October 1993

Building a Large Annotated Corpus of English: The Penn Treebank

Mitchell Marcus
University of Pennsylvania, mitch@cis.upenn.edu

Beatrice Santorini
University of Pennsylvania

Mary Ann Marcinkiewicz
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-93-87.

Building a Large Annotated Corpus of English: The Penn Treebank

Abstract

In this paper, we review our experience with constructing one such large annotated corpus--the Penn Treebank, a corpus consisting of over 4.5 million words of American English. During the first three-year phase of the Penn Treebank Project (1989-1992), this corpus has been annotated for part-of-speech (POS) information. In addition, over half of it has been annotated for skeletal syntactic structure.

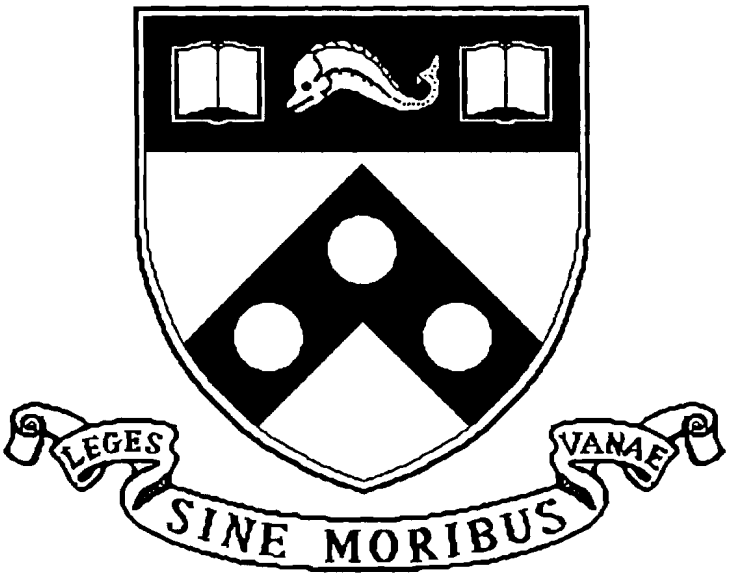
Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-93-87.

Building A Large Annotated Corpus of English: The Penn Treebank

MS-CIS-93-87
LINC LAB 260

Mitchell P. Marcus
Beatrice Santorini
Mary Ann Marcinkiewicz



University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

October 1993

Building a large annotated corpus of English: the Penn Treebank

Mitchell P. Marcus* Beatrice Santorini† Mary Ann Marcinkiewicz*

*Department of Computer and Information Sciences
University of Pennsylvania
Philadelphia, PA 19104

†Department of Linguistics
Northwestern University
Evanston, IL 60208

1 Introduction

There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora.[1] Such corpora are beginning to serve as an important research tool for investigators in natural language processing, speech recognition, and integrated spoken language systems, as well as in theoretical linguistics. Annotated corpora promise to be valuable for enterprises as diverse as the automatic construction of statistical models for the grammar of the written and the colloquial spoken language, the development of explicit formal theories of the differing grammars of writing and speech, the investigation of prosodic phenomena in speech, and the evaluation and comparison of the adequacy of parsing models.

In this paper, we review our experience with constructing one such large annotated corpus—the Penn Treebank, a corpus[2] consisting of over 4.5 million words of American English. During the first three-year phase of the Penn Treebank Project (1989–1992), this corpus has been annotated for part-of-speech (POS) information. In addition, over half of it has been annotated for skeletal syntactic structure. These materials are available to members of the Linguistic Data Consortium; for details, see section 5.1.

The paper is organized as follows. Section 2 discusses the POS tagging task. After outlining the considerations that informed the design of our POS tagset and presenting the tagset itself, we describe our two-stage tagging process, in which text is first assigned POS tags automatically and

[1]The work reported here was partially supported by DARPA grant No. N0014-85-K0018, by DARPA and AFOSR jointly under grant No. AFOSR-90-0066 and by ARO grant No. DAAL 03-89-C0031 PRI. Seed money was provided by the General Electric Corporation under grant No. J01746000. We gratefully acknowledge this support. We would also like to acknowledge the contribution of the annotators who have worked on the Penn Treebank Project: Florence Dong, Leslie Dossey, Mark Ferguson, Lisa Frank, Elizabeth Hamilton, Alissa Hinckley, Chris Hudson, Karen Katz, Grace Kim, Robert MacIntyre, Mark Parisi, Britta Schasberger, Victoria Tredinnick and Matt Waters; in addition, Rob Foye, David Magerman, Richard Pito and Steven Shapiro deserve our special thanks for their administrative and programming support. We are grateful to AT&T Bell Labs for permission to use Kenneth Church's PARTS part-of-speech labeller and Donald Hindle's Fidditch parser. Finally, we would like to thank Sue Marcus for sharing with us her statistical expertise and providing the analysis of the time data of the experiment reported in section 3. The design of that experiment is due to the first two authors; they alone are responsible for its shortcomings.

[2]A distinction is sometimes made between a *corpus* as a carefully structured set of materials gathered together to jointly meet some design principles, as opposed to a *collection*, which may be much more opportunistic in construction. We acknowledge that from this point of view, the raw materials of the Penn Treebank form a collection.

then corrected by human annotators. Section 3 briefly presents the results of a comparison between entirely manual and semi-automated tagging, with the latter being shown to be superior on three counts: speed, consistency, and accuracy. In section 4, we turn to the bracketing task. Just as with the tagging task, we have partially automated the bracketing task: the output of the POS tagging phase is automatically parsed and simplified to yield a skeletal syntactic representation, which is then corrected by human annotators. After presenting the set of syntactic tags that we use, we illustrate and discuss the bracketing process. In particular, we will outline various factors that affect the speed with which annotators are able to correct bracketed structures, a task which—not surprisingly—is considerably more difficult than correcting POS-tagged text. Finally, section 5 describes the composition and size of the current Treebank corpus, briefly reviews some of the research projects that have relied on it to date, and indicates the directions that the project is likely to take in the future.

2 Part-of-speech tagging

2.1 A simplified POS tagset for English

The POS tagsets used to annotate large corpora in the past have traditionally been fairly extensive. The pioneering Brown Corpus distinguishes 87 simple tags ([Francis 1964]), [Francis and Kučera 1982]) and allows the formation of compound tags; thus, the contraction *I'm* is tagged as PPSS+BEM (PPSS for "non-3rd person nominative personal pronoun" and BEM for "am, 'm".[3] Subsequent projects have tended to elaborate the Brown Corpus tagset. For instance, the Lancaster-Oslo/Bergen (LOB) Corpus uses about 135 tags, the Lancaster UCREL group about 165 tags, and the London-Lund Corpus of Spoken English 197 tags.[4] The rationale behind developing such large, richly articulated tagsets is to approach "the ideal of providing distinct codings for all classes of words having distinct grammatical behaviour" ([Garside et al 1987, 167]).

2.1.1 Recoverability

Like the tagsets just mentioned, the Penn Treebank tagset is based on that of the Brown Corpus. However, the stochastic orientation of the Penn Treebank and the resulting concern with sparse data led us to modify the Brown Corpus tagset by paring it down considerably. A key strategy in reducing the tagset was to eliminate redundancy by taking into account both lexical and syntactic information. Thus, whereas many POS tags in the Brown Corpus tagset are unique to a particular lexical item, the Penn Treebank tagset strives to eliminate such instances of lexical redundancy. For instance, the Brown Corpus distinguishes five different forms for main verbs: the base form is tagged VB, and forms with overt endings are indicated by appending D for past tense, G for present participle/gerund, N for past participle and Z for third person singular present. Exactly the same paradigm is recognized for the *have*, but *have* (regardless of whether it is used as an auxiliary or a main verb) is assigned its own base tag HV. The Brown Corpus further distinguishes three forms

[3] Counting both simple and compound tags, the Brown Corpus tagset contains 187 tags.

[4] A useful overview of the relation of these and other tagsets to each other and to the Brown Corpus tagset is given in Appendix B of [Garside et al 1987].

of *do*—the base form (DO), the past tense (DOD), and the third person singular present (DOZ),[5] and eight forms of *be*—the five forms distinguished for regular verbs as well as the irregular forms *am* (BEM), *are* (BER) and *was* (BEDZ). By contrast, since the distinctions between the forms of VB on the one hand and the forms of BE, DO and HV on the other are lexically recoverable, they are eliminated in the Penn Treebank, as shown in Table 1.[6]

Table 1:
Elimination of lexically recoverable distinctions

sing/VB	be/VB	do/VB	have/VB
sings/VBZ	is/VBZ	does/VBZ	has/VBZ
sang/VBD	was/VBD	did/VBD	had/VBD
singing/VBG	being/VBG	doing/VBG	having/VBG
sung/VBN	been/VBN	done/VBN	had/VBN

A second example of lexical recoverability concerns those words that can precede articles in noun phrases. The Brown Corpus assigns a separate tag to pre-qualifiers (*quite*, *rather*, *such*), pre-quantifiers (*all*, *half*, *many*, *nary*) and *both*. The Penn Treebank, on the other hand, assigns all of these words to a single category PDT (predeterminer). Further examples of lexically recoverable categories are the Brown Corpus categories PPL (singular reflexive pronoun) and PPLS (plural reflexive pronoun), which we collapse with PRP (personal pronoun), and the Brown Corpus category RN (nominal adverb), which we collapse with RB (adverb).

Beyond reducing lexically recoverable distinctions, we also eliminated certain POS distinctions that are recoverable with reference to syntactic structure. For instance, the Penn Treebank tagset does not distinguish subject pronouns from object pronouns even in cases where the distinction is not recoverable from the pronoun's form, as with *you*, since the distinction is recoverable on the basis of the pronoun's position in the parse tree in the parsed version of the corpus. Similarly, the Penn Treebank tagset conflates subordinating conjunctions with prepositions, tagging both categories as IN. The distinction between the two categories is not lost, however, since subordinating conjunctions can be recovered as those instances of IN that precede clauses, whereas prepositions are those instances of IN that precede noun phrases or prepositional phrases. We would like to emphasize that the lexical and syntactic recoverability inherent in the POS-tagged version of the Penn Treebank corpus allows end users to employ a much richer tagset than the small one described in section 2.2 if the need arises.

2.1.2 Consistency

As noted above, one reason for eliminating a POS tag such as RN (nominal adverb) is its lexical recoverability. Another important reason for doing so is consistency. For instance, in the Brown

[5]The gerund and the participle of *do* are tagged VBG and VBN in the Brown Corpus, respectively—presumably because they are never used as auxiliary verbs.

[6]The irregular present tense forms *am* and *are* are tagged as VBP in the Penn Treebank (see section 2.1.3), just like any other non-third person singular present tense form.

Corpus, the deictic adverbs *there* and *now* are always tagged RB (adverb), whereas their counterparts *here* and *then* are inconsistently tagged as RB (adverb) or RN (nominal adverb)—even in identical syntactic contexts, such as after a preposition. It is clear that reducing the size of the tagset reduces the chances of such tagging inconsistencies.

2.1.3 Syntactic function

A further difference between the Penn Treebank and the Brown Corpus concerns the significance accorded to syntactic context. In the Brown Corpus, words tend to be tagged independently of their syntactic function.[7] For instance, in the phrase *the one*, *one* is always tagged as CD (cardinal number), whereas in the corresponding plural phrase *the ones*, *ones* is always tagged as NNS (plural common noun), despite the parallel function of *one* and *ones* as heads of their noun phrase. By contrast, since one of the main roles of the tagged version of the Penn Treebank corpus is to serve as the basis for a bracketed version of the corpus, we encode a word's syntactic function in its POS tag whenever possible. Thus, *one* is tagged as NN (singular common noun) rather than as CD (cardinal number) when it is the head of a noun phrase. Similarly, while the Brown Corpus tags *both* as ABX (pre-quantifier, double conjunction), regardless of whether it functions as a prenominal modifier (*both the boys*), a postnominal modifier (*the boys both*), the head of a noun phrase (*both of the boys*) or part of a complex coordinating conjunction (*both boys and girls*), the Penn Treebank tags *both* differently in each of these syntactic contexts—as PDT (predeterminer), RB (adverb), NNS (plural common noun) and coordinating conjunction (CC), respectively.

There is one case in which our concern with tagging by syntactic function has led us to bifurcate Brown Corpus categories rather than to collapse them: namely, in the case of the uninflected form of verbs. Whereas the Brown Corpus tags the bare form of a verb as VB regardless of whether it occurs in a tensed clause, the Penn Treebank tagset distinguishes VB (infinitive or imperative) from VBP (non-third person singular present tense).

2.1.4 Indeterminacy

A final difference between the Penn Treebank tagset and all other tagsets we are aware of concerns the issue of indeterminacy: both POS ambiguity in the text and annotator uncertainty. In many cases, POS ambiguity can be resolved with reference to the linguistic context. So, for instance, in Katherine Hepburn's witty line *Grant can be outspoken—but not by anyone I know*, the presence of the *by*-phrase forces us to consider *outspoken* as the past participle of a transitive derivative of *speak—outspeak*—rather than as the adjective *outspoken*. However, even given explicit criteria for assigning POS tags to potentially ambiguous words, it is not always possible to assign a unique tag to a word with confidence. Since a major concern of the Treebank is avoid requiring annotators to make arbitrary decisions, we allow words to be associated with more than one POS tag. Such multiple tagging indicates either that the word's part of speech simply cannot be decided or that the annotator is unsure which of the alternative tags is the correct one. In principle, annotators can

[7] An important exception is *there*, which the Brown Corpus tags as EX (existential *there*) when it is used as a formal subject and as RB (adverb) when it is used as a locative adverb. In the case of *there*, we did not pursue our strategy of tagset reduction to its logical conclusion, which would have implied tagging existential *there* as NN (common noun).

tag a word with any number of tags, but in practice, multiple tags are restricted to a small number of recurring two-tag combinations: JJ|NN (adjective or noun as prenominal modifier), JJ|VBG (adjective or gerund/present participle), JJ|VBN (adjective or past participle), NN|VBG (noun or gerund), and RB|RP (adverb or particle).

2.2 The POS tagset

The Penn Treebank tagset is given in Table 2. It contains 36 POS tags and 12 other tags (for punctuation and currency symbols). A detailed description of the guidelines governing the use of the tagset is available in [Santorini 1990].[8]

Table 2:
The Penn Treebank POS tagset

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	`	Left open single quote
22.	RBS	Adverb, superlative	46.	``	Left open double quote
23.	RP	Particle	47.	'	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	''	Right close double quote

[8] In versions of the tagged corpus distributed before November 1992, singular proper nouns, plural proper nouns and personal pronouns were tagged as "NP", "NPS" and "PP", respectively. The current tags "NNP", "NNPS" and "PRP" were introduced in order to avoid confusion with the syntactic tags "NP" (noun phrase) and "PP" (prepositional phrase) (see Table 3).

2.3 The POS tagging process

The tagged version of the Penn Treebank corpus is produced in two stages, using a combination of automatic POS assignment and manual correction.

2.3.1 Automated stage

During the early stages of the Penn Treebank project, the initial automatic POS assignment was provided by PARTS ([Church 1988]), a stochastic algorithm developed at AT&T Bell Labs. PARTS uses a modified version of the Brown Corpus tagset close to our own and assigns POS tags with an error rate of 3–5%. The output of PARTS was automatically tokenized[9] and the tags assigned by PARTS were automatically mapped onto the Penn Treebank tagset. This mapping introduces about 4% error, since the Penn Treebank tagset makes certain distinctions that the PARTS tagset does not.[10] A sample of the resulting tagged text, which has an error rate of 7–9%, is shown in Figure 1.

Figure 1:
Sample tagged text—**before** correction

> Battle-tested/NNP industrial/JJ managers/NNS here/RB always/RB buck/VB up/IN nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
> "/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/, "/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN Mitsui/NNS group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.

More recently, the automatic POS assignment is provided by a cascade of stochastic and rule-driven taggers developed on the basis of our early experience. Since these taggers are based on the Penn Treebank tagset, the 4% error rate introduced as an artefact of mapping from the PARTS tagset to ours is eliminated, and we obtain error rates of 2–6%.

2.3.2 Manual correction stage

The result of the first, automated stage of POS tagging is given to annotators to correct. The annotators use a mouse-based package written in GNU Emacs Lisp, which is embedded within the

[9] In contrast to the Brown Corpus, we do not allow compound tags of the sort illustrated above for *I'm*. Rather, contractions and the Anglo-Saxon genitive of nouns are automatically split into their component morphemes, and each morpheme is tagged separately. Thus, *children's* is tagged "children/NNS 's/POS" and *won't* is tagged "wo-/MD n't/RB".

[10] The two largest sources of mapping error are that the PARTS tagset distinguishes neither infinitives from the non-third person singular present tense forms of verbs, nor prepositions from particles in cases like *run up a hill* and *run up a bill*.

GNU Emacs editor ([Lewis et al 1990]). The package allows annotators to correct POS assignment errors by positioning the cursor on an incorrectly tagged word and then entering the desired correct tag (or sequence of multiple tags). The annotators' input is automatically checked against the list of legal tags in Table 2 and, if valid, appended to the original word-tag pair separated by an asterisk. Appending the new tag rather than replacing the old tag allows us to easily identify recurring errors at the automatic POS assignment stage. We believe that the confusion matrices that can be extracted from this information should also prove useful in designing better automatic taggers in the future. The result of this second stage of POS tagging is shown in Figure 2. Finally, in the distribution version of the tagged corpus, any incorrect tags assigned at the first, automatic stage are removed.

Figure 2:
Sample tagged text—**after** correction

Battle-tested/NNP*/JJ industrial/JJ managers/NNS here/RB always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ./. a/DT boatload/NN of/IN samurai/NNS*/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

"/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/."/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.

The learning curve for the POS tagging task takes under a month (at 15 hours a week), and annotation speeds after a month exceed 3,000 words per hour.

3 Two modes of annotation - an experiment

To determine how to maximize the speed, inter-annotator consistency and accuracy of POS tagging, we performed an experiment at the very beginning of the project to compare two alternative modes of annotation. In the first annotation mode ("tagging"), annotators tagged unannotated text entirely by hand; in the second mode ("correcting"), they verified and corrected the output of PARTS, modified as described above. This experiment showed that manual tagging took about twice as long as correcting, with about twice the inter-annotator disagreement rate and an error rate that was about 50% higher.

Four annotators, all with graduate training in linguistics, participated in the experiment. All completed a training sequence consisting of fifteen hours of correcting, followed by six hours of tagging. The training material was selected from a variety of nonfiction genres in the Brown Corpus. All the annotators were familiar with GNU Emacs at the outset of the experiment. Eight 2,000 word samples were selected from the Brown Corpus, two each from four different genres (two fiction, two nonfiction), none of which the annotators had encountered in training. The texts for the correction task were automatically tagged as described in section 2.3. Each annotator first

manually tagged four texts and then corrected four automatically tagged texts. Each annotator completed the four genres in a different permutation.

A repeated measures analysis of annotation speed with annotator identity, genre and annotation mode (tagging vs. correcting) as classification variables showed a significant annotation mode effect ($p = .05$). No other effects or interactions were significant. The average speed for correcting was more than twice as fast as the average speed for tagging: 20 minutes vs. 44 minutes per 1,000 words. (Median speeds per 1,000 words were 22 vs. 42 minutes.)

A simple measure of tagging consistency is inter-annotator disagreement rate, the rate at which annotators disagree with one another over the tagging of lexical tokens, expressed as a percentage of the raw number of such disagreements over the number of words in a given text sample. For a given text and n annotators, there are $\binom{n}{2}$ disagreement ratios (one for each possible pair of annotators). Mean inter-annotator disagreement was 7.2% for the tagging task and 4.1% for the correcting task (with medians 7.2% and 3.6%, respectively). Upon examination, a disproportionate amount of disagreement in the correcting case was found to be caused by one text that contained many instances of a cover symbol for chemical and other formulas. In the absence of an explicit guideline for tagging this case, the annotators had made different decisions on what part of speech this cover symbol represented. When this text is excluded from consideration, mean inter-annotator disagreement for the correcting task drops to 3.5%, with the median unchanged at 3.6%.

Consistency, while desirable, tells us nothing about the validity of the annotators' corrections. We therefore compared each annotator's output not only with the output of each of the others, but also with a benchmark version of the eight texts. This benchmark version was derived from the tagged Brown Corpus by (1) mapping the original Brown Corpus tags onto the Penn Treebank tagset and (2) carefully hand-correcting the revised version in accordance with the tagging conventions in force at the time of the experiment. Accuracy was then computed as the rate of disagreement between each annotator's results and the benchmark version. The mean accuracy was 5.4% for the tagging task (median 5.7%) and 4.0% for the correcting task (median 3.4%). Excluding the same text as above gives a revised mean accuracy for the correcting task of 3.4%, with the median unchanged.

We obtained a further measure of the annotators' accuracy by comparing their error rates to the rates at which the raw output of Church's PARTS program—appropriately modified to conform to the Penn Treebank tagset—disagreed with the benchmark version. The mean disagreement rate between PARTS and the benchmark version was 9.6%, while the corrected version had a mean disagreement rate of 5.4%, as noted above.[11] The annotators were thus reducing the error rate by about 4.2%.

[11] We would like to emphasize that the percentage given for the modified output of PARTS does not represent an error rate for PARTS. It reflects not only true mistakes in PARTS performance, but also the many and important differences in the usage of Penn Treebank POS tags and the usage of tags in the original Brown Corpus material on which PARTS was trained.

4 Bracketing

4.1 Basic methodology

The methodology for bracketing the corpus is completely parallel to that for tagging—hand correction of the output of an errorful automatic process. Fidditch, a deterministic parser developed by Donald Hindle first at the University of Pennsylvania and subsequently at AT&T Bell Labs ([Hindle 1983], [Hindle 1989]), is used to provide an initial parse of the material. Annotators then hand correct the parser's output using a mouse-based interface implemented in GNU Emacs Lisp. Fidditch has three properties that make it ideally suited to serve as a preprocessor to hand correction:

- Fidditch always provides exactly one analysis for any given sentence, so that annotators need not search through multiple analyses.
- Fidditch never attaches any constituent whose role in the larger structure it cannot determine with certainty. In cases of uncertainty, Fidditch chunks the input into a string of trees, providing only a partial structure for each sentence.
- Fidditch has rather good grammatical coverage, so that the grammatical chunks that it does build are usually quite accurate.

Because of these properties, annotators do not need to rebracket much of the parser's output—a relatively time-consuming task. Rather, the annotators' main task is to "glue" together the syntactic chunks produced by the parser. Using a mouse-based interface, annotators move each unattached chunk of structure under the node to which it should be attached. Notational devices allow annotators to indicate uncertainty concerning constituent labels, and to indicate multiple attachment sites for ambiguous modifiers. The bracketing process is described in more detail in section 4.3.

4.2 The syntactic tagset

Table 3 shows the set of syntactic tags and null elements that we use in our skeletal bracketing. More detailed information on the syntactic tagset and guidelines concerning its use are to be found in [Santorini and Marcinkiewicz 1991].

Table 3:
The Penn Treebank syntactic tagset

	Tags	
1.	ADJP	Adjective phrase
2.	ADVP	Adverb phrase
3.	NP	Noun phrase
4.	PP	Prepositional phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by subordinating conjunction or *0* (see below)
7.	SBARQ	Direct question introduced by *wh*-word or *wh*-phrase
8.	SINV	Declarative sentence with subject-aux inversion
9.	SQ	Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase
10.	VP	Verb phrase
11.	WHADVP	*Wh*-adverb phrase
12.	WHNP	*Wh*-noun phrase
13.	WHPP	*Wh*-prepositional phrase
14.	X	Constituent of unknown or uncertain category
	Null elements	
1.	*	"Understood" subject of infinitive or imperative
2.	0	Zero variant of *that* in subordinate clauses
3.	T	Trace—marks position where moved *wh*-constituent is interpreted
4.	NIL	Marks position where preposition is interpreted in pied-piping contexts

Although different in detail, our tagset is similar in delicacy to that used by the Lancaster Treebank Project, except that we allow null elements in the syntactic annotation. Because of the need to achieve a fairly high output per hour, it was decided not to require annotators to create distinctions beyond those provided by the parser. Our approach to developing the syntactic tagset was highly pragmatic and strongly influenced by the need to create a large body of annotated material given limited human resources. Despite the skeletal nature of the bracketing, however, it is possible to make quite delicate distinctions when using the corpus by searching for combinations of structures. For example, an SBAR containing the word *to* immediately before the VP will necessarily be infinitival, while an SBAR containing a verb or auxiliary with a tense feature will necessarily be tensed. To take another example, so-called *that*-clauses can be identified easily by searching for SBARs containing the word *that* or the null element *0* in initial position.

As can be seen from Table 3, the syntactic tagset used by the Penn Treebank includes a variety of null elements, a subset of the null elements introduced by Fidditch. While it would be expensive to insert null elements entirely by hand, it has not proved overly onerous to maintain and correct those that are automatically provided. We have chosen to retain these null elements because we believe that they can be exploited in many cases to establish a sentence's predicate-argument structure;

at least one recipient of the parsed corpus has used it to bootstrap the development of lexicons for particular NLP projects and has found the presence of null elements to be a considerable aid in determining verb transitivity (Robert Ingria, personal communication). While these null elements correspond more directly to entities in some grammatical theories than in others, it is not our intention to lean toward one or another theoretical view in producing our corpus. Rather, since the representational framework for grammatical structure in the Treebank is a relatively impoverished flat context-free notation, the easiest mechanism to include information about predicate-argument structure, although indirectly, is by allowing the parse tr to contain explicit null items.

4.3 Sample bracketing output

Below, we illustrate the bracketing process for the first sentence of our sample text. Figure 3 shows the output of Fidditch (modified slightly to include our POS tags).

Figure 3:
Sample bracketed text—full structure provided by Fidditch

```
( (S
    (NP (NBAR (ADJP (ADJ "Battle-tested/JJ")
                    (ADJ "industrial/JJ"))
              (NPL "managers/NNS")))
    (? (ADV "here/RB"))
    (? (ADV "always/RB"))
    (AUX (TNS *))
    (VP (VPRES "buck/VBP")))
    (? (PP (PREP "up/RP")
           (NP (NBAR (ADJ "nervous/JJ")
                     (NPL "newcomers/NNS")))))
    (? (PP (PREP "with/IN")
           (NP (DART "the/DT")
               (NBAR (N "tale/NN"))
                     (PP of/PREP
                         (NP (DART "the/DT")
                             (NBAR (ADJP
                                   (ADJ "first/JJ")))))))))
    (? (PP of/PREP
           (NP (PROS "their/PP$")
               (NBAR (NPL "countrymen/NNS"))))
    (? (S (NP (PRO *))
              (AUX to/TNS)
              (VP (V "visit/VB")
                  (NP (PNP "Mexico/NNP")))))
    (? (MID ",/,"))
    (? (NP (IART "a/DT")
           (NBAR (N "boatload/NN"))
                 (PP of/PREP
                     (NP (NBAR
                         (NPL "warriors/NNS"))))
                 (VP (VPPRT "blown/VBN")
                     (? (ADV "ashore/RB"))
                     (NP (NBAR (CARD "375/CD")
                               (NPL "years/NNS"))))))
    (? (ADV "ago/RB"))
    (? (FIN "./.")))
```

As Figure 3 shows, Fidditch leaves very many constituents unattached, labeling them as "?", and its output is perhaps better thought of as a string of tree fragments than as a single tree structure.

Fidditch only builds structure when this is possible for a purely syntactic parser without access to semantic or pragmatic information, and it always errs on the side of caution. Since determining the correct attachment point of prepositional phrases, relative clauses, and adverbial modifiers almost always requires extrasyntactic information, Fidditch pursues the very conservative strategy of *always* leaving such constituents unattached, even if only one attachment point is syntactically possible. However, Fidditch does indicate its best guess concerning a fragment's attachment site by the fragment's depth of embedding. Moreover, it attaches prepositional phrases beginning with *of* if the preposition immediately follows a noun; thus, *tale of* ... and *boatload of* ... are parsed as single constituents, while *first of* ... is not. Since Fidditch lacks a large verb lexicon, it cannot decide whether some constituents serve as adjuncts or arguments and hence leaves subordinate clauses such as infinitives as separate fragments. Note further that Fidditch creates adjective phrases only when it determines that more than one lexical item belongs in the ADJP. Finally, as is well known, determining the scope of conjunctions and other coordinate structures can only be determined given the richest forms of contextual information; here again, Fidditch simply turns out a string of tree fragments around any conjunction. Because all decisions within Fidditch are made locally, all commas (which often signal conjunction) must disrupt the input into separate chunks.

The original design of the Treebank called for a level of syntactic analysis comparable to the skeletal analysis used by the Lancaster Treebank, but a limited experiment was performed early in the project to investigate the feasibility of providing greater levels of structural detail. While the results were somewhat unclear, there was evidence that annotators could maintain a much faster rate of hand correction if the parser output was simplified in various ways, reducing the visual complexity of the tree representations and eliminating a range of minor decisions. The key results of this experiment were:

- Annotators take substantially longer to learn the bracketing task than the POS tagging task, with substantial increases in speed occurring even after two months of training.
- Annotators can correct the full structure provided by Fidditch at an average speed of approx. 375 words per hour after three weeks, and 475 words per hour after six weeks.
- Reducing the output from the full structure shown in Figure 3 to a more skeletal representation similar to that used by the Lancaster UCREL Treebank Project increases annotator productivity by approx. 100–200 words per hour.
- It proved to be very difficult for annotators to distinguish between a verb's arguments and adjuncts in all cases. Allowing annotators to ignore this distinction when it is unclear (attaching constituents high) increases productivity by approx. 150–200 words per hour. Informal examination of later annotation showed that forced distinctions cannot be made consistently.

As a result of this experiment, the originally proposed skeletal representation was adopted, without a forced distinction between arguments and adjuncts. Even after extended training, performance varies markedly by annotator, with speeds on the task of correcting skeletal structure without requiring a distinction between arguments and adjuncts ranging from approx. 750 words per hour to well over 1,000 words per hour after three or four months experience. The fastest annotators work in bursts of well over 1,500 words per hour alternating with brief rests. At an

average rate of 750 words per hour, a team of five part-time annotators annotating three hours a day should maintain an output of about 2.5 million words a year of "treebanked" sentences, with each sentence corrected once.

It is worth noting that experienced annotators can proofread previously corrected material at very high speeds. A parsed subcorpus of over one million words was recently proofread at an average speed of approx. 4,000 words per annotator per hour. At this rate of productivity, annotators are able to find and correct gross errors in parsing, but do not have time to check, for example, whether they agree with all prepositional phrase attachments.

The process that creates the skeletal representations to be corrected by the annotators simplifies and flattens the structures shown in Figure 3 by removing POS tags, non-branching lexical nodes and certain phrasal nodes, notably NBAR. The output of the first automated stage of the bracketing task is shown in Figure 4.

Figure 4:
Sample bracketed text--**after** simplification, **before** correction

```
( (S
    (NP (ADJP Battle-tested industrial)
        managers)
    (? here)
    (? always)
    (VP buck))
    (? (PP up
           (NP nervous newcomers)))
    (? (PP with
           (NP the tale
               (PP of
                   (NP the
                       (ADJP first))))))
    (? (PP of
           (NP their countrymen)))
    (? (S (NP *)
           to
          (VP visit
              (NP Mexico))))
    (? ,)
    (? (NP a boatload
           (PP of
               (NP warriors))
           (VP blown
               (? ashore)
               (NP 375 years))))
    (? ago)
    (? .))
```

Annotators correct this simplified structure using a mouse-based interface. Their primary job is to "glue" fragments together, but they must also correct incorrect parses and delete some structure. Single mouse clicks perform the following tasks, among others. The interface correctly reindents the structure whenever necessary.

- Attach constituents labeled *?*. This is done by pressing down the appropriate mouse button on or immediately after the *?*, moving the mouse onto or immediately after the label of the intended parent and releasing the mouse. Attaching constituents automatically deletes their *?* label.
- Promote a constituent up one level of structure, making it a sibling of its current parent.

- Delete a pair of constituent brackets.
- Create a pair of brackets around a constituent. This is done by typing a constituent tag and then sweeping out the intended constituent with the mouse. The tag is checked to assure that it is a legal label.
- Change the label of a constituent. The new tag is checked to assure that it is legal.

The bracketed text after correction is shown in Figure 5. The fragments are now connected together into one rooted tree structure. The result is a *skeletal* analysis in that much syntactic detail is left unannotated. Most prominently, all internal structure of the NP up through the head and including any single-word post-head modifiers is left unannotated.

Figure 5
Sample bracketed text—**after** correction

```
( (S
   (NP Battle-tested industrial managers
       here)
    always
   (VP buck
       up
       (NP nervous newcomers)
       (PP with
           (NP the tale
               (PP of
                   (NP (NP the
                           (ADJP first
                                 (PP of
                                     (NP their countrymen)))
                           (S (NP *)
                              to
                              (VP visit
                                  (NP Mexico))))
                       ,
                       (NP (NP a boatload
                               (PP of
                                   (NP (NP warriors)
                                       (VP-1 blown
                                           ashore
                                           (ADVP (NP 375 years)
                                                 ago)))))
                           (VP-1 *pseudo-attach*)))))))))
  .)
```

As noted above in connection with POS tagging, a major goal of the Treebank project is to allow annotators only to indicate structure of which they were certain. The Treebank provides two notational devices to ensure this goal: the *X* constituent label and so-called "pseudo-attachment". The *X* constituent label is used if an annotator is sure that a sequence of words is a major constituent but is unsure of its syntactic category; in such cases, the annotator simply brackets the sequence and labels it *X*. The second notational device, pseudo-attachment, has two primary uses. On the one hand, it is used to annotate what Kay has called *permanent predictable ambiguities*, allowing an annotator to indicate that a structure is globally ambiguous even given the surrounding context (annotators always assign structure to a sentence on the basis of its context). An example of this use of pseudo-attachment is shown in Figure 5, where the participial phrase *blown ashore 375 years ago* modifies either *warriors* or *boatload*, but there is no way of settling the question—both attachments mean exactly the same thing. In the case at hand, the pseudo-attachment notation indicates that the annotator of the sentence thought that `VP-1` is most likely a modifier of *warriors*, but that it is also possible that it is a modifier of *boatload*.[12] A second use of pseudo-attachment is to allow annotators to represent the "underlying" position of extraposed elements; in addition to being attached in its superficial position in the tree, the extraposed constituent is pseudo-attached within the constituent to which it is semantically related. Note that except for the device of pseudo-attachment, the skeletal analysis of the Treebank is entirely restricted to simple context-free trees.

The reader may have noticed that the ADJP brackets in Figure 4 have vanished in Figure 5. For the sake of the overall efficiency of the annotation task, we leave all ADJP brackets in the simplified structure, with the annotators expected to remove many of them during annotation. The reason for this is somewhat complex, but provides a good example of the considerations that come into play in designing the details of annotation methods. The first relevant fact is that Fidditch only outputs ADJP brackets within NPs for adjective phrases containing more than one lexical item. To be consistent, the final structure must contain ADJP nodes for all adjective phrases within NPs or for none; we have chosen to delete all such nodes within NPs under normal circumstances. (This does not affect the use of the ADJP tag for predicative adjective phrases outside of NPs.) In a seemingly unrelated guideline, all coordinate structures are annotated in the Treebank; such coordinate structures are represented by Chomsky-adjunction when the two conjoined constituents bear the same label. This means that if an NP contains coordinated adjective phrases, then an ADJP tag will be used to tag that coordination even though simple ADJPs within NPs will not be bear an APJP tag. Experience has shown that annotators can delete pairs of brackets extremely quickly using the mouse-based tools, whereas creating brackets is a much slower operation. Because the coordination of adjectives is quite common, it is more efficient to leave in ADJP labels, and delete them if they are not part of a coordinate structure, than to reintroduce them if necessary.

5 Progress to date

5.1 Composition and size of corpus

Table 4 shows the output of the Penn Treebank project at the end of its first phase.

[12] This use of pseudo-attachment is identical to its original use in Church's parser ([Church 1980]).

Table 4:
Penn Treebank
(as of 11/92)

Description	Tagged for Part-of-Speech (Tokens)	Skeletal Parsing (Tokens)
Dept. of Energy abstracts	231,404	231,404
Dow Jones Newswire stories	3,065,776	1,061,166
Dept. of Agriculture bulletins	78,555	78,555
Library of America texts	105,652	105,652
MUC-3 messages	111,828	111,828
IBM Manual sentences	89,121	89,121
WBUR radio transcripts	11,589	11,589
ATIS sentences	19,832	19,832
Brown Corpus, retagged	1,172,041	1,172,041
Total:	4,885,798	2,881,188

All the materials listed above are available on CD-ROM to members of the Linguistic Data Consortium.[13] About 3 million words of POS-tagged material and a small sampling of skeletally parsed text are available as part of the first Association for Computational Linguistics/Data Collection Initiative CD-ROM, and a somewhat larger subset of materials is available on cartridge tape directly from the Penn Treebank Project. For information, contact the first author of this paper or send email to treebank@unagi.cis.upenn.edu.

Some comments on the materials included:

- Dept. of Energy abstracts are scientific abstracts from a variety of disciplines.
- All of the skeletally parsed Dow Jones Newswire materials are also available as digitally recorded read speech as part of the DARPA WSJ-CSR1 corpus, available through the Linguistic Data Consortium.
- The Dept. of Agriculture materials include short bulletins such as when to plant various flowers and how to can various vegetables and fruits.
- The Library of America texts are 5,000-10,000 word passages, mainly book chapters, from a variety of American authors including Mark Twain, Henry Adams, Willa Cather, Herman Melville, W.E.B. Dubois, and Ralph Waldo Emerson.
- The MUC-3 texts are all news stories from the Federal News Service about terrorist activities in South America. Some of these texts are translations of Spanish news stories or transcripts of radio broadcasts. They are taken from training materials for the 3rd Message Understanding Conference.

[13] Contact The Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305 or send email to ldc@unagi.cis.upenn.edu for more information.

- The Brown Corpus materials were completely retagged by the Penn Treebank project starting from the untagged version of the Brown Corpus ([Francis 1964]).
- The IBM sentences are taken from IBM computer manuals; they are chosen to contain a vocabulary of 3,000 words, and are limited in length.
- The ATIS sentences are transcribed versions of spontaneous sentences collected as training materials for the DARPA Air Travel Information System project.

The entire corpus has been tagged for POS information, at an estimated error rate of approx. 3%. The POS-tagged version of the Library of America texts and the Department of Agriculture bulletins have been corrected twice (each by a different annotator), and the corrected files were then carefully adjudicated; we estimate the error rate of the adjudicated version at well under 1%. Using a version of PARTS retrained on the entire preliminary corpus and adjudicating between the output of the retrained version and the preliminary version of the corpus, we plan to reduce the error rate of the final version of the corpus to approx. 1%. All the skeletally parsed materials have been corrected once, except for the Brown materials, which have been quickly proofread an additional time for gross parsing errors.

5.2 Future directions

A large number of research efforts, both at the University of Pennsylvania and elsewhere, have relied on the output of the Penn Treebank Project to date. A few examples already in print: A number of projects investigating stochastic parsing have used either the POS-tagged materials ([Magerman and Marcus 1990], [Brill et al 1990], [Brill 1991]) or the skeletally parsed corpus ([Weischedel et al 1991], [Pereira and Schabes 1992]). The POS-tagged corpus has also been used to train a number of different POS taggers including [Meteer et al 1991], and the skeletally parsed corpus has been used in connection with the development of new methods to exploit intonational cues in disambiguating the parsing of spoken sentences ([Veilleux and Ostendorf 1992]). The Penn Treebank has been used to bootstrap the development of lexicons for particular applications (Robert Ingria, personal communication) and is being used as a source of examples for linguistic theory and psychological modelling (e.g. [Niv 1991]). To aid in the search for specific examples of grammatical phenomena using the Treebank, Richard Pito has developed `tgrep`, a tool for very fast context-free pattern matching against the skeletally parsed corpus, which is available through the Linguistic Data Consortium.

While the Treebank is being widely used, the annotation scheme employed has a variety of limitations. Many otherwise clear argument/adjunct relations in the corpus are not indicated due to the current Treebank's essentially context-free representation. For example, there is at present no satisfactory representation for sentences in which complement noun phrases or clauses occur after a sentential level adverb. Either the adverb is trapped within the VP, so that the complement can occur within the VP where it belongs, or else the adverb is attached to the S, closing off the VP and forcing the complement to attach to the S. This "trapping" problem serves as a limitation for groups that currently use Treebank material to semiautomatically derive lexicons for particular applications. For most of these problems, however, solutions are possible on the basis of mechanisms already used by the Treebank Project. For example, the pseudo-attachment notation

can be extended to indicate a variety of crossing dependencies. We have recently begun to use this mechanism to represent various kinds of dislocations, and the Treebank annotators themselves have developed a detailed proposal to extend pseudo-attachment to a wide range of similar phenomena.

A variety of inconsistencies in the annotation scheme used within the Treebank have also become apparent with time. The annotation schemes for some syntactic categories should be unified to allow a consistent approach to determining predicate-argument structure. To take a very simple example, sentential adverbs attach under VP when they occur between auxiliaries and predicative ADJPs, but attach under S when they occur between auxiliaries and VPs. These structures need to be regularized.

As the current Treebank has been exploited by a variety of users, a significant number have expressed a need for forms of annotation richer than provided by the project's first phase. Some users would like a less skeletal form of annotation of surface grammatical structure, expanding the essentially context-free analysis of the current Penn Treebank to indicate a wide variety of non-contiguous structures and dependencies. A wide range of Treebank users now strongly desire a level of annotation which makes explicit some form of predicate-argument structure. The desired level of representation would make explicit the logical subject and logical object of the verb, and would indicate, at least in clear cases, which subconstituents serve as arguments of the underlying predicates and which serve as modifiers.

During the next phase of the Treebank project, we expect to provide both a richer analysis of the existing corpus and to provide a parallel corpus of predicate-argument structures. This will be done by first enriching the annotation of the current corpus, and then automatically extracting predicate-argument structure, at the level of distinguishing logical subjects and objects, and distinguishing arguments from adjuncts for clear cases. Enrichment will be achieved by automatically transforming the current Penn Treebank into a level of structure close to the intended target, and then completing the conversion by hand.

References

[Brill 1991] Brill, Eric, 1991. Discovering the lexical features of a language. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA.*

[Brill et al 1990] Brill, Eric; Magerman, David; Marcus, Mitchell P.; and Santorini, Beatrice, 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop, June 1990*, pages 275–282.

[Church 1980] Church, Kenneth W., 1980. Memory limitations in natural language processing, MIT LCS Technical Report 245. Master's thesis, Massachusetts Institute of Technology.

[Church 1988] Church, Kenneth W., 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of*

the Second Conference on Applied Natural Language Processing. 26th Annual Meeting of the Association for Computational Linguistics, pages 136–143.

[Francis 1964] Francis, W. Nelson, 1964. *A standard sample of present-day English for use with digital computers. Report to the U.S Office of Education on Cooperative Research Project No. E–007.* Brown University, Providence.

[Francis and Kučera 1982] Francis, W. Nelson and Kučera, Henry, 1982. *Frequency analysis of English usage. Lexicon and grammar.* Houghton Mifflin, Boston.

[Garside et al 1987] Garside, Roger; Leech, Geoffrey; and Sampson, Geoffrey, 1987. *The computational analysis of English. A corpus-based approach.* Longman, London.

[Hindle 1983] Hindle, Donald, 1983. *User manual for Fidditch.* Technical memorandum 7590–142, Naval Research Laboratory.

[Hindle 1989] Hindle, Donald, 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.*

[Lewis et al 1990] Lewis, Bil; LaLiberte, Dan; and the GNU Manual Group, 1990. *The GNU Emacs Lisp reference manual.* Free Software Foundation, Cambridge.

[Magerman and Marcus 1990] Magerman, David and Marcus, Mitchell P., 1990. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90.*

[Meteer et al 1991] Meteer, Marie; Schwartz, Richard; and Weischedel, Ralph, 1991. Studies in part of speech labelling. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop, February 1991, Draft version.*

[Niv 1991] Niv, Michael, 1991. Syntactic disambiguation. In *The Penn Review of Linguistics 14*, pages 120–126.

[Pereira and Schabes 1992] Pereira, Fernando and Schabes, Yves, 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.*

[Santorini 1990] Santorini, Beatrice, 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project.* Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

[Santorini and Marcinkiewicz 1991] Santorini, Beatrice and Marcinkiewicz, Mary Ann, 1991. Bracketing guidelines for the Penn Treebank Project. Ms., Department of Computer and Information Science, University of Pennsylvania.

[Veilleux and Ostendorf 1992] Veilleux, N. M. and Ostendorf, Mari, 1992. Probabilistic parse scoring based on prosodic features. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop, February 1992.*

[Weischedel et al 1991] Weischedel, Ralph; Ayuso, Damaris; Bobrow, R.; Boisen, Sean; Ingria, Robert; and Palmucci, Jeff, 1991. Partial parsing: a report of work in progress. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop, February 1991, Draft version.*