressing is that of the reader who is at least reading critically and possibly is even antagonistic with regard to the material being presented. The parentheticals are an attempt to head off any anticipated criticisms in advance.

(52) Various machine-readable versions of monolingual and bilingual dictionaries are more or less readily available for NLP research and development (eg. from Longman, Collins, Oxford University Press, Larousse, Bibliograf etc.), and provide *(more or less explicitly and comprehensively)* morphological, syntactic, collocational and semantic category information.

(53) There are over 8100 verbs *(not including auxiliary verbs)* that make up almost 9000 entries in the database.

(54) The second is to make the parser as efficient as possible, and to produce the derivations in rank order based on certain preference heuristics *(assuming that no semantic information is available)*.

(55) This parse strategy allows us to more easily incorporate statistical information about the likelihood of two categories combining into the parser so as to minimize syntactic and derivational ambiguity. *(While a space-efficient CCG algorithm has been proposed by Vijay-Shanker and Weir (1994), it is incompatible with our feature system. In brief, feature co-indexation introduces dependencies among constituents of complex categories which are incompatible with the use of pointers to maximize efficiency in category representation.)*

6.3.3 Other uses

References and cross-references

(56) The distinction between competence and performance popularized by Chomsky *(1965)* is fundamental to modern linguistics.

(57) While researchers in NLU have made great progress in extracting lexical information automatically from machine-readable versions of dictionaries *(eg. (Wilks, Slator, and Guthrie, 1996; Richardson, 1997))*

Abbreviations

(58) In this paper I look at this issue in relation to one particular NLP task, Information Extraction *(hereafter IE)*, and one subtask for which both lexical and general knowledge are required, Word Sense Disambiguation *(WSD)*.

(59) ...structures consisting of a relative clause *(RC)* and a sentential complement *(SC)*...

6.4 Discussion

The two texts we have looked at here, fighter plane maintenance manuals and computational linguistics research articles, are highly dissimilar. One is instructional, the other persuasive; one holds clarity at a premium, the other esteems impenetrability; one hopes for an attentive readership, the other anticipates contentious readers. Not surprisingly, we find that some of the ways they use parentheses to further these ends are rather different. But Nunberg's two classes of parentheticals are quite descriptively adequate, perhaps in part because they are so general. Both genre of text show that a finer-grained classification is useful. In addition, there are uses which do not fit well into either class. In the first section, I gave several examples with quoted speech in parentheses (examples (12)-(14)), which cannot be characterized in any simple way as alternative-introducing or context-restricting. In the F16 corpus, there are the parentheticals described in Section 6.2.3 which further elaborate descriptions rather than providing alternative descriptions.

Chapter 7

Conclusion

My aim in undertaking this research was to find out how feasible it was to handle a sizable core of punctuation phenomena at the level of the sentence grammar, without either adversely impacting the existing grammar or deriving analyses which would be incompatible with later levels of processing, in particular at the discourse level. The major contributions of this work are (1) the extension of an already large English grammar to cover a wider range of texts "as is" and (2) a detailed analysis of three sets of constructions where punctuation plays a crucial role.

My results confirm that punctuation can be used in analyzing sentences to increase the coverage of the grammar, reduce the ambiguity of certain word sequences and facilitate discourse-level processing of the texts. I have implemented quite an extensive grammar for punctuation which has been incorporated into the XTAG English Grammar, and found that the punctuation rules do indeed improve the coverage of the existing grammar with no negative impact on the rest of the grammar. In analyzing the class of reported speech constructions, I have shown that they have both sentence grammar and discourse grammar realizations; how they use punctuation is one the crucial distinguishing features of the two classes. I furthermore show that the LTAG analysis of the text adjunct variant is fully compatible with a discourse grammar of the sort proposed by Webber and Joshi [1998]. Consideration of the role of punctuation in a class of constructions which superficially resemble NP-appositives finds that those constructions with NP-level modification and punctuation are non-restrictive in meaning, while those which are either at the N level or do not involve punctuation are in a restrictive relationship. Non-restrictive modifiers have been argued by Safir [1986] among others to be processed at a later point than restrictive modifiers, which would place them squarely on the border between the sentence grammar and the discourse grammar. Also, punctuation marks serve here to delimit noun sequences which would otherwise be highly ambiguous. Finally, punctuation gives us some of the additional constraints needed to license constructions which would previously have been too unconstrained. The parentheticals I have discussed are just such a case; they occur in a wide-range of positions and can be

of any syntactic category. In general, we do not want to have arbitrary constituents freely licensed at arbitrary locations in a sentence, but we can license them in the presence of parentheses.

7.1 Future work

As with all aspects of grammar development, there is always more data to be looked at, yielding more constructions to be handled. One particularly challenging case is the use of ellipsis marks (...). Some instances are easily handled, as in (1), but other instances occur in the context of syntactic ellipsis and are more challenging. Example (2) has ellipses between two adjectival phrases and (3) has them between a noun phrase and a complete clause. The latter cases, where the ellipsis marks indicated where quoted material has been reduced, are discussed in grammar books; there is no explicit discussion of uses like that in (1).

(1) You look around at professional ballplayers or accountants . . . and nobody blinks an eye. [wsj0049]

(2) Mitsubishi's investment in Free State is "very small . . . less than $4 million," Mr. Wakui says. [wsj0083]

(3) Eighty-three years ago, William James wrote to H.G. Wells: "The moral flabbiness born of the exclusive worship of the bitch goddess success . . . that, with the squalid cash interpretation put on the word success, is our national disease." [wsj0937]

More work remains to be done on the joint work with Hockey discussed in 1.4.2. We need to complete the acoustic analysis of the data we have already collected, and also plan to do a second phase of the experiment. We think there are two sets of texts which will help us answer the questions that interest us about the connections between punctuation and prosody. The first is texts which are written to be read silently, in this case the Wall Street Journal news we used in the first phase of the experiment. The second is texts written to be read aloud, for which we plan to use news radio scripts. We feel it is important to look at both types of texts, to try get at the issue of whether writers do anything different when they are writing things intended to be read aloud.

I am very much interested in seeing how well the analysis presented here would connect with a discourse grammar like that of [Webber and Joshi1998], and in seeing such an integrated system implemented. Integrating the two accounts would serve to assess the validity of both, in that we would learn whether the approach to constructing a discourse grammar was a viable one, and whether the LTAG sentence grammar really provided the discourse grammar with the right sorts of information.

It might also be a way of accomplishing a more general evaluation of the punctuation rules which I was unable to execute in parsing with the entire large grammar. Additionally, there seem to be interesting parallels between the way that punctuation marks the boundaries information units and the way this can be done with prosody. Bierner et al. [1998] look at integrating a treatment of prosodic cues about information structure into and LTAG, and it would be very interesting to look more closely at how that work could be connected to the analysis of punctuation presented here.

Since the XTAG grammar only handles syntactic aspects of analysis, it does not provide the ideal environment for evaluating the semantic and pragmatic aspects of punctuation. However, the generation system presented in [Stone and Doran1996; Stone and Doran1997] provides an very promising environment for exploring these issues. SPUD (**S**entence **P**lanning **U**sing **D**escription) is a sentence planning system which uses LTAG as the grammatical specification, and description as the underlying paradigm. SPUD uses flat semantic representations and a rich discourse model which provide information on definiteness, discourse status of entities and propositions, etc. to generate contextually appropriate sentences. It would be a natural extension to incorporate punctuation into the system, allowing it to generate sentences with contextually appropriate punctuation. Knowledge about definiteness, perspective (e.g. with quoted speech) and the discourse status of entities (relative clauses) could all be brought to bear in choosing suitable punctuation.

Bibliography

[Adams et al.1992] B. Adams, C. Clifton, and D. Mitchell. 1992. Syntactic guidance in sentence processing. Poster presented at Psychonomic Society.

[Baldwin et al.1997] Breckenridge Baldwin, Christine Doran, Jeffrey Reynar, Michael Niv, B. Srinivas, and Mark Wasson. 1997. EAGLE: An Extensible Architecture for General Linguistic Engineering. In *Proceedings of RIAO97*, Montreal.

[Bamgboṣe1986] Ayọ Bamgboṣe. 1986. Reported Speech in Yoruba. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Beeferman et al.1998] Doug Beeferman, Adam Berger, and John Lafferty. 1998. Cyberpunc: A Lightweight Punctuation Annotation System for Speech. To appear at ICASSP '98.

[Bierner et al.1998] Gann Bierner, Anoop Sarkar, and Aravind Joshi. 1998. Deriving Information Structure from Prosodically Marked Text with Lexicalized Tree Adjoining Grammars. Manuscript, University of Pennsylvania.

[Birner1992] Betty Birner. 1992. *The Discourse Function of Inversion in English*. Ph.D. thesis, Northwestern University.

[Bolinger1972] Dwight Bolinger. 1972. *That's That*. Mouton, The Hague.

[Briscoe and Carroll1995] Ted Briscoe and John Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the Fourth International Workshop on Parsing Technologies (IWPT'95)*, pages 48–57, Prague/Karlovy Vary, Czech Republic, September.

[Briscoe1994] Ted Briscoe. 1994. Parsing (with) Punctuation etc. Technical Report MLTT-TR-002, Rank Xerox Research Centre, Grenoble, France.

[Briscoe1996] Ted Briscoe. 1996. The Syntax and Semantics of Punctuation and its Use in Interpretation. In *Proceedings of the SIGPARSE96*, Santa Cruz, California, June.

[Chafe1988] Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written Communication*, 5(4):395–426, October.

[Chandrasekar and Srinivas1997] R. Chandrasekar and B. Srinivas. 1997. Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10:183–190.

[Chandrasekar et al.1996] R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark, August.

[Chi1982] 1982. Chicago Manual of Style. The University of Chicago Press. 13th Edition.

[Clifton1995] Charles Clifton, Jr. 1995. Thematic Roles in Sentence Parsing. In Henderson et al., editor, *Reading and Language Processing*. Laurence Erlbaum Associates.

[Collins and Branigan1996] Chris Collins and Phil Branigan. 1996. Quotative Inversion. *Natural Language and Linguistic Theory*, 14.

[Coulmas1986] Florian Coulmas. 1986. *Direct and Indirect Speech*. Mouton de Gruyter.

[Dale1991] 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 110–20, T. U. Berlin.

[Davidson1984] Donald Davidson. 1984. On Saying That. In *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford.

[de Beaugrande1984] Robert de Beaugrande. 1984. *Text Production: Toward a Science of Composition*, volume XI of *Advances in Discourse Processes*. ABLEX Publishing, Norwood,NJ.

[Delorme and Dougherty1972] Evelyn Delorme and Ray Dougherty. 1972. Appositive NP Constructions. *Foundations of Language*, 8:2–29.

[Doran and Hockey1998] Christine Doran and Beth Ann Hockey. 1998. When Do Punctuation and Prosody Coincide? Manuscript, University of Pennsylvania.

[Doran et al.1994] Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. XTAG System - A Wide Coverage Grammar for English. In *Proceedings of the* 17^{th} *International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan, August.

[Doran et al.1997] Christine Doran, Michael Niv, Breckenridge Baldwin, Jeffrey Reynar, and B. Srinivas. 1997. Mother of Perl: A Multi-tier Pattern Description Language. In *Proceedings of the International Workshop on Lexically Driven Information Extraction*, Frascati, Italy, July.

[Doran1994] Christine Doran. 1994. Parsing Multi-Clausal Sentences. Unpublished manuscript.

[Emonds1973] Joseph Emonds. 1973. Parenthetical Clauses. In *You Take the High Node and I'll Take the Low Node: Papers from the Comparative Syntax Festival, The Differences between Main and Subordinate Clauses*. Chicago Linguistic Society, Chicago.

[Emonds1976] Joseph Emonds. 1976. *A Transformational Approach to English Syntax*. Academic Press, New York.

[Emonds1979] Joseph Emonds. 1979. Appositive Relatives Have No Properties. *Linguistic Inquiry*, 10(2):211–243.

[Espinal1991] M. Teresa Espinal. 1991. The Representation of Disjunct Constituents. *Language*, 67(4):726–761.

[Fónagy1986] Ivan Fónagy. 1986. Reported Speech in French and Hungarian. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Frank1992] Robert Frank. 1992. *Syntactic locality and Tree Adjoining Grammar: grammatical, acquisition and processing perspectives*. Ph.D. thesis, University of Pennsylvania,IRCS-92-47.

[Gardent and Webber1998] Clair Gardent and Bonnie Webber. 1998. Underspecification in Discourse Structure and Semantics. Manuscript, University of Pennsylvania.

[Gardent1997] Clair Gardent. 1997. Discourse TAG. Manuscript, University of Saarbrücken.

[Haberland1986] Hartmut Haberland. 1986. Reported Speech in Danish. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Hand1991] Michael Hand. 1991. On Saying That Again. *Linguistics and Philosophy*, 14:349–365.

[Hand1993] Michael Hand. 1993. Parataxis and Parentheticals. *Linguistics and Philosophy*, 16:495–507.

[Hewitt and Crisp1986] B.G. Hewitt and S. R. Crisp. 1986. Speech reporting in the Caucasus. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Hill and Murray1999] Robin L. Hill and Wayne S. Murray. 1999. Commas and Spaces: The Point of Punctuation. Poster presented at the 11th Annual CUNY Conference on Human Sentence Processing.

[Hollenbach1983] Barbara Hollenbach. 1983. Apposition and X-Bar Rules. In *Exploring Language: Linguistic Heresies From the Desert*, volume 4 of *Coyote Papers, Working Papers in Linguistics from A-Z*, University of Arizona, Tucson.

[Jesperson1966] Otto Jesperson. 1966. *Essentials of English Grammar*. University of Alabama Press.

[Jones1994] Bernard E. M. Jones. 1994. Exploring the Role of Punctuation in Parsing Natural Text. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan, August.

[Jones1996a] Bernard Jones. 1996a. Towards a Syntactic Account of Punctuation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, August.

[Jones1996b] Bernard Jones. 1996b. *What's the Point? A (Computational) Theory of Punctuation*. Ph.D. thesis, University of Edinburgh.

[Joshi et al.1975] Aravind K. Joshi, L. Levy, and M. Takahashi. 1975. Tree Adjunct Grammars. *Journal of Computer and System Sciences*.

[Joshi1985] Aravind K. Joshi. 1985. Tree Adjoining Grammars: How much context Sensitivity is required to provide a reasonable structural description. In D. Dowty, I. Karttunen, and A. Zwicky, editors, *Natural Language Parsing*, pages 206–250. Cambridge University Press, Cambridge, U.K.

[Kiefer1986] Ferenc Kiefer. 1986. Some semantic aspects of indirect speech in hungarian. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Kroch and Joshi1985] Anthony S. Kroch and Aravind K. Joshi. 1985. The Linguistic Relevance of Tree Adjoining Grammars. Technical Report MS-CIS-85-16, Department of Computer and Information Science, University of Pennsylvania.

[Larson1985] Richard K. Larson. 1985. Bare-NP Adverbs. *Linguistic Inquiry*, 16(4):595–621.

[Lasersohn1986] Peter Lasersohn. 1986. The Semantics of Appositive and Pseudo-Appositive NP's. In *Proceedings of ESCOL '86*.

[Lee1995] Sherman Lee. 1995. A Syntax and Semantics for Text Grammar. Master's thesis, Cambridge University.

[LePore and Loewer1989] Ernest LePore and Barry Loewer. 1989. You Can Say *That* Again. In *Midwest Studies in Philosophy, XIV*.

[Li1986] Charles N. Li. 1986. Direct and Indirect Speech: A Functional Study. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Liberman1975] Mark Liberman. 1975. *The Intonation System of English.* Ph.D. thesis, MIT, Boston, MA.

[Massamba1986] David P. B. Massamba. 1986. Reported speech in Swahili. In Florian Coulmas, editor, *Direct and Indirect Speech.* Mouton de Gruyter.

[Maynare1986] Senko K. Maynare. 1986. The particle -o and content-oriented indirect speech in japanese. In Florian Coulmas, editor, *Direct and Indirect Speech.* Mouton de Gruyter.

[McCawley1982] James D. McCawley. 1982. Parentheticals and Discontinuous Constituent Structure. *Linguistic Inquiry*, 13(1):91–106.

[McCawley1995] James D. McCawley. 1995. An overview of "appositive" constructions. In *Proceedings of ESCOL '95.*

[Meyer1987] Charles F. Meyer. 1987. *A Linguistic Study of American Punctuation*, volume 5 of *American University Studies, Series XIII - Linguistics.* Peter Lang, New York.

[Meyer1992] Charles F. Meyer. 1992. *Apposition in Contemporary English.* Studies in English Language. Cambridge University Press, Cambridge.

[Mugarza1992] Miren Itziar Laka Mugarza. 1992. *Negation in Syntax: On the Nature of Functional Categories and Projections.* Ph.D. thesis, Massachusetts Institute of Technology.

[Nunberg1990] Geoffrey Nunberg. 1990. *The Linguistics of Punctuation.* CSLI Lecture Notes, No. 18. Center for the Study of Language and Information, Stanford.

[Parkes1993] M. B. Parkes. 1993. *Pause and Effect: An Introduction to the History of Punctuation in the West.* University of California Press.

[Penhallurick1984] John Penhallurick. 1984. Full-Verb Inversion in English. *Australian Journal of Linguistics*, 4:33–56.

[Pierrehumbert1980] Janet Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, MIT, Boston, MA.

[Prevost and Steedman1993] Scott Prevost and Mark Steedman. 1993. Generating contextually appropriate intonation. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics.*

[Prince1986] Ellen Prince. 1986. On the Syntactic Marking of Presupposed Open Propositions. In *Proceedings of the 22nd Annual Meeting of the Chicago Linguistic Society*, pages 208–222, Chicago. CLS.

[Quirk et al.1985] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

[Reynar and Ratnaparkhi1997] Jeff Reynar and Adwait Ratnaparkhi. 1997. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., April.

[Rice1986] Keren D. Rice. 1986. Some remarks on direct and indirect speech in Slave (Northern Athapaskan). In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Ross1973] John Robert Ross. 1973. Slifting. In *The Formal Analysis of Natural Languages: Proceedings of the First International Conference*. Mouton, The Hague.

[Sabin1996] William A. Sabin. 1996. *The Gregg Reference Manual*. Glencoe/McGraw Hill, 8^{th} edition.

[Safir1986] Ken Safir. 1986. Relative Clauses in a Theory of Binding and Levels. *Linguistic Inquiry*, 17(4).

[Sampson1992] Geoffrey Sampson. 1992. Review of Geoff Nunberg: *The Linguistics of Punctuation. Linguistics*, 30(2):467–475.

[Sampson1995] Geoffrey Sampson. 1995. *English for the Computer:The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford.

[Sarkar and Joshi1996] Anoop Sarkar and Aravind Joshi. 1996. Coordination in Tree Adjoining Grammars: Formalization and Implementation. In *Proceedings of the* 18^{th} *International Conference on Computational Linguistics (COLING '94)*, Copenhagen, Denmark, August.

[Say and Akman1995] Bilge Say and Varol Akman. 1995. Dashes as Cues to Discourse Structure. Manuscript, Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey.

[Say and Akman1996a] Bilge Say and Varol Akman. 1996a. An Information-Based Approach to Punctuation. Manuscript, Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey. Available at `http://www.cs.bilkent.edu.tr/~say/bilge.html`.

[Say and Akman1996b] Bilge Say and Varol Akman. 1996b. An Information-Based Treatment of Punctuation. In *IInd ICML (International Conference on Mathematical Linguistics) Abstracts*, number 7/96 in Technical Report, pages 93–95, Tarragona, Spain.

[Schabes et al.1988] Yves Schabes, Anne Abeillé, and Aravind K. Joshi. 1988. Parsing strategies with 'lexicalized' grammars: Application to Tree Adjoining Grammars. In *Proceedings of the* 12^{th} *International Conference on Computational Linguistics (COLING'88)*, Budapest, Hungary, August.

[Schmidt1995] Mark Schmidt. 1995. *Acoustic Correlates of Encoded Prosody in Written Conversation*. Ph.D. thesis, The University of Edinburgh.

[Scott and de Souza1990] Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the Message Across in RST-based Text Generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, Cognitive Science Series. Academic Press.

[Segal and Speas1986] Gabriel Segal and Margaret Speas. 1986. On Saying *That*. *Mind and Language*, 1(2):124–132.

[Sevald and Trueswell1997] Christine A. Sevald and John C. Trueswell. 1997. Speakers Cooperate with Listeners: Prosody to Help Sidestep the Garden Path. Poster presented at the 10th Annual Meeting of the CUNY Conference on Sentence Processing.

[Shieber and Schabes1990] Stuart Shieber and Yves Schabes. 1990. Synchronous Tree Adjoining Grammars. In *Proceedings of the* 13^{th} *International Conference on Computational Linguistics (COLING'90)*, Helsinki, Finland.

[Smith1969] Carlotta Smith. 1969. Determiners and Relative Clauses in a Generative Grammar of English. In D. Reibel and S. Schane, editors, *Modern Studies in English*, pages 247–263.

[Srinivas1997] B. Srinivas. 1997. *Complexity of Lexical Descriptions and its Relevance to Partial Parsing*. Ph.D. thesis, Department of Computer and Information Sciences, University of Pennsylvania.

[Stone and Doran1996] Matthew Stone and Christine Doran. 1996. Paying Heed to Collocations. In *Proceedings of the Eighth International Workshop on Natural Language Generation (INLG-96)*, Herstmonceux, Sussex, UK.

[Stone and Doran1997] Matthew Stone and Christine Doran. 1997. Sentence Planning as Description using Tree Adjoining Grammar. In *Proceedings of the* 35^{th} *Annual Meeting of the Association for Computational Linguistics and the* 8^{th} *Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL97)*, Madrid.

[Tannen1986] Deborah Tannen. 1986. Introducing constructed dialogue in Greek and American conversation and Literary Narrative. In Florian Coulmas, editor, *Direct and Indirect Speech*. Mouton de Gruyter.

[Thomson and Mulac1991a] Sandra A. Thomson and Anthony Mulac. 1991a. A Quantitative Perspective on the Grammaticization of Epistemic Parentheticals. In E. Traugott and B. Heine, editors, *Approaches to Grammaticization: Vol II, Focus on Types of Grammatical Markers*. John Benjamins.

[Thomson and Mulac1991b] Sandra A. Thomson and Anthony Mulac. 1991b. The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics*, 15:237–251.

[Trueswell and Tanenhaus1994] John Trueswell and Mike Tanenhaus. 1994. Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In K. Rayner C. Clifton and L. Frazier, editors, *Perspectives on sentence processing*. Lawrence Erlbaum Associates, Hillsdale, NJ.

[Trueswell1993] John Trueswell. 1993. *The Use of Verb-Based Subcategorization and Thematic Role Information in Sentence Processing*. Ph.D. thesis, University of Rochester.

[Urmson1963] J. O. Urmson. 1963. Parenthetical Verbs. In Charles E. Caton, editor, *Philosophy and Ordinary Language*. University of Illinois Press.

[Vijay-Shanker and Joshi1991] K. Vijay-Shanker and Aravind K. Joshi. 1991. Unification Based Tree Adjoining Grammars. In J. Wedekind, editor, *Unification-based Grammars*. MIT Press, Cambridge, Massachusetts.

[Ward and Prince1991] Gregory Ward and Ellen Prince. 1991. On the topicalization of indefinite NPs. *Journal of Pragmatics*, 15(8):338–351.

[Ward1985] Gregory Ward. 1985. *The Semantics and Pragmatics of Preposing*. Ph.D. thesis, University of Pennsylvania. Published 1988 by Garland.

[Webber and Joshi1998] Bonnie Webber and Aravind Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. Manuscript, University of Pennsylvania.

[White1995] Michael White. 1995. Presenting Punctuation. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, pages 107–125, Leiden, The Netherlands.

[XTAG-Group1995] The XTAG-Group. 1995. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 95-03, University of Pennsylvania.