PROFILING FOR BELIEF ACQUISITION FROM REPORTED SPEECH

Monia Doandes

A thesis
in
The Department
of
Computer Science

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Computer Science
Concordia University
Montréal, Québec, Canada

April 2003

Abstract

Profiling for Belief Acquisition from Reported Speech

Monia Doandes

We are in the Internet era.

Now, more than ever, there is an excessive amount of documents available online, making it harder and harder to find the ones that are relevant to our searches. The new systems, which use belief models to improve the understanding of what the user really wants, need a practical text representation model that can make the extraction of beliefs easier. Profiling is an appropriate method for creating such a text representation.

In this thesis we develop the details for a process of building basic profiles from *reported speech in newspaper articles.* This process recognizes and extracts reported speech, and determines the internal structure of reported speech clauses, fitting it within basic profile frames. In order to accomplish this, three building blocks are used: *recognition of reporting verbs, punctuation analysis* and *the structure of the reported speech.*

We implemented a fully automated system which extracts basic profiles and analyzes them, differentiating between circumstantial information and primary information. Inside the primary information, a differentiation is made between direct quotes and paraphrases. The source, reporting verb and other circumstantial information are extracted. The reporting verb is analyzed, determining its tense, aspect, modality, mood, voice and polarity. The semantic dimensions of the reporting verb are also analyzed, paving the path for a future belief analysis based on reporting verbs.

Acknowledgments

A heartfelt thanks to Dr. Sabine Bergler, my thesis advisor, who guided me from my very first baby steps in the field of Computational Linguistics and who has been very supportive and patient throughout my learning process. She is a wonderful teacher, adviser and person who, despite her extremely tight schedule, has always found the time to listen to me and to help me stay on the good path.

A grand Thank You to Dr. Bill Atwood, whose words of encouragement have helped me go through hard times for my confidence.

Big thanks to René Witte for sharing his experience on what it means to be a graduate student and for patiently answering all my questions and to Frank Rudzicz for being my ERS code guide and for his help and kindness. Thank you to the whole CLaC group without whom this thesis would have not been possible.

Special thanks go to Bari, who has taken care of me all this time, who has been next to me when angst was setting in and helped me cope with all the stress, sharing his words of wisdom with me. He has been very supportive and patient, despite my morosity, depression and the many times when I seemed to not see anything around me but the computer screen.

I wish to thank my parents for pushing me to go through this and for being there every time my motivation was taking a hit. I thank my mother for transmitting me the love for foreign languages and my father for being the one who was always there for me when my engines needed fuel.

Caroline is a great friend who helped me tremendously and who shared her home with me on many occasions. Thank you ever so much Caroline!

My sincere thanks to all my friends at Concordia University who have made my stay and make my visits to Montreal enjoyable. The wonderful people at the International Students Office have always helped me in many ways and I would have not been able to finish this thesis withtout them. Special thanks to Amin, Angie, Claire, Claudette, Edwina, Galina, Halina, Hirut, Joel, Larry, The Three Michaels, Monica, Pat, Paul, Raymond, Stan, Stephanie and Stuart. You all have a special place in my heart.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are now in the Internet era. Every year, Bill Cheswick's Map of the Internet[1] becomes more and more complex, showing how servers all over the world form a world-wide web of networks, interconnecting and exchanging information. This means that now, more than ever, there is an excessive amount of digital data available. If we are to look just at text data, the Google crawler[2] is searching 3,083,324,652 web pages [www.google.com, March 2003], and the crawler does not consider all the data from dynamic web pages, does not consider web pages that do not have enough links to them or that have an anti-robot request, it might have problems considering data inside frames and so on. In their News section, Google searches over 4,500 selected online newspapers which offer their articles free of charge and through searchable web pages. That is a huge amount of data to search, and often the relevant information is buried within a pile of irrelevant data. The current search techniques are all based on keywords, and irrelevant documents which contain just a reference to the word or phrase searched will come up in the search, clogging the list of results the user is looking at.

Some systems have started using models of the user to improve the search results, but these models are solely based on either feedback from the user for each document in the training phase, either on browsing patterns. Browsing patterns are not very accurate, since the user cannot know fully what a page really contains before actually looking at the page. Also, usually these models do not have processes in place to treat the change in the user's points of view, and their matching strategy is still based on

[1] http://research.lumeta.com/ches/map/

[2] http://www.google.com/

keywords [Gerard, 2000].

For a search to be truly successful, a system should have a representation model of the user's beliefs, what is she looking for and why is she looking for it. If the user reads an ad about the XYZ diet plan and wants to know about the drawbacks of the plan, she would want to see opposing points of view to the diet. Similarly, a user might want to know the arguments on both sides of the war issue, with the parties clearly identified to ease the user's understanding of the argumentative. A model of the user's beliefs would make such searches more successful by being able to better determine which documents to display in the results list.

ViewGen [Ballim, 1992] is a belief maintenance system which implements such a user model for dialogue understanding. The ViewGen user representation model has been adapted by Gerard, who developed a system (Percolator) envisaging the users are readers of newspaper articles [Gerard, 2000]. Percolator dynamically simulates how readers acquire beliefs while reading newspaper articles, based on criteria such as the confidence of the reader in the reporter and the original source, the confidence of the reporter in the original source and the credibility of the information, and the beliefs that the user already has about a particular topic. This model can be used to predict which articles the reader would want to consult. Out of curiosity, I manually tried Gerard's model on research papers in order to find opposing points of view to the Discourse Representation Theory, and points of view about how it can be improved. I also used the help of opinion verbs such as *argue*, *propose* or *claim*. The writer of the article can be considered as the original source and the journal in which it has been published as criteria for the confidence in the reporter. The confidence of the reporter in the source cannot be established from reporting verbs in this case, but the fact that the paper has been published in a journal determines this criterion, its value depending on the editorial process the journal has. With these slight modifications, Percolator can also be used to determine the relevance of research papers.

Such systems need certain beliefs as input, and these can be extracted from the text. For newspaper articles, the beliefs of the reporter are encoded within the description of the source and the reporting verbs used [Bergler, 1992; Bergler, 1993; Bergler, 1995a]. If the reporter describes a source as "a trusted official", he is encoding the fact that he trusts the source. Likewise, if the reporter is using the verb

concede, the reported information can be trusted, since a source would not say something that is bad for them if that information is not true.

Bergler developed a method for representing newspaper articles to extract these types of beliefs [Bergler, 1992; Bergler, 1995a].

1.1 What is profiling?

Profiling is a method of representing the information in newspaper articles. A profile is a collection of all the information pertaining to a source in a newspaper article [Bergler, 1992]. A profile contains such information as the lexicalization of the source, the reporting verbs used, other circumstances of the original speech acts and all the corresponding speech acts uttered by the original source.

In order to better understand profiles, let us consider the following newspaper article extract as a short example (Wall Street Journal 10/27/89, full text available in Appendix C):

> Senate Majority Leader George Mitchell (D., Maine) said he intends to use Senate procedures to force advocates of the tax cut to come up with at least 60 votes before they can address the issue.
>
> "The 60-vote requirement will be there and they don't have the 60 votes," Sen. Mitchell said.
>
> [...]
>
> Sen. Bob Packwood (R., Ore.), the leading Republican proponent of the tax cut, had previously said he would be able to find the requisite 60 votes eventually.
>
> Yesterday, Sen. Packwood acknowledged, "We don't have the votes for cloture today."
>
> [...]
>
> The Republicans contend that they can garner a majority in the 100-member Senate for a capital-gains tax cut. They accuse the Democrats of unfairly using Senate rules to erect a 60-vote hurdle. Democrats counter that the Republicans have often used the same rules to suit their own ends.
>
> [...]

> Democrats asserted that the proposal, which also would create a new type of individual retirement account, was fraught with budget gimmickry that would lose billions of dollars in the long run. Republicans countered that long-range revenue estimates were unreliable.

As previously stated, profiles contain information about the sources. In the short example given here, we can find four different sources: two named sources and two mass sources. The named sources are George Mitchell and Bob Packwood and the mass sources are the Democrats and the Republicans. In this extract, the reporter conveys two speech acts by George Mitchell, two by Bob Packwood, two by the Democrats as a group, and three by the Republicans as a group. The reporter also uses certain verbs to report the original speech acts, some of the verbs having argumentative connotations, such as the verbs *accuse, contend* and *counter*. Also, at some points in the article, the reporter explains a few things about the sources and the original situation, for example, explaining that Bob Packwood is "the leading Republican proponent of the tax cut" and that the timeframe for his acknowledgement of not having the votes is "yesterday."

We have four sources, so we have four profiles, one for each source. For Bob Packwood, the profile will contain his name, the reporter's explanations about his position towards the topic, the information about the timeframe and the two reported speech acts, each preceded by the reporting verb which the reporter chose to use. As shown by Bergler, this profile can be represented graphically as the box below.

Sen. Bob Packwood (R., Ore.) — the leading Republican proponent of the tax cut — Yesterday

- had previously said he would be able to find the requisite 60 votes eventually
- acknowledged "We don't have the votes for cloture today."

This is what a profile is. Profiles can be embedded into each other based on their member–of or part–of relationships. For example, from the text, we can also discern that Bob Packwood is part of the mass source of the Republicans, as the "R." in his description shows and as the reporter explains. This part–of relationship can be represented graphically using nested boxes, as Bergler showed in her dissertation [Bergler, 1992].

The Republicans

Sen. Bob Packwood (R., Ore.) — the leading Republican proponent of the tax cut — Yesterday

- had previously said he would be able to find the requisite 60 votes eventually
- acknowledged "We don't have the votes for cloture today."

- contend that they can garner a majority in the 100-member Senate for a capital-gains tax cut
- accuse the Democrats of unfairly using Senate rules to erect a 60-vote hurdle
- countered that long-range revenue estimates were unreliable

The example chosen being an argumentative article, the profiles can also be separated by opposing groups: on one side the Democrats and on the other side the Republicans.

This thesis uses the concept of *basic profiles*, which are profiles at the sentence level instead of at the whole article level. The profiles of sources within the complete news article will be called *full profiles*. Full profiles are built from the basic profiles by grouping the basic profiles by source.

1.2 What is belief?

Belief is the mental acceptance of a truth, actuality, or validity of something, without necessarily having evidential personal knowledge of that something. Beliefs, as used by Bergler [Bergler, 1992] and Gerard [Gerard, 2000], make the distinction between the reader *knowing* what he just read in a newspaper article and the reader *believing* the information he just read. The reader may or may not believe the information, the decision being made based on the evidence presented by the reporter and the sources.

Beliefs are different from facts. While facts are defined as *justifiable true beliefs*, beliefs may not always be true or justifiable [Gerard, 2000]. When a reader peruses a newspaper article, the reader acquires beliefs about what she believes the reporter believes the sources believe about the topic. These beliefs may not be true, the reader

might have read too fast and not completely understood the article, the reporter might have twisted the information around, or the source might have been lying about the topic or just simply been misled. If the reader re–reads the newspaper later on, or an errata is posted by the reporter, or the source apologizes for lying or acknowledges it was misled, these beliefs might change. Even if none of these mishaps happen, beliefs can change, such is the nature of the human mind.

Beliefs can also conflict with each other, a decision having to be made about which one to keep. For example, when a reader has acquired a belief from a previous article and then reads a better article which makes her believe a totally opposite thing, these two beliefs will conflict.

Beliefs can depend on other beliefs and if a belief needs to be changed, all the beliefs depending on this belief need to be updated too. A belief may also need to be removed instead of changed. It is possible that a belief is proved wrong, but there is no alternative belief to replace it.

Belief logics have been extensively studied and many theoretical or practical systems to maintain beliefs have been developed, however, for this thesis, we are only interested in two systems, ViewGen and Percolator, which are briefly presented in the second chapter.

Belief extraction from newspaper articles is facilitated by the profile representation. The description of the source the reporter gives, the reporting verb used by the reporter and the other circumstantial information provided can all be used to extract beliefs of the reporter about the source and about the reported information.

1.3 The challenge

In order to build profiles, we need first to **determine the reported speech** within the sentences in the newspaper article and to be able to **distinguish between the different parts of a reported speech clause**, such as the source, the reporting verb, other circumstantial information, and the primary information, and whether the primary information was transmitted through the use of direct quotes or paraphrases. A partial parser is needed to find the reported speech occurrences and build frames that will later be used to build profiles. These frames should keep enough information from the initial text to be able to determine coreferences between noun phrases (i.e.

belonging to the same profile) and to be able to determine the supporting and the opposing groups in the future. **No information should be lost** since it is needed for the evidential analysis and the noun phrase resolution used to build full profiles.

It should be possible to extract properties of the verb such as tense, aspect, voice, modality, polarity and so on. Hence the parser needed should be able to **find verb clusters** containing all auxiliaries and adverbs connected to the main verb. As mentioned before, information should not be lost in the process. Simplifications such as grouping all verb forms together into one single entity are not acceptable for this project because this would mean that information such as tense, aspect or modality will be extremely difficult, if not impossible, to extract from this type of structure.

It is also important to take into consideration the **punctuation**, since direct reported speech usually makes use of quotation marks, and, as previous studies have shown, parenthetical clauses can be usually extracted from the comma usage in the text. As Quirk shows, the reported clause can be a parenthetical within the matrix clause and vice–versa [Quirk *et al.*, 1991]. The parser should at minimum treat quotation marks and commas.

Another goal is to manage a good coverage of the reported speech structure without doing a full text semantic analysis. That is, manage to build a rule–based grammar which only needs to look at the immediate context for determining the basic profiles. This is the goal of this thesis.

Finally, a system which builds profiles has to be **scalable** and there has to be a concern for the time it takes this system to give results (**runtime**).

This thesis addresses these concerns and provides a proof–of–concept implementation for the building of basic profiles. The steps from basic profiles to full profiles are described. This implementation can be built on and improved in order to combine with other parts, such as Percolator, to form a complex system which could be used for searches, summarization and many other Natural Language processing tasks.

1.4 The solution

As outlined in the previous section, there are many pieces which need to be considered. In this section, I will present a short overview on how they are solved and how all these pieces fit together into a big picture.

First, we need a solution for the recognition and extraction of reported speech clauses. There have been very few implementations which can do this properly, and none which retain enough information to be able to further analyze the structure of the reported speech. One such recognition implementation was done by Doran in the context of a Lexicalized Tree Adjoining Grammar and based on the punctuation usage within the text [Doran, 1998]. However, this is not enough for our goals because there is no treatment of circumstantial information and the implementation only treats direct quotes.

This thesis determines that there are at least three elements needed for recognizing, extracting, and determining the internal structure of reported speech clauses without a full text analysis. These are the *recognition of the reporting verb, the punctuation analysis* and *the structure of the reported speech.* Using these three elements, we are able to recognize, extract and determine the internal structure of most reported speech clauses by just looking at the text, one sentence at a time and not needing any other information.

The reporting verb not only has to be recognized, but also analyzed in order to have its tense, aspect, voice and other properties determined, since this will be needed for future work. A *verb cluster* is defined as a coherent group of modal auxiliaries, normal auxiliaries, a main verb, and any adverbs that may be present. For example, the italicized text in these sentences represents verb clusters: "The photo *will not have actually been being seen*"; "He *really did not want* that hat." Chapter 4 describes the grammar built to extract verb clusters and the algorithms for the verb cluster analysis. These algorithms are for the extraction of the tense, aspect, modality, mood, voice and polarity of the verb cluster. The verb cluster grammar and the verb cluster analyzer are applied to reporting verb clusters for the purpose of extracting and analyzing reported speech. A *reporting verb cluster* is a verb cluster which contains a reporting verb as a main verb. Reporting verbs are determined by pattern matching on a list of reporting verb forms. Finally, the reporting verbs are matched against a semantic dictionary [Gerard, 2000] to determine their semantic dimensions. Reporting verbs and semantic dimensions are described in detail in [Bergler, 1992] and are summarized in Chapter 3.

The second element needed is the punctuation analysis. This is used to determine direct quotes and paraphrases within the reported speech clause, as well as the

placement of circumstantial information other than source and reporting verb.

Direct quotes are defined as a chunk of reported speech text included within quotation marks, while *paraphrases* are defined as reported speech not included within quotation marks. Syntactically, they form the subordinate clause and belong to the *primary information* of the reported speech, as defined by Bergler [Bergler, 1992].

The *other circumstantial information*, is part of the matrix clause and is not the source, nor the reporting verb. It forms a description of the source provided by the reporter, a patient, or a temporal or spatial adverbial clause. While the lexicalization of the source and the reporting verb as well as the rest of the matrix information all provide circumstantial information, we separate these constituents as to make it easier to build basic profiles, and later on full profiles. The use of the punctuation by the reporter is shown here to be able to define syntactic differences which can help determine if a chunk of text is a paraphrase or circumstantial information other source or reporting verb. Section 5.2 discusses this issue and describes other methods which could be used in the future in order to better differentiate between paraphrases and other circumstantial information.

The reporting verb, the source of the original speech act, the punctuation usage, the primary information and the other circumstantial information form the ordered *structure of the reported speech sentence.* The source is a noun phrase and is recognized as a source by matching the sentence in which the source appears to the structure of the reported speech sentence. The noun phrase occurring in a position which can be occupied by the source is considered to be the source. Noun phrases are extracted using the NPE-0 system based on [Bergler, 1997]. The structure of the reported speech sentence is also used to disambiguate between other circumstantial information and paraphrases.

With the reported speech extracted and the internal structure defined, we build basic profiles by simply fitting the reported speech structure within the basic profile frame.

A fully automated implementation is provided and it is explained in the last chapter. The implementation uses a postprocessor, a grammar, and the profiling from reported speech and the verb cluster analyzers described in Chapters 5 and 4, respectively. The postprocessor was introduced to improve the runtime of the application, as it was faster to change a few tags once and use them in the grammar, rather

than having to test for the text of all the words, relevant or not. The implementation also reuses modules from the NPE–0 system based on [Bergler, 1997]: the Brill tagger [Brill, 1994b], a preprocessor, a noun phrase extractor and analyzer, and an Earley parser [Earley, 1970]. The implementation of the Earley parser used is based on the Scheme implementation of the Earley parser tools developed by Marc Feeley, available from the Scheme Repository[3].

1.5 Thesis summary

This thesis develops the implementation process of building basic profiles from reported speech in a newspaper article and presents a proof–of–concept fully automated system which extracts basic profiles from newspaper text and analyzes them, differentiating between source, reporting verb, other circumstantial information and primary information. Inside the primary information, a differentiation is made between direct quotes and paraphrases. The reporting verb is analyzed, determining its grammatical properties as well as its semantic dimensions, paving the path for a future belief analysis based on reporting verbs.

The second chapter presents a brief literature review on Reported Speech and different implementations of taggers, parsers and other NLP tools and a belief attribution system called Percolator [Gerard, 2000], which is an imagined user of the frames encoded in this thesis.

Chapter 3 offers a short overview on Reported Speech and Bergler's Evidential Analysis [Bergler, 1992], as a motivation for this thesis.

Chapter 4 describes the process of building a verb cluster extractor and analyzer and defines a few terms related to verb clusters as used in this thesis. The verb cluster extractor and analyzer is used to find the reporting verb clusters and to analyze them in order to facilitate their evidential and belief analysis. The following grammatical properties of the verb are extracted: tense, aspect, modality, mood, voice and polarity. Semantic dimensions are included in the frame of the verb cluster and information about auxiliaries and adverbs is separated from the main verb. The verb cluster extractor and analyzer has been developed for the general use on any verb and as such can also be used on reporting verbs in newspaper articles, which usually have

[3]http://www.cs.indiana.edu/scheme-repository/code.lang.html

much less complex structures than verb clusters used in fiction. It can also be used to analyze the non–reporting verbs which are inside the quoted material, or any verb in general in any other domain than newspaper articles.

Chapter 5 describes in detail the implementation process of profiling and the different pieces needed in order to build basic profiles. Chapter 5 rests on material from Chapter 4, on material from [Bergler, 1992], which is summarized in Chapter 3, as well as new material relating to the syntactic analysis of primary and circumstantial information and the way they fit together to define the structure of reported speech. This chapter shows that basic profiles can be built without a full analysis of the text, the information contained in the local sentence being enough.

Chapter 6 offers a short overview of the proof–of–concept system developed following the definitions and information presented in the previous chapters. An annotated example is presented, showing how the output from the basic profiler builder looks and describing the fields retained in order to present the basic profiles in an easy to see manner which is also complete enough to be used for future work. The evaluation of the system is also described in this chapter.

Possible improvements and future work are defined in Chapters 4 to 6, within their corresponding contexts.

Chapter 2

Literature review

2.1 Reported speech

While the definition of reported speech depends on which paper one reads, it is generally acknowledged that reported speech is information about someone's act of speaking, relayed at a later time. Some papers consider reported speech to be synonymous with indirect reported speech, but the general consensus is that reported speech refers to both indirect and direct reported speech.

Reported speech has been studied both within the domain of speech recognition and within the domain of written text analysis. While speech analysis is also very interesting, this thesis concerns itself with the study of reported speech in written texts[1].

Research on reported speech in written text has been traditionally part of larger fields [Bergler, 1992] such as Illocutionary Logic and Speech Acts [Searle and Vanderveken, 1985], Propositional Attitudes (Mental States) [Asher and Lascarides, 1994], and Points of View and Belief Maintenance [Ballim and Wilks, 1992].

Bach and Harnish's study of speech acts [Bach and Harnish, 1979] represents an important cornerstone for the field of Illocutionary Logic and Speech Acts. They present a theory of communication through defined speech acts and show that communication can be inferred given an utterance. The path of inference that enables the communication from a given text is presented in a Speech Act Schema. Gazdar [Gazdar, 1981] asserts that the truth conditions of the speech acts depend on the

[1] In particular, newspaper articles in the Wall Street Journal corpus

context and presents the notion of commitment of belief of a speaker, relative to the context.

Asher [Asher and Lascarides, 1994] asserts that research on speech acts should be integrated with theories of discourse structure and that there should be a separation between reasoning about mental states and reasoning about discourse structure. The framework for the discourse structure analysis used in their paper is the Segmented Discourse Representation Theory (SDRT) [Asher, 1993] which was developed on the foundations built by the Discourse Representation Theory (DRT) [Kamp and Reyle, 1993]. The prepositional attitudes used to extend the SDRT are *believes, wants* and *intends.*

In his highly philosophical book [Quine, 1960], Quine discusses, among other things, propositional attitudes and their relation to verbs of cognition[2]. These verbs introduce an intensional context from subject to complement, but the problem with intensional contexts is that two expressions referring to the same underlying entities may not be substituted for one another if they are referring expressions in complement clauses. Quine calls this phenomenon *opacity* [Quine, 1960].

For example, if we take Frege's [Frege, 1892] famous *morning star paradox* as presented in [Bergler, 1992]:

(1) (a) John expects the Evening Star to rise in five minutes.

(b) John expects the Morning Star to rise in five minutes.

Even though both the Evening Star and the Morning Star both represent the planet Venus, we cannot necessarily infer (1b) from (1a) because it is possible that in John's world there might be a belief that the Evening Star is different from the Morning Star, and also because (1a) would occur in the evening while (1b) would occur in the morning.

A few different methods of dealing with opacity within quotes have been proposed, such as the name theory [Quine, 1951], the description theory [Geach, 1972], the demonstrative theory [Davidson, 2001], the modal Russellian theory [Yagisawa, 1997] and so on.

Quine proposed that the quotation is considered as a name for the expression inside quotation marks [Quine, 1951]:

[2] *believe, expect, think, doubt, imagine,* etc.

> The name of a name or other expression is commonly formed by putting the named expression in single quotation marks; the whole, called a quotation, denotes its interior. To mention Boston we use 'Boston' or a synonym, and to mention 'Boston' we use "Boston" or a synonym. "Boston" contains six letters and just one pair of quotation marks; 'Boston' contains six letters and no quotation marks; and Boston contains some 800,000 people.

Geach presented quotation as a structure of concatenated ordered individual words [Geach, 1972]:

> I maintain that the quotation 'man is mortal' is rightly understood only if we read it as describing the quoted expression in terms of the expressions it contains and their order.

Davidson presented another theory, the *demonstrative theory of quotation* [Davidson, 2001]. He asserts that quoted material is semantically irrelevant and that it can be easily replaced with a demonstrative pronoun and a gesture indicating the replaced material. For the demonstrative theory, quotation is just a device for pointing to utterances.

Takashi Yagisawa gives an example of one of the problems that the demonstrative theory encounters and why the demonstrative theory is not satisfactory [Yagisawa, 1997].

> As an indexical word, the demonstrative pronoun 'that' shifts its reference from context to context, and its reference is determined by contextually perspicuous of contextually understood parameters. Suppose I utter to you "The proposition that snow is white is famous." As I utter it, I hold up a **carrot** in front of your eyes and do everything within my power to attract your undivided attention to the **carrot**, and succeed. If I am uttering a sentence with a hidden demonstrative in its logical structure, as the Demonstrative Theory says, then since the carrot is the demonstratum, the context of my utterance clearly determines the referent of that demonstrative to be the carrot, or some entity that bears a contextually obvious relation to the carrot. Hence my utterance will be true in the

> Demonstrative theory if an only if the carrot, or the entity that bears the contextually obvious relation to the carrot, is a proposition and is famous. This is not the right truth condition for my utterance. No matter how strongly the context may present the carrot as the demonstratum for any potential occurrence of the demonstrative 'that' in that context, the carrot remains utterly irrelevant to the truth condition of my utterance. The Demonstrative Theory is incapable of explaining why.

Further on, Yagisawa goes on to present his own theory he calls REMORT [Yagisawa, 1997], short for Revised Modal Russellian theory, based on Russell's principle which holds that the objects appearing in a proposition must be known to the speaker.

However, as Bergler notes [Bergler, 1992], all these theories ignore the lexicalization and linguistic structure of reported or quoted material. In her thesis, Bergler proposes that quoted material should get the same semantic treatment as indirect reported speech [Bergler, 1992]. This approach is also supported by other linguists and philosophers of language, such as Cappelen and Lepore in their paper on the varieties of quotation [Cappelen and Lepore, 1997].

[Bergler, 1992] defines *primary information* as the information that belongs to the main story and *circumstantial information* as the meta-information which presents the circumstances of the reported utterance. The reporting verbs play an important role in the evaluation of the original speech act that a reporter is trying to convey to the reader. In [Bergler, 1992], Bergler describes a detailed analysis of reported speech in newspaper articles and presents an evidential analysis of reported speech. The reporting verb used encodes a wealth of information pertaining to the manner in which the original speech act was made, the confidence the reporter has in the source, the audience in front of which the original speech act was performed and so on. For example, *thunder* and *mutter* will denote two different levels of voice tonality, while *claim* and *announce* encode different reliabilities, as well as different types of audiences and speech acts.

This thesis keeps all the internal structure of both direct quotations and paraphrases, which can both be analyzed semantically according to Bergler's theory.

Recent work on punctuation for reported speech has been done by Doran [Doran, 2000; Doran, 1998; Doran, 1996] and Say [Say, 1998], among others.

Doran analyzes the punctuation structure of reported speech in the context of

a Lexicalized Tree Adjoining Grammar and patterns direct and indirect reported speech as parenthetical clauses [Doran, 1998]. The corresponding implementation was designed to find direct quotes relevant to the topic of templates provided. For the implementation, a Supertagger [Bangalore, 1997] was used to detect direct quotes and verbs taking clausal complements. The Supertagger was trained on 200,000 hand-annotated words of the Wall Street Journal corpus and tested on over 2000 sentences of the same corpus to determine its accuracy on parenthetical verbs of saying. Out of 192 instances of verbs taking clausal complements, 162 (84%) were assigned the correct parenthetical Supertag. In the reported speech detection testing, there were 25 non–parenthetical reported speech out of 602 (4%) which were incorrectly tagged as parenthetical and 4 non–reported speech sentences which were tagged as reported speech.

Reported speech is also mentioned in a few papers on event recognition, however, reported speech seems to be considered irrelevant by researchers. For example, in his dissertation, Macdonald Crowe [Crowe, 1997] argues that reported speech is largely irrelevant to the event recognition process that he was studying. He defines reported speech as typically providing "elaborative information and commentary" and as such not crucial to the event recognition task. In his opinion, reported speech would only trigger the analyzer with events which are not central to the text, but are only referred to for comparison or commentary. Although he is trying to recognize reported speech in order to ignore it, he makes the interesting observation that clause fragments which contain a reporting verb and an instance of the word "that" make up 95% of the reported speech occurrences in his tests. The corpus used was the Latin American terrorist attack corpus that constitutes the Message Understanding Conferences 3 and 4 (MUC–3/4) development and test corpus.

Finally, reported speech has also been studied in the field of Point of View and Belief Maintenance [Wiebe, 1990; Chalupsky, 1996], where many researchers have tried to build an appropriate model of the beliefs of the agents. Points of view as used in [Gerard, 2000; Ballim and Wilks, 1992] and this thesis, represent the viewpoints of an agent relative to its beliefs.

In his ViewGen[3] system for dialogue understanding, Ballim defines two types of

[3]ViewGen and Ballim's ViewFinder dissertation are available from
ftp://crl.nmsu.edu/pub/non-lexical/ViewFinder
ViewGen's website is http://www.dcs.shef.ac.uk/nlp/viewgen/

environments: point of view environments and topic environments [Ballim, 1992]. The point of view environments contain the beliefs ascribed to an agent, where agent is defined as people or intelligent programs interacting with each other. An agent can have beliefs about another agent's beliefs, in which case there would be nested environments. A topic environment contains all the beliefs of an agent about that topic. The process of *belief ascription* ascribes beliefs from one view point to another one without direct evidence, based on the idea that unless there is counter-evidence to a belief, principles of commonality can be used to ascribe the belief [Ballim, 1992]. *Belief percolation* occurs when an agent has a belief which is not held by the overall system and the agent is "reliably informed," in which case the system will acquire the agent's belief.

Gerard adapts Ballim's nested environments representation by introducing the differentiation between *potential beliefs* and *held beliefs* [Gerard, 2000]. Initially, a belief is a potential belief, i.e. the agent (in our case, the reader) has knowledge about that belief, but has not yet incorporated it as a belief of his own. For example, if the reader sees the beginning of an article about a new kind of matter that has been allegedly discovered, the reader now knows about the alleged discovery of a new matter, but has not yet enough reason to believe it is true; the reader holds a *potential belief.* However, as the interested reader looks at the continuation of the article, which enlists the opinions of famous sources in the know and other proofs that the new matter really exists, the belief the reader has about the existence of this new matter becomes stronger and stronger and eventually becomes a belief of his own, a *held belief.*

The Percolator system developed by Gerard [Gerard, 2000] is presented in the next section.

There has been much more research going on in the belief and points of view area, but the papers of interest to this thesis have been outlined above. For an extensive, recent literature review on research about points of view and beliefs, see the Belief Revision chapter in [Gerard, 2000].

2.2 Implementations

2.2.1 Percolator

Gerard [Gerard, 2000] presents the Percolator system for building and maintaining ViewGen-type environments [Ballim and Wilks, 1992], representing the view points of the agents in newspaper articles relative to the topic of the articles. Percolator differentiates between *potential beliefs* and *held beliefs*, where potential beliefs are beliefs which the reader is uncertain about and held beliefs are beliefs which the reader has accepted. The system developed by Gerard implements a few important mechanisms such as the percolation mechanism, the decomposition mechanism and the promotion mechanism, using three main elements: a default percolation rule, a belief reliability heuristic and a source list.

The percolation mechanism defined by Gerard is implemented as the reverse of Ballim's belief ascription and the *default percolation rule* is to percolate beliefs as potential beliefs when there is reason to do so. This is done through the process of attribution from the original source, to the reporter, to the reader, while updating a *source list* which keeps track of this process.

For example [Gerard, 2000], in an initial viewpoints state, Dr. X is the only agent having a belief concerning the common cold, this belief being `cure(Dr. X, common_cold)`. Once this belief is percolated to the reporter's viewpoint, the reporter's topic environment for common_cold will contain `cure(Dr. X, common_cold)`. When this belief is percolated to the reader's viewpoint, it initially belongs to the `potential belief` side of his topic environment on common cold. After applying a *belief heuristic*, if the above belief is considered reliable, it will become a `held belief`.

A source list is a list of frame-like structures, containing two slots for sources and reporter (`source name list` and `reporter name`) as well as other slots for information about confidence criteria[4]. The `source name list` is updated every time there is an exposure to a given belief. The *source list* is used to trace the sources of percolated beliefs such that, if need be, the belief can be tracked back to the original source. The source list is very useful in the decomposition mechanism, because it

[4]The confidence criteria are: the reader's confidence in the source, the reader's confidence in the reporter, the reporter's confidence in the source and the reporter's confidence in the credibility of the reported information.

will show where a belief came from, even after the environments which are not useful anymore are removed (decomposed) once a belief becomes a potential belief in the reader's environment. In the previous common cold example, once the reader has `cure(Dr. X, common_cold)` as a potential belief, we might consider that the environments of Dr. X and the reporter are not needed anymore and remove them. Only the *source list* would be kept containing the source Dr. X, the reporter's name and the confidence criteria which will be used for the belief heuristic.

Gerard defines a *belief reliability heuristic* which is applied to transform basic profiles of reported speech sentences into belief environments[5].

This thesis aims to produce the required input for Percolator. While the reader's confidence in the reporter and in the source is something that a system would have to acquire from the reader or from the reader's behavior, the reporter's confidence in the source and the reporter's confidence in the credibility of the reported information can be acquired from the lexicalization of the source and of the reporting verb used, as described in [Bergler, 1992; Bergler, 1993; Bergler, 1995a].

2.2.2 Taggers

Part-of-speech *tags* are widely used as the basis for text analysis. The process which assigns tags to the words in a corpus or other text is called *tagging*. A *tagger* is the system which accomplishes this. A tagger will try to assign tags from a given *tagset*, one tag per word, however, this is not a trivial task for most languages as there are words that are homonyms, such as the English noun *bear*, which is an animal, and the verb *bear*, which can mean "carry", for example. Some of the most common words are highly ambiguous and even phrases not containing ambiguous words can be hard to parse, even human inter-annotator agreement being less than 100% [Marcus *et al.*, 1993].

There are a few different methods of disambiguating this process. One method, used by rule-based taggers, is to work with a large database of handwritten disambiguation rules, such as a rule to determine that if a word is ambiguous between noun and verb and that particular word appears after a modal, the word should be tagged as a verb. For example, if "would bear" is encountered, "would" is the modal

[5]See [Gerard, 2000] for details.

and "bear" is the ambiguous word, which in this case is a verb, since nouns cannot appear directly after modals. This was the original method used when research started in this direction and it is called rule–based tagging, or linguistic tagging. One such tagger is the English Constraint Grammar tagger (EngCG–2) [Voutilainen, 1997]. EngCG–2 consists of three modules: a tokenizer, a morphological analyzer and a rule–based disambiguator. The *tokenizer* identifies words, punctuation and some multiword expressions. The *morphological analyzer* is composed of a two level lexicon and morphology (over 90000 entries) and a "guesser," a rule-based heuristic analyzer for unknown words. Finally, the *rule–based disambiguator* uses some 1150 constraint rules to remove alternative analyses on the basis of context conditions expressed in these rules.

Another methodology is to make use of statistical information in order to decide between ambiguous uses for a word. These taggers need to be trained on hand annotated corpora in order to construct a language model. Most tagging algorithms based on statistical models use approaches such as Hidden Markov Models, Memory–Based Learning, Maximum Entropy and so on.

The most widely used statistical taggers are the N–gram part of speech taggers, more specifically trigram taggers using Hidden Markov Models (also called Maximum Likelihood Taggers). This algorithm tries to maximize the formula shown in (2) below, where N is 2 in the case of trigram taggers[6]. A HMM tagger will try to choose the best sequence of tags for an entire sentence [Jelinek and Mercer, 1980], [Steetskamp, 1995], [DeRose, 1988].

(2) Probability(word | tag)*Probability(tag | previous N tags)

In a comparison study [Samuelsson and Voutilainen, 1997] between the EngCG–2 tagger and a stochastic trigram tagger using Hidden Markov Models [DeRose, 1988], the rule–based tagger achieved an accuracy rate of 99.3%. The statistical tagger's

[6]In plain English, it tries to maximize the product of the probability that a given tag corresponds to the word we are looking at, and the probability of succession of a given tag given the previous N tags. For example, if N=1 and we are trying to disambiguate between the animal "bear" and the verb "bear", and the previous word in the text is tagged as TO (as in "to bear"), it will look at the probability that a word tagged as a VERB can be "bear", based on a tagged corpus (assuming a corpus which contains articles about wildlife, the NOUN tag would probably have a higher probability) and at the probability that a VERB can follow a TO (compared to a NOUN following a TO). It multiplies them for each tag, and whichever product is higher is selected.

accuracy was trained on 357000 words of the Brown corpus and achieved an accuracy of 96.49%. The tests were done on a subset of 55000 words of the Brown corpus.

Another type of stochastic algorithm is Memory–Based Learning [Stanfill and Waltz, 1986], in which a set of learnt context cases is kept in memory, the tagger extrapolating the part–of–speech tags from the most similar cases that it has seen before. This model is computationally expensive because each word has to be kept in memory together with its context cases and each new word has to be compared to all the patterns which are kept in memory. However, Walter Daelemans shows that this can be improved by using a heuristic case base compression formalism [Daelemans *et al.*, 1996].

The Maximum Entropy Model [Ratnaparkhi, 1996] is another statistical approach for part–of–speech tagging and it seems to achieve very good accuracy rates. The Maximum Entropy tagger assumes maximum ignorance and the model's entropy is lowered with each observed event, unknown events being considered of maximum probability. The model tries to predict tags by maximizing the entropy, given the history available when predicting a given tag. This approach also uses a combination of several "features" which are much like the grammar rules used by the rule–based taggers. It was mainly developed as an alternative to deal with sparse data. The MXPOST [Ratnaparkhi, 1996] tagger[7] developed by Ratnaparkhi follows this model.

Some models have chosen to combine rule–based approaches with statistical information, such as the Transformation Based Learning model [Brill, 1993; Brill, 1995]. This is one of the best known taggers in the field, known as transformation–based tagger (TBT) or Brill tagger [Brill, 1992; Brill, 1994b]. The rules are induced stochastically from a tagged corpus through a supervised learning algorithm. The Brill tagger will initially assign the most likely tag to each word and then use a linguistic model to apply rules in a decreasing granularity fashion, changing some of the original tags. For example, if a verb in future tense is mistagged as a noun initially, the tag will be corrected when the rule *Change a tag from NOUN to VERB if the previous tag is a MODAL* is applied. The tagger looks at a window of three words [Brill and Wu, 1998].

Taggers are often evaluated by comparing them to the *Gold Standard*, or by comparing them with other taggers. The *Gold Standard* is a test set which is labeled

[7]Available from ftp://ftp.cis.upenn.edu/pub/adwait/jmx/jmx.tar.gz

by humans, the accuracy of a candidate tagger being measured against this set, in percentages of correct tags. A note on human tagging [Jurafsky and Martin, 2000], annotators usually agreed only on 96–97% of the tags in the Penn Treebank II [Marcus *et al.*, 1993], however in a later study, the humans were allowed to discuss the tags and in this case they agreed 100% [Voutilainen, 1995].

Similarly, when taggers are allowed to "discuss", a higher accuracy is obtained. This is a relatively new model which combines multiple existing taggers and uses a voting system to determine the most likely tag. The simplest voting system is to use majority voting between a number of taggers, in case of tie picking the tagger with the highest accuracy.

In a study [Brill and Wu, 1998] on four taggers, a simple unigram tagger (assigning the most likely tag regardless of context), a trigram HMM tagger using deleted interpolation [Jelinek and Mercer, 1980], the Brill tagger [Brill, 1994b] and a maximum entropy tagger [Ratnaparkhi, 1996], the accuracy obtained by voting between the four of them was of 97.2%, while the individual accuracies were 93.26% for the unigram tagger, 96.36% for the trigram tagger, 96.61% for the Brill tagger and 96.83% for the maximum entropy tagger. The taggers were trained on 80% of the Wall Street Journal corpus (1.1 million words) and tested on 20% of the corpus (265000 words). The maximum accuracy was obtained when the context was used to pick a tagger out of four, based on which tagger performed best on that type of context [Brill and Wu, 1998].

In a similar study [van Halteren *et al.*, 1998] (on a different corpus), a trigram stochastic tagger [Steetskamp, 1995], the Brill tagger [Brill, 1994b], a memory–based learning tagger [Daelemans *et al.*, 1996] and a maximum entropy tagger [Ratnaparkhi, 1996] were combined, the accuracy obtained by voting between the four of them being 97.92%, while the individual accuracies were 96.08% for the trigram tagger, 96.46% for the Brill tagger, 96.95% for the memory–based tagger and 97.43% for the maximum entropy tagger. The taggers were trained on 80% of the Lancaster–Oslo–Bergen corpus (931062 tokens) and tested on 10% (115101 tokens), the leftover 10% of the corpus was used to develop the voting system (Tune). The maximum accuracy was obtained when every tagger was allowed to vote with a predefined probability based on tag pairs observed on the Tune set [van Halteren *et al.*, 1998].

There is currently an ongoing effort to standardize the tagging process according to

the ISO standards for software quality [EAGLES, 1996]. The Expert Advisory Group on Language Engineering Standards is trying to find a common standard for all the existing tagging systems. It is of particular interest that the guidelines they propose include tags for main verbs as well as Mood, Tense, Person and other categories.

2.2.3 Parsers and Natural Language Processing tools

Daniel Jurafsky defines parsing as recognizing an input string and assigning some structure to it [Jurafsky and Martin, 2000]. More specifically, a parser is an algorithm, implemented as a computer program, designed to determine the syntactic and/or semantic structure of a natural language text, building structures which organize this information in such a way as to make it easy to use and analyze. The basic parsers are generally implemented as a search process.

There are a few different methods of assigning a structure to a natural language text, each approach having its strengths and weaknesses.

The performance of most parsers and tools presented in this section is shown using ParsEval measures. These are standard techniques for evaluating parsers and grammars, proposed in 1991 by Black and his colleagues [Black *et al.*, 1991]. These measures are *Precision* and *Recall*[8]. *Labeled recall* (LR) is expressed as the percentage of correct constituents in a candidate parse of a given sentence compared to the correct constituents in a treebank parse of the same sentence, measuring comprehensiveness. *Labeled precision* (LP) is expressed as a percentage of the correct constituents in a candidate parse of a given sentence compared to the number of total constituents in the candidate parse, measuring accuracy. In my opinion, as long as systems don't use the same treebank and same tasks, the ParsEval values can only be used for a rough comparison, as the results will obviously depend on these variables.

One way of parsing a text is by using grammar rules written according to the grammar of the human language in which the text to be analyzed is written in, this type of parser being a rule–based or grammar based parser. A very known parsing algorithm which is often used with simple rule–based grammars is Earley's algorithm [Earley, 1970], which has also been shown to be possible to be extended to compute the probability of its parses in order to find the most likely parse [Stolcke, 1995]. The

[8] A new metric has been added called *crossing brackets*. This will measure the number of constituents which have a different bracketing than the treebank, such as (A (B C)) compared to ((A B) C)

XTAG English Grammar [Joshi *et al.*, 2001; Bangalore *et al.*, 1998] and LOLITA [Garigliano *et al.*, 1998; Callaghan, 1998] are examples of NLP tools using rule–based parsers.

Earley's algorithm [Earley, 1970] uses a dynamic programming approach in order to implement a parallel top–down search, reducing an exponential time search to the polynomial one in which the worst case is a O(N3) [Jurafsky and Martin, 2000], where N is the number of words in the input text. The parser analyzes the text in a single left to right sweep and it fills an array (called a chart) in which for each word it will associate a list of states representing the partial parse trees that have been generated up to then. The algorithm moves forward only, applying one of the predictor, scanner or completer operators to process states from the chart [Earley, 1970]. The predictor will create new states based on expectations, the scanner will try to figure out what part of speech is the next word while the completer will be applied when a successful grammatical category is found and will advance all previous states waiting for this category to the current position in the input.

The XTAG English Grammar [Joshi *et al.*, 2001] is a rule–based grammar that can deal with a wide range of syntactic phenomena. The XTAG English Grammar can find past and present verb tenses but it will attach them to the first verb in a verbal sequence, no matter if this verb is a main verb or an auxiliary. For aspect, it uses markers for each possibility such as progressive or perfect, while it groups modality and mode under a single feature that can have the following values: base, ger, ind, inf, imp, nom, ppart, prep, sbjunct. This specialized formalism is particular to XTAG and does not fit in with other environments. The XTAG English Grammar achieved 83.94% precision and 82.15% recall for the NP chunking and 76.43% precision and 74.51% recall for the Verb Group chunking, on 940 sentences of length 15 words or less from sections 17–23 of the Penn Treebank [Bangalore *et al.*, 1998].

LOLITA (Large–scale, Object–based, Linguistic Interactor, Translator and Analyzer) [Garigliano *et al.*, 1998], [Callaghan, 1998] was designed to provide a core system which can be used to build other NLP systems and was coded in Haskell. The system is organized around a Semantic Network containing a knowledge base and into which different modules plug in, such as a preprocessor module, morphology module, parser module and other specialized modules [Callaghan, 1998]. The parsing module uses Tomita's algorithm [Tomita, 1986] to create a very large "forest" of all possible

parses and picks one of the trees based on heuristics and preferring deeper trees. If the forest is too large and the parsing overloads the system, it will be abandoned and a new attempt will be made using the Brill tagger and then trying a reduced grammar. At this point, specialized modules can take the information and add further features to it, such as coreferences, translations, etc. In terms of accuracy at the MUC-7 competition, LOLITA achieved overall a precision of 67% and a recall of 63% [Garigliano *et al.*, 1998], where the parser failed in 20% of the cases. A new parser is under development.

Another approach to parsing uses statistics to determine structures based on the probability of such structures occurring in an annotated text in a given human language. These parsers need to be trained on annotated corpora in order to gather the information needed to build probabilistic-based grammars. The probabilistic parsers' performance depends highly on the corpora on which the parser was trained and it usually takes a long time to train such a parser well. An example of a probabilistic parser algorithm, also called a stochastic parser, is the CYK parser (Cocke-Younger-Kasami parser) [Ney, 1991; Collins, 1999]. The Apple Pie parser is a another well known example of an implementation of a probabilistic parser [Sekine, 1996; Sekine and Grishman, 1995].

The Cocke-Younger-Kasami [Ney, 1991] parser uses a bottom-up approach to parsing Probabilistic Context Free Grammars (PCFG) [Booth, 1969], also called Stochastic Context Free Grammars (SCFG) [Jurafsky and Martin, 2000]. Like the Earley parser, it uses dynamic programming and stores the information in a chart holding the maximum probability constituent that spans a set of words. If there are two constituents spanning the same words and have the same properties, but one constituent has a lower probability, this last one will be discarded. The goal is to find the maximum probability parse tree for each sentence. Collins [Collins, 1999] implements CYK and uses a grammar with many non-terminals, such as S, NP, VP, PP, SBAR, NP-C. Collins' implementation was trained on 40000 sentences of the Penn Treebank and tested on around 2500 sentences also from the Penn Treebank, achieving overall 85.7% precision and 85.3% recall [Collins, 1999].

The Apple Pie Parser [Sekine, 1996] is part of the Proteus Project[9] at New York

[9]http://nlp.cs.nyu.edu/

University. It is implemented as a bottom–up probabilistic chart parser, using a best–first search algorithm and only two non-terminals: S and NP. Because it stores the grammar as a finite state automaton, the parse trees can become unmanageable for long sentences. For this case, it uses a process called "fitted parsing" which builds complete trees from partial trees which could not be joined. The Apple Pie Parser's grammar [Sekine and Grishman, 1995] was automatically extracted from 96% Penn Tree Bank corpus, however, a few performance related simplifications were introduced, such as category merging, for example (VBP, VBZ, VD) become VBX, (NNP, NNPS) become NNPX, etc. The system does not currently offer tense/aspect information and it does not treat reported speech. It achieves 71.04% precision and 70.33% recall on 1989 sentences of the Wall Street Journal corpus [Sekine, 1996].

Some systems have chosen a mixed approach, starting with a rule–based analysis and using probabilistic analysis to resolve ambiguous structures. TRAINS [Allen, 1996] is mainly a rule–based parser, but it also uses probabilistic evidence to make final corrections.

James F. Allen's TRAINS parser [Allen, 1996] is a rule–based bottom–up parser. TRAINS was designed to analyze and understand spoken natural language and converse with a human being in order to evaluate freight train shipping plans. While the domain is very narrow, this parser was designed following the rules described in Allen's Natural Language Understanding book [Allen, 1995], which is one of the best books in the field and as such this parser has a high educational value. TRAINS is mainly a rule–based parser, but it can also make use of probabilities in certain cases presenting ambiguities [Ferguson *et al.*, 1996]. No ParsEval metrics are available for this system, to my knowledge.

Sometimes, especially for information extraction, it is not necessary to extract the full syntactic structure of a text. In these cases, a partial parse (also called a shallow parse) could be used. A way to do this is to make use of FSA (Finite State Automata). An example of a Finite State Automaton parser is FASTUS [Hobbs *et al.*, 1993; Hobbs *et al.*, 1996], developed at the Stanford Research Institute International Artificial Intelligence Center.

FASTUS, also known as Finite State Automaton Text Understanding System [Hobbs *et al.*, 1993], is a finite state automata–based text–understanding system, and it works as a cascaded, nondeterministic finite–state automaton, starting by

recognizing smaller linguistic objects such as names or fixed form expressions and working its way up to complex phrases and structures. FASTUS finds verb groups, but they are only tagged as Active, Passive, Gerunds and Infinitive and most adverbs and predicate adjectives are ignored altogether [Hobbs *et al.*, 1996]. During the MUC-5, FASTUS achieved a recall of 34% and a precision of 56% [Hobbs *et al.*, 1996]. It should be noted that FASTUS only performs a partial parse for information extraction.

Amit Bagga's dissertation [Bagga, 1998] presents a good overview of current trends in NLP tools for text extraction and understanding, especially NLP tools presented at the Message Understanding Conferences (MUC).

Chapter 3

Reported speech

This chapter is intended as an overview of reported speech and Bergler's analysis [Bergler, 1992], needed to understand the following chapters. I will start by giving a brief tour of reported speech, its basic syntactic and semantic structure and an overview of Bergler's evidential analysis. The evidential analysis of reported speech motivates the work presented here.

3.1 Overview

Since humans started communicating, they had relayed not only information that they acquired by themselves, but also information that they learned from their fellows. In today's English, the sharing of information that someone else has told us is accomplished through reported speech, in a direct or indirect manner[1] [Quirk *et al.*, 1991]. In direct reported speech, the current speaker presents the knowledge that was learned from another person using the exact words that that person has used, in writing this is usually encased in quotes, as in:

> Caroline said, "It is really cold in Montreal these days."

In indirect speech, the current speaker paraphrases the original speaker, often introducing the paraphrase by using the word *that* [Quirk *et al.*, 1991], as in:

[1] While in newspaper articles, direct speech and indirect speech prevail as methods of reporting original speech acts, in fictional works there are two other types of reported speech that can appear: free indirect speech and free direct speech. These are usually used to represent a person's stream of thought and do not have any obvious indications that it is reported speech, other than the backshift or forwardshift of the verb tense [Quirk *et al.*, 1991]. Free indirect speech and free direct speech will not be studied here.

Caroline said that the weather in Montreal was really cold.

Reported speech appears frequently in newspaper articles, where the reporter will usually use it to evaluate the information presented [Bergler, 1992]. For example, in a recent article[2], the reporter writes about the launch of a Russian cargo rocket carrying fuel and food to the International Space Station:

> "The launch has gone ahead as planned. So far, everything is fine," said a spokesman at ground control just outside Moscow.

Here, by quoting the spokesman for ground control, the reporter shows the evidence that he has to the claim that the cargo rocket did not have problems at launch.

When perusing a newspaper article, two types of information are conveyed to the reader: the main, primary information which is the main story, and the circumstantial information which gives the bits of information a context and puts them within a perspective [Bergler, 1992]. The primary information in newspaper articles is usually provided by sources[3] which were contacted and interviewed by the reporter, such as the source *spokesman at the ground control just outside Moscow* above, and shows the **what** of the event. The circumstantial information is usually provided by the reporter writing the story, explaining the **who, where, when** and **how** of the reported utterances, such as in the Russian cargo rocket example, the who being the spokesman and the where of ground control being just outside Moscow.

The primary information which the source has presented to the reporter, as well as the circumstantial information, is assumed to be relayed in an accurate manner, according to the rules of journalistic integrity.

While the evaluation of the evidence is mostly left to the reader, a reporter can convey their own feelings or beliefs through the subtle usage of language. An evidential analysis [Bergler, 1992] of the reported speech can extract a lot of cues that are encoded within such use of language, for example, the description of the original source given by the reporter can convey more or less authority, while the reporting verb chosen to relay the original speech act can encode the reporter's trust in the original source, the type of audience at the time of the original speech act, as well as other cues[4].

[2]Reuters, 02 Feb 2003 16:06, *Russian cargo rocket blasts off for ISS.*

[3]By sources, it is meant the original speakers from whom the reporter has the information that is being reported

[4]See Table 3.3 on page 33

3.2 Structure

Bergler [Bergler, 1995a] describes newspaper style[5] as employing complex constructions, such as embedded sentences and heavy noun phrases. The reason reporters use heavy noun phrases is because it is the most compact way of encoding information about entities and sources. Newspaper articles also make use of reported speech much more often than other genres, in fact, many newspaper articles are composed of up to 90% reported speech sentences [Bergler, 1992]. Also, reporting verbs are a quantifiable subgroup of all verbs.

A reported speech sentence consists of two parts, a matrix clause and a subordinated, or complement clause. The reported speech matrix clause contains at least two constituents: the source and the reporting verb. The source of the original speech act appears in subject position and is often a complex noun phrase. The reporting verb links the source to the reported utterance. It is most often a finite verb, but not necessarily, as in the following example:

> Caroline said "It is cold in Montreal these days", *adding*, "I almost froze yesterday."

In the chapter "*Complex Sentence, Reporting the Language of Others*," Quirk et al mention the positions in which the reported clause can appear, such as before, within or after the reported utterance.

The semantics of the reported speech have been defined by Bergler [Bergler, 1992] as presented in (3), where A is the source, B is the utterance, B_{org} is the original utterance and C is the reporter. The meaning of this equation is that the reporter's rendition of the original source is assumed to be an accurate interpretation of what the original source said, including the original context in which the original speech act was made, and that it is assumed that by default the reporter witnessed the original speech act given by the source.

$$paraphrase\text{-}of(B_{org}, B) \;\&\; utter(A, B_{org}) \;\&\; utter(C, utter(A, B)) \;\&\; default : witness(C, utter(A, B_{org})) \tag{3}$$

[5] As a side note, for the funny side of newspaper style, see *Banned For Life*, a website keeping track of the clichés "so hackneyed and insufferable that they should be forever banned from the nation's news reports," as the author, Tom Mangan, puts it.
http://tom.mangan.com/banned.htm

3.3 Representation and evidential analysis

As mentioned before, reported speech is used as evidence to the facts presented by the reporter. The way the reporter presents the source of the original speech act encodes a certain level of reliability and authority of the source and the reporting verb used may encode the context of the original speech act, such as manner or intention.

Bergler [Bergler, 1992] notes that a reader must accomplish two tasks while reading a newspaper:

1. interpret the newspaper article as presented by the reporter, using the reporter's encoded evaluation of the situation or context in which the original speech act occurred
2. use his own background knowledge to interpret the original situation or context, asses his own beliefs and points of view and compare those to the assumed points of view of the reporter and the newspaper

Of particular interest to this thesis is the representation of the viewpoint of the reporter, which is accomplished through the lexicalization of the reporting verb and the source.

3.3.1 Lexicalization of reporting verbs

It is important to analyze the reporting verbs, as the choice of verb made by the reporter reflects the reporter's assessment of the context of the original speech act. Bergler [Bergler, 1992; Bergler, 1995a; Bergler, 1991; Bergler, 1993] has researched reporting verbs in depth and identified meanings components of reporting verbs and defined a paradigm which can be used to reconstruct the original context in a computational manner.

vocalization	neutral	contextual	pragmatic intent
cry	say	allude	accuse
call	tell	argue	admit
mumble	report	contend	claim
mutter	relate	insist	deny
scream	announce	reiterate	promise
shout	release	reply	joke
stammer	state	dispute	assure
stutter	add	ask	pledge

Table 3.1: Reporting verbs encoding the dimension manner

Table 3.1 shows some examples of how the manner of the original situation can be partially reconstructed from the choice of reporting verb[6]. For example, if the reporter used the reporting verb "joke", it means that the intention of the original speaker was not one of seriousness.

Another example of dimension reconstruction is shown in Table 3.2, in this case the dimension being that of textual status of the reported material within the context of the original speech act[7]. For example, if the reporter has chosen to use the reporting verb "announce", it means that the information provided by the original speaker was not mentioned before.

new	previously mentioned or implied	
	confirming	contrasting
announce	agree	deny
report	confirm	insist
tell	consent	refute

Table 3.2: Reporting verbs encoding the dimension textual status

Bergler [Bergler, 1993; Bergler, 1992] distinguishes and classifies reporting verbs using nine *semantic dimensions*: voice quality, explicitness, formality, audience, polarity, presupposition, speech act, affectedness and strength. The first four semantic dimensions pertain to how the original speech act was perceived in its surroundings, the next four semantic dimensions pertain to the relationship between the source and the speech act and the last semantic dimension pertains to the relationship between the reporter and the original context. Thus they encode: situation of the original utterance, attitudes towards the complement clause and strength of the complement.

These semantic structures are explained in detail in [Bergler, 1993] and [Bergler, 1992]. Table 3.3 summarizes the semantic dimensions and their values and defaults[8]. Table 3.4 shows a few examples of reporting verbs and their semantic dimensions[9]. Semantic dimensions will be used for inferences later on.

3.3.2 Lexicalization of the source

Bergler [Bergler, 1992] also presents the lexical semantics of the source which is important in conveying the level of reliability, credibility and authority of the original

[6]This table corresponds to the Figure 3.4 on page 46 of [Bergler, 1992].
[7]This table corresponds to the Figure 3.5 on page 47 of [Bergler, 1992].
[8]This table corresponds to the Figure 4 on page 9 of [Bergler, 1993].
[9]This table corresponds to the Figure 5 on page 10 of [Bergler, 1993].

Course Structure	Semantic Dimension	Values	Default
Situation of the Original Utterance	voice quality explicitness formality audience	range: high, low, clear, ... explicit - implicit formal - informal public - private	unmarked explicit unmarked unmarked
Attitude to Complement	polarity presupposition speech act affectedness	positive - negative new - presupposed range: request, question, ... positive - negative	positive unmarked inform unmarked
Strength of Complement	strength	high - low	unmarked

Table 3.3: The semantic field of reporting verbs

Verb	Voice Quality	Explicitness	Formality	Audience	Polarity	Presupposition	Speech Act	Affectedness	Strength
thunder	loud	explicit	-	-	pos.	-	inform	-	-
say	-	explicit	-	-	pos.	-	inform	-	-
claim	-	explicit	-	-	pos.	-	inform	-	low
affirm	-	explicit	-	-	pos.	presup.	inform	-	high
concede	-	explicit	-	-	pos.	presup.	inform	neg.	-
deny	-	explicit	-	-	neg.	presup.	inform	neg.	-
allude	-	explicit	-	-	pos.	new	inform	-	-
announce	-	explicit	-	public	pos.	new	announce	-	high
declare	-	explicit	formal	public	pos.	new	announce	-	high

Table 3.4: Semantic dimensions of some reporting verbs

speaker, as well as, by extension, the reliability of the information relayed through the original speech act. The reporter uses the source description to inform about the situation of the original speech act, and often will also encode personal beliefs about the source.

The lexicalization of the source is not analyzed in this thesis. The sources (subject NPs) are extracted from the text using the NPE–0 system based on [Bergler, 1997] and are stored in the basic profile frame as *source*.

3.3.3 Evidential analysis

Evidential analysis [Bergler, 1992; Bergler, 1995a] results in a general purpose representation which is domain independent and can be used by different systems with varying belief sets. This representation can also be used for summarization and information retrieval.

Bergler defines evidential scope as the perception of the reported speech in a given set of contexts (original, current and temporal). The original context (OC) is shown

by the source, the reporting verb and the other circumstantial information accompanying them. The current context (CC) represents the structure of the text. The temporal context (TC) contains temporal information corresponding to the primary and circumstantial information. These three context variables form a *localized context*. Evidential analysis empowers the reader of the article to judge the context, the complement being valid proportionally to the reliability of the context. Each clause can be partially evaluated in isolation in its own localized context, as shown in (4).

$$SAY(Source, \phi\prime) \ is \ true \ \Rightarrow \phi\prime[OC, CC, TC] \ is \ true \tag{4}$$

In other words, if the reader believes that the source, reporting verb, other circumstantial information, temporal information and coherence cues are reliable, then the user will believe that the primary information is true.

For a very clear walk–through example of evidential analysis, see [Bergler, 1992], pages 61–67.

Evidential analysis is a two step process. First, the process needs to analyze the text in order to determine the evidential scope, and then it needs to assign a reliability to the information that has been reported.

For the first step, Bergler [Bergler, 1992] defines MTR, a *Minimal Text Representation*, which is a scheme that can be used to adequately and concisely represent the linguistic structure of a text in general, and a newspaper article in particular. MTR provides three devices for the representation of text: coherence structure, trace and profiles. The coherence structure corresponds to the current context (CC), the trace corresponds to the temporal context (TC), and the profile corresponds to the original context (OC).

The purpose of this thesis is to develop the profile side of the MTR in correspondence with the information reported in newspaper articles.

Based on the information provided by the MTR scheme, an analysis can be done in order to assign reliability to the reported information. Gerard [Gerard, 2000] implemented a belief percolation system (Percolator) which computes the likelihood of belief of the readers in newspaper articles, based on simulated original contexts with semantic dimensions and source reliability information.

Chapter 4

Verb cluster extraction and analysis

This chapter describes the steps and research involved in the creation of a grammar used to parse a text and extract verb clusters as well as the corresponding code used to analyze these clusters and provide information such as modal auxiliary, other auxiliaries, main verb, tense, aspect, voice, modality, mood, polarity and semantic dimensions of reporting verbs.

While the final purpose of the Verb Cluster extractor and analyzer was to detect and analyze reporting verb clusters in the domain of newspaper articles, the research has been done with generalization in mind and we believe that the current model is easy to apply to other domains.

4.1 Introduction

In order to be able to do profiling, there was a need for a tool that would identify reported speech and its internal structure. The AETNA Group, now merged with the CLaC Lab[1], has developed ERS, a noun phrase coreference resolution system, which contains a tagger[2] and a noun phrase extractor module. The work of this thesis was developed to fit into this environment. This document describes the steps and research involved in the development of a tool that extracts verb clusters from a text and analyses their tense, aspect and modality. Any other information that appears

[1] http://www.cs.concordia.ca/CLAC/

[2] Brill in the version used here

from the contents of a verb cluster is also retained. Examples of such information are voice, polarity and the presence of subject–operator inversion.

The next section starts by presenting a short overview of verb clusters that is the basis of the choices made later in the thesis. A definition of the term "verb cluster" is provided later on and the steps leading to the development of the grammar used to parse the verb clusters are described.

Later sections will present overviews and implementations for some of the fields of the verb cluster frame, which can appear for all verb clusters, reporting or not.

Finally, the last sections will present the application to the reporting verbs and the field of semantic dimensions and how this relates to reported speech. The verb extractor and analyzer tool can be used for any verb cluster, including reporting verb clusters, which will have the extra field for semantic dimensions. For other verbs, this field will be left empty.

4.2 Verb cluster

This section will define the term *verb cluster* and how it differs from *verb phrase*, present a short overview of the grammar categories used to build the verb cluster grammar and show the steps involved in building the aforementioned grammar and the final grammar used to parse verb clusters.

The notion of *verb cluster* used here has to be distinguished from the linguistic notion of *verb phrase*.

Verb phrases consist of one or more verb constituents: a main verb and/or up to four auxiliaries as well as other arguments such as object complements. Auxiliaries can be a modal auxiliary verb (*can, could*, etc.) or other auxiliary verbs (*have, be, do*) [Quirk *et al.*, 1991]. Some verbs have an intermediate status between main verbs and auxiliary verbs, these being presented later on in the *intermediate auxiliaries* section. A verb phrase can be divided into operator and predication. The *operator* is the first auxiliary and it has an important role for the formation of questions as well as for negation. A yes–no question can be formed by reversing the order between the operator and the subject of the clause. By inserting "not" right after the operator, the polarity of the sentence becomes negative. The *predication* is defined as what is left of the predicate (verb phrase) after the operator is removed [Quirk *et al.*, 1991].

The notion of *verb cluster* differs from that of verb phrase in that only the verb and adverb elements belong to the verb cluster. In the example *John could have been badly hurt by that rock* the verb phrase is *could have been badly hurt by that rock* while the verb cluster is *could have been badly hurt.* The verb cluster is a subset of the verb phrase. The notions of operator, auxiliaries and main verb are the same for both verb phrase and verb cluster.

A verb cluster can be finite or nonfinite. The verbs in a *finite verb cluster* follow the ABCD[3] pattern rules shown in [Quirk *et al.*, 1991] and the operator cannot be a base form, present participle or past participle. Finite verb phrases can front predicates in independent clauses. Verb clusters that have a base form, present participle or past participle as an operator are *nonfinite verb clusters.* Nonfinite verb clusters have no tense or modality distinctions and cannot occur accompanied by a subject. Since at this point we are interested in extracting verb clusters that belong to the predicate in independent clauses and which are accompanied by a subject, only finite verb phrases will be treated here. In a later section on reporting verbs, simple non-finite gerund reporting verbs will be considered.

Modal auxiliaries

The first auxiliary in a finite verb phrase can be a modal auxiliary verb. There are nine modal auxiliary verbs in English:

`can`	`could`	`may`	`might`	`must`
`shall`	`should`	`will`	`would`	

Modal auxiliaries indicate the modality of a verb cluster, carrying meanings such as *possibility, volition, obligation,* etc. A special modal auxiliary verb is "will", which is used as a *future* tense marker.

Modal auxiliaries correspond to the position A in the ABCD pattern structure of the finite verb phrase.

[3]Quirk's ABCD pattern rules are presented in [Quirk *et al.*, 1991] on page 152, and state that in a complex finite verb cluster, each auxiliary has a specific position denoted by the letter A, B, C or D. Verb clusters can only be formed by combining the four positions in alphabetical order, with or without gaps (AB, ABC, ACD, AD, CD, C, etc.) The modal auxiliaries always occur before any other auxiliary, in what is called *position* or *type A*, and are always followed by an infinitive verb form. The *position* or *type B* corresponds to the auxiliary *have*, while the *positions* or *types C* and *D* correspond to forms of the auxiliary *be*, the difference being made on what verb form follows the *be* form.

Other auxiliaries

Three verbs function mainly as auxiliaries, and Quirk calls them *primary verbs* [Quirk *et al.*, 1991]. These three verbs can also function as main verbs.

```
have        be        do
```

Have, when used as an auxiliary, denotes a perfective[4] aspect (position B) and is always followed by a past participle verb.

Be, when used as an auxiliary, can be a progressive[5] aspect auxiliary (position C), which is always followed by a present participle verb, or a passive[6] voice auxiliary (position D), which is always followed by a past participle verb.

Do, also called a "dummy auxiliary", doesn't have any influence on aspect, tense or modality and is used as a filler operator when there are no semantic reasons for other operators to be present. The main uses of *do* as an auxiliary are for negating the polarity of a clause containing a simple verb, for providing an operator for inversion of subject and operator when the predicate is a simple verb in a question and for emphasizing simple verbs [Quirk *et al.*, 1991].

Main verbs

The last verb in a verb cluster, when there is no ellipsis, is the main verb. The main verb carries the lexical information of the predicate and describes what the subject is experiencing or doing. The main verb is very important in the analysis of reported speech, as this is the reporting verb for which the semantic dimensions will be analyzed.

Intermediate auxiliaries

This is just a short overview on intermediate auxiliaries that can be used to further improve the analysis and extraction of verb clusters. These verbs' status is somewhat intermediate between main verbs and full-blown auxiliaries or modal auxiliaries. This status is defined on the basis of the criteria for auxiliary verbs defined by Quirk [Quirk *et al.*, 1991]. In the rest of this section I will present summarized descriptions

[4]Referring to anterior actions or events relative to the time defined through the use of tense.
[5]Referring to actions in progress at the given time defined through the use of tense.
[6]Expressing an action of an agent on the subject.

of *marginal modals, idiomatic modals, semi-auxiliaries* and *catenative verbs.* These intermediate auxiliaries are described in detail in [Quirk *et al.*, 1991], pages 136–148.

Marginal modals are defined as verbs that closely resemble the modal auxiliaries. The four verbs falling into this category are *dare, need, ought to and used to* [Quirk *et al.*, 1991].

Idiomatic modals contain multiword verbs and can serve the role of a modal auxiliary. Examples of idiomatic modals are: *had better, would rather, have got to, be to, would sooner, may as well, might as well, had best, had rather*, etc. [Quirk *et al.*, 1991]

A *semi-auxiliary* consists of a combination of verbs and other particles introduced by one of the primary auxiliary verbs *have* or *be*, and which expresses modal or aspectual meaning. Often, the boundaries of this category are not clear and some of the idioms can easily fall under the idiomatic modals category. Examples of semi-auxiliaries: *be about to, be bound to, be supposed to, be obliged to, be willing to*, etc. [Quirk *et al.*, 1991].

Catenative verb constructions have meanings related to aspect or modality, but are closer to main verbs than to semi–auxiliaries, since they take *do–support.* Catenative verbs can be concatenated to form chain like structures. Examples of catenative constructions are: *appear to, come to, fail to, get to, happen to, manage to, seem to, tend to, turn out to* [Quirk *et al.*, 1991].

4.2.1 Verb cluster overview

A verb cluster (or verb group [Allen, 1995]) is composed of an optional modal auxiliary, other optional auxiliaries, a main verb and optional adverbial elements. The verb cluster is a part of the verb phrase. For the analysis of verb clusters, subject noun phrases which were inverted with the operator are also considered.

As explained by Quirk [Quirk *et al.*, 1991], the noun phrase taking part in the inversion can only appear between the operator and the verb following it. Most of the time such inversion will indicate an interrogative mood, but there are special cases of formal usage in which it can occur in subordinate clauses of condition and concession. These cases will not be treated here.

Most of the time adverbs will appear at the beginning of the finite verb cluster, right after the operator, right before the main verb or right after the main verb [Baker,

1. CC Coordinating conjunction	19. PP$ Possessive pronoun
2. CD Cardinal number	20. RB Adverb
3. DT Determiner	21. RBR Adverb, comparative
4. EX Existential there	22. RBS Adverb, superlative
5. FW Foreign word	23. RP Particle
6. IN Preposition or subordinating conjunction	24. SYM Symbol
7. JJ Adjective	25. TO to
8. JJR Adjective, comparative	26. UH Interjection
9. JJS Adjective, superlative	27. VB Verb, base form
10. LS List item marker	28. VBD Verb, past tense
11. MD Modal	29. VBG Verb, gerund or present participle
12. NN Noun, singular or mass	30. VBN Verb, past participle
13. NNS Noun, plural	31. VBP Verb, non-3rd person singular present
14. NP Proper noun, singular	32. VBZ Verb, 3rd person singular present
15. NPS Proper noun, plural	33. WDT Wh-determiner
16. PDT Predeterminer	34. WP Wh-pronoun
17. POS Possessive ending	35. WP$ Possessive wh-pronoun
18. PP Personal pronoun	36. WRB Wh-adverb

Figure 4.1: Penn Treebank tag set

1989]. However, an adverb can possibly appear anywhere inside the verb cluster, as shown by Quirk [Quirk *et al.*, 1991]. This implementation will follow Quirk's approach and define the adverbs' position as anywhere inside the verb cluster.

4.2.2 Verb cluster grammar

This section will describe how the grammar used to extract the verb clusters was built and the reasoning behind it. The parser uses the Brill part of speech tags, described below. A full description of all the possible simple and complex verb cluster types will be provided, creating the basis for the grammar described at the end of this section.

Part of speech tags overview

The verb cluster extraction and analysis tool is based on NPE–0 which is based on ideas in [Bergler, 1997]. It makes use of the Brill tagger [Brill, 1994b]. The part of speech tags used by the Brill Tagger are based on the Penn Treebank tag set [http://www.cis.upenn.edu/~treebank], displayed in Figure 4.1.

In order to speed the processing of the verb cluster, a few extra tags were defined in the postprocessing phase of the implementation. The extra tags are presented in Figure 4.2.

Be verb forms:	Have verb forms:	Do verb forms:
BE_VB	HAVE_VB	DO_VB
BE_VBZ	HAVE_VBZ	DO_VBZ
BE_VBP	HAVE_VBP	DO_VBP
BE_VBD	HAVE_VBD	DO_VBD
BE_VBG	HAVE_VBG	
BE_VBN	HAVE_VBN	

Figure 4.2: Additions to the Penn Treebank tag set

Possible simple and complex finite verb clusters

Tables 4.1, 4.2 and 4.3 show all the possible forms for simple and complex finite verb cluster types and were constructed based on Figure 3.55 from Quirk [Quirk *et al.*, 1991], Figures 2.7 and 2.8 from Allen [Allen, 1995] and two tables on the "English Grammar on the Web" [Byrd *et al.*, 2002] website. These tables do not treat the case of ellipsis[7].

Table 4.2 shows the constructions starting with a modal other than "will" for which tense, aspect and voice are not analyzed in this thesis. Quirk [Quirk *et al.*, 1991] considers modals as having an abnormal time reference and the Penn Treebank tags do not differentiate between the different forms for modal verbs. Modals cannot be placed on the time axis as clearly as other verb forms.

These tables form the basis for the development of the verb cluster grammar presented in the next section. To my knowledge, this is the most clear and complete graphical representation of the simple and complex finite verb clusters and their tense, aspect and voice.

Implementation

The resulting grammar is shown in a simplified form[8] in Appendix A.

[7] Verb ellipsis is a very complex issue, as there are many exceptions to the possibility of identification with an antecedent. Perusing the ACL proceedings, one can encounter many papers discussing ellipsis, the rules and the exceptions, as for example [Hardt, 1992], which points out three cases in which an antecedent cannot be found. Since ellipsis is extremely rare in newspaper articles, I will not study it here.

[8] In this simplified form, the ADVERBS have been removed. ADVERBS can appear anywhere within the verb cluster. The ADVERBS definition has been kept in this simplified representation.

modal	auxiliaries			main	tense	aspect	voice	Traditional grammar notation	ABCD
				VBP	present	indefinite	active	VBP	
				VBZ	present	indefinite	active	VBZ	
	HAVE_VBP			VBN	present	perfect	active	have+VBN	B
	HAVE_VBZ			VBN	present	perfect	active	has+VBN	B
	HAVE_VBP	BE_VBN		VBG	present	perfect progressive	active	have+been+VBG	BC
	HAVE_VBZ	BE_VBN		VBG	present	perfect progressive	active	has+been+VBG	BC
	HAVE_VBP	BE_VBN	BE_VBG	VBN	present	perfect progressive	passive	have+been+being+VBN	BCD
	HAVE_VBZ	BE_VBN	BE_VBG	VBN	present	perfect progressive	passive	has+been+being+VBN	BCD
	HAVE_VBP		BE_VBN	VBN	present	perfect	passive	have+been+VBN	BD
	HAVE_VBZ		BE_VBN	VBN	present	perfect	passive	has+been+VBN	BD
		BE_VBP		VBG	present	progressive	active	am/are+VBG	C
		BE_VBZ		VBG	present	progressive	active	is+VBG	C
		BE_VBP	BE_VBG	VBN	present	progressive	passive	am/are+being+VBN	CD
		BE_VBZ	BE_VBG	VBN	present	progressive	passive	is+being+VBN	CD
			BE_VBP	VBN	present	indefinite	passive	am/are+VBN	D
			BE_VBZ	VBN	present	indefinite	passive	is+VBN	D
				VBD	past	indefinite	active	VBD	
	HAVE_VBD			VBN	past	perfect	active	had+VBN	B
	HAVE_VBD	BE_VBN		VBG	past	perfect progressive	active	had+been+VBG	BC
	HAVE_VBD	BE_VBN	BE_VBG	VBN	past	perfect progressive	passive	had+been+being+VBN	BCD
	HAVE_VBD		BE_VBN	VBN	past	perfect	passive	had+been+VBN	BD
		BE_VBD		VBG	past	progressive	active	was/were+VBG	C
		BE_VBD	BE_VBG	VBN	past	progressive	passive	were/was+being+VBN	CD
			BE_VBD	VBN	past	indefinite	passive	was/were+VBN	D
WILL_MD				VB	future	indefinite	active	will+VB	A
WILL_MD	HAVE_VB			VBN	future	perfect	active	will+have+VBN	AB
WILL_MD	HAVE_VB	BE_VBN		VBG	future	perfect progressive	active	will+have+been+VBG	ABC
WILL_MD	HAVE_VB	BE_VBN	BE_VBG	VBN	future	perfect progressive	passive	will+have+been+ being+VBN	ABCD
WILL_MD	HAVE_VB		BE_VBN	VBN	future	perfect	passive	will+have+been+VBN	ABD
WILL_MD		BE_VB		VBG	future	progressive	active	will+be+VBG	AC
WILL_MD		BE_VB	BE_VBG	VBN	future	progressive	passive	will+be+being+VBN	ACD
WILL_MD			BE_VB	VBN	future	indefinite	passive	will+be+VBN	AD

Table 4.1: Simple and complex finite verb clusters

The parser uses this grammar to fill the *modal* (modal auxiliary), *auxs* (auxiliaries), *mainverb* (main verb) and *adverbs* verb cluster frame fields[9]. The *tense*, *aspect*, *voice*, *modality*, *mood* and *polarity* verb cluster frame fields are filled starting from the information provided through the parse.

[9]Implemented as a Gambit structure. A frame is a representation of a verb cluster like a template containing its properties.

modal	auxiliaries			main verb	ABCD
MD				VB	A
MD	HAVE_VB			VBN	AB
MD	HAVE_VB	BE_VBN		VBG	ABC
MD	HAVE_VB	BE_VBN	BE_VBG	VBN	ABCD
MD	HAVE_VB		BE_VBN	VBN	ABD
MD		BE_VB		VBG	AC
MD		BE_VB	BE_VBG	VBN	ACD
MD			BE_VB	VBN	AD

Table 4.2: Verb clusters with modality irrealis

DO_VB	VB
DO_VBP	VB
DO_VBZ	VB
DO_VBD	VB

Table 4.3: Do support

4.3 Tense

This section will define the term "tense" and present a short overview of the current trends on the meanings of tense and different types of tense. A choice will be made as to which of these types will be used for the purpose of this thesis. These types will be analyzed syntactically and semantically from the traditional English grammar point of view and then the corresponding programming algorithm for the extraction of tenses will be described.

4.3.1 Tense overview

Tense positions an event or action on the time axis. Most languages attach tense to the verb expressing the action or occurrence. Time is considered as a continuous dimension in which events occur in succession from the past, through the present and into the future. If we were to stand on the time axis, whatever events happened behind us would be in the past, events happening where we are standing would be in the present and events that did not happen yet would be in front of us and in the future. As we move forward on the time axis, events that were once in the future would become present events.

The American Heritage Dictionary of the English Language [American Heritage Dictionary, 2000] derives tense from the inflected form of a verb, and considers three tenses: past, present and future.

The *Webster's Revised Unabridged Dictionary* [Webster Revised, 1998] defines tense slightly differently, but in the end it also considers tense as indicating the time of the action or the event, and counts three tenses: past, present and future, with the annotation that this list may admit modifications.

Other definitions[10] will also show differences as to what exactly the types of tenses are and what can be considered a tense, but semantically, they all agree to the fact that tense positions an event or action on the time axis.

Reichenbach [Reichenbach, 1947] considers tense as placing an event or action on the time axis, but also as representing the time of speech and the reference time. Tenses are defined semantically. For this, he defines three time points, namely the *point of speech* (S), the *point of the event* (E), and the *point of reference* (R). By varying the temporal placement of the three points, Reichenbach defines tenses as being combinations between (*past, present, future*) and (*simple, anterior, posterior*), such as simple past, anterior past, etc. Past is defined as R<S, present as R=S and future as R>S. Simple is defined as E=R, anterior is defined as E<R and posterior is defined as E>R. Reichenbach accounts for the progressive forms by giving E the possibility to stretch over a period of time, appending *extended* to the tense. This theory combines tense and aspect in a unified model.

James Allen [Allen, 1995] adopts the same theory, but describes tenses in a different way, considering the following 12 tenses: simple present, simple past, simple future, present perfect, future perfect, past perfect, present progressive, past progressive, future progressive, present perfect progressive, future perfect progressive and past perfect progressive. These tenses are differentiated by their syntactic structure. This tense system also takes into consideration the relativity of the position of the event or action in time with respect to other event or actions. Allen's *perfect* corresponds to Reichenbach's *anterior* and *progressive* corresponds to *extended*, but there is no correspondence to *posterior*.

A similar tense system is presented in [Byrd *et al.*, 2002], except the future tense is replaced with a Modal tense (Modal + Base form, Modal Progressive, Modal Perfect

[10][Quirk *et al.*, 1991], [Allen, 1995], [Aarts *et al.*, 2000], [Byrd *et al.*, 2002], [Loos *et al.*, 2002] etc.

and Modal Perfect Progressive).

Randolph Quirk [Quirk *et al.*, 1991] defines tense on a referential level, as well as on a semantic and syntactic level. While the future tense is present on the referential level and on the semantic level, it is not considered anymore on a grammatical level based on the definition preferred by the author that the tense is to be defined strictly as a category realized by verb inflexion. The future tense is hence replaced with the semantic category of future time. The author also mentions that some grammarians prefer to call the present tense "nonpast" to avoid the confusion between present tense and present time.

While every approach has its merits, I prefer Quirk's usage of tense, with the rest of referential temporal information stored in aspect rather than in tense. I do believe though that it is important to analyze the future time as a tense for the purpose of this thesis. Hence, the tenses analyzed and discussed here are: past, present and future, as defined grammatically, rather than semantically.

4.3.2 Tense implementation

According to Quirk [Quirk *et al.*, 1991], an operator is the first auxiliary/modal of a complex predicate. Grammatically, the tense can be extracted from the operator or from the main verb in the case of simple verb clusters which do not contain auxiliaries or modals. The tense of the whole verb cluster will be the same as the tense of the operator or of the main verb. In both cases, this represents the first verb in the cluster as read from left to right, as illustrated in the table Table 4.4.

The past tense is defined as a tense that is formed by having the first verb of the verb cluster (operator or main verb if operator is missing) expressed in the simple past form. The simple past form is represented as VBD in the Brill tagger. Semantically, the past tense will not always represent with certainty a semantic past time, as for example in the use of the hypothetical past or attitudinal past [Quirk *et al.*, 1991].

Examples of verb clusters in the *past tense*:

He *ate* all the cookies.

She *was already leaving* the house.

I *had already eaten* the apples.

We *had been going there* for months.

				VBP	present	
				VBZ	present	
			BE_VBP	VBN	present	D
			BE_VBZ	VBN	present	D
		BE_VBP		VBG	present	C
		BE_VBZ		VBG	present	C
		BE_VBP	BE_VBG	VBN	present	CD
		BE_VBZ	BE_VBG	VBN	present	CD
	HAVE_VBP			VBN	present	B
	HAVE_VBZ			VBN	present	B
	HAVE_VBP	BE_VBN		VBG	present	BC
	HAVE_VBZ	BE_VBN		VBG	present	BC
	HAVE_VBP		BE_VBN	VBN	present	BD
	HAVE_VBZ		BE_VBN	VBN	present	BD
	HAVE_VBP	BE_VBN	BE_VBG	VBN	present	BCD
	HAVE_VBZ	BE_VBN	BE_VBG	VBN	present	BCD
				VBD	past	
			BE_VBD	VBN	past	D
		BE_VBD		VBG	past	C
		BE_VBD	BE_VBG	VBN	past	CD
	HAVE_VBD			VBN	past	B
	HAVE_VBD	BE_VBN		VBG	past	BC
	HAVE_VBD		BE_VBN	VBN	past	BD
	HAVE_VBD	BE_VBN	BE_VBG	VBN	past	BCD
WILL_MD				VB	future	A
WILL_MD			BE_VB	VBN	future	AD
WILL_MD		BE_VB		VBG	future	AC
WILL_MD		BE_VB	BE_VBG	VBN	future	ACD
WILL_MD	HAVE_VB			VBN	future	AB
WILL_MD	HAVE_VB		BE_VBN	VBN	future	ABD
WILL_MD	HAVE_VB	BE_VBN		VBG	future	ABC
WILL_MD	HAVE_VB	BE_VBN	BE_VBG	VBN	future	ABCD

Table 4.4: Tense illustration

The newspaper *was left* on the steps.

The books *were not being read.*

The managers *had been demoted.*

The feast *had been being eaten.*

The present tense is defined as a tense that is formed by having the first verb of the verb cluster expressed in the simple present form. The Brill tagger uses the VBZ tag for the 3rd person singular of the simple present and VBP for the rest of the simple present forms. Semantically, the present tense will not always represent with certainty a semantic present time, as for example in the use of the historic present or future present [Quirk *et al.*, 1991].

Examples of verb clusters in the *present tense*:

He *sits there.*

I *sit here.*

She *is leaving* the house.

I *am eating* the chocolate chip cookies.

We *had driven* for a long time.

I *have been eating* like a pig lately.

The bell *is rung.*

The horse *is being ridden.*

The homework *has been done.*

The glass *has been being broken.*

The future tense is defined as a tense that is formed by having the first verb of the verb cluster be a future tense marker, as for example "will." All modals, including "will," are labeled as MD by the Brill tagger [Brill, 1994b]. For now, we look at the modals inside the code to determine if we have a future marker or not. A possible improvement is to define additional tags for modals, which would include a future marker tag. Quirk [Quirk *et al.*, 1991] notes that semantically, the future tense will not always represent with certainty a semantic future time since it is not possible to attach an event or action to the future time in an ascertainable manner. However, it is possible to assume that the event will happen.

Examples of verb clusters in the *future tense*:

He *will visit* his friends tonight.

She *will be driving* to San Francisco.

We *will have left* by then.

The kid *will have been eating* the cookies.

The disk *will be thrown.*

The dress *will be being worn.*

The feast *will have been eaten.*

The photo *will have been being seen.*

The previous definitions for present, past and future tense can be easily implemented by taking the first verb (operator) in a verb cluster parse and looking up the tense[11] of this first verb from the corresponding Brill tag. Compound future markers, such as "going to," are not treated as of yet.

4.4 Aspect

This section will define the term "aspect" and overview the different ways of analyzing and categorizing aspect. As with tense, a decision will be made as to how exactly aspect will be treated for the purpose of the programming implementation of the verb cluster analysis. Syntactic and semantic definitions will be provided, as well as a simple heuristic method of extracting the aspect from the verb cluster.

4.4.1 Aspect overview

While tense positions an event on the time axis, *aspect* describes the relation of the event or action with respect to the passage of time [Quirk *et al.*, 1991]. Aspect is attached to the verb expressing the action or occurrence by means of inflection or use of auxiliary verbs. In general, aspect indicates whether an event or action is completed or not with respect to a certain point in time. In some cases, it is not possible to decide if an event or action is completed or not, making the occurrence unmarked for aspect, or indefinite.

The *American Heritage Dictionary of the English Language* [American Heritage Dictionary, 2000] defines grammatical aspect as follows:

> A category of the verb designating primarily the relation of the action to the passage of time, especially in reference to completion, duration, or repetition.

The *Webster's Encyclopedic Unabridged Dictionary* [Webster Encyclopedic, 2001] defines grammatical aspect as:

> A category or interrelated set of categories for which the verb is inflected in some languages, typically to indicate the duration, repetition, completion, or quality of the action or state denoted by the verb.

[11] VBP and VBZ are present, VBD is past and MD "will" is future.

The HyperGrammar [Megginson *et al.*, 2002] of the University of Ottawa defines aspect as the nature of the action described by the verb and mentions three aspects: indefinite, complete and continuing. The HyperGrammar also offers alternative denominations: simple aspect, perfect aspect and progressive aspect respectively. The indefinite aspect is defined as an aspect which is used when the beginning or ending of an action, an event, or condition is unknown or unimportant to the meaning of the sentence, or to indicate a habitual or repeated action. The complete aspect is defined as an aspect which is used to indicate the fact that the end of the action, event, or condition is known and is used to emphasize the fact that the action is complete, be it in the past, present or future. The continuing aspect is defined as an aspect which indicates that the action, event, or condition is ongoing in the past, present or future. The aspects above can be combined to show that an action was in progress at one point and then was completed [Megginson *et al.*, 2002].

Klein [Klein, 1994] follows Reichenbach's [Reichenbach, 1947] semantic theory in the sense that he is using notations equivalent to the points of reference, but he separates the two references of **TT to TU** and **TT to Tsit** into tense and aspect, respectively. In Klein's theory, the three points are *time of the utterace (TU)*, similar to Reichenbach's *point of speech (S)*, *situation time (Tsit)*, similar to Reichenbach's *point of the event (E)*, and *topic time (TT)*, similar to Reichenbach's *point of reference (R)*.

Randolph Quirk defines aspect as being the grammatical category which reflects the way in which the verb action is regarded or experienced with respect to time and considers two kinds of aspects: perfective aspect and progressive aspect. Syntactically, aspects are defined based on the ABCD structure of the verb phrase [Quirk *et al.*, 1991], the type B (HAVE + -ed participle) corresponding to the perfective aspect and the type C (BE + -ing participle) corresponding to the progressive aspect. The author also treats perfective progressive and simple aspects [Quirk *et al.*, 1991].

Some sources combine aspect and tense under the definition of tense [Reichenbach, 1947; Allen, 1995], but while tense and aspect are closely related, here we choose to follow theories which separate the two. As Quirk notes [Quirk *et al.*, 1991], there are two kinds of realization: the morphological realization of tense and the syntactic realization of aspect. In practical terms, there are two different algorithms, one for analyzing tense based on the morphological features of the operator verb, and one for

analyzing aspect based on the syntactic features of the verb cluster. Thus, we prefer the non-unified theory of tense and aspect.

The syntactic definitions used for the purpose of the verb cluster analysis implementation are based on Quirk's work [Quirk *et al.*, 1991]. There are three types of aspects: indefinite, perfect and progressive and one complex type, perfect progressive.

4.4.2 Aspect implementation

Based on Quirk's definitions [Quirk *et al.*, 1991], any verb cluster containing a *have* (type B) auxiliary, is considered as having a perfective aspect, any verb cluster containing a *be* auxiliary followed by a gerundive (type C) as having a progressive aspect, any verb cluster containing both a *have* auxiliary (type B) and a *be* auxiliary followed by a gerundive (type C) as having a perfect progressive aspect and the verb clusters not containing a type B or C auxiliary as having an unmarked aspect, or indefinite, as shown in the table Table 4.5.

				VBP	indefinite	
				VBD	indefinite	
WILL_MD				VB	indefinite	A
				VBZ	indefinite	
			BE_VBP	VBN	indefinite	D
			BE_VBZ	VBN	indefinite	D
			BE_VBD	VBN	indefinite	D
WILL_MD			BE_VB	VBN	indefinite	AD
	HAVE_VBP			VBN	perfect	B
	HAVE_VBZ			VBN	perfect	B
WILL_MD	HAVE_VB			VBN	perfect	AB
	HAVE_VBD			VBN	perfect	B
	HAVE_VBP		BE_VBN	VBN	perfect	BD
	HAVE_VBZ		BE_VBN	VBN	perfect	BD
	HAVE_VBD		BE_VBN	VBN	perfect	BD
WILL_MD	HAVE_VB		BE_VBN	VBN	perfect	ABD
	HAVE_VBP	BE_VBN		VBG	perfect progressive	BC
	HAVE_VBZ	BE_VBN		VBG	perfect progressive	BC
	HAVE_VBD	BE_VBN		VBG	perfect progressive	BC
WILL_MD	HAVE_VB	BE_VBN		VBG	perfect progressive	ABC
	HAVE_VBP	BE_VBN	BE_VBG	VBN	perfect progressive	BCD
	HAVE_VBD	BE_VBN	BE_VBG	VBN	perfect progressive	BCD
WILL_MD	HAVE_VB	BE_VBN	BE_VBG	VBN	perfect progressive	ABCD
	HAVE_VBZ	BE_VBN	BE_VBG	VBN	perfect progressive	BCD
		BE_VBP		VBG	progressive	C
		BE_VBZ		VBG	progressive	C
		BE_VBD		VBG	progressive	C
WILL_MD		BE_VB		VBG	progressive	AC
		BE_VBP	BE_VBG	VBN	progressive	CD
		BE_VBZ	BE_VBG	VBN	progressive	CD
		BE_VBD	BE_VBG	VBN	progressive	CD
WILL_MD		BE_VB	BE_VBG	VBN	progressive	ACD

Table 4.5: Aspect illustration

We remark that the aspect table can be converted into a Venn diagram as in Figure 4.3.

As seen on the diagram, any verb cluster containing a form of present participle (VBG) will have the word "progressive" in it's aspect denomination, any verb cluster of type B, containing a form of the HAVE auxiliary will have the word "perfect" in it's aspect denomination. If none of these forms was encountered, and the verb cluster is finite (has a tense), then the aspect is indefinite.

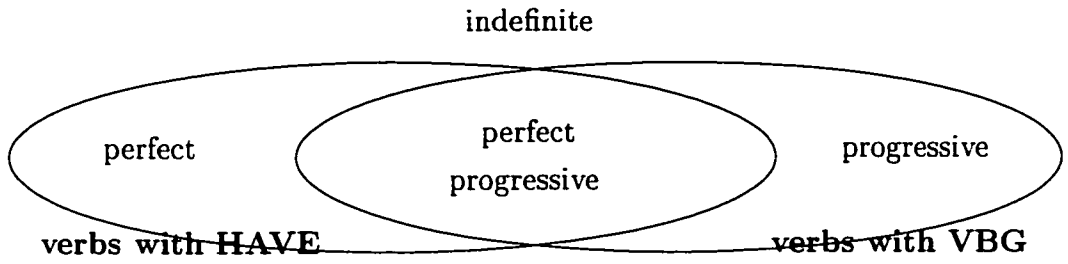

Figure 4.3: Venn diagram describing verb aspect

4.5 Modality

This section will define the term "modality" and overview a few different ways of analyzing and categorizing this grammatical category. As with the previous terms, a decision will be made as to how exactly modality will be treated for the purpose of the programming implementation of the verb cluster analysis. Syntactic and semantic definitions will be provided.

4.5.1 Modality overview

Modality, Mood and Mode are used in the same sense by some grammarians, while others differentiate between these grammatical categories. In general, the grammarians who differentiate between Modality and Mood consider Modality as a grammatical category used to express the factuality, nonfactuality, possibility, impossibility, necessity, etc.

The *American Heritage Dictionary of the English Language* [American Heritage Dictionary, 2000] defines modality as follows:

> The classification of propositions on the basis of whether they assert or deny the possibility, impossibility, contingency, or necessity of their content. Also called mode.

The *Oxford Companion to the English Language* [McArthur, 1992] defines modality as:

> In syntactic and semantic analysis, a term chiefly used to refer to the way in which the meaning of a sentence or clause may be modified through the use of a modal auxiliary, such as may, can, will, must. In a wider sense, the term is used to cover linguistic expression of these concepts other than through the modal auxiliaries: 'It will possibly rain later this evening'; 'I am sure that the plane has landed by now' [...] Adverbs such as possibly, perhaps, probably, certainly have been called modal adverbs, and such adjectives as possible have been called modal adjectives. The term is also extended to include the subjunctive mood and the past verb forms used to express hypothetical meaning [...] In case grammar, modality refers to one of the two underlying constituents of sentence structure (the other being proposition). The modality includes those features that relate to the sentence as a whole, such as tense and negation.

The Summer Institute of Linguistics [Loos *et al.*, 2002], defines modality as a facet of illocutionary force that expresses the illocutionary point or general intent of the speaker or the speaker's degree of commitment to the expressed proposition's believability, obligatoriness, desirability, or reality. SIL defines five types of modality: irrealis, realis, alethic, deontic and epistemic.

Biber [Biber *et al.*, 1999] classifies modality in three groups, showing a correspondence between each group and a group of modal verbs. One group denotes permission, possibility or ability and is expressed by the use of the modal verbs *can, could, may or might.* A second group denotes obligation or necessity and is expressed by the use of the modal verbs *must or should*, or the modal idioms *had better, have to, have got to, ought to* or *be supposed to.* A third group denotes volition or prediction and is

expressed by the use of the modal verbs *shall, will or would*, or the modal idiom be *going to*.

Quirk defines modality as the manner in which the meaning of a clause is qualified so as to reflect the speaker's judgment of the likelihood of the proposition it expresses being true and it serves to indicate in general terms the function which the modal auxiliary verbs perform in the language [Quirk *et al.*, 1991]. The author splits modality into two types: intrinsic and extrinsic. Intrinsic modality implies a meaning such as permission, obligation and volition which involves some kind of intrinsic human control of events. The extrinsic modality implies a meaning such as possibility, necessity and prediction, which do not primarily involve human control of events, but do typically involve human judgment of what is or is not likely to happen. He mentions that due to the overlapping between the intrinsic and extrinsic senses of modal verbs it is very difficult to categorize them.

Other sources define modality as simply the fact that a modal auxiliary verb is present, usually indicating that the clause does not present real facts [Allen, 1995], [Megginson *et al.*, 2002], etc.

For implementation purposes, modality is defined as being irrealis if a modal verb other than *will* is present.

4.5.2 Modality field implementation

It is possible that in some future work modality will be treated in more depth than here, using one of the definitions presented in the previous section. For the current analysis of verb clusters, I am only interested in the fact that the verb cluster contains a modal auxiliary marking an irrealis modality. If a verb cluster contains a modal auxiliary that is not a future marker, its modality frame field is marked as "irrealis".

4.6 Mood

This section will define the term "mood" and give a short overview of the different ways of analyzing and categorizing this grammatical category. A decision will be made as to how exactly mood will be treated for the purpose of the programming implementation of the verb cluster analysis.

4.6.1 Mood overview

As mentioned in the previous chapter, many grammarians mix or combine the terms mood, modality and mode. The mood field was added to the frame of the verb cluster to show that the verb cluster has an inversion between the operator and the object noun phrase, which usually denotes an interrogative mood. While the current implementation does not look at mood more in detail, future implementations treating wh–questions, imperatives and other moods will use this field. Below, a short overview of the grammatical category *mood.*

The *American Heritage Dictionary of the English Language* [American Heritage Dictionary, 2000] defines grammatical mood as follows:

> A set of verb forms or inflections used to indicate the speaker's attitude toward the factuality or likelihood of the action or condition expressed. In English the indicative mood is used to make factual statements, the subjunctive mood to indicate doubt or unlikelihood, and the imperative mood to express a command. [Alteration of mode.]

The *Oxford Companion to the English Language* [McArthur, 1992] defines grammatical mode and mood as synonyms:

> As a synonym of mood. In traditional grammar, a term for a form of the verb that affects the general meaning of the sentence and for the sentence or clause type in which it occurs. Three moods are customarily recognized for English: the indicative (God helps us); the imperative (Help us); and the subjunctive (God help us).

In the *Modern English Grammar HyperBook* by Daniel Kies [Kies, 2002], the Mood system is described as being divided into four subcategories: indicative, interrogative, imperative and subjunctive. The indicative mood indicates that the speaker is making a statement, the interrogative mood indicates that the speaker is asking a question, the imperative mood indicates that the speaker commands, exhorts or requests and the subjunctive mood indicates that the speaker is stating that the action is unlikely (a wish, prayer, hope, etc.)

Allen [Allen, 1995] defines mood as the way in which a sentence is used, such as asserting that a sentence is true, asking whether a sentence is true, or command

someone to do something described in the sentence. The author specifies four basic sentence moods: declarative (assertion), yes/no question, wh–question and imperative (command).

SIL [Loos *et al.*, 2002] defines mood as a set of distinctive forms which are used to signal modality and specifies quite a few types of moods: assumptive mood, declarative mood, deductive mood, deliberative mood, directive mood, dubitative mood, hypothetical mood, immediate imperative mood, imperative mood, imprecative mood, interrogative mood, jussive mood, obligative mood, optative mood, permissive mood, precative mood, prohibitive mood, speculative mood and subjunctive mood.

The HyperGrammar [Megginson *et al.*, 2002] defines three moods: indicative mood, imperative mood, and subjunctive mood. The indicative mood is used to express facts and opinions or to make inquiries. The imperative mood is used to give orders or to make requests. The subjunctive mood is used in dependent clauses to express unreal conditions, in dependent clauses following verbs of wishing or requesting, in dependent clauses attached to an independent clause that uses specific verbs ("ask", "demand", etc.) or in a dependent clause attached to an independent clause that uses an adjective of urgency such as "essential", "crucial" etc.

In the current implementation, the mood field does not have all the values defined yet, for now only the interrogative mood (question) is recognized. Moods added in the future could include indicative, imperative and subjunctive.

4.6.2 Mood field implementation

This field was created to hold the information that the verb cluster presents a noun phrase object and operator inversion. If an inversion is present, the mood is marked as "interrogative". In the future, when the whole sentence will be analyzed, this field can also be marked with other information as "imperative", "subjunctive", etc.

4.7 Voice

This section will give a short overview of the two voices present in the English language and describe a simple algorithm for the implementation of voice analysis of the verb cluster.

4.7.1 Voice overview

Voice is the only category where all the texts consulted agreed on its types: active and passive.

The *Webster's Revised Unabridged Dictionary* [Webster Revised, 1998] defines grammatical voice as follows:

> A particular mode of inflecting or conjugating verbs, or a particular form of a verb, by means of which is indicated the relation of the subject of the verb to the action which the verb expresses.
>
> Active voice: that form of the verb by which its subject is represented as the agent or doer of the action expressed by it.
>
> Passive voice: a verb, or form of a verb, which expresses the effect of the action of some agent.

The *Oxford Companion to the English Language* [McArthur, 1992] defines grammatical voice as:

> A category that involves the relationship of subject and object in a sentence or clause. In English, the contrast is between active voice and passive voice, affecting both the structure of the sentence and the form of the verb: Susan chose the furniture is an active sentence whose corresponding passive is The furniture was chosen by Susan. The two sentences have the same truth value, though there are differences in style and emphasis, in that passives are usually more formal than actives and the end of a sentence or clause tends to have the greatest emphasis. The by–phrase is often omitted from the passive sentence, especially in technical writing, producing an agentless passive. The passive verb adds auxiliary be to the corresponding active verb, and is followed by the passive participle of the main verb (–ed in regular verbs). The distinction between active and passive applies only when the verb is transitive, since only a transitive verb can be accompanied by an object. Typically, the active subject is the doer of an action (Ted in Ted was repairing the computer), whereas the passive subject (like the active object) is the person or thing affected by the action (the computer).

Quirk [Quirk *et al.*, 1991] gives a syntactic definition of voice as: a verb cluster containing a construction of a form of *be* followed by the main verb in form of a past participle (type D) is passive, otherwise it is active. However, the definition of passive voice being any verb cluster that contains a BE form + a -ed past participle (VBN Brill tab) would let constructions without a passive agent be considered as passive. Quirk defines a passive gradient to take care of the cases in which the syntactic definition of voice does not correspond to a semantic definition of voice, the gradients being: central (true) passive, semi-passives and pseudo-passives.

This implementation will ignore Quirk's finer semantic distinctions and consider only active and passive voice, defined syntactically.

4.7.2 Voice implementation

As defined by Quirk [Quirk *et al.*, 1991], we consider as passive any construction that contains an auxiliary BE form and a past participle (VBN), as seen in the table Table 4.6. Any other construction is considered as active.

Implementation-wise, this translates to an AND test on a BE auxiliary and a VBN main verb. A BE auxiliary marker can be set once one is encountered, the final test being done once we arrive at the end of the parse (main verb), if there is a BE auxiliary marker and the main verb is in the VBN form, the voice will be set to "passive", otherwise it will be set to "active".

The above paragraph describes the algorithm used to determine the voice of a verb cluster.

4.8 Polarity

The polarity field was added to the frame of the verb cluster in order to hold the information that there are adverbs connected to the cluster that denote a negation. The current implementation looks at simple adverbs such as "not" and "n't" and does not go into more depth regarding the negativity of a sentence. Quirk [Quirk *et al.*, 1991] and Baker [Baker, 1989] provide great syntactic definitions of negation formation that could be used to improve the treatment of polarity in a future version.

In the current implementation, if any of the adverbs in the verb cluster has a negative polarity ("not", "n't"), the polarity is set to negative.

				VBZ	active	
				VBP	active	
				VBD	active	
WILL_MD				VB	active	A
WILL_MD		BE_VB		VBG	active	AC
WILL_MD	HAVE_VB	BE_VBN		VBG	active	ABC
	HAVE_VBP	BE_VBN		VBG	active	BC
	HAVE_VBZ	BE_VBN		VBG	active	BC
	HAVE_VBD	BE_VBN		VBG	active	BC
		BE_VBP		VBG	active	C
		BE_VBZ		VBG	active	C
		BE_VBD		VBG	active	C
	HAVE_VBZ			VBN	active	B
	HAVE_VBP			VBN	active	B
	HAVE_VBD			VBN	active	B
WILL_MD	HAVE_VB			VBN	active	AB
			BE_VBP	VBN	passive	D
			BE_VBZ	VBN	passive	D
			BE_VBD	VBN	passive	D
		BE_VBP	BE_VBG	VBN	passive	CD
		BE_VBZ	BE_VBG	VBN	passive	CD
		BE_VBD	BE_VBG	VBN	passive	CD
	HAVE_VBP		BE_VBN	VBN	passive	BD
	HAVE_VBZ		BE_VBN	VBN	passive	BD
	HAVE_VBD		BE_VBN	VBN	passive	BD
	HAVE_VBP	BE_VBN	BE_VBG	VBN	passive	BCD
	HAVE_VBZ	BE_VBN	BE_VBG	VBN	passive	BCD
	HAVE_VBD	BE_VBN	BE_VBG	VBN	passive	BCD
WILL_MD			BE_VB	VBN	passive	AD
WILL_MD		BE_VB	BE_VBG	VBN	passive	ACD
WILL_MD	HAVE_VB		BE_VBN	VBN	passive	ABD
WILL_MD	HAVE_VB	BE_VBN	BE_VBG	VBN	passive	ABCD

Table 4.6: Voice illustration

4.9 Reporting verbs

Reported speech clauses can be recognized properly only if the verb clusters containing a main verb which is a reporting verb are recognized. While quoted direct reported speech clauses could be easily found without knowing beforehand which of the verb clusters contains a reporting verb, this becomes quasi impossible for unquoted indirect reported speech clauses, as the punctuation used is not enough to determine whether the sentence is a reported speech sentence or another type of sentence. The verb clusters containing reporting verbs are the most important ones for the treatment of reported speech, and while other verb clusters could be ignored during the analysis stage, reporting verbs must be analyzed. The AETNA Group has compiled a list of reporting verbs.

4.10 Semantic dimensions

As described in [Bergler, 1992; Bergler, 1993], reporting verbs are characterized by their semantic dimensions. Semantic dimensions are an important step for the belief analysis of the reported speech. There are nine semantic dimensions: voice quality, explicitness, formality, audience, polarity, presupposition, speech act, affectedness and strength, as defined by Sabine Bergler.

Semantic dimensions are looked up in a dictionary in order to fill the semantic dimensions field. Christine Gerard [Gerard, 2000] has built a lexicon for reporting verbs in newspaper articles, which also contains their semantic dimensions.

4.11 Summary

The verb cluster frame (template) built by the analyser encodes information such as tense, aspect, modality, mode, voice, polarity and semantic dimensions which can be used for the evidential analysis of reported speech.

Reporting verbs are identified through pattern matching on a list of reporting verb forms and the verb cluster grammar built can be integrated with the existing noun phrase grammar based on [Bergler, 1997] to extract the reported speech sentences from newspaper articles and build the corresponding basic profiles. Mechanisms have been defined to extract the information for the fields of the verb cluster.

Chapter 5

Profiling

In this chapter I will first present a brief overview of profiles as defined in [Bergler, 1992]. I will then incrementally present the detailed analysis steps which are needed for the implementation of basic profiles from reported speech in newspaper articles. The structure of reported speech is studied in detail. Once the reported speech is analyzed, it is but an easy step to build basic profiles. I then show how the information stored in basic profiles is enough to build full and embedded profiles.

5.1 Overview

Profiling is the process of transforming reported speech sentences into a structured representation called a *profile*.

This method of representing text was developed by Bergler [Bergler, 1992], where a *profile* is defined as:

> A profile contains a list of all properties the text asserts or implies about a particular discourse entity. Distinct discourse entities have separate profiles.
>
> Profiles are a collection of information that has to do with one specific discourse entity: attributes and features of that entity, such as being blond or being a CEO; expressed thoughts and beliefs attributed to the entity, such as being concerned about the environment or believing that the recession will end soon; and generally all utterances made by that entity, such as announcing a joint venture or saying "Oh, no!".

Profiling discourse entities from a newspaper article is very useful, the information being all organized by entities, facilitating the process of summarizing, discourse analysis, text understanding and so on, and we as humans do profiling without even thinking too much. When we read a newspaper article and one of the sources we are interested in is quoted, we always put this quote in the context of previous speech acts from the same source, especially if they are contradicting.

When a source is a member of another source, profiles can be embedded into each other, such as the *Chair of a Department* embedded into the profile of the *Department*[1]. Bergler uses nested boxes, such as the ones used by Ballim and Wilks [Ballim and Wilks, 1992] to represent nested beliefs.

Another way of linking profiles is by grouping them according to their viewpoints about a certain subject [Bergler, 1992]. Depending on their stance vis-a-vis the matter described, discourse entities can agree, disagree or be neutral. Depending on their agreements, sources can be grouped in *supporting or opposing group structures.* A supporting group is when different sources agree, supporting a same issue. When a newspaper article presents an argument, there can be groups that support an issue, or oppose that issue.

For example[2], when a reporter describes a situation where two different entities have gone to court over an issue, there will be a plaintiff side, of all the discourse entities that support the party that instituted the court suit, there will be a defendant side, of all the discourse entities that support the party against which the action was brought, and there will be the court side, which is supposed to be neutrally subjective in listening to both parties and which is usually the judge and/or the jury. Each of these sides has a corresponding supporting group of discourse entities, the plaintiff supporting group being in contradiction with the defendant supporting group, which makes them opposing supporting groups.

Grouping profiles requires a semantic analysis of the reported material for each source and of the text, and possibly a full text analysis in the worst case scenario, however, a partial analysis is possible [Bergler, 1992].

[1] A source can also be a corporation or a group of people which is given human properties for the purpose of reported speech. For example: *Microsoft said that they are working on fixing the security flaws.* or *The Department announced the mailboxes will be moved to a different location.*

[2] For a much more detailed example describing this situation, including the graphical representation of supporting and opposing groups, see [Bergler, 1992], pages 186–189.

Another way of organizing information that will be briefly mentioned here is temporal. Based on verb tenses and other temporal modifiers, the information in a text can be partially organized chronologically, using trace fragments [Bergler and Pustejovsky, 1990]. A *trace* shows the chronological sequence in which events can be placed on the narrative time line and is constructed in a bottom–up manner by trace composition and trace unification. Trace is a different representation of text compared to profiles.

When complex profiles are built, the text is transformed into a non linear representation of the original text. A *coherence structure* can be used to hold the information about the linear structure of the original text, as well as the relationships between the sentences in the original text [Bergler, 1992; Bergler, 1995a].

These three methods of representing text — profiles, traces and coherence structure — form a *tripartite representation* which provides a complete and concise description of the original textual matter, a *Minimal Text Representation* scheme[3].

5.2 Components of the structure of reported speech

As Dooley and Levinsohn put it [Dooley and Levinsohn, 2000], reported speech tends to not follow the normal structure of narratives. Especially for newspaper articles, reported speech has a much more complex structure than normal sentences, with heavier noun phrases and use of embedded information [Bergler, 1992].

As described in [Bergler, 1992] and summarized in Chapter 3, reported speech can be analyzed as a binary structure of primary information and circumstantial information. The primary information is usually provided by one or more original sources and conveyed by the reporter through the use of direct or indirect reported speech. The circumstantial information is provided by the reporter in order to identify the source, location, time or other information about the circumstances of the original speech act. For the purpose of the implementation, we need to split the circumstantial information into source, reporting verb, and other circumstantial information.

In the next few sections, I will discuss the different parts which can form a reported speech sentence: the source, the reporting verb, the primary information and the other circumstantial information. The primary information can be conveyed through the

[3]For a detailed definition of MTR, see [Bergler, 1992].

use of direct quotes or paraphrases, or a mix of the two. As noted by Doran [Doran, 1998; Doran, 1996], the analysis of punctuation is pivotal, and the rules defined in the following sections take punctuation into account.

Finally, a short section will look at how all these parts fit together to form the structure of reported speech sentences.

The analysis is done for American newspaper usage, where terminal commas and end-of-sentence punctuation marks are moved inside the quotes. Doran [Doran, 1998] notes that in an analysis of 2.5 million words of the Wall Street Journal corpus, there were only 30 instances of punctuation in the British English style.

5.2.1 Source

The source is the person, personified entity or document which supplies the primary information the reporter is conveying to the reader. A personified entity is an entity such as a corporation, organization, or other group of people which can be personified in order to attribute speech acts to them.

In newspaper style, the source is usually a heavy noun phrase, in which the reporter encoded information about the source. The implementation uses NPE-0 based on [Bergler, 1997] to extract source noun phrases.

5.2.2 Reporting verb cluster

The reporting verb links the Source to the original speech act presented through the reporter. A reporting verb is always the main verb of a verb cluster[4].

As discussed in [Bergler, 1992; Bergler, 1993], the choice of reporting verb made by the reporter conveys information about the reporter's beliefs[5] about the original speech act, as well as the reporter's trust in the source and other information related to the conditions of the original speech act. The tense of the reporting verb provides a temporal point for the information relayed and the aspect provides the temporal frame.

[4] A detailed study of verb clusters is presented in Chapter 4.

[5] See [Bergler, 1992; Bergler, 1993] for a framework for decoding the reporter's beliefs and cues, through the use of semantic dimensions.

Currently, reporting verbs are matched against a list, and exceptions to the heuristics are very few and far between. It is very rare[6] that a reporting verb form appears in a position in which it is not a direct reporting verb, such as in *The FCC "didn't say he couldn't do it again," Uhr said.* In this example, "didn't say" is not a direct reporting verb, but it belongs to a reported clause embedded in another reported clause: the reporter reports that Uhr reported that the FCC didn't say he couldn't do it again. The current grammar does not treat nested reported speech, however, it might be interesting to consider it in future research.

5.2.3 Primary information

Bergler [Bergler, 1992] calls *primary information* the main information that the reporter is conveying to the reader, which helps move the story along. In newspaper articles, the primary information is provided by the sources which the reporter consulted and provides the **what** of the article. The reporter may choose to report this information as a direct reported speech act or as a paraphrase or summary of the original speech act. We differentiate between direct quotes (text embeded within quotation marks[7]) and paraphrases. These usually correspond to direct reported speech and indirect reported speech, but there can be exceptions.

It is important to differentiate between direct quotes and paraphrases, because the fact that the reporter used quotation marks shows that the reported clause is a literal excerpt of the original speech act.

Direct reported speech (direct quote)

Direct reported speech is usually marked by the use of quotation marks. Depending on the language in which the text is written, as well as the domain in which the text appears, direct quotes can be embedded between a few different punctuation marks.

[6]Less than 3% of reported speech sentences.

[7]Note that there is no way to know if the quotation punctuation around a short text attached to a reported clause are scare quotes or quotation marks without looking at the general context or without having a way to recognize coinages and other expressions which are usually used within scare quotes. Even for a reader, it can be difficult to determine if the reporter used scare quotes or quoted something that the original source said. The Text Encoding Initiative (TEI) guidelines [Sperberg-McQueen and Burnard, 2002; TEI Consortium *et al.*, 2003] propose a <soCalled> tag which can be used to differentiate between quotation marks and scare quotes, which would replace quotes in the markup of an electronic text. However, the Wall Street Journal corpus does not follow these guidelines and for now I will not study the issue.

In her analysis of punctuation in direct and indirect speech, Christine Doran [Doran, 1998] gives a list of possible punctuation marks that can be used for quoting, i.e. guillemets, dashes, colons, single apostrophes, double apostrophes, double quotes, in raised or lowered position. In the Wall Street Journal corpus, the standard direct reported speech uses double quotes (" "). The directly quoted text may or may not be a completely formed sentence.

Relative to the direct quote, the source and reporting verb can appear in initial position as in (1), medial position[8] (2), (3) or final position (4), (5). When the source and reporting verb appear in initial position, subject-verb inversion does not usually appear in newspaper articles[9]. If the source and reporting verb occur in medial (3) or final position (5), inversion can occur if the verb is in the simple present or simple past. When the source/reporting verb are in initial position, the direct quote may be introduced by the word "that". The case where the direct quote is presented before the source and the reporting verb is by far the most common in the Wall Street Journal corpus.

(1) Mr. Felten said, "We got what amounted to a parking ticket, and by complaining about it, we ended up with a sizable fine and suspension." (WSJ 11/02/89)

(2) "By the end of the 1990s," he said, "we want to be producing roughly two vehicles overseas for every vehicle that we export from Japan." (WSJ 10/30/89)

(3) "There may be sticker-shock reaction initially," said Mr. Pratt, "but as the wine is talked about and starts to sell, they eventually get excited and decide it's worth the astronomical price to add it to their collection." (WSJ 11/02/89)

[8] A variant of the medial position is when there is a sentence with a syntactic structure of type 4/5 or 9/10 (figures 5.1 and 5.2) followed by a hanging direct quote which is presented in a different sentence, such as in this modified example: *"There is incredible negative psychology building in the market," said Donna Avedisian. "People are very concerned about who is going to step up to the plate and buy municipal bonds in the absence of institutional buyers."* In this case, the hanging direct quote inherits the subject and the reporting verb from the previous full reported speech sentence. This however has very few occurences in the Wall Street Journal corpus, reporters seeming to prefer the fully formed reporting sentences. It does appear more often in other sources though, such as CNN.

[9] Technically, it is possible to encounter inversion of subject-verb in initial position when using a colon to introduce a direct quote.

(4) "We look upon this as a great opportunity to prove the fact that we have a tremendous management team," he said. (WSJ 11/02/89)

(5) "This is the first of many rumors we expect to hear during the sale's process," said a Bloomingdale's spokesman. (WSJ 10/26/89)

In the case of two adjoining reported sentences, it is possible to have omission of the subject, which is inferred from the adjoined previous reported sentence. In this case, only the reporting verb will be present and it will be positioned before the direct quote (6).

(6) He said Sassy will keep its irreverent tone, but added, "We will keep a close watch on the editorial content of the magazine." (WSJ 10/19/89)

Another possible instance of source omission is when using a non–finite reporting verb to introduce another direct quote (7). The non–finite verb is always followed by a comma or "that". As in the previous case, the source is inferred from the previous reported sentence.

(7) "I never agreed to it," Mr. Alexander says, adding that "it's not necessary to pay these nuisance settlements." (WSJ 10/17/89)

The possible syntactic structures[10] stemming from the possible relative positions between the direct quote and the reporting subject/verb, and from the possibility or not of subject–verb inversion are displayed in Figure 5.1. In the grammar, the Source can be a complex NP, the Reporting verb can be a simple reporting verb or a verb cluster whose main verb is a reporting verb, and the direct quote can be any text or punctuation inside quotation marks[11]. The square brackets [ ] mark optionality.

[10]The syntactic structures 1, 2 and 3 have been adapted from [Bergler, 1992]. Structures 4 and 5 have been adapted from [Gerard, 2000]. The structure numbers correspond to the example numbers.

[11]I am currently not looking at tense backshift in verbs inside the direct quotes, but the current implementation makes this possible if we want to analyze backshift in future work

1. <Source> <reporting verb> [that ∨ ,] "<direct quote>"
2. "<Part of direct quote>," <source> <reporting verb>, "<rest of direct quote>"
3. "<Part of direct quote>," <reporting verb> <source>, "<rest of direct quote>"
4. "<Direct quote>," <source> <reporting verb>
5. "<Direct quote>," <reporting verb> <source>
6. <coordinating conjunction> <reporting verb> [that ∨ ,] "<direct quote>"
7. <non-finite reporting verb> <that ∨ ,> "<direct quote>"

Figure 5.1: Syntactic structures for reported speech with direct quotes

Indirect reported speech (paraphrase)

As mentioned before, indirect reported speech is defined syntactically by the absence of quotation marks, even if it's not always a restatement of the original speech act in different words from the original. Throughout the thesis this may be also called a *paraphrase*. Paraphrases are very difficult to analyze syntactically, as they can take many different forms, and often can be mistaken for other types of clauses. Here, I will attempt to give a rough syntactic description of the paraphrase, corresponding to the most frequent possible occurrences.

Usually, in indirect speech, the reported clause is in final position and takes the form of a that-clause [Quirk *et al.*, 1991] following a subject and a reporting verb or reporting verb cluster (8a), but it is also possible to have a paraphrase which is not introduced by any special word that we could recognize, such as the second example in (8b). Inversion of subject-verb cannot occur in this case.

(8) (a) Premier Shamir denied that Israel provided any technology to Pretoria. (WSJ 10/27/89)

(b) Mr. Ward has previously denied any wrongdoing. (WSJ 10/12/89)

A paraphrase can also come in initial position, in which case it has to be separated

by a comma from the source and the reporting verb (9), (10). Inversion of subject-verb can occur, such as in example (10).

(9) An Asian bloc isn't intended, he said. (WSJ 11/01/89)

(10) Nixdorf, Bull and others will also sell versions of the machine, said Mips President Robert Miller. (WSJ 11/01/89)

The subject and reporting verb or reporting verb cluster can be inserted in medial position (11), (12). In this case, subject–verb inversion is possible (12) and the source and reporting verb have to be surrounded by commas.

(11) For the past two years, he said, he and the exchange's research department have been working on the new natural gas contract, seeking a good delivery site and studying the natural gas market. (WSJ 10/30/89)

(12) Large, heterogeneous election districts would encourage good government, said Madison, because a representative would be compelled to serve the interests of all his constituents and be servile to none. (WSJ 10/25/89)

As with direct quotes, it is possible to have source omission if it can be inferred from an adjoined previous reporting clause, such as two sentences adjoined by a coordinating conjunction (13) or the use of a comma followed by a non-finite reporting verb to introduce another sentence (14). The paraphrase may be introduced by the word "that".

(13) Mr. Dell said he doesn't expect a loss in either the third or fourth quarter, but said third-quarter earnings could be as low as four cents a share. (WSJ 10/20/89)

(14) The fourth quarter is usually the strongest for that business, he said, adding that the turnaround of Case isn't complete. (WSJ 10/12/89)

The syntactic structures[12] corresponding to the cases presented above are displayed in Figure 5.2. In the grammar, the Source can be a complex NP, the reporting

[12]The syntactic structure 8 in Figure 5.2 is adapted from [Bergler, 1992]. The structure numbers correspond to the example numbers.

8. <Source> <reporting verb> [that] <paraphrase>
9. <Paraphrase>, <source> <reporting verb>
10. <Paraphrase>, <reporting verb> <source>
11. <Part of paraphrase>, <source> <reporting verb>, <rest of paraphrase>
12. <Part of paraphrase>, <reporting verb> <source>, <rest of paraphrase>
13. <coordinating conjunction> <reporting verb> [that] <paraphrase>
14. <non-finite reporting verb> [that] <paraphrase>

Figure 5.2: Syntactic structures for reported speech with paraphrases

verb can be a simple reporting verb or a verb cluster whose main verb is a reporting verb and the paraphrase can be any text which is not inside quotation marks and does not belong to the source, reporting verb or reporting verb cluster, and is not other circumstantial information. The square brackets [ ] mark optionality.

Mixed direct quotes and paraphrases

Reported speech will not always be limited to only direct quotes or only paraphrases, the two forms not being exclusive. A reporter may choose to summarize most of the original speech act, but quote directly any portion of it. However, none of the procedures or inference mechanisms for direct quotes apply, thus mixed quotes are a subclass of paraphrases.

As with direct quotes and de facto paraphrases, the source noun phrase and the reporting verb cluster can occur in initial position, final position, or, in rare cases, medial position. For final and medial positions, inversion of subject-verb may be encountered.

When the source and the reporting verb are in initial position, they can be followed by a direct quote, preceded or not by the word "that", and followed by a paraphrase of the rest of the original speech act (15). The paraphrase and direct quote can also appear in reverse order, a paraphrase first and the directly quoted rest of the original

speech act afterwards (16). The reporter may also choose to direct quote only a few words in the middle of the original speech act, in which case there will be a paraphrase of the beginning of the original speech act, a direct quote of a part of the speech act and another paraphrase of the rest of the original speech act (17).

(15) U.S. Banknote said "there can be no assurance" a sale agreement would be concluded. (WSJ 10/17/89)

(16) Mrs. Hills said that the U.S. is still concerned about "disturbing developments in Turkey and continuing slow progress in Malaysia." (WSJ 11/02/89)

(17) The chancellor of the exchequer said he was leaving his cabinet post because "the successful conduct of economic policy" wasn't possible so long as Sir Alan Walters remained as her economic adviser. (WSJ 10/27/89, shortened.)

When the source and reporting verb appear in final position, there are three usual structures that can be encountered in newspaper articles. A direct quote of a part of the original speech act, a paraphrase of the rest of the speech act, a comma and the source and reporting verb in final (18) or second to final (19) position. A paraphrase of a part of the original speech act, followed by a direct quote of the rest of the original speech act, followed by the source and the reporting verb, in normal (20) or inverted (21) order. A paraphrase of a part of the original speech act, followed by a direct quote of another part, followed by a paraphrase of the rest of the original speech act and ending with the source/reporting verb group which can be in normal (22) or inverted (23) order.

(18) "I don't think there's a lot in the wings" in other sectors of the economy to keep growth above 1%, he said. (WSJ 10/30/89)

(19) "The Georgia-Pacific bid may open the door to a new era of consolidation" in the paper industry, said Mark Devario of Shearson Lehman Hutton Inc. (WSJ 11/01/89)

(20) The buy-back "is really a comfort to those who want to buy the stock that there is a price floor," he said. (WSJ 10/25/89)

(21) Proceeds from the planned sale of the 250,000-square-foot building "will help reduce the debt incurred as a result of our July 1988 recapitalization," said a USG official. (WSJ 10/31/89)

(22) Costa Rica also would be able to pay overdue interest on its still-outstanding loans at "more favorable terms" than regular debtors, other U.S. bankers said. (WSJ 10/27/89)

(23) The Bear Stearns order that marked the late-day turnaround caused a "massive buying effort" as UAL jumped $20 a share to $170 in the last half hour, said Mr. Bates. (WSJ 10/25/89)

Finally, the last of the cases presented here is when the source noun phrase and the reporting verb cluster occur between direct quotes and paraphrases. The possible structures would be direct quote, followed by the matrix clause, followed by comma and a paraphrase (24), (25) or reversed: a paraphrase, followed by comma, followed by the matrix clause, followed by comma and by a direct quote (26), (27). Inversion of subject–verb may occur (25), (27).

(24) "We had underfunded Maalox for a year," he said, because the company was concentrating on research and development and promoting other drugs. (WSJ 10/20/89)

(25) "The Japanese apple market is very keyed to high quality," says David Lane, and so apples are more of a delicacy there than a big food commodity. (WSJ 10/25/89, shortened.)

(26) Britain's economic fundamentals, he said, "don't look very bright." (WSJ 10/18/89)

(27) Then, says Dr. Levy, "she woke up paralyzed." (WSJ 11/01/89)

The cases presented above show the frequent syntactic structures for mixed direct quotes and paraphrases. We do not claim these are all the possible occurences. More research needs to be done to determine a complete model of mixed direct quotes and paraphrases which takes into consideration the punctuation usage and the possible presence of other circumstantial information.

5.2.4 Circumstantial information other than source or reporting verb

While the primary information is usually provided by the sources and relayed by the reporter, the circumstantial information is provided by the reporter and contains meta–information which describes the circumstances in which the original speech act was made, or properties of the source [Bergler, 1992].

Bergler [Bergler, 1992] defines *circumstantial information* as the matrix clause and *primary information* as the complement clause. The circumstantial information is composed of source, reporting verb and other circumstantial information. However simple this might sound, in some cases, it is extremely difficult to differentiate at the sentence level between when a piece of text is other circumstantial information and when it is a paraphrase. Take for example the following reported speech sentences.

(28) (a) In the new show, he said, "we're going to spend $60,000 building a start-up house" for a young couple. (WSJ 10/26/89)

(b) And in a subsequent television interview, he said, "I think it is right that advisers do not talk or write in public." (WSJ 10/27/89 — note that this is a different article than the one above)

Even a human reader might need some extra thinking to figure out if some of the information is reported speech or a description of temporal / spatial / etc. circumstances by the reporter. It is even more difficult for a computer. At the sentence level, without the context and without a semantic analysis, the (a) and (b) sentences cannot be easily differentiated.

(28a) and (28b) are both unmodified sentences from the Wall Street Journal. For the first sentence, the context (not shown here) contains a sentence where the source criticizes his old show, making it clear that in (28a) the source is talking about his new show. For the second sentence, a human reader can see that there would be no coherence if the "false paraphrase" and the direct quote would be joined to form the reported speech, hence the first part of the sentence is other circumstantial information.

While this type of ambiguity is very hard to resolve computationally and would require a full treatment of the text, luckily it happens very rarely, as successful reporters usually try to disambiguate their articles, and editors work against ambiguity too.

The only problem that we might encounter is a "false circumstantial information" which would appear inside a paraphrase and belong to the original speech act.

Obviously, an in depth research of the syntactic differences between paraphrases and circumstantial information (other than source or reporting verb) needs to be done. This problem is mostly called prepositional phrase attachment and is known to be very difficult. We have defined a few simple heuristics which cover other circumstantial information to some extent, differentiate it from paraphrases and seemed to be successful on most of the random newspaper text we tested them on. However, a more detailed study needs to be done in the future, such as for example the fronting of circumstantial adverbials by certain spatial and temporal adverbs, as well as WH–words.

Other circumstantial information can also be treated as parenthetical information and there are some existing studies on parenthetical expressions, but none to my knowledge that looks at them with the level of granularity needed to differentiate between paraphrases and circumstantial information other than source or reporting verb. As a few studies mention, the analysis of comma positioning within the text can be used to point out parenthetical expressions. But as William Strunk [Strunk Jr. and White, 1979] acknowledges in his chapter describing the usage of commas, parenthetical expressions represent a difficult issue. He notes that parenthetic expressions should be enclosed between commas, but the commas may be omitted if the interruption to the flow of the sentence is not significant or in the case of restrictive relative clauses. In the case of newspaper articles, when a parenthetical expression is not between commas, it is usually because it is a simple temporal placement[13] such as *yesterday, during the briefing*, etc. For the second case, I assert that in a parser which can detect maximal length noun phrases the restrictive clauses should be joined to the simple noun phrase representing the source. Without the restrictive information, we can not reduce the sources domain in order to clearly identify the original source.

However, punctuation cannot be used alone to determine the differences I am looking for. How reliable can a comma be when Oscar Wilde himself is known to have spent a lot of time taking out a comma and then adding it again [Keyes, 1996]?

For now, based on a study of a part of the Wall Street Journal corpus, more specifically observed clichés in the WSJ style, I will define a few heuristics which I believe

[13]The current implementation does not do a semantic analysis to find temporal clauses. This case will be ignored for now.

cover most occurrences other circumstantial information encountered in newspaper articles in the Wall Street Journal corpus.

By far the most observed cliché is to introduce source describing information between the source and the reporting verb, separated by commas, as in: Paul Keough, *acting regional director for the Environmental Protection Agency in Boston*, said, "We hope that other large companies follow McDonald's lead to undertake similar programs." (WSJ 10/27/89) This case can only appear in the non inverted subject–verb situations. Most of the time, source describing circumstantial information in medial position are verb–less subordinates, or presenting an adjectival non–finite verb, such as "acting" in *acting regional director for the Environmental Protection Agency in Boston.* This type corresponds to the expression below.

<source>, <other circumstantial information>, <reporting verb>

This type of other circumstantial information also appears following a comma after a reporting verb–source formation when in final position within the sentence[14], as in: "This instant gratification overcomes the usual bias against savings," says John Skinner, *a University of Virginia economist.* (WSJ 10/27/89) This type corresponds to the expression below.

<reporting verb> <source>, <other circumstantial information>.

Just the above two observations cover an estimated 60% of the occurrences of other circumstantial information within newspaper articles.

Another observation is that if the reporter presents the matrix clause between two paraphrases, he will always treat the matrix clause as a parenthetical and enclose it between commas. One cannot have

* <paraphrase>, <source> <reporting verb> <that> <rest of paraphrase>.

If we encounter this formation, the first one is other circumstantial information reporting to the matrix clause.

[14]Except in sentences containing complex mixed direct quotes and paraphrases.

<other circumstantial information>, <source> <reporting verb> <that> <paraphrase>.

This observation generally covers the cases of parenthetical temporal and spatial circumstantial adverbials, without the need to analyze them semantically. This covers an estimated 20% of the occurrences of other circumstantial information within newspaper articles.

Also, spatial/temporal circumstantial adverbials might appear whithout a comma after a reporting verb when in final position within the sentence. In this case, the circumstantial information other than source or reporting verb is usually fronted by markers such as "at" or "during" (*he said at the conference, he said during the meeting*). There is also the case of other circumstantial information presenting the audience which can appear as a simple noun phrase in the same formation as the spatial/temporal circumstantial adverbials I just mentioned, such as in: "We're in the process of discussing an amended plan with the creditors and anticipate filing that amended plan shortly," Mr. Lorenzo told *reporters.* (WSJ 10/18/89)

There are other observations that we made, but we could find examples or think of examples when the same syntactic structure could appear as a paraphrase, so they will not be described here.

5.2.5 Structure of reported speech sentences

By now, we have the source, we have the reporting verb, we have the rest of the circumstantial information and we have the primary information (differentiating between direct quotes and paraphrases). The rules and information presented in the previous sections fit together to define the structure of reported speech in newspaper articles.

This structure can be extracted without a full text analysis in most of the cases. While this was defined for reported speech in newspaper articles and the structural study of some constructs has been based on newspaper articles, I am certain that the rules defined can also cover reported speech within other domains which make use of quotation marks for direct reported speech and of reporting verbs.

For a more thorough coverage, further research is needed on the subject of other

circumstantial information—paraphrase differentiation, which may include WH–word fronting and determining temporal cues (*last month, yesterday*, etc.).

In the next section I will show how the extraction of the structure of reported speech can be used for the basic profile building process.

5.3 Extracting the basic profiles

Bergler [Bergler, 1995a] defines *profile* as "a collection of all properties that a text asserts or implies about a particular discourse entity. In particular, the profile contains all statements made by or attributed to the entity."

A *basic profile* is a list of all attributes and features of a source, as well as all expressed thoughts and beliefs attributed to that source *in one reported speech sentence.*

With the rules and descriptions presented in the previous sections, the computer is able to extract the source, the reporting verb cluster, the direct quotes or paraphrases, and to a certain extent the other circumstantial information. Once all this data is linked to each other based on the appartenance to the same reported speech sentence and the source and reporting verb are analyzed, it forms a basic profile. In the basic profile, we might discern the source, the way the original speech act was made through the use of the reporting verb, the expressed thoughts and beliefs attributed to the source through the use of direct quotes or paraphrases and other attributes and features of the source and circumstances through the use of other circumstantial information.

For example, in the Wall Street Journal 10/27/89 (modified for exemplification purposes)

> (S1) *"The editorial side is complicated," Ms. Salembier added during the briefing,* (S2) *saying that editorial layoffs will be decided later.*

we have two basic profiles, one for each reported speech sentence.

The first basic profile[15] corresponds to (S1).

[15]Here, I only want to show the basic profile, without going into the details of the whole frame. Whenever the details are ommitted, the text "[detail omitted]" will appear.

```
#s(rs (text "\"/\" The editorial side is complicated ,/, \"/\" Ms. Salembier added during
 the briefing")
      (textsource "Ms. Salembier ")
      (textverb "added ")
      (reported_speech "\"/\" The editorial side is complicated ,/, \"/\" ")
      (textcirc "during the briefing ")
      (textdirq "\"/\" The editorial side is complicated ,/, \"/\" ")
      (textindirq "")
      [detail omitted]
```

This can be represented graphically using nested boxes, as described in [Bergler, 1992].

Ms. Salembier — during the briefing

- added "The editorial side is complicated"

The second basic profile corresponds to (S2).

```
#s(rs (text "saying that editorial layoffs will be decided later")
      (textsource "")
      (textverb "saying ")
      (reported_speech "that editorial layoffs will be decided later ")
      (textcirc "")
      (textdirq "")
      (textindirq "that editorial layoffs will be decided later ")
      (rep_source ())
      [detail omitted]
```

Note that being in the same sentence, this profile does not have an explicit source, the source being inferred from the previous reported clause. This will be clear in a later version when full profiles will be built.

... (points to the previous box)

- saying that editorial layoffs will be decided later

5.4 Building full profiles

Full profiles contain all attributes and features of a source, as well as all expressed thoughts and beliefs attributed to that source *in the whole newspaper article.* These profiles can be embedded in each other based on membership, representation, employment or part-of relations between sources [Bergler, 1995a].

As future work, full profiles can be built from basic profiles by using the information in the `coreferences` field for the source which will point to all other basic profiles which belong to the same full profile. The basic profiles for which the source is inferred belong to the last augmented full profile. Circumstantial information describing the source would be appended to the whole profile while temporal and spatial circumstantial adverbials will remain with the particular reported speech to which they referred. The source could be represented by the full name of the source[16], if encountered, eventually being prefixed by titles such as Mr., Ms., Mrs., Dr., etc., if encountered.

For a simple example of full embedded profiles, let us look at a text from the Wall Street Journal of 10/27/89, modified[17] for exemplification purposes:

> In late September, the Post, a well established newspaper, announced it was canceling its Sunday edition.
>
> Valerie Salembier, president of the Post, said the Sunday circulation has reached only about 250,000.
>
> "In any other city, 250,000 would be considered great, but it just wasn't enough in New York," added Peter Kalikow, owner and publisher of the Post.
>
> Ms. Salembier said about 30 people in circulation, ad sales and other business departments would lose their jobs. "What we don't know about is the number of layoffs on the editorial side," she said. "The editorial side is complicated," Ms. Salembier added during the briefing, saying that editorial layoffs will be decided later.

The corresponding graphical representation follows.

[16]Usually, the most descriptive noun phrase for a source is the one that appears the first in the newspaper article, subsequent noun phrases used for the same source tending to be simplified. Gender title information may appear in the first noun phrase, but more often it is introduced in a later one.

[17]The full unmodified article is reproduced in Appendix B.

the Post — a well established newspaper

- — in late September — announced it was canceling its Sunday edition

Valerie Salembier — president of the Post

- said the Sunday circulation has reached only about 250,000
- — during the briefing — said about 30 people in circulation, ad sales and other business departments would lose their jobs
- — during the briefing — said "What we don't know about is the number of layoffs on the editorial side"
- — during the briefing — added "The editorial side is complicated"
- — during the briefing — saying that editorial layoffs will be decided later

Peter Kalikow — owner and publisher of the Post

- added "In any other city, 250,000 would be considered great, but it just wasn't enough in New York"

In the previous nested boxes representation, one glance is enough to determine that there are three sources: the Post, Valerie Salembier and Peter Kalikow; and that Valerie Salembier and Peter Kalikow are part of the Post. The representation shows that the Post is a well established newspaper, that Valerie Salembier is its president and that Peter Kalikow is the owner and publisher. It also shows that the timeframe when the announcement was made is late September and that there was a briefing. And finally, for each source, all the reported speech acts which mirror the thoughts and beliefs of the sources, and the reporting verbs associated with them, which encode the thoughts and beliefs of the reporter.

Full embedded profiles can be grouped to form the highest form of organization of the text: supporting groups. For a detailed description of supporting groups, see [Bergler, 1992].

Chapter 6

Description of the system

6.1 Architecture

This fully automated implementation consists of several modules which are called sequentially, but can possibly be skipped in the case of the parsing of texts that are already annotated, or if one only wants the annotated text but not the analysis results. The flow chart is presented in Figure 6.1. On the flow chart, the preprocessor, the Brill tagger, the Earley parser and the noun phrase analyser are part of NPE-0 as of September 24, 2002. NPE-0 was developed by Dr. Bergler's research group.

The first module that is run on a newspaper article is a preprocessor developed by Bergler's research group, part of NPE-0 based on [Bergler, 1997].

The tagger used is the transformation-based Brill tagger [Brill, 1994b], which is available on the web[1] and which uses the Penn Treebank tagset shown in Figure 4.1 on page 40. The contextual rules [Brill, 1993; Brill, 1994a] for the verbs need improvement, as in some occurrences the Brill tagger will mislabel between the base form of a verb (VB) and the non-third person singular (VBP), as in the case of "have" (*I could have gone*) and "have" (*I have been going*), or between the simple past tense (VBD) and the past participle (VBN), as in the case of "said" (*I said*) and "said" (*I had said*). New contextual rules should look if one of these forms is in a context that implies non-finiteness or finiteness for the verb, and retag accordingly.

[1] http://www.cs.jhu.edu/~brill

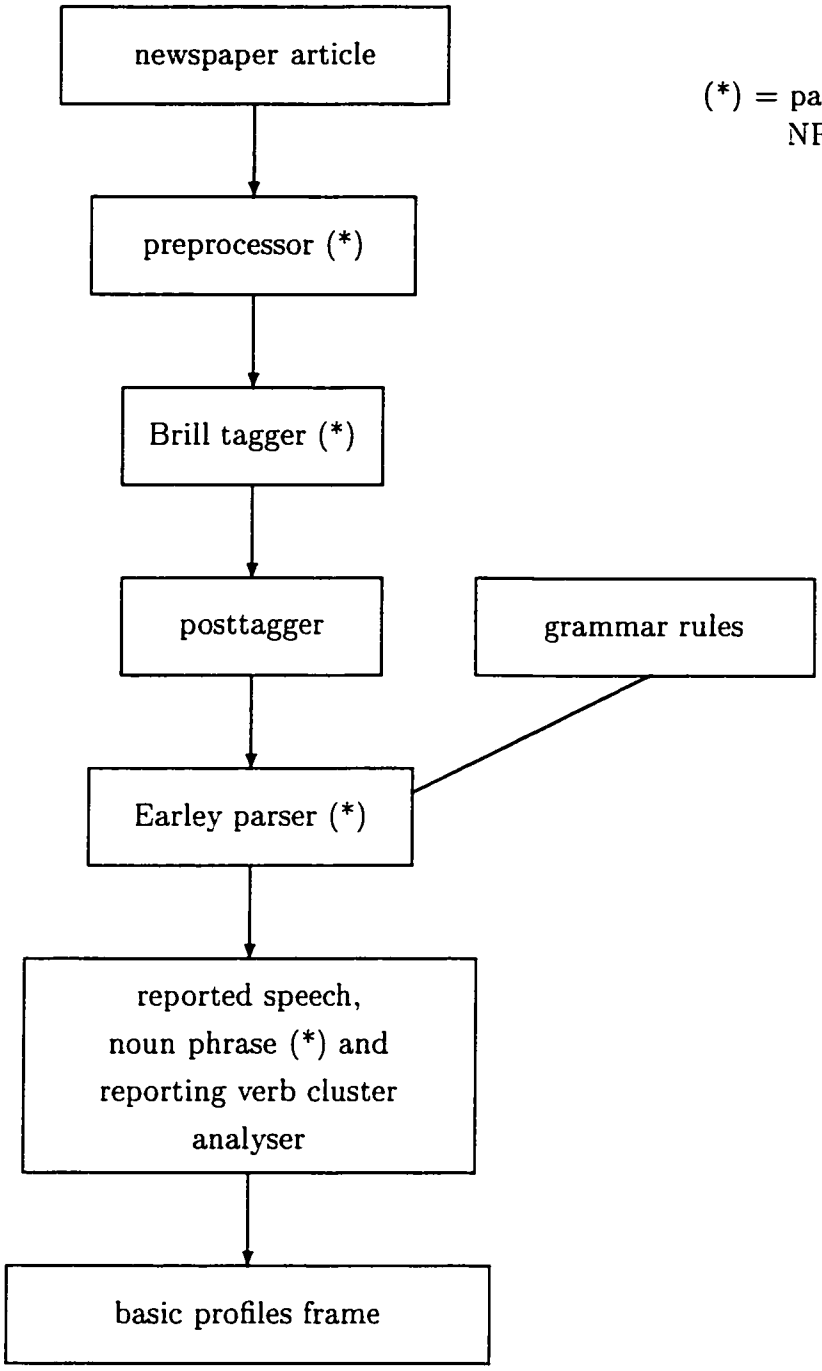

Figure 6.1: Flow chart

The postprocessor (posttagger) is an extension to the tagger, needed for this project. Some of the tags are changed to show if the verb is a possible auxiliary verb, such as a "have", "be" or "do" form, or to show that a verb is a form of a reporting verb, which can be used in reported speech. This was done in order to speed up the runtime of the implementation, as the change will be done only once instead of testing for the text of the verb for every verb in every parse.

The reannotated text will then be passed to an implementation of the Earley parser which is based on the Scheme implementation of the Earley parser tools developed by Marc Feeley, available from the Scheme Repository[2]. The Earley parser is used with a grammar for NPs, for verb clusters and for reported speech structures to yield partial parses.

The parse trees are analyzed by the implementation of the reported speech sentence, noun phrase and verb cluster analyzer[3], which will build basic profile frames.

A basic profile frame (structure) has the following fields: the text of the whole reported speech sentence corresponding to the basic profile, an index number for the basic profile, where it starts and where it ends in the text of the newspaper article, frames for the analyzed noun phrase representing the original source and the analyzed reporting verb cluster and the trees corresponding to the direct reported speech, indirect reported speech and other circumstantial information. For reading clarity, text representations of the source of the original speech act, reporting verb cluster, all reported speech, direct reported speech, indirect reported speech and other circumstantial information have also been added to the frame.

The source noun phrase frame contains fields for determiners, modifiers, the head of the noun phrase, coreferences and classification, such as number, gender or person. The analysis of coreferences and classification is worked on in another project[4].

The frame corresponding to the reporting verb cluster contains fields for a modal verb, auxiliaries, the main reporting verb, adverb modifiers to the cluster, tense, aspect, modality, mood and polarity as described in Chapter 4 and a list of semantic dimensions as defined in [Bergler, 1993; Bergler, 1995b].

The current performance is affected by compromises needed to be made to deal

[2]http://www.cs.indiana.edu/scheme-repository/code.lang.html

[3]Technically, the analyzer is composed of three modules: a noun phrase analyzer, a verb cluster analyzer and a reported speech clause analyzer and basic profile builder.

[4]ERSS (February 2003), a summarization system built on ERS-0, participated in DUC 2003.

with errors of tagging and the fact that the parser chooses minimal NPs and sometimes minimal verb clusters. On texts that are correctly tagged, all defined reported speech occurrences are found. The very few occurrences in which there were analysis errors were problematic because of the ambiguity between other circumstantial information and indirect reported speech. More work is needed to analyze in depth the syntactic differences between the presentation of other circumstantial information and the presentation of paraphrases of the original speech act, as well as a more in depth analysis of the syntactics of reported speech clauses.

6.2 Annotated example

An annotated example of a basic profile frame as created by this fully automated implementation is presented below. The example text has been extracted from the Wall Street Journal corpus, 10/27/89 and has been modified for exemplification purposes. The full unmodified text can be found in WSJ 891027-0054 or in [Bergler, 1992] on pages 176–177.

This is an example made of two different sentences which were adjoined to show one of the most complex cases that can appear in reported speech, also using a complex verb cluster. The example text is (S1) *Leonard Bickwit, a Washington attorney for Lincoln, has actually conceded that some memos had been written after the fact,* (S2) *adding, "its management never authorized it".*

The first basic profile found corresponds to (S1). As it can be seen in the output, the source of the original speech act is *Leonard Bickwit*, the reporting verb cluster is *has actually conceded*, the reporter chose to paraphrase the original utterance instead of using direct reported speech and has also given extra circumstantial information explaining who Leonard Bickwit is.

A more detailed annotation of the example follows next to some of the fields. The output is presented in `small fonts`. The output is fully automated and all the output presented in this example is continuous when the program is run. Each basic profile is stored in one frame connecting all the elements. Some subtrees have been removed for the purpose of exemplification, examples of full outputs can be found in Appendix D.

```
#s(rs (text "Leonard Bickwit ,/, a Washington attorney for Lincoln ,/, has actually conceded
           that some memos had been written after the fact")
```

```
(textsource "Leonard Bickwit ")
(textverb "has actually conceded ")
(reported_speech "some memos had been written after the fact ")
(textcirc ",/, a Washington attorney for Lincoln ,/, ")
(textdirq "")
(textindirq "some memos had been written after the fact ")
```

The analysis of the noun phrase representing the source of the original speech is shown below. Coreference information will be used in the future to build full profiles for each source.

```
(rep_source
  #s(np (npno 1)
        (det ((DET 6)))
        (mod ((MOD 34)))
        (head ((HEAD 4)
               ((PR_NAME 3)
                ((NNP 0)
                 #s(token (text "Leonard")
                          (classes ("Leonard" *item* NNP))
                          (location 1)))
                ((PR_NAME 1)
                 ((NNP 0)
                  #s(token (text "Bickwit")
                           (classes ("Bickwit" *item* NNP))
                           (location 2)))))))
        (coreferences ())
        [detail omitted]
```

The analysis of the reporting verb cluster is shown below. The reporting verb chosen by the reporter is *conceded*, which is the main verb in this cluster and is in present tense, perfect aspect, active voice. There is one auxiliary and an adverb modifier. The semantic dimensions for the reporting verb show an unmarked strength, a negative affectedness, a speech act of informing, the information is presupposed, a positive polarity, an unmarked audience, an unmarked formality, the information was relayed in a explicit manner and the voice of the original source was unmarked.

Note that in this example we don't have the markers of a negation, such as the adverb modifier "not", but would such a marker appear in the cluster, the polarity would be set to "negative". This is important for the evidential analysis, as the polarity of the verb cluster can combine with the semantic dimension polarity to change the general polarity of the reporting verb. For example, if the source *didn't deny* the allegations, it means that there is a possible belief that the source accepted

the allegations, which has a positive polarity, while the reporting verb *deny* carries a negative polarity.

```
(rep_verb
  $s(vc (vcno 1)
        (modal ())
        (auxs ((AUX_HAVE_FINITE 3)
               ((HAVE_VBZ 0)
                $s(token (text "has")
                         (classes ("has" *item* HAVE_VBZ))
                         (location 10)))))
        (mainverb
          ((RS_MAIN_VERB_VBN 1)
           ((RS_VBN 0)
            $s(token (text "conceded")
                     (classes ("conceded" *item* RS_VBN))
                     (location 12)))))
        (adverbs ((ADVERBS 3)
                  ((RB 0)
                   $s(token (text "actually")
                            (classes ("actually" *item* RB))
                            (location 11)))))
        (tense present)
        (aspect perfect)
        (voice active)
        (modality ())
        (mood ())
        (polarity ())
        (semantic_dimensions
          (STR_unmarked
            AFF_negative
            SPA_inform
            PRESP_presupposed
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
```

The `othercirc` field stores the provided circumstantial information other than source or reporting verb, in this case only one, *a Washington attorney for Lincoln*, but it is possible to have multiple other circumstantial information. As above, the information inside `othercirc` is not analyzed, but nothing is lost so if there is a need for an analysis, the analyzer could be run on the other circumstantial information subtree.

```
(othercirc
     [detail omitted])
```

The reporter didn't use direct reported speech, choosing to paraphrase the source instead. The information inside direct quotes or paraphrases is partially parsed but not analyzed in detail. Nothing is lost, so if in the future there is a need to analyze the noun phrases and verb clusters inside the reported speech, the analyzer can be run on the direct or indirect speech parse subtree.

```
(dirquots ())
(indirquots
      [detail omitted])
```

Next, a second basic profile is found, which corresponds to (S2). The reporter used the reporting verb *adding* by itself and directly quoted the original speech act. Note that there is no source attached to this basic profile, which means that when constructing full profiles, this profile should be combined with the previous encountered profile (basic profile (no - 1)). The analysis process is done the same as for the previous profile. No other circumstantial information is provided by the reporter with this basic profile.

```
#s(rs (text "adding ,/, \"/\" its management never authorized it \"/\"")
      (textsource "")
      (textverb "adding ")
      (reported_speech "\"/\" its management never authorized it \"/\" ")
      (textcirc "")
      (textdirq "\"/\" its management never authorized it \"/\" ")
      (textindirq "")
      (rep_source ())
      (rep_verb
        #s(vc (vcno 2)
              (modal ())
              (auxs ())
              (mainverb
                ((RS_MAIN_VERB_VBG 1)
                 ((RS_VBG 0)
                  #s(token (text "adding")
                           (classes ("adding" *item* RS_VBG))
                           (location 23)))))
              (adverbs ())
              (tense no_tense)
              (aspect progressive)
              (voice active)
              (modality ())
              (mood ())
              (polarity ())
              (semantic_dimensions
                (STR_unmarked
```

```
            AFF_unmarked
            SPA_inform
            PRESP_unmarked
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
(othercirc ())
(dirquots
      [detail omitted])
(indirquots ()))
```

In the previous example, two basic profiles were found and analyzed by the automated system. The information retained can be used for building more complex profiles and for profile grouping and belief analysis.

6.3 Evaluation

The development work has been done on a training corpus of 65739 sentences from the Wall Street Journal corpus (sections 1-12, except 3 and 11).

The machine we tested on is a Pentium III 1266MHz with 1.5G memory, running Linux. The system was tested on 6260 sentences from a contiguous chunk of the Wall Street Journal corpus of 10/27/89 and 10/30/89 (section 3, 285 texts), as well as on a few random articles extracted from the websites of CNN, Associated Press and Reuters (through Google news).

The performance of finding basic profiles was influenced by tagging errors, as well as by the length of some of the sentences. A few errors also occured because of untreated punctuation or rarely used reporting verbs which were not treated in this implementation[5] (ex: *answered*, which is in the second 50% of the list of corpus words, in terms of frequency). Future work could complete the list of treated reporting verbs. There were also a few problems with the parsing process, which made complex verb clusters and heavy noun phrases be seen as a few smaller constituents, and as such not fit properly into the reported speech structures. Future work is needed to define a process through which the best parse tree is selected.

[5]The testing corpus contains all the sentences, even if the reporting verb used in a sentence is not treated yet. This influences recall.

The system has been evaluated against a hand built Gold Standard. The correctness is judged on the correctness of the internal structure of the basic profile frames as well as on the fact that what we found is indeed a basic profile.

For example, the correct internal structure for the basic profile frame in the the following sentence does not match the way the system analyzed it. The analyzed basic profile frame is incomplete, missing the text "fully discounted."

He added that talk of strike settlements at producing mines has been fully discounted. (WSJ 10/30/89)

Correct frame:

```
#s(rs (text "He added that talk of strike settlements at producing mines has been fully discounted")
      (textsource "He ")
      (textverb "added ")
      (reported_speech
        "talk of strike settlements at producing mines has been fully discounted ")
      (textcirc "")
```

Incorrect frame, as built by the system:

```
#s(rs (text "He added that talk of strike settlements at producing mines has been")
      (textsource "He ")
      (textverb "added ")
      (reported_speech
        "talk of strike settlements at producing mines has been ")
      (textcirc "")
```

An example of non basic profile occurs in the sentence *December silver added 7.7 cents to close at $5.237 an ounce* (WSJ 10/27/89) which is analysed as shown below. However, this is not an instance of reported speech and hence does not contain a basic profile.

Not a basic profile:

```
#s(rs (text "December silver added 7.7 cents to close at $5.237 an ounce")
      (textsource "December silver ")
      (textverb "added ")
      (reported_speech "7.7 cents to close at $5.237 an ounce ")
      (textcirc "")
```

The implementation achieves a precision of 91.58% on the internal structure of the basic profile frames. 69 of the basic profiles found did not have a correct internal structure. 12 out of these 69 were not basic profiles. The major factors influencing precision are:

1. Incomplete parsetree for the basic profile (36.23%) - out of which 6 are heavy NPs which were split in smaller NPs
2. Other circumstantial information and paraphrase mix ups (31.88%)
3. Verb patterned as reporting verb in a non–reporting situation (17.39%)
4. Heavy noun phrase split up, including the 6 from incomplete parses (14.49%)
5. Untreated punctuation (4.35%)

The implementation found a total of 819 basic profiles, out of which 750 were correct, for a recall of 43.60%. The total number of basic profiles in the Gold standard is 1720, including 129 basic profiles which were found but could not be analyzed due to hardware limitations. The major factors influencing recall are:

1. Source as split up heavy noun phrase (21.42%)
2. Untreated reporting verb (20.20%)
3. Hardware limitations (14.32%)
4. Tagging errors (7.55%)
5. Phrasal reporting verbs (6.22%)

6.3.1 Major factors influencing precision

Incomplete parsetree

In 25 instances, the parsetree for the basic profile was incomplete. For example, in the sentence *Philadelphia-based Deb Shops said it saw little significance in Mr. Petrie selling his stock to Petrie Stores* (WSJ 10/30/89), the source found was *Deb Shops*, the adjective connected to it being left out of the parsetree. A parsetree selector could help fix this problem by selecting the longest parse.

Other circumstantial information and paraphrases

As discussed in Chapter 5, the task of differentiating between other circumstantial information and paraphrases is a very difficult one. Paraphrases might look like circumstantial parentheticals at times, but since they convey something the source said, they are not part of the circumstantial information.

Reporters and editors subconsciously avoid the ambiguous structures, even if the context would disambiguate them. The ambiguous structures which cannot be resolved through the syntactic analysis used by the implementation amount to less than 2% of the total found basic profiles in the test corpus.

An example of non-ambiguous circumstantial information which was wrongly considered as part of the paraphrase is when the reporting verb *tell* is used in the middle of the sentence, as follows.

```
#s(rs (text "Mr. Guttman told one person familiar with the New York exchanges during the search
            for a replacement that he was looking for a president who would be \"/\" responsive
            to the needs of the membership and the board \"/\"")
      (textsource "Mr. Guttman ")
      (textverb "told ")
      (reported_speech
        "one person familiar with the New York exchanges during the search for a replacement
        that he was looking for a president who would be \"/\" responsive to the needs of the
        membership and the board \"/\" ")
      (textcirc "")
```

Here (WSJ 10/30/89), *told* requires both an agent and a patient, the patient being part of the circumstantial information. Also, in this same example, we have a temporal circumstantial adverbial which is not separated from the rest of the sentence by commas. The only clue here is the word *that* which marks the beginning of the paraphrase, but since a paraphrase can also start without markers, this basic profile analysis gives a wrong internal structure.

More research is needed to define the paraphrase–other circumstantial information differentiation from a finer granularity point of view.

Verb patterned as reporting verb in a non–reporting situation

During the testing phase, we encountered 12 instances in which normal sentences were patterned as basic profiles when they were not. In every case, this happened because a verb which can be a reporting verb has been used as a non-reporting verb in a sentence with a structure which could be patterned as reporting speech sentence.

One such example is the sentence *Only time will tell if Mr. Boyd can restore to the Alley the acclaim it received when its founder, Nina Vance, was at the height of her powers* (WSJ 10/27/89) which is analysed as follows.

```
#s(rs (text "time will tell if Mr. Boyd can restore to the Alley the acclaim it received when
            its founder ,/, Nina Vance ,/, was at the height of her powers")
      (textsource "time ")
      (textverb "will tell ")
      (reported_speech
        "if Mr. Boyd can restore to the Alley the acclaim it received when its founder ,/,
         Nina Vance ,/, was at the height of her powers ")
      (textcirc "")
```

Source as heavy NP

In this thesis, we did not study the recognition of full heavy noun phrases. Future work is needed to add the syntactic structures of sources as heavy noun phrases to the current grammar or to implement a source parsetree selector.

In ten cases, the sources have not been recognized properly as we do not extract maximal noun phrases right now. One such example occurs in the sentence *People involved in the negotiations said the accord will reduce Costa Rica's foreign debt principal by more than 60%.* (WSJ 10/27/89)

```
#s(rs (text "the negotiations said the accord will reduce Costa Rica 's foreign debt")
      (textsource "the negotiations ")
      (textverb "said ")
      (reported_speech "the accord will reduce Costa Rica 's foreign debt ")
      (textcirc "")
```

Since *People involved in the negotiations* is split into two noun phrases, the parser patterns only the second one as the source. If we would have more complex definitions for the sources, as well as a parsetree selector, the correct source could be found.

Untreated punctuation

In three instances, untreated punctuation caused the basic profile frames to be incorrect.

For example, the sentence below was parsed, but because of the parentheses which are not treated correctly yet, the internal structure is wrong.

Senate Majority Leader George Mitchell (D. , Maine) said he intends to use Senate procedures to force the issue. (WSJ 10/27/89, primary information was shortened)

```
#s(rs (text "<p> <s> Senate Majority Leader George Mitchell (D. ,/, Maine) said he intends
            to use Senate procedures to force the issue")
      (textsource "Maine) ")
      (textverb "said ")
      (reported_speech
        "he intends to use Senate procedures to force the issue ")
      (textcirc "<p> <s> Senate Majority Leader George Mitchell (D. ,/, ")
```

6.3.2 Major factors influencing recall

Source as heavy NP

The source lexicalization in reported speech often takes the form of heavy noun phrases. In most of the cases, these were recognized as such, but there were 193 occurrences when a heavy noun phrase would be split in two or more simpler noun phrases which didn't fit within any of the reported speech structures, so no reported speech sentence parse tree would be found.

For example, in the sentence *The maker of hand-held computers and computer systems said the personnel changes were needed to improve the efficiency of its manufacturing operation* (WSJ 10/30/89), the source is *The maker of hand-held computers and computer systems*, but it is composed of three minimal NPs which do not fit any pattern defined.

The same as for the factors influencing precision, defining rules for finding maximal NPs and having a parsetree selector could help. Maximal NPs can be defined on a syntactic level starting from the existing NPE system which finds minimal NPs.

Rarely used reporting verbs

Reporting verbs are recognized based on pattern matching on a defined list of reporting verbs. The reporting verbs treated in the current implementation are a subset of the more complete list compiled by the AETNA Group, the subset being chosen based on a word frequency tally which was run on the training corpus. The list we built for the implementation contains 33 reporting verbs right now.

Since the testing corpus is a contiguous section of the Wall Street Journal corpus, it also contains basic profiles centered around reporting verbs which are not treated in the current implementation. The reporting verbs which are not treated are not yet recognized as reporting verbs, and in the cases in which they are used, the basic profiles will not be found, unless they appear in an embedded reported speech sentence. Examples of untreated reporting verbs are *blame, attack, demur*, etc.

A more complete list of over 100 reporting verbs has been built and can be used for future improvements. Some of the rarely used reporting verbs appear more often in non–reporting situations, so it will be important to make sure that they will only be recognized as reporting verbs in the right context. Corpus research in this area will be needed to determine if this can be done without a semantic analysis.

Long sentences

Long sentences[6] tended to slow down the total runtime and affect the accuracy of the system, as they do for most systems. This usually happens because as a sentence grows longer, there are more possible parse trees. Some systems choose to limit the length of the sentences they analyze. For example, The XTAG English Grammar was only tested on sentences of length 15 or less [Bangalore *et al.*, 1998]. Even for taggers, the length of the sentence influences the performance. The Brill tagger was tested by Eric Brill on sentences of up to a length of 25, finding the best performance for when the limit was lowered to 15 [Brill, 1993].

In our tests, we did not limit the length of the sentences. In average, it seems that performance decreases dramatically (by two orders of magnitude) once a sentence goes over a length of 25, and some sentences of an average length of over 35 need more memory than our computer could offer.

During our tests, 129 of the long sentences could not be completely analyzed due to hardware limitations. The average length of these sentences is of 36. These sentences would probably yield correct basic profile frames on a more powerful machine, using the current grammar, however, reducing the grammar's ambiguity is the best approach.

[6]The length of a sentence is counted in number of words. Sentences with a length over 25 are considered long.

The current grammar is very ambiguous due to the treatment of commas, treatment of *that*, paraphrases and other circumstantial information having similar rules, etc. For example, commas can appear both as a delimiter or as belonging to a paraphrase or to other circumstantial information, and a paraphrase can be introduced by *that* or by nothing when there is an absence of comma. Also, *that* can appear inside the paraphrase too, all this causing multiple possibilities for the paraphrase non–terminal. A better comma treatment could reduce ambiguity for comma fronted paraphrases, for example.

In the future, efficiency could be increased by reducing ambiguity. If long sentences will still pose problems, we could limit the length of the sentences the profile builder would look at, but I believe that there might be a better solution, involving grouping of words in a preprocessing phase. More research is needed in this area.

Tagging errors

The Brill tagger sometimes mislabels words, in which case the parsing will fail on sentences which contain the mistagged word. During the testing, we had 68 cases of mistags which happened inside reported speech sentences. Verbs sometimes get labeled as nouns (ex: *charges*) or vice–versa (ex: *hours*), or as the wrong tense (ex: *said*–past instead of past participle), in which case verb clusters expecting these verbs cannot be built. Other ambiguous words are also mislabeled at times (ex: *back*–noun mislabeled as *back*–adverb), which would also cause a failed parse. Following is an example of a verb with the wrong tag, and how the basic profile might have looked if it was tagged correctly.

> Previously he had said he would be able to find the requisite 60 votes eventually. (WSJ 10/27/89)

<S> Previously/RB he/PRP **had/HAVE_VBD said/RS_VBD** he/PRP would/MD be/BE_VB able/JJ to/TO find/VB the/DT requisite/JJ 60/CD votes/NNS eventually/RB ./.*end-of-sentence*</S>

The Brill tagger labeled "said" as a VBD (past tense), but as Quirk's [Quirk *et al.*, 1991] ABCD structure shows, we can only have a VBN (past participle) after a "have" auxiliary. Since "said" is mislabeled, it cannot fit within the verb cluster structure and the parse fails. To show how this parse might have looked if "said" was

labeled correctly, I replaced "said" with "announced," which is labeled as VBN by the Brill tagger in this context. Note that while the second example parses, the temporal adverbial is dropped by the parser. This is part of a possible problem influencing precision, which was presented in a previous section.

> Previously he had announced he would be able to find the requisite 60 votes eventually. (WSJ 10/27/89, modified)

<S> Previously/RB he/PRP **had/HAVE_VBD announced/RS_VBN** he/PRP would/MD be/BE_VB able/JJ to/TO find/VB the/DT requisite/JJ 60/CD votes/NNS eventually/RB ././*end-of-sentence*</S>

```
#s(rs (text "he had announced he would be able to find the requisite 60 votes")
      (textsource "he ")
      (textverb "had announced ")
      (reported_speech "he would be able to find the requisite 60 votes ")
      (textcirc ""))
```

Most mislabels can be fixed by defining better contextual rules in the tagger, which would be a cleaner solution than treating the mislabels inside the grammar, but the grammar solution is also a possibility.

Semi–auxiliaries, modal idioms, phrasal verbs, etc.

The current grammar is not dealing with semi–auxiliaries, modal idioms or phrasal verbs. While this doesn't have much of an impact on the performance of building basic profiles, there were 56 cases in which phrasal verbs or other grammatical constructs were used by the reporter to convey the primary information. By far the most common occurence (60%) was the usage of the complex proposition *according to*, as in *Fifteen joint ventures have already signed on as clients, according to Robert W. Cox, chairman of the firm.* Other examples are *blurt out, point out*, etc.

If we decide to analyze the verb clusters inside reported speech, semi–auxiliaires and modal idioms can also appear. This can be solved by treating these cases in the postprocessor and by adding them to the grammar rules. A lexicon could also be used.

6.4 Conclusion

This thesis is a development on previous research presented by Dr. Sabine Bergler. This thesis used a bottom–up approach to describe the details needed to build a practical fully–automated system which implements the extraction of profiles from a newspaper article.

While more research is needed to improve the current system, the implemented system shows that a sentence level syntactic analysis is enough to find basic profiles in most cases.

This thesis also developed a verb cluster extractor and analyzer which can be used as a stand–alone system. Due to time constraints, the research for the verb cluster extractor and analyzer had to be limited to finite verb clusters containing simple verbs. The cases of phrasal verbs (ex: *blurt out, think over, get over with*) have been ignored for now, for example, but they can be introduced later without modifications to the current system.

Because of the complexity of the task, our research did not address all the possible details needed for a complete profile–builder system. Where appropriate, future work has been defined.

Bibliography

[Aarts *et al.*, 2000] Bas Aarts, Judith Broadbent, Jonathan White, Gerald Nelson, and Justin Buckley. *The Internet Grammar of English - Website.* The Survey of English Usage, University College London, 2000.
URL: http://www.ucl.ac.uk/internet-grammar.

[Allen, 1995] James F. Allen. *Natural Language Understanding.* The Benjamin Cummings Publishing Company, Inc., Second edition, 1995.

[Allen, 1996] James F. Allen. *The TRAINS Parsing System Version 4.0, a User's Manual - Website*, 1996.
URL: http://www.cs.rochester.edu/u/james/ParserManual.html.

[American Heritage Dictionary, 2000] *The American Heritage Dictionary of the English Language.* Houghton Mifflin Company, Boston, Fourth edition, 2000.

[Asher and Lascarides, 1994] Nicholas Asher and Alex Lascarides. Intentions and Information in Discourse. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 34–41, New Mexico State University, Las Cruces, New Mexico, USA, 1994.

[Asher, 1993] Nicholas Asher. *Reference to Abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy.* Kluwer Academic Publishers, Dordrecht, Holland, 1993.

[Bach and Harnish, 1979] Kent Bach and Robert M. Harnish. *Linguistic Communication and Speech Acts.* MIT Press, Cambridge, MA, 1979.

[Bagga, 1998] Amit Bagga. *Coreference, Cross-document Coreference, and Information Extraction Methodologies.* PhD thesis, Duke University, 1998.

[Baker, 1989] Carl L. Baker. *English Syntax.* MIT Press, 1989.

[Ballim and Wilks, 1992] Afzal Ballim and Yorick Wilks. *Artificial Believers — The Ascription of Belief.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

[Ballim, 1992] Afzal Ballim. *ViewFinder: A Framework for Representing, Ascribing and Maintaining Nested Beliefs of Interacting Agents.* PhD thesis, University of Geneva, 1992.

[Bangalore *et al.*, 1998] Srinivas Bangalore, Anoop Sarkar, Christine Doran, and Beth Ann Hockey. Grammar and Parser Evaluation in the XTAG Project. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[Bangalore, 1997] Srinivas Bangalore. *Complexity of Lexical Descriptions and its Relevance to Partial Parsing.* PhD thesis, University of Pennsylvania, 1997.

[Bergler and Pustejovsky, 1990] Sabine Bergler and James Pustejovsky. Temporal reasoning from lexical semantics. In Brigitte Endres-Niggemeyer, Thomas Herrmann, Alfred Kobsa, and Dietmar Rösner, editors, *Interaktion und Kommunikation mit dem Computer*, pages 145–154. Springer, Berlin, Heidelberg, 1990.

[Bergler, 1991] Sabine Bergler. The Semantics of Collocational Patterns for Reporting Verbs. In *Proceedings of the Fifth European Conference of the Association for Computational Linguistics*, pages 216–221, Berlin, Germany, 1991.

[Bergler, 1992] Sabine Bergler. *Evidential Analysis of Reported Speech.* PhD thesis, Brandeis University, 1992.

[Bergler, 1993] Sabine Bergler. Semantic Dimensions in the Field of Reporting Verbs. In *Making Sense of Words, Proceedings of the Ninth Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, Oxford, U.K., 1993.

[Bergler, 1995a] Sabine Bergler. From lexical semantics to text analysis. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*, pages 98–124. Cambridge University Press, Cambridge, England, 1995.

[Bergler, 1995b] Sabine Bergler. Generative Lexicon Principles for Machine Translation: A Case for Meta-Lexical Structure. *Machine Translation*, 9(3), 1995.

[Bergler, 1997] Sabine Bergler. Towards reliable partial anaphora resolution. In Ruslan Mitkov and Branimir Boguraev, editors, *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 62–66, Madrid, Spain, 1997.

[Biber *et al.*, 1999] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English.* Pearson ESL, 1999.

[Black *et al.*, 1991] Ezra Black, Steven Abney, Dan Flickenger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Fred Jelinek, Judith Klavans, Mark Liberman, Mitch Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proceedings of the Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, CA, 1991. Morgan Kaufmann.

[Booth, 1969] Taylor L. Booth. Probabilistic representation of formal languages. In *IEEE Conference Record of 1969 Tenth Annual Symposium on Switching and Automata Theory*, pages 74–81, Waterloo, Ontario, Canada, 1969.

[Brill and Wu, 1998] Eric Brill and Jun Wu. Classifier Combination for Improved Lexical Disambiguation. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL*, pages 191–195, Montréal, Canada, 1998.

[Brill, 1992] Eric Brill. A Simple Rule-based Part-of-speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy, 1992.

[Brill, 1993] Eric Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, 1993.

[Brill, 1994a] Eric Brill. A Report of Recent Progress in Transformation-Based Error-Driven Learning. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ, 1994.

[Brill, 1994b] Eric Brill. Some Advances in Rule–based Part–of–speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI*, pages 722–727, 1994.

[Brill, 1995] Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[Byrd *et al.*, 2002] Pat Byrd, Tom McKlin, and Susan Jones. *English Grammar on the Web - Website.* Department of Applied Linguistics and ESL at Georgia State University, 2002.
URL: http://www.gsu.edu/~wwwesl/egw.

[Callaghan, 1998] Paul Callaghan. *An Evaluation of LOLITA and Related Natural Language Processing Systems.* PhD thesis, University of Durham, 1998.

[Cappelen and Lepore, 1997] Herman Cappelen and Ernest Lepore. Varieties of Quotation. *Mind*, 106:429–450, Jul 1997.

[Chalupsky, 1996] Hans Chalupsky. *Simba: Belief Ascription by Way of Simulative Reasoning.* PhD thesis, State University of New York at Buffalo, 1996.

[Collins, 1999] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing.* PhD thesis, University of Pennsylvania, 1999.

[Crowe, 1997] Jeremy David Macdonald Crowe. *Constraint Based Event Recognition for Information Extraction.* PhD thesis, University of Edinburgh, 1997.

[Daelemans *et al.*, 1996] Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. MBT: A Memory–Based Part–of–speech Tagger Generator. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 14–27, Copenhagen, Denmark, 1996.

[Davidson, 2001] Donald Davidson. *Inquiries into Truth and Interpretation, Philosophical Essays.* Oxford University Press, Second edition, 2001.

[DeRose, 1988] Steven J. DeRose. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, 14:31–39, 1988.

[Dooley and Levinsohn, 2000] Robert A. Dooley and Stephen H. Levinsohn. *Analyzing Discourse: A Manual of Basic Concepts*, 2000.

[Doran, 1996] Christine Doran. Punctuation in Quoted Speech. In *Proceedings of SIGPARSE 96 - Punctuation in Computational Linguistics*, 1996.

[Doran, 1998] Christine D. Doran. *Incorporating Punctuation into the Sentence Grammar: A Lexicalized Tree Adjoining Grammar Perspective.* PhD thesis, University of Pennsylvania, 1998.

[Doran, 2000] Christine Doran. Punctuation in a Lexicalized Grammar. In *Proceedings of the 5th Workshop on Tree-Adjoining Grammars and Related Formalisms (TAG+5)*, Paris, May 2000.

[EAGLES, 1996] EAGLES. *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages.* 1996. In Progress.

[Earley, 1970] Jay Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[Ferguson *et al.*, 1996] George M. Ferguson, James F. Allen, Brad W. Miller, and Eric K. Ringger. The Design and Implementation of the TRAINS-96 System: A Prototype Mixed-Initiative Planning Assistant. Technical Report TN96-5, The University of Rochester, 1996.

[Frege, 1892] Gottlob Frege. Über Sinn und Bedeutung (On Sense and Reference). *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892. (Translated in Peter T. Geach and Max Black, editors, Translations from the philosophical writings of Gottlob Frege, pages 56–78, Blackwell, 1960).

[Garigliano *et al.*, 1998] Roberto Garigliano, Agnieszka Urbanowicz, and David J. Nettleton. Description of the LOLITA system as used in MUC-7. Technical report, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham, 1998.

[Gazdar, 1981] Gerald Gazdar. Speech Act Assignment. In Aravind Joshi, Bonnie Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 64–83, Cambridge, 1981. Cambridge University Press.

[Geach, 1972] Peter Thomas Geach. Quotation and Quantification. In *Logic Matters*. Blackwell Publishers, Oxford, U.K., 1972.

[Gerard, 2000] Christine Gerard. Modelling Readers of News Articles Using Nested Beliefs. Master's thesis, Concordia University, 2000.

[Hardt, 1992] Daniel Hardt. Some Problematic Cases of VP Ellipsis. In *Proceedings to the 30th Annual Meeting of the Association for Computational Linguistics*, pages 276–278, University of Delaware, Newark, Delaware, USA, 1992. Association for Computational Linguistics.

[Hobbs *et al.*, 1993] Jerry R. Hobbs, Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. FASTUS: A System for Extracting Information from Text. In *Proceedings of the Workshop on Human Language Technology*, pages 133–137, Princeton, New Jersey, 1993. Morgan Kaufmann.

[Hobbs *et al.*, 1996] Jerry R. Hobbs, Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. *FASTUS: A Cascaded Finite-state Transducer for Extracting Information from Natural-language Text.* MIT Press, Cambridge, MA, 1996.

[Jelinek and Mercer, 1980] Frederick Jelinek and Robert Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In Edzard S. Gelsema and Laveen N. Kanal, editors, *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies and Hybrid Systems*, pages 381–402, Amsterdam, North-Holland, 1980. Elsevier.

[Joshi *et al.*, 2001] Aravind Joshi, Tonia Bleam, Chung-hye Han, Rashmi Prasad, Carlos Prolo, Anoop Sarkar, Matt Feldman, Martin Kappus, Seth Kulick, Virginie Nanta, Moses Kimanzi, Eric Kowey, William Schuler, and Fei Xia. A Lexicalized Tree Adjoining Grammar for English. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, 2001.

[Jurafsky and Martin, 2000] Daniel Jurafsky and James H. Martin. *Speech and Language Processing.* Prentice Hall, 2000.

[Kamp and Reyle, 1993] Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Kluwer Academic Publishers, Dordrecht, Holland, 1993.

[Keyes, 1996] Ralph Keyes. *The Wit and Wisdom of Oscar Wilde: A Treasury of Quotations, Anecdotes, and Repartee.* Harper Collins, 1996.

[Kies, 2002] Daniel Kies. *Modern English Grammar - Website.* Department of English, College DuPage, 2002.
URL: http://papyr.com/hypertextbooks/engl_126.

[Klein, 1994] Wolfgang Klein. *Time in Language.* Routledge, New York, NY, 1994.

[Loos *et al.*, 2002] Eugene E. Loos, Susan Anderson, Dwight H. Day Jr., Paul C. Jordan, and J. Douglas Wingate. *Glossary of Linguistic Terms - Website.* Summer Institute of Linguistics, 2002.
URL: http://www.sil.org/linguistics/GlossaryOfLinguisticTerms.

[Marcus *et al.*, 1993] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[McArthur, 1992] Tom McArthur, editor. *The Oxford Companion to the English Language.* Oxford University Press, 1992.

[Megginson *et al.*, 2002] David Megginson, Heather MacFadyen, Frances Peck, and Dorothy Turner. *HyperGrammar - Website.* Department of English, University of Ottawa, 2002.
URL: http://www.uottawa.ca/academic/arts/writcent/hypergrammar.

[Ney, 1991] Hermann Ney. Dynamic Programming Parsing for Context–Free Grammars in Continuous Speech Recognition. *IEEE Transactions on Signal Processing*, 39(2):336–340, 1991.

[Quine, 1951] Willard Van Orman Quine. *Mathematical Logic.* Harvard University Press, Second edition, 1951.

[Quine, 1960] Willard Van Orman Quine. *Word and Object.* MIT Press, Cambridge, MA, 1960.

[Quirk *et al.*, 1991] Randolph Quirk, Jan Svartvik, Geoffrey Leech, and Sidney Greenbaum. *A Comprehensive Grammar of the English Language.* Longman Group Limited, 1991.

[Ratnaparkhi, 1996] Adwait Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[Reichenbach, 1947] Hans Reichenbach. *Elements of Symbolic Logic.* Macmillan, New York, NY, 1947.

[Samuelsson and Voutilainen, 1997] Christer Samuelsson and Atro Voutilainen. Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, Madrid, Spain, July 1997.

[Say, 1998] Bilge Say. *An Information-Based Approach to Punctuation.* PhD thesis, Bilkent University, 1998.

[Searle and Vanderveken, 1985] John R. Searle and Daniel Vanderveken. *Foundations of Illocutionary Logic.* Cambridge University Press, Cambridge, England, 1985.

[Sekine and Grishman, 1995] Satoshi Sekine and Ralph Grishman. A Corpus-based Probabilistic Grammar with Only Two Non-terminals. In *Proceedings of the Fourth International Workshop on Parsing Technologies*, 1995. Prague, Czech Republic.

[Sekine, 1996] Satoshi Sekine. *Manual of Apple Pie Parser*, 1996.

[Sperberg-McQueen and Burnard, 2002] C. Michael Sperberg-McQueen and Lou Burnard. *Guidelines for Electronic Text Encoding and Interchange: TEI P4, XML-compatible edition - Website.* TEI Consortium, 2002.
URL: http://www.tei-c.org/P4X/.

[Stanfill and Waltz, 1986] Craig Stanfill and David Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.

[Steetskamp, 1995] Rene Steetskamp. An implementation of a probabilistic tagger. Master's thesis, TOSCA Research Group, University of Nijmegen, Nijmegen, The Netherlands, 1995.

[Stolcke, 1995] Andreas Stolcke. An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities. In *Computational Linguistics*, volume 21, pages 165–202. MIT Press, 1995.

[Strunk Jr. and White, 1979] William Strunk Jr. and E. B. White. *The Elements of Style.* MacMillan Publishing Company, Inc., Third edition, 1979.

[TEI Consortium *et al.*, 2003] TEI Consortium, C. Michael Sperberg-McQueen, and Lou Burnard. *Guidelines for Electronic Text Encoding and Interchange: Volumes 1 and 2: P4.* University Press of Virginia, 2003.

[Tomita, 1986] Masaru Tomita. *Efficient Parsing of Natural Language: A Fast Algorithm for Practical Systems.* Kluwer Academic Publishers, Dordrecht, Holland, 1986.

[van Halteren *et al.*, 1998] Hans van Halteren, Jakub Zavrel, and Walter Daelemans. Improving Data Driven Wordclass Tagging by System Combination. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL*, pages 491–497, Montréal, Canada, 1998.

[Voutilainen, 1995] Atro Voutilainen. Morphological disambiguation. In Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila, editors, *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*, pages 165–284, Berlin and New York, 1995. Mouton de Gruyter.

[Voutilainen, 1997] Atro Voutilainen. Using English Constraint Grammar as a Syntactic Parser's Preprocessor. In Tom Brondsted and Inger Lytje, editors, *Sprog og Multimedier (Language and Multimedia)*, pages 20–38, Aalborg, Denmark, 1997. Aalborg Universitetsforlag.

[Webster Encyclopedic, 2001] *Webster's Encyclopedic Unabridged Dictionary of the English Language.* Random House Value Publishing and Thunder Bay Press, 2001.

[Webster Revised, 1998] *Webster's Revised Unabridged Dictionary.* Micra Inc., 1998.

[Wiebe, 1990] Janyce M. Wiebe. Identifying Subjective Characters in Narrative. In *Proceedings of the 13th Conference on Computational Linguistics (COLING)*, volume 2, pages 401–406, Helsinki, Finland, 1990.

[Yagisawa, 1997] Takashi Yagisawa. A Somewhat Russellian Theory of Intensional Contexts. In James E. Tomberlin, editor, *Mind, Causation, and World*, volume 11 of *Philosophical Perspectives Series*, pages 43–82, Oxford, U.K., 1997. Blackwell Publishers.

Appendix A

Verb cluster grammar

The notations used are presented in Table A.1. While these long notations can make the grammar look somewhat unwieldy, they should provide a good level of maintenance support for later revisions.

Notation	Description
VC	verb cluster
AUX_HAVE_VB	auxiliary *have* in base form
AUX_HAVE_FINITE	auxiliary *have* in present or past form
AUX_BE_VB	auxiliary *be* in base form
AUX_BE_FINITE	auxiliary *be* in present or past form
AUX_BE_VBN	auxiliary *be* in past participle form
AUX_BE_VBG	auxiliary *be* in present participle form
AUX_DO	auxiliary *do* in base, present or past form
MAIN_VERB_VB	main verb in base form
MAIN_VERB_FINITE	main verb in present or past form
MAIN_VERB_VBN	main verb in past participle form
MAIN_VERB_VBG	main verb in present participle form
ADVERBS	adverbs cluster
INV_OBJ	inverted object
NP	noun phrase as defined in NPE

Table A.1: Notations used

The verb cluster grammar is shown below, in a form simplified for display purposes. In this simplified form, the ADVERBS have been removed from the main rules. In the full grammar, ADVERBS can appear anywhere within the verb cluster. The ADVERBS definition is shown in this simplified representation.

```
(VC
      (MD INV_OBJ MAIN_VERB_VB)                                    ; A
      (MD INV_OBJ AUX_HAVE_VB MAIN_VERB_VBN)                       ; AB
      (MD INV_OBJ AUX_HAVE_VB AUX_BE_VBN MAIN_VERB_VBG)            ; ABC
      (MD INV_OBJ AUX_HAVE_VB AUX_BE_VBN AUX_BE_VBG MAIN_VERB_VBN) ; ABCD
      (MD INV_OBJ AUX_HAVE_VB AUX_BE_VBN MAIN_VERB_VBN)            ; ABD
      (MD INV_OBJ AUX_BE_VB MAIN_VERB_VBG)                         ; AC
      (MD INV_OBJ AUX_BE_VB AUX_BE_VBG MAIN_VERB_VBN)              ; ACD
      (MD INV_OBJ AUX_BE_VB MAIN_VERB_VBN)                         ; AD

      (AUX_HAVE_FINITE INV_OBJ MAIN_VERB_VBN)                      ; B
      (AUX_HAVE_FINITE INV_OBJ AUX_BE_VBN MAIN_VERB_VBG)           ; BC
      (AUX_HAVE_FINITE INV_OBJ AUX_BE_VBN AUX_BE_VBG MAIN_VERB_VBN) ; BCD
      (AUX_HAVE_FINITE INV_OBJ AUX_BE_VBN MAIN_VERB_VBN)           ; BD

      (AUX_BE_FINITE INV_OBJ MAIN_VERB_VBG)                        ; C
      (AUX_BE_FINITE INV_OBJ AUX_BE_VBG MAIN_VERB_VBN)             ; CD

      (AUX_BE_FINITE INV_OBJ MAIN_VERB_VBN)                        ; D

      (MAIN_VERB_FINITE)
      (AUX_DO INV_OBJ MAIN_VERB_VB)    ; dummy DO
)

(INV_OBJ
      (NP)                    ; for questions with inversion
      ()
)
(ADVERBS
      (RB CC RB)                      ; does absolutely and truly mean
      (RB RB)                         ; does not actually mean
      (RB)                            ; does not mean
      ()
)
;;;;;; have auxiliaries
(AUX_HAVE_VB
      (HAVE_VB)
)
(AUX_HAVE_FINITE
      (HAVE_VBD)
      (HAVE_VBP)
      (HAVE_VBZ)
)
;;;;;; be auxiliaries
(AUX_BE_VB
      (BE_VB)
)
(AUX_BE_FINITE
      (BE_VBD)
      (BE_VBP)
      (BE_VBZ)
```

```
)
(AUX_BE_VBN
      (BE_VBN)
)
(AUX_BE_VBG
      (BE_VBG)
)
;;;;;; DO support - dummy DO
(AUX_DO
      (DO_VB)                 ;do
      (DO_VBP)                ;does
      (DO_VBZ)                ;do
      (DO_VBD)                ;did
)
;;;;;; main verbs
(MAIN_VERB_VB
      (VB)
      (HAVE_VB)
      (BE_VB)
      (DO_VB)
)
(MAIN_VERB_FINITE
      (VBD)
      (VBP)
      (VBZ)
      (HAVE_VBD)
      (HAVE_VBP)
      (HAVE_VBZ)
      (BE_VBD)
      (BE_VBP)
      (BE_VBZ)
      (DO_VBD)
      (DO_VBP)
      (DO_VBZ)
)
(MAIN_VERB_VBN
      (VBN)
      (HAVE_VBN)
      (BE_VBN)
)
(MAIN_VERB_VBG
      (VBG)
      (HAVE_VBG)
      (BE_VBG)
)
```

Appendix B

WSJ891027-0042

Wall Street Journal 10/27/89

The New York Post plans to close down its Sunday edition after launching it just eight months ago with much fanfare.

Peter Kalikow, owner and publisher of the Post, announced that he would discontinue the edition, which he initially started in the hopes of luring upscale readers and retail advertisers that shunned the more sensational tabloid in favor of the New York Times and the Daily News.

Industry observers and analysts still are unsure whether folding the Sunday Post will guarantee that the paper will survive. When Rupert Murdoch sold the paper to Mr. Kalikow, it was posting losses of $10 million a year by conservative estimates. In its first 10 weeks, the Sunday edition alone was estimated to have recorded a deficit of as much as $7 million.

Mr. Kalikow and Editor Jerry Nachman assembled employees in the paper's newsroom in lower Manhattan yesterday afternoon to break the news, which had been rumored in recent weeks. Post executives had denied earlier reports that the Sunday edition was closing down.

Mr. Kalikow said the death of the Sunday paper is no indication that the Post itself is in danger of closing down, as some industry observers have speculated. By closing down the Sunday Post, he said in a statement, the Post "foresees a profitable year in 1990." The Post's last Sunday edition will be Nov. 26.

Mr. Kalikow blamed the demise of the Sunday Post on the soft ad market in the New York area and on low circulation. "The economics just aren't there," he said.

While there has been a wave of new Sunday newspapers launched in small to medium-sized cities in the last decade, the Post was by far the largest newspaper to launch a Sunday edition in years. While Post executives had expected circulation would reach 500,000, Valerie Salembier, president of the Post, said the Sunday circulation has reached only about 250,000.

"In any other city, 250,000 would be considered great, but it just wasn't enough in New York," said Ms. Salembier.

Ms. Salembier said about 30 people in circulation, ad sales and other business departments would lose their jobs. "What we don't know about is the number of layoffs on the editorial side," she said. "The editorial side is more complicated," she added, saying that editorial layoffs will be announced later.

The paper has been particularly hurt by declines in the level of real estate and retail advertising, and by the fact that the Sunday paper could not attract any coupons from national companies. In late September, the paper announced its was canceling its weekly real estate section. The section had shrunk from more than 20 pages to six pages recently as a result of declines in advertising.

Months earlier, Mr. Kalikow had already begun to scale back the Sunday Post. In May, the Post said it would scratch all supplements in the Sunday edition, including USA Weekend magazine, and cut the $1 price to 40 cents.

Ms. Salembier said closing the Sunday paper was a "big disappointment" but the right move. "We were losing a lot of money. While Peter Kalikow has deep pockets, it makes sense to put our resources into our Monday through Saturday product," she said.

Appendix C

WSJ891027-0074

Wall Street Journal 10/27/89

Democratic leaders have bottled up President Bush's capital-gains tax cut in the Senate and may be able to prevent a vote on the issue indefinitely.

Senate Majority Leader George Mitchell (D., Maine) said he intends to use Senate procedures to force advocates of the tax cut to come up with at least 60 votes before they can address the issue. And neither Democrats nor Republicans are predicting that the capital-gains forces can produce enough votes.

"The 60-vote requirement will be there and they don't have the 60 votes," Sen. Mitchell said. "They don't have the votes to get it passed."

Sen. Bob Packwood (R., Ore.), the leading Republican proponent of the tax cut, didn't disagree. "I'm not sure what's going to happen," he said. Previously he had said he would be able to find the requisite 60 votes eventually.

Sen. Packwood has offered his capital-gains-cut package as an amendment to a bill, now pending in the Senate, that would authorize aid to Poland and Hungary. Democrats are holding up a vote on the amendment by threatening a filibuster, or extended debate. For a cloture vote to stop the filibuster, Republicans must muster at least 60 votes.

Yesterday, Sen. Packwood acknowledged, "We don't have the votes for cloture today."

The Republicans show no sign of relenting. GOP leaders continued to press for a vote on the amendment to the Eastern Europe aid measure. And they threatened to try to amend any other revenue bill in the Senate with the capital-gains provision.

"This is serious business; we're serious about a capital-gains reduction," said Kansas Sen. Robert Dole, the Senate's Republican leader. "The strategy is 'Let's vote.' "

The Republicans contend that they can garner a majority in the 100-member Senate for a capital-gains tax cut. They accuse the Democrats of unfairly using Senate rules to erect a 60-vote hurdle. Democrats counter that the Republicans have often used the same rules to suit their own ends.

The two sides also traded accusations about the cost of the Packwood plan. Democrats asserted that the proposal, which also would create a new type of individual retirement account, was fraught with budget gimmickry that would lose billions of dollars in the long run. Republicans countered that long-range revenue estimates were unreliable.

The Packwood proposal would reduce the tax depending on how long an asset was held. It also would create a new IRA that would shield from taxation the appreciation on investments made for a wide variety of purposes, including retirement, medical expenses, first-home purchases and tuition.

A White House spokesman said President Bush is "generally supportive" of the Packwood plan.

Appendix D

Output examples

Wall Street Journal, 10/27/89

(S1) Fujitsu Ltd. posted a modest 6.7% gain in unconsolidated pretax profit in the first half ended Sept. 30, (S2) the company said.

(NR1) Pretax profit for the Tokyo computer maker rose to 50.161 trillion yen ($354.4 million) from 47.013 trillion yen a year ago. (NR2) Net income rose 18% to 28.461 trillion yen, or 15.87 yen a share, from 24.213 trillion yen, or 13.81 yen a share.

(S3) The company said falling sales in its communications division hurt results. (NR3) Sales from the division slipped 7.5% to 161.080 trillion yen. (S4) A Fujitsu spokesman said the weaker performance reflected slipping sales to Nippon Telegraph & Telephone Corp., Japan's domestic telephone monopoly.

(NR4) Total sales rose 4.7% to 966.471 trillion yen.

(NR5) Sales of electrical devices rose about 19% to 147.874 trillion yen, while sales in the information systems division gained 5.4% to 657.517 trillion yen.

(S1)-(S4) are reported speech instances, corresponding to four basic profiles. (NR1)-(NR5) are non-reported speech. (S1) is embedded into (S2). The system found four basic profiles, but it did not embed (S1) into (S2).

```
<S> <RS1> Fujitsu/NNP Ltd./NNP posted/RS_VBD a/DT modest/JJ 6.7%/CD gain/NN in/IN
    unconsolidated/JJ pretax/JJ profit/NN in/IN the/DT first/JJ half/NN ended/VBN Sept./NNP
    30/CD </RS> ,/,/*comma* <RS2> the/DT company/NN said/RS_VBD </RS>
    ././*end-of-sentence*</S>
#s(rs (text "Fujitsu Ltd. posted a modest 6.7% gain in unconsolidated pretax profit in the
            first half ended Sept. 30")
      (no 1)
      (start 2)
      (end 19)
      (textsource "Fujitsu Ltd. ")
      (textverb "posted ")
      (reported_speech
        "a modest 6.7% gain in unconsolidated pretax profit in the first half ended Sept. 30 ")
      (textcirc "")
      (textdirq "")
      (textindirq
        "a modest 6.7% gain in unconsolidated pretax profit in the first half ended Sept. 30 ")
```

```
(rep_source
  #s(np (npno 1)
        (det ((DET 6)))
        (mod ((MOD 34)))
        (head ((HEAD 4)
               ((PR_NAME 3)
                ((NNP 0)
                 #s(token (text "Fujitsu")
                          (classes ("Fujitsu" *item* NNP))
                          (location 2)))
                ((PR_NAME 1)
                 ((NNP 0)
                  #s(token (text "Ltd.")
                           (classes ("Ltd." *item* NNP))
                           (location 3)))))))
        (coreferences ())
        (classification ())
        (comma ())))
(rep_verb
  #s(vc (vcno 1)
        (modal ())
        (auxs ())
        (mainverb
          ((RS_MAIN_VERB_FINITE 1)
           ((RS_VBD 0)
            #s(token (text "posted")
                     (classes ("posted" *item* RS_VBD))
                     (location 4)))))
        (adverbs ())
        (tense past)
        (aspect indefinite)
        (voice active)
        (modality ())
        (mood ())
        (polarity ())
        (semantic_dimensions
          (STR_high
            AFF_unmarked
            SPA_announce
            PRESP_new
            POL_positive
            AUD_public
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
(othercirc ())
(dirquots ())
(indirquots
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 1)
```

```
 ((DT 0)
  #s(token (text "a") (classes ("a" *item* DT)) (location 5))))
((MOD 14)
 ((ADJM 5)
  ((JJ 0)
   #s(token (text "modest")
            (classes ("modest" *item* JJ))
            (location 6)))
  ((ADJM 8)
   ((CD 0)
    #s(token (text "6.7%")
             (classes ("6.7%" *item* CD))
             (location 7))))))
((HEAD 1)
 ((NN 0)
  #s(token (text "gain")
           (classes ("gain" *item* NN))
           (location 8)))))
((INDIRQUOT 1)
 ((ANY 5)
  ((IN 0)
   #s(token (text "in") (classes ("in" *item* IN)) (location 9))))
 ((MANY 1)
  ((ANY 8)
   ((ADJM 1)
    ((JJ 0)
     #s(token (text "unconsolidated")
              (classes ("unconsolidated" *item* JJ))
              (location 10)))))
  ((MANY 2)))
((INDIRQUOT 2)
 ((NP 1)
  ((DET 6))
  ((MOD 14)
   ((ADJM 1)
    ((JJ 0)
     #s(token (text "pretax")
              (classes ("pretax" *item* JJ))
              (location 11)))))
  ((HEAD 1)
   ((NN 0)
    #s(token (text "profit")
             (classes ("profit" *item* NN))
             (location 12)))))
 ((INDIRQUOT 1)
  ((ANY 5)
   ((IN 0)
    #s(token (text "in") (classes ("in" *item* IN)) (location 13))))
  ((MANY 1)
   ((ANY 4)
    ((NON_FINITE_VC 4)
```

```
                ((NP 1)
                 ((DET 1)
                  ((DT 0)
                   $s(token (text "the")
                            (classes ("the" *item* DT))
                            (location 14))))
                 ((MOD 14)
                  ((ADJM 1)
                   ((JJ 0)
                    $s(token (text "first")
                             (classes ("first" *item* JJ))
                             (location 15)))))
                 ((HEAD 1)
                  ((NN 0)
                   $s(token (text "half")
                            (classes ("half" *item* NN))
                            (location 16)))))
                ((VBN 0)
                 $s(token (text "ended")
                          (classes ("ended" *item* VBN))
                          (location 17)))))
               ((MANY 2)))
              ((INDIRQUOT 2)
               ((NP 1)
                ((DET 6))
                ((MOD 34))
                ((HEAD 4)
                 ((PR_NAME 1)
                  ((NNP 0)
                   $s(token (text "Sept.")
                            (classes ("Sept." *item* NNP))
                            (location 18))))))
               ((INDIRQUOT 1)
                ((ANY 8)
                 ((ADJM 4)
                  ((CD 0)
                   $s(token (text "30")
                            (classes ("30" *item* CD))
                            (location 19)))))
                ((MANY 2))
                ((INDIRQUOT 4)))))))))
$s(rs (text "the company said")
      (no 2)
      (start 21)
      (end 23)
      (textsource "the company ")
      (textverb "said ")
      (reported_speech "")
      (textcirc "")
      (textdirq "")
      (textindirq "")
```

```
(rep_source
  #s(np (npno 2)
        (det ((DET 1)
              ((DT 0)
               #s(token (text "the")
                        (classes ("the" *item* DT))
                        (location 21)))))
        (mod ((MOD 34)))
        (head ((HEAD 1)
               ((NN 0)
                #s(token (text "company")
                         (classes ("company" *item* NN))
                         (location 22)))))
        (coreferences ())
        (classification ())
        (comma ())))
(rep_verb
  #s(vc (vcno 2)
        (modal ())
        (auxs ())
        (mainverb
          ((RS_MAIN_VERB_FINITE 1)
           ((RS_VBD 0)
            #s(token (text "said")
                     (classes ("said" *item* RS_VBD))
                     (location 23)))))
        (adverbs ())
        (tense past)
        (aspect indefinite)
        (voice active)
        (modality ())
        (mood ())
        (polarity ())
        (semantic_dimensions
          (STR_unmarked
            AFF_unmarked
            SPA_inform
            PRESP_unmarked
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
  (othercirc ())
  (dirquots ())
  (indirquots ()))
<S> Pretax/JJ profit/NN for/IN the/DT Tokyo/NNP computer/NN maker/NN rose/VBD to/TO 50.161/CD
    trillion/CD yen/NN ($354.4/CD million)/NN from/IN 47.013/CD trillion/CD yen/NN a/DT year/NN
    ago/RB ./.*end-of-sentence*</S>
<S> Net/JJ income/NN rose/VBD 18%/CD to/TO 28.461/CD trillion/CD yen/NN ,/,/*comma* or/CC
    15.87/CD yen/NN a/DT share/NN ,/,/*comma* from/IN 24.213/CD trillion/CD yen/NN ,/,/*comma*
```

```
    or/CC 13.81/CD yen/NN a/DT share/NN ././*end-of-sentence*</S>
<S> <RS3> The/DT company/NN said/RS_VBD falling/VBG sales/NNS in/IN its/PRP$ communications/NNS
    division/NN hurt/VBD results/NNS </RS> ././*end-of-sentence*</S>
#s(rs (text "The company said falling sales in its communications division hurt results")
      (no 3)
      (start 83)
      (end 93)
      (textsource "The company ")
      (textverb "said ")
      (reported_speech
        "falling sales in its communications division hurt results ")
      (textcirc "")
      (textdirq "")
      (textindirq "falling sales in its communications division hurt results ")
      (rep_source
        #s(np (npno 3)
              (det ((DET 1)
                    ((DT 0)
                     #s(token (text "The")
                              (classes ("The" *item* DT))
                              (location 83)))))
              (mod ((MOD 34)))
              (head ((HEAD 1)
                     ((NN 0)
                      #s(token (text "company")
                               (classes ("company" *item* NN))
                               (location 84)))))
              (coreferences ())
              (classification ())
              (comma ())))
      (rep_verb
        #s(vc (vcno 3)
              (modal ())
              (auxs ())
              (mainverb
                ((RS_MAIN_VERB_FINITE 1)
                 ((RS_VBD 0)
                  #s(token (text "said")
                           (classes ("said" *item* RS_VBD))
                           (location 85)))))
              (adverbs ())
              (tense past)
              (aspect indefinite)
              (voice active)
              (modality ())
              (mood ())
              (polarity ())
              (semantic_dimensions
                (STR_unmarked
                  AFF_unmarked
                  SPA_inform
```

```
            PRESP_unmarked
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
(othercirc ())
(dirquots ())
(indirquots
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 6))
    ((MOD 14)
     ((ADJM 3)
      ((VBG 0)
       #s(token (text "falling")
                (classes ("falling" *item* VBG))
                (location 86)))))
    ((HEAD 2)
     ((NNS 0)
      #s(token (text "sales")
               (classes ("sales" *item* NNS))
               (location 87)))))
   ((INDIRQUOT 1)
    ((ANY 5)
     ((IN 0)
      #s(token (text "in") (classes ("in" *item* IN)) (location 88))))
    ((MANY 2))
    ((INDIRQUOT 2)
     ((NP 1)
      ((DET 4)
       ((PRP$ 0)
        #s(token (text "its")
                 (classes ("its" *item* PRP$))
                 (location 89))))
      ((MOD 34))
      ((HEAD 2)
       ((NNS 0)
        #s(token (text "communications")
                 (classes ("communications" *item* NNS))
                 (location 90)))))
     ((INDIRQUOT 2)
      ((NP 1)
       ((DET 6))
       ((MOD 34))
       ((HEAD 1)
        ((NN 0)
         #s(token (text "division")
                  (classes ("division" *item* NN))
                  (location 91)))))
      ((INDIRQUOT 3)
```

```
            ((VC 16)
             ((ADVERBS 4))
             ((MAIN_VERB_FINITE 1)
              ((VBD 0)
               $s(token (text "hurt")
                        (classes ("hurt" *item* VBD))
                        (location 92)))))
            ((INDIRQUOT 2)
             ((NP 1)
              ((DET 6))
              ((MOD 34))
              ((HEAD 2)
               ((NNS 0)
                $s(token (text "results")
                         (classes ("results" *item* NNS))
                         (location 93)))))
             ((INDIRQUOT 4)))))))))
<S> Sales/NNS from/IN the/DT division/NN slipped/VBD 7.5%/CD to/TO 161.080/CD trillion/CD
    yen/NN ././*end-of-sentence*</S>
<S> <RS4> A/DT Fujitsu/NNP spokesman/NN said/RS_VBD the/DT weaker/JJR performance/NN
    reflected/VBD slipping/VBG sales/NNS to/TO Nippon/NNP Telegraph/NNP &/CC Telephone/NNP
    Corp./NNP ,/,/*comma* Japan/NNP 's/POS domestic/JJ telephone/NN monopoly/NN </RS>
    ././/*end-of-sentence*</S>
$s(rs (text "A Fujitsu spokesman said the weaker performance reflected slipping sales to
            Nippon Telegraph & Telephone Corp. ,/, Japan 's domestic telephone monopoly")
      (no 4)
      (start 110)
      (end 131)
      (textsource "A Fujitsu spokesman ")
      (textverb "said ")
      (reported_speech
        "the weaker performance reflected slipping sales to Nippon Telegraph & Telephone
         Corp. ,/, Japan 's domestic telephone monopoly ")
      (textcirc "")
      (textdirq "")
      (textindirq
        "the weaker performance reflected slipping sales to Nippon Telegraph & Telephone
         Corp. ,/, Japan 's domestic telephone monopoly ")
      (rep_source
        $s(np (npno 4)
              (det ((DET 1)
                    ((DT 0)
                     $s(token (text "A")
                              (classes ("A" *item* DT))
                              (location 110)))))
              (mod ((MOD 17)
                    ((PR_NAME 1)
                     ((NNP 0)
                      $s(token (text "Fujitsu")
                               (classes ("Fujitsu" *item* NNP))
                               (location 111))))))
```

```
        (head ((HEAD 1)
               ((NN 0)
                #s(token (text "spokesman")
                         (classes ("spokesman" *item* NN))
                         (location 112)))))
        (coreferences ())
        (classification ())
        (comma ())))
 (rep_verb
   #s(vc (vcno 4)
         (modal ())
         (auxs ())
         (mainverb
           ((RS_MAIN_VERB_FINITE 1)
            ((RS_VBD 0)
             #s(token (text "said")
                      (classes ("said" *item* RS_VBD))
                      (location 113)))))
         (adverbs ())
         (tense past)
         (aspect indefinite)
         (voice active)
         (modality ())
         (mood ())
         (polarity ())
         (semantic_dimensions
           (STR_unmarked
             AFF_unmarked
             SPA_inform
             PRESP_unmarked
             POL_positive
             AUD_unmarked
             FRM_unmarked
             EXP_explicit
             V_unmarked))))
 (othercirc ())
 (dirquots ())
 (indirquots
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       #s(token (text "the") (classes ("the" *item* DT)) (location 114))))
     ((MOD 14)
      ((ADJM 2)
       ((JJR 0)
        #s(token (text "weaker")
                 (classes ("weaker" *item* JJR))
                 (location 115)))))
     ((HEAD 1)
      ((NN 0)
```

```
   #s(token (text "performance")
            (classes ("performance" *item* NN))
            (location 116)))))
((INDIRQUOT 3)
 ((VC 16)
  ((ADVERBS 4))
  ((MAIN_VERB_FINITE 1)
   ((VBD 0)
    #s(token (text "reflected")
             (classes ("reflected" *item* VBD))
             (location 117)))))
 ((INDIRQUOT 2)
  ((NP 1)
   ((DET 6))
   ((MOD 14)
    ((ADJM 3)
     ((VBG 0)
      #s(token (text "slipping")
               (classes ("slipping" *item* VBG))
               (location 118)))))
   ((HEAD 2)
    ((NNS 0)
     #s(token (text "sales")
              (classes ("sales" *item* NNS))
              (location 119)))))
  ((INDIRQUOT 1)
   ((ANY 6)
    ((TO 0)
     #s(token (text "to") (classes ("to" *item* TO)) (location 120))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 6))
     ((MOD 34))
     ((HEAD 4)
      ((PR_NAME 3)
       ((NNP 0)
        #s(token (text "Nippon")
                 (classes ("Nippon" *item* NNP))
                 (location 121)))
       ((PR_NAME 1)
        ((NNP 0)
         #s(token (text "Telegraph")
                  (classes ("Telegraph" *item* NNP))
                  (location 122)))))))
    ((INDIRQUOT 1)
     ((ANY 7)
      ((CC 0)
       #s(token (text "&") (classes ("&" *item* CC)) (location 123))))
     ((MANY 2))
     ((INDIRQUOT 2)
```

```
((NP 1)
 ((DET 6))
 ((MOD 34))
 ((HEAD 4)
  ((PR_NAME 3)
   ((NNP 0)
    #s(token (text "Telephone")
             (classes ("Telephone" *item* NNP))
             (location 124)))
   ((PR_NAME 1)
    ((NNP 0)
     #s(token (text "Corp.")
              (classes ("Corp." *item* NNP))
              (location 125)))))))
((INDIRQUOT 1)
 ((ANY 20)
  ((*comma* 0)
   #s(token (text ",/,")
            (classes (",/," *comma*))
            (location 126))))
 ((MANY 2))
 ((INDIRQUOT 2)
  ((NP 1)
   ((DET 6))
   ((MOD 34))
   ((HEAD 4)
    ((PR_NAME 7)
     ((NNP 0)
      #s(token (text "Japan")
               (classes ("Japan" *item* NNP))
               (location 127)))
     ((POS 0)
      #s(token (text "'s")
               (classes ("'s" *item* POS))
               (location 128))))))
  ((INDIRQUOT 1)
   ((ANY 26)
    ((ADJFORM 1)
     ((JJ 0)
      #s(token (text "domestic")
               (classes ("domestic" *item* JJ))
               (location 129)))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 6))
     ((MOD 34))
     ((HEAD 1)
      ((NN 0)
       #s(token (text "telephone")
                (classes ("telephone" *item* NN))
```

```
                    (location 130)))))
                ((INDIRQUOT 2)
                 ((NP 1)
                  ((DET 6))
                  ((MOD 34))
                  ((HEAD 1)
                   ((NN 0)
                    #s(token (text "monopoly")
                             (classes ("monopoly" *item* NN))
                             (location 131)))))
                 ((INDIRQUOT 4)))))))))))))))
<S> Total/JJ sales/NNS rose/VBD 4.7%/CD to/TO 966.471/CD trillion/CD yen/NN
    ././*end-of-sentence*</S>
<S> Sales/NNS of/IN electrical/JJ devices/NNS rose/VBD about/IN 19%/CD to/TO 147.874/CD
    trillion/CD yen/NN ,/,/*comma* while/IN sales/NNS in/IN the/DT information/NN systems/NNS
    division/NN gained/VBD 5.4%/CD to/TO 657.517/CD trillion/CD yen/NN ././*end-of-sentence*</S>
```

D.1 Example of problematic output

Wall Street Journal, 10/27/89

(S1) A potentially safer whooping cough vaccine made by novel genetic engineering techniques was described by a team of Italian, U.S. and Japanese scientists.

(S2) The team reported they managed to induce bacteria to produce a non-toxic version of the poisons produced by the bacterium that causes whooping cough. (S3) Laboratory tests showed that non-toxic versions of the poisons are capable of inducing an immunity to whooping cough, the researchers reported in this week's issue of the journal Science.

(NR1) The current vaccine for whooping cough, or pertussis, is part of the "DPT" (for diphtheria, pertussis, tetanus) shot given most infants and young children. (NR2) The vaccine is effective in preventing a disease that still afflicts about 60 million children a year world-wide, causing an estimated one million deaths.

(NR3) The vaccine, however, causes allergic reactions that can be fatal. (NR4) The reactions stem from the fact that the vaccine contains multiple copies of the whole Bordetella pertussis bacterium, which causes whooping cough. (NR5) This bacterium produces a toxin that, if used as a vaccine, can induce immunity to whooping cough. (NR6) Unfortunately, the toxin is also poisonous.

(S4) The Italian-led scientific team said they had succeeded in getting bacteria to produce a non-toxic version of the pertussis toxin, which could be used as a safe vaccine. (S5) The researchers reported they have been able to pluck the five genes that produced the toxin out of the pertussis bacterium. (NR7) It turned out that although it took all five genes to produce the toxin, only one was responsible for the toxin's virulence.

(NR8) Ordinarily in genetic engineering each of these genes, minus the one that caused the virulence, would have been transferred to another bacterium, called E. coli, which would then produce a nonvirulent version of the toxin. (S6) The researchers said they did this, but the toxin didn't induce immunity to whooping cough.

(NR9) The scientists then took the five toxin genes and triggered a mutation in the one gene that caused virulence. (NR10) Then, using a new technique (called homologous recombination) for introducing genes into cells, they transferred all five genes to bacteria closely related to the pertussis organism. (NR11) These bacterial "cousins" ordinarily don't make the toxin. (NR12) But the genes were accompanied by a piece of DNA, called a promoter, that turns the genes on.

(NR13) The new bacteria recipients of the genes began producing pertussis toxin which, because of the mutant virulence gene, was no longer toxic. (S7) Experiments showed that the new, non-virulent toxin is capable of inducing immunity, according to the researchers from the Selavo Research Center in Siena, Italy, the Medical College of Wisconsin in Milwaukee and the Japanese National Institutes of Health.

(S1)-(S7) represent reported speech instances, corresponding to seven basic profiles. (NR1)-(NR13) are non-reported speech.

(S1) sentence reports what the reporters announced by using the verb "described," which is not considered as a reported verb in the implementation. Hence, the basic profile corresponding to it is not found.

(S3) presents the other circumstantial information "in this week's issue of the journal Science." The preposition "in" was not defined in the grammar as fronting other circumstantial information, due to its very high frequency while used in other situations. Hence, the basic profile corresponding to (S3) is not found.

(S6) contains the determiner "this" as an object. There is no rule in the grammar to account for determiners appearing alone. Hence, the profile corresponding to it is not found. The NP grammar for ERSS, as presented to the DUC 2003 conference, accounts for this case.

(S7) uses the phrasal verb "according to" to report what the source said. Phrasal verbs were not treated in this thesis, so they are not found. Hence, the profile corresponding to (S7) is not found.

The sytem found three profiles in this "bad" example. Only the profiles that were found are presented below.

```
<S> <RS1> The/DT team/NN reported/RS_VBD they/PRP managed/VBD to/TO induce/VB bacteria/NNS
    to/TO produce/VB a/DT non-toxic/JJ version/NN of/IN the/DT poisons/NNS produced/VBN by/IN
    the/DT bacterium/NN that/IN causes/NNS whooping/JJ cough/NN </RS> ./.*end-of-sentence*</S>
$s(rs (text "The team reported they managed to induce bacteria to produce a non-toxic version
            of the poisons produced by the bacterium that causes whooping cough")
     (no 1)
     (start 31)
     (end 54)
     (textsource "The team ")
     (textverb "reported ")
     (reported_speech
       "they managed to induce bacteria to produce a non-toxic version of the poisons
        produced by the bacterium that causes whooping cough ")
     (textcirc "")
     (textdirq "")
     (textindirq
       "they managed to induce bacteria to produce a non-toxic version of the poisons
        produced by the bacterium that causes whooping cough ")
     (rep_source
       $s(np (npno 1)
             (det ((DET 1)
                   ((DT 0)
                    $s(token (text "The")
```

```
                    (classes ("The" *item* DT))
                    (location 31)))))
        (mod ((MOD 34)))
        (head ((HEAD 1)
               ((NN 0)
                #s(token (text "team")
                         (classes ("team" *item* NN))
                         (location 32)))))
        (coreferences ())
        (classification ())
        (comma ())))
 (rep_verb
   #s(vc (vcno 1)
         (modal ())
         (auxs ())
         (mainverb
           ((RS_MAIN_VERB_FINITE 1)
            ((RS_VBD 0)
             #s(token (text "reported")
                      (classes ("reported" *item* RS_VBD))
                      (location 33)))))
         (adverbs ())
         (tense past)
         (aspect indefinite)
         (voice active)
         (modality ())
         (mood ())
         (polarity ())
         (semantic_dimensions
           (STR_unmarked
             AFF_unmarked
             SPA_inform
             PRESP_new
             POL_positive
             AUD_unmarked
             FRM_unmarked
             EXP_explicit
             V_unmarked))))
 (othercirc ())
 (dirquots ())
 (indirquots
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 6))
     ((MOD 34))
     ((HEAD 3)
      ((PRP 0)
       #s(token (text "they")
                (classes ("they" *item* PRP))
                (location 34)))))
   ((INDIRQUOT 3)
```

```
((VC 16)
 ((ADVERBS 4))
 ((MAIN_VERB_FINITE 1)
  ((VBD 0)
   #s(token (text "managed")
            (classes ("managed" *item* VBD))
            (location 35)))))
((INDIRQUOT 1)
 ((ANY 4)
  ((NON_FINITE_VC 1)
   ((TO 0)
    #s(token (text "to") (classes ("to" *item* TO)) (location 36)))
   ((VB 0)
    #s(token (text "induce")
             (classes ("induce" *item* VB))
             (location 37)))))
 ((MANY 2))
 ((INDIRQUOT 2)
  ((NP 1)
   ((DET 6))
   ((MOD 34))
   ((HEAD 2)
    ((NNS 0)
     #s(token (text "bacteria")
              (classes ("bacteria" *item* NNS))
              (location 38)))))
  ((INDIRQUOT 1)
   ((ANY 4)
    ((NON_FINITE_VC 1)
     ((TO 0)
      #s(token (text "to") (classes ("to" *item* TO)) (location 39)))
     ((VB 0)
      #s(token (text "produce")
               (classes ("produce" *item* VB))
               (location 40)))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       #s(token (text "a") (classes ("a" *item* DT)) (location 41))))
     ((MOD 14)
      ((ADJM 1)
       ((JJ 0)
        #s(token (text "non-toxic")
                 (classes ("non-toxic" *item* JJ))
                 (location 42)))))
     ((HEAD 1)
      ((NN 0)
       #s(token (text "version")
                (classes ("version" *item* NN))
```

```
          (location 43)))))
((INDIRQUOT 1)
 ((ANY 5)
  ((IN 0)
   #s(token (text "of")
            (classes ("of" *item* IN))
            (location 44))))
 ((MANY 2))
 ((INDIRQUOT 1)
  ((ANY 4)
   ((NON_FINITE_VC 4)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       #s(token (text "the")
                (classes ("the" *item* DT))
                (location 45))))
     ((MOD 34))
     ((HEAD 2)
      ((NNS 0)
       #s(token (text "poisons")
                (classes ("poisons" *item* NNS))
                (location 46)))))
    ((VBN 0)
     #s(token (text "produced")
              (classes ("produced" *item* VBN))
              (location 47)))))
  ((MANY 1)
   ((ANY 5)
    ((IN 0)
     #s(token (text "by")
              (classes ("by" *item* IN))
              (location 48))))
   ((MANY 2)))
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 1)
     ((DT 0)
      #s(token (text "the")
               (classes ("the" *item* DT))
               (location 49))))
    ((MOD 34))
    ((HEAD 1)
     ((NN 0)
      #s(token (text "bacterium")
               (classes ("bacterium" *item* NN))
               (location 50)))))
   ((INDIRQUOT 1)
    ((ANY 5)
     ((IN 0)
      #s(token (text "that")
```

```
                    (classes ("that" *item* IN))
                    (location 51))))
                ((MANY 2))
                ((INDIRQUOT 2)
                 ((NP 1)
                  ((DET 6))
                  ((MOD 34))
                  ((HEAD 2)
                   ((NNS 0)
                    #s(token (text "causes")
                             (classes ("causes" *item* NNS))
                             (location 52)))))
                 ((INDIRQUOT 1)
                  ((ANY 26)
                   ((ADJFORM 1)
                    ((JJ 0)
                     #s(token (text "whooping")
                              (classes ("whooping" *item* JJ))
                              (location 53)))))
                  ((MANY 2))
                  ((INDIRQUOT 2)
                   ((NP 1)
                    ((DET 6))
                    ((MOD 34))
                    ((HEAD 1)
                     ((NN 0)
                      #s(token (text "cough")
                               (classes ("cough" *item* NN))
                               (location 54)))))
                   ((INDIRQUOT 4)))))))))))))))))
<S> <RS2> The/DT Italian-led/JJ scientific/JJ team/NN said/RS_VBD they/PRP had/HAVE_VBD
    succeeded/VBN in/IN getting/VBG bacteria/NNS to/TO produce/VB a/DT non-toxic/JJ version/NN
    of/IN the/DT pertussis/NN toxin/NN ,/,/*comma* which/WDT could/MD be/BE_VB used/VBN as/IN
    a/DT safe/JJ vaccine/NN </RS> ././*end-of-sentence*</S>
#s(rs (text "The Italian-led scientific team said they had succeeded in getting bacteria to
             produce a non-toxic version of the pertussis toxin ,/, which could be used as
             a safe vaccine")
      (no 2)
      (start 232)
      (end 260)
      (textsource "The Italian-led scientific team ")
      (textverb "said ")
      (reported_speech
        "they had succeeded in getting bacteria to produce a non-toxic version of the
         pertussis toxin ,/, which could be used as a safe vaccine ")
      (textcirc "")
      (textdirq "")
      (textindirq
        "they had succeeded in getting bacteria to produce a non-toxic version of the
         pertussis toxin ,/, which could be used as a safe vaccine ")
      (rep_source
```

```
  #s(np (npno 2)
        (det ((DET 1)
              ((DT 0)
               #s(token (text "The")
                        (classes ("The" *item* DT))
                        (location 232)))))
        (mod ((MOD 14)
              ((ADJM 5)
               ((JJ 0)
                #s(token (text "Italian-led")
                         (classes ("Italian-led" *item* JJ))
                         (location 233)))
               ((ADJM 1)
                ((JJ 0)
                 #s(token (text "scientific")
                          (classes ("scientific" *item* JJ))
                          (location 234)))))))
        (head ((HEAD 1)
               ((NN 0)
                #s(token (text "team")
                         (classes ("team" *item* NN))
                         (location 235)))))
        (coreferences ())
        (classification ())
        (comma ())))
(rep_verb
  #s(vc (vcno 2)
        (modal ())
        (auxs ())
        (mainverb
          ((RS_MAIN_VERB_FINITE 1)
           ((RS_VBD 0)
            #s(token (text "said")
                     (classes ("said" *item* RS_VBD))
                     (location 236)))))
        (adverbs ())
        (tense past)
        (aspect indefinite)
        (voice active)
        (modality ())
        (mood ())
        (polarity ())
        (semantic_dimensions
          (STR_unmarked
            AFF_unmarked
            SPA_inform
            PRESP_unmarked
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
```

```
          V_unmarked))))
(othercirc ())
(dirquots ())
(indirquots
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 6))
    ((MOD 34))
    ((HEAD 3)
     ((PRP 0)
      #s(token (text "they")
               (classes ("they" *item* PRP))
               (location 237)))))
   ((INDIRQUOT 3)
    ((VC 9)
     ((ADVERBS 4))
     ((AUX_HAVE_FINITE 1)
      ((HAVE_VBD 0)
       #s(token (text "had")
                (classes ("had" *item* HAVE_VBD))
                (location 238))))
     ((ADVERBS 4))
     ((INV_OBJ 2))
     ((ADVERBS 4))
     ((MAIN_VERB_VBN 1)
      ((VBN 0)
       #s(token (text "succeeded")
                (classes ("succeeded" *item* VBN))
                (location 239)))))
    ((INDIRQUOT 1)
     ((ANY 5)
      ((IN 0)
       #s(token (text "in") (classes ("in" *item* IN)) (location 240))))
     ((MANY 2))
     ((INDIRQUOT 1)
      ((ANY 8)
       ((ADJM 3)
        ((VBG 0)
         #s(token (text "getting")
                  (classes ("getting" *item* VBG))
                  (location 241)))))
      ((MANY 2))
      ((INDIRQUOT 2)
       ((NP 1)
        ((DET 6))
        ((MOD 34))
        ((HEAD 2)
         ((NNS 0)
          #s(token (text "bacteria")
                   (classes ("bacteria" *item* NNS))
                   (location 242)))))
```

```
((INDIRQUOT 1)
 ((ANY 4)
  ((NON_FINITE_VC 1)
   ((TO 0)
    #s(token (text "to")
             (classes ("to" *item* TO))
             (location 243)))
   ((VB 0)
    #s(token (text "produce")
             (classes ("produce" *item* VB))
             (location 244)))))
 ((MANY 2))
 ((INDIRQUOT 2)
  ((NP 1)
   ((DET 1)
    ((DT 0)
     #s(token (text "a")
              (classes ("a" *item* DT))
              (location 245))))
   ((MOD 14)
    ((ADJM 1)
     ((JJ 0)
      #s(token (text "non-toxic")
               (classes ("non-toxic" *item* JJ))
               (location 246)))))
   ((HEAD 1)
    ((NN 0)
     #s(token (text "version")
              (classes ("version" *item* NN))
              (location 247)))))
  ((INDIRQUOT 1)
   ((ANY 5)
    ((IN 0)
     #s(token (text "of")
              (classes ("of" *item* IN))
              (location 248))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       #s(token (text "the")
                (classes ("the" *item* DT))
                (location 249))))
     ((MOD 11)
      ((CNM 1)
       ((NN 0)
        #s(token (text "pertussis")
                 (classes ("pertussis" *item* NN))
                 (location 250)))))
     ((HEAD 1)
```

```
  ((NN 0)
   $s(token (text "toxin")
            (classes ("toxin" *item* NN))
            (location 251)))))
((INDIRQUOT 1)
 ((ANY 20)
  ((*comma* 0)
   $s(token (text ",/,")
            (classes (",/," *comma*))
            (location 252))))
 ((MANY 1)
  ((ANY 14)
   ((WDT 0)
    $s(token (text "which")
             (classes ("which" *item* WDT))
             (location 253))))
  ((MANY 2)))
 ((INDIRQUOT 3)
  ((VC 8)
   ((ADVERBS 4))
   ((MD 0)
    $s(token (text "could")
             (classes ("could" *item* MD))
             (location 254)))
   ((ADVERBS 4))
   ((INV_OBJ 2))
   ((ADVERBS 4))
   ((AUX_BE_VB 1)
    ((BE_VB 0)
     $s(token (text "be")
              (classes ("be" *item* BE_VB))
              (location 255))))
   ((ADVERBS 4))
   ((MAIN_VERB_VBN 1)
    ((VBN 0)
     $s(token (text "used")
              (classes ("used" *item* VBN))
              (location 256)))))
  ((INDIRQUOT 1)
   ((ANY 5)
    ((IN 0)
     $s(token (text "as")
              (classes ("as" *item* IN))
              (location 257))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       $s(token (text "a")
                (classes ("a" *item* DT))
```

```
                              (location 258))))
                        ((MOD 14)
                         ((ADJM 1)
                          ((JJ 0)
                           $s(token (text "safe")
                                    (classes ("safe" *item* JJ))
                                    (location 259)))))
                        ((HEAD 1)
                         ((NN 0)
                          $s(token (text "vaccine")
                                   (classes ("vaccine" *item* NN))
                                   (location 260)))))
                       ((INDIRQUOT 4))))))))))))))))
<S> <RS3> The/DT researchers/NNS reported/RS_VBD they/PRP have/HAVE_VBP been/BE_VBN able/JJ
    to/TO pluck/VB the/DT five/CD genes/NNS that/WDT produced/VBD the/DT toxin/NN out/IN
    of/IN the/DT pertussis/NN bacterium/NN </RS> ./.*end-of-sentence*</S>
$s(rs (text "The researchers reported they have been able to pluck the five genes that
            produced the toxin out of the pertussis bacterium")
      (no 3)
      (start 264)
      (end 284)
      (textsource "The researchers ")
      (textverb "reported ")
      (reported_speech
        "they have been able to pluck the five genes that produced the toxin out of the
         pertussis bacterium ")
      (textcirc "")
      (textdirq "")
      (textindirq
        "they have been able to pluck the five genes that produced the toxin out of the
         pertussis bacterium ")
      (rep_source
        $s(np (npno 3)
              (det ((DET 1)
                    ((DT 0)
                     $s(token (text "The")
                              (classes ("The" *item* DT))
                              (location 264)))))
              (mod ((MOD 34)))
              (head ((HEAD 2)
                     ((NNS 0)
                      $s(token (text "researchers")
                               (classes ("researchers" *item* NNS))
                               (location 265)))))
              (coreferences ())
              (classification ())
              (comma ())))
      (rep_verb
        $s(vc (vcno 3)
              (modal ())
              (auxs ())
```

```
        (mainverb
          ((RS_MAIN_VERB_FINITE 1)
           ((RS_VBD 0)
            #s(token (text "reported")
                     (classes ("reported" *item* RS_VBD))
                     (location 266)))))
        (adverbs ())
        (tense past)
        (aspect indefinite)
        (voice active)
        (modality ())
        (mood ())
        (polarity ())
        (semantic_dimensions
          (STR_unmarked
            AFF_unmarked
            SPA_inform
            PRESP_new
            POL_positive
            AUD_unmarked
            FRM_unmarked
            EXP_explicit
            V_unmarked))))
(othercirc ())
(dirquots ())
(indirquots
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 6))
    ((MOD 34))
    ((HEAD 3)
     ((PRP 0)
      #s(token (text "they")
               (classes ("they" *item* PRP))
               (location 267)))))
   ((INDIRQUOT 3)
    ((VC 9)
     ((ADVERBS 4))
     ((AUX_HAVE_FINITE 2)
      ((HAVE_VBP 0)
       #s(token (text "have")
                (classes ("have" *item* HAVE_VBP))
                (location 268))))
     ((ADVERBS 4))
     ((INV_OBJ 2))
     ((ADVERBS 4))
     ((MAIN_VERB_VBN 3)
      ((BE_VBN 0)
       #s(token (text "been")
                (classes ("been" *item* BE_VBN))
                (location 269)))))
```

```
((INDIRQUOT 1)
 ((ANY 8)
  ((ADJM 1)
   ((JJ 0)
    #s(token (text "able")
             (classes ("able" *item* JJ))
             (location 270)))))
 ((MANY 2))
 ((INDIRQUOT 1)
  ((ANY 4)
   ((NON_FINITE_VC 1)
    ((TO 0)
     #s(token (text "to") (classes ("to" *item* TO)) (location 271)))
    ((VB 0)
     #s(token (text "pluck")
              (classes ("pluck" *item* VB))
              (location 272)))))
  ((MANY 2))
  ((INDIRQUOT 2)
   ((NP 1)
    ((DET 1)
     ((DT 0)
      #s(token (text "the")
               (classes ("the" *item* DT))
               (location 273))))
    ((MOD 28)
     ((AMOUNT 1)
      ((CD 0)
       #s(token (text "five")
                (classes ("five" *item* CD))
                (location 274)))))
    ((HEAD 2)
     ((NNS 0)
      #s(token (text "genes")
               (classes ("genes" *item* NNS))
               (location 275)))))
   ((INDIRQUOT 1)
    ((ANY 14)
     ((WDT 0)
      #s(token (text "that")
               (classes ("that" *item* WDT))
               (location 276))))
    ((MANY 2))
    ((INDIRQUOT 3)
     ((VC 16)
      ((ADVERBS 4))
      ((MAIN_VERB_FINITE 1)
       ((VBD 0)
        #s(token (text "produced")
                 (classes ("produced" *item* VBD))
                 (location 277)))))
```

```
((INDIRQUOT 2)
 ((NP 1)
  ((DET 1)
   ((DT 0)
    #s(token (text "the")
             (classes ("the" *item* DT))
             (location 278))))
  ((MOD 34))
  ((HEAD 1)
   ((NN 0)
    #s(token (text "toxin")
             (classes ("toxin" *item* NN))
             (location 279)))))
 ((INDIRQUOT 1)
  ((ANY 5)
   ((IN 0)
    #s(token (text "out")
             (classes ("out" *item* IN))
             (location 280))))
  ((MANY 2))
  ((INDIRQUOT 1)
   ((ANY 5)
    ((IN 0)
     #s(token (text "of")
              (classes ("of" *item* IN))
              (location 281))))
   ((MANY 2))
   ((INDIRQUOT 2)
    ((NP 1)
     ((DET 1)
      ((DT 0)
       #s(token (text "the")
                (classes ("the" *item* DT))
                (location 282))))
     ((MOD 34))
     ((HEAD 1)
      ((NN 0)
       #s(token (text "pertussis")
                (classes ("pertussis" *item* NN))
                (location 283)))))
    ((INDIRQUOT 2)
     ((NP 1)
      ((DET 6))
      ((MOD 34))
      ((HEAD 1)
       ((NN 0)
        #s(token (text "bacterium")
                 (classes ("bacterium" *item* NN))
                 (location 284)))))
     ((INDIRQUOT 4))))))))))))))))
```