July 1990

First Steps Towards an Annotated Database of American English

Mitchell P. Marcus
University of Pennsylvania

Beatrice Santorini
University of Pennsylvania

David Magerman
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-46.

First Steps Towards an Annotated Database of American English

Abstract
This paper reports on one of the first steps in building a very large annotated database of American English. We present and discuss the results of an experiment comparing manual part-of-speech tagging with manual verification and correction of automatic stochastic tagging. The experiment shows that correcting is superior to tagging with respect to speed, consistency and accuracy.

Comments
University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-46.

First Steps Towards An Annotated Database of American English

MS-CIS-90-46
LINC LAB 175

Mitchell P. Marcus
Beatrice Santorini
David Magerman

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

July 1990

Contents

First Steps Towards an Annotated Database of American English

Mitchell P. Marcus, Beatrice Santorini and David Magerman
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

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 material and by attempting to automatically extract information about language from large corpora of natural language. Such corpora, appropriately and judiciously annotated, provide a new and important research tool with major impact upon investigators in natural language processing, speech recognition, integrated spoken language systems and linguistics. Such data bases are of value for enterprises as diverse as the automatic construction of statistical models for the grammar of both the written and 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 self evaluation of the adequacy of parsing models, the various formal syntactic theories embedded in those parsers, and the particular grammars of English encoded within those theories.

Ultimately, a corpus of at least 100 million words of annotated text is desirable for such research. As a first step towards this goal, we have begun to develop and test techniques for such annotation, with an expected output of about 5 million words of annotated text and 1 million words of annotated speech over a three-year period. In particular, we have:

- designed and implemented a highly portable annotators' workstation, written as a subsystem of GNU EMACS,
- designed a tagset of parts of speech for English, and
- carried out an experiment comparing two alternative modes of annotation.

In this paper, we report the results of our experiment. 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, the automatic stochastic tagging algorithm described in Church 1988. The purpose of the experiment was to provide evidence concerning the relative merits of tagging vs. correcting. As we will show in detail in Section 2.3, correcting turns out to be superior to tagging on three counts—speed, consistency and accuracy. Briefly, the results of the experiment are as follows. For correcting vs. tagging,

- median annotation speed is about twice as fast (~20 minutes vs. ~40 minutes per 1,000 words),
- disagreement rates among annotators are about half as high (~3.5% vs. ~7%), and
- error rates are also about half as high (~3% vs. ~6%).

Based on the results of the experiment, we have gone on to correct over 400,000 words of automatically tagged text. Current mean net annotation speed, based on a total of 408,402 words, is 18 minutes per 1,000 words, corresponding to 3,365 words per hour. The corresponding raw figures (including idle time) are 21 minutes per 1,000 words, which is equivalent to 2,794 words per hour.

2 Tagging vs. correcting: an experiment

2.1 Method

2.1.1 Training and prior experience

Four annotators participated in the experiment (JF, DM, MM, BS). All four completed a training sequence consisting of fifteen hours of correcting, followed by six hours of tagging. The training material was selected from the Standard Sample of Present-Day English (Francis 1964), henceforth referred to as the Brown Corpus, and came from a variety of nonfiction genres. In addition to this training, all four annotators had completed at least one year of graduate study in linguistics, and one (BS) enjoyed the double advantage of having developed the tagset that was used and of already being familiar with GNU EMACS at the outset of the experiment.

2.1.2 Material

Like the training material, the material for the experiment came from the Brown Corpus. We used material from two categories—fiction and nonfiction. Each category was in turn divided into two genres. For the tagging task, we selected a judgment sample of four texts (2 categories × 2 genres), subject to the constraint that no text could come from a genre that the annotators had encountered in training. For the correction task, we selected the four texts that follow the texts selected for the tagging task. This second set of texts was automatically tagged using Church's PARTS program and then run through a conversion utility that mapped Church's tags onto the Penn Treebank tagset wherever the correspondences between the two tagsets were predictable. In sum, the material to be annotated consisted of eight texts (2 categories × 2 genres × 2 annotation modes), each containing ~2,000 words, broken down as shown in Table 1.

Table 1:
Breakdown of materials by annotation mode, category and genre

Annotation mode:		Tagging	Correcting
Category:	Nonfiction		
	Genre:		
	Government documents	H03	H04
	Natural sciences	J03	J04
Category:	Fiction		
	Genre:		
	Science fiction novels	M01	M02
	Humorous novels	R01	R02

2.2 Procedure

Each annotator first manually tagged the set of four texts in the first column of Table 1 and then corrected the automatically tagged set of four texts in the second column. In order to eliminate potential priming

effects, we had the annotators complete the four genres in a different order. The order for each annotator, which was held constant across annotation mode, is shown in Table 2.

Table 2:
Order of annotation

JF:	J	R	H	M
DM:	R	M	J	H
MM:	M	H	R	J
BS:	H	J	M	R

2.3 Results

In this subsection, we present the results of the experiment just described, comparing the annotators' performance on the tagging task to their performance on the correction task on three counts—speed, consistency and accuracy.

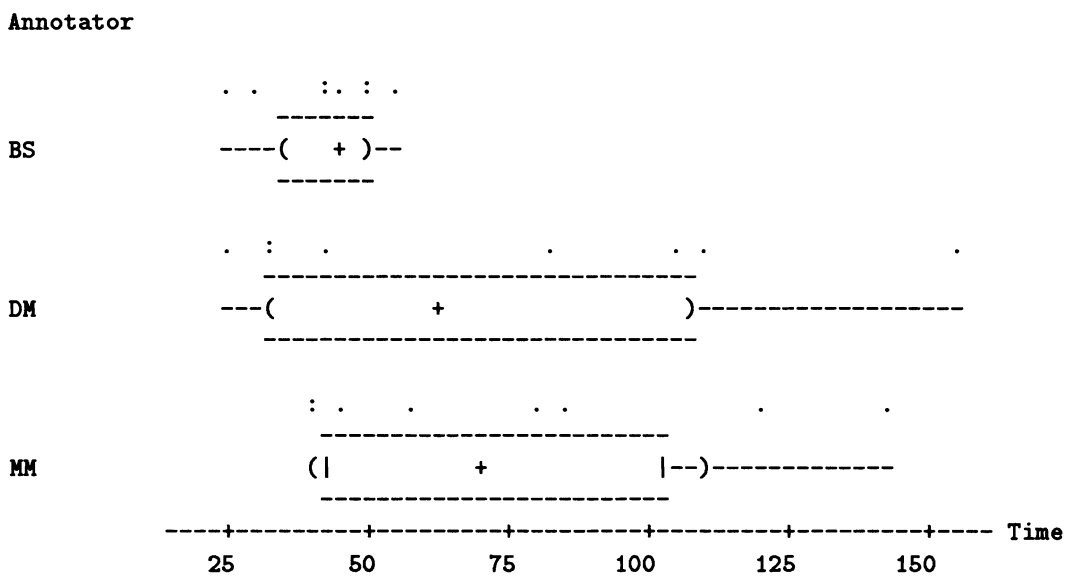
2.3.1 Speed

We analyzed the effect of four variables on annotation time per text sample (~2,000 words): (1) annotator, (2) text category (fiction vs. nonfiction), (3) genre and (4) annotation mode. We present our results in the form of dot and box plots. The box plots are to be interpreted as follows. The box itself, the ends of which are marked by a vertical bar (|), indicates the range of data points within the second and third quartile, while the "whiskers" extending beyond the box to the left and right indicate the range of data points within the first and fourth quartile, respectively. The median is marked by a plus sign (+), and its 90% confidence interval is enclosed in parentheses. If one of the bounds of the 90% confidence interval is identical to one of the sides of the box, then the parenthesis supplants the vertical bar. Time is measured in minutes.[1]

[1] The results concerning annotation times, in contrast to those concerning consistency and accuracy, do not reflect the performance of JF, who was not yet done annotating when the time data were being calculated.

Annotation time by individual annotator: Table 3 shows the time that it took each annotator to annotate a text.

Table 3:
Annotation time by annotator

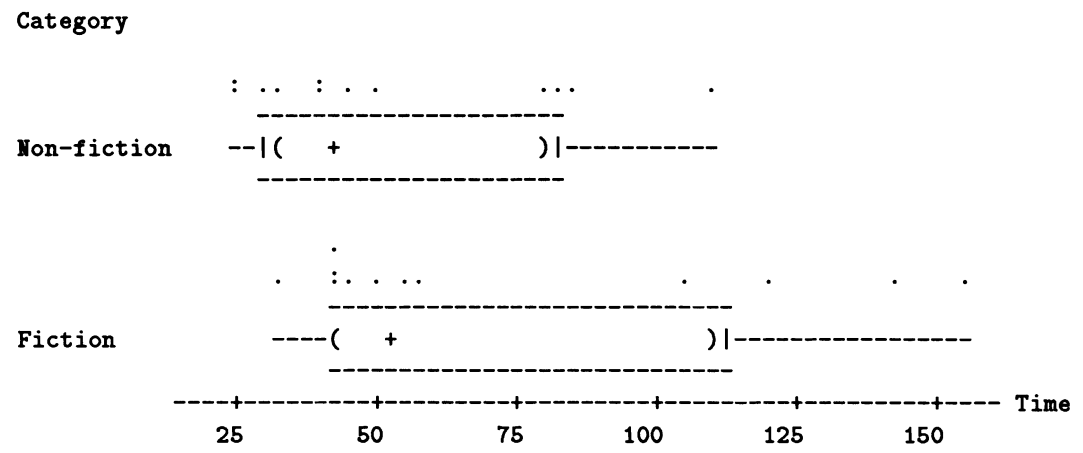


Correlation of time and annotator = 0.372

The results for DM and MM are very similar, while the median annotation time of BS and the variability of her performance are appreciably lower. We attribute this to BS's greater familiarity with both the tagset and the text editor. Nevertheless, we conclude from the complete overlap of her 90% confidence interval with those of the other annotators and the low correlation that there is no significant effect of individual annotator on annotation time.

Annotation time by category: Table 4 shows the effect of the category of a text (fiction vs. nonfiction) on annotation time.

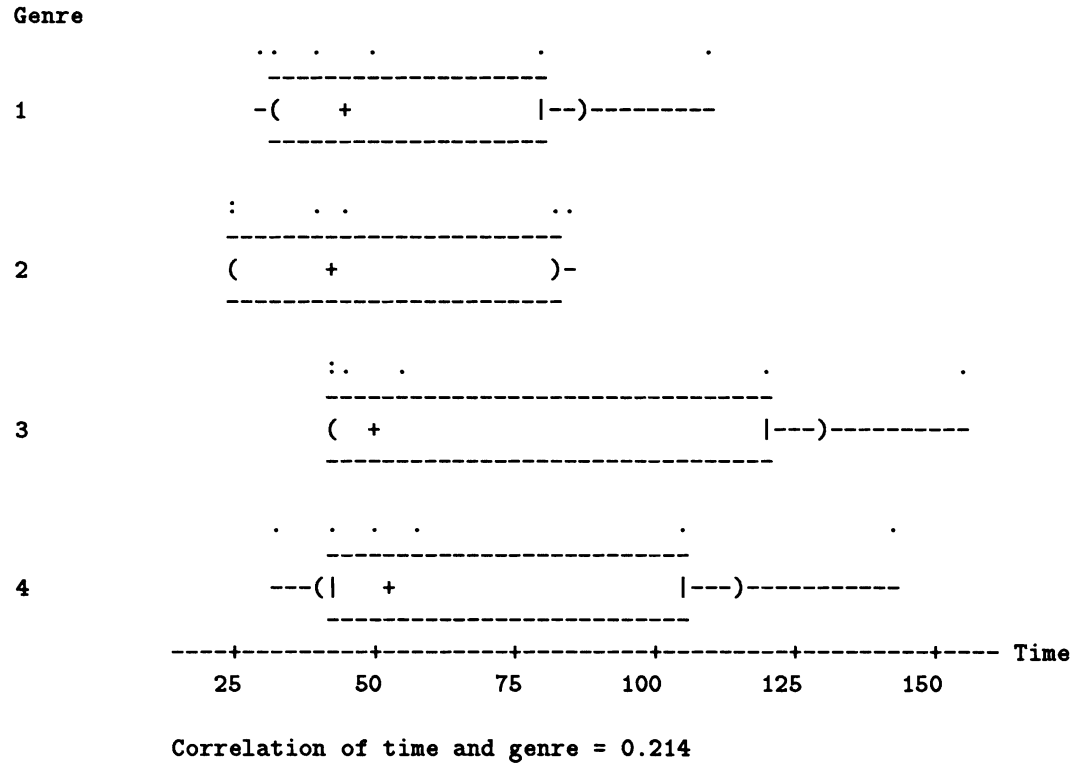
Table 4:
Annotation time by category



While fiction takes somewhat more time to annotate than non-fiction, presumably due to the greater linguistic complexity of fiction, we again conclude from the closeness of the medians, the considerable overlap between the two confidence intervals and the low correlation that there is no appreciable effect of category on annotation time.

Annotation time by genre: Table 5 shows the effect of genre on annotation time.

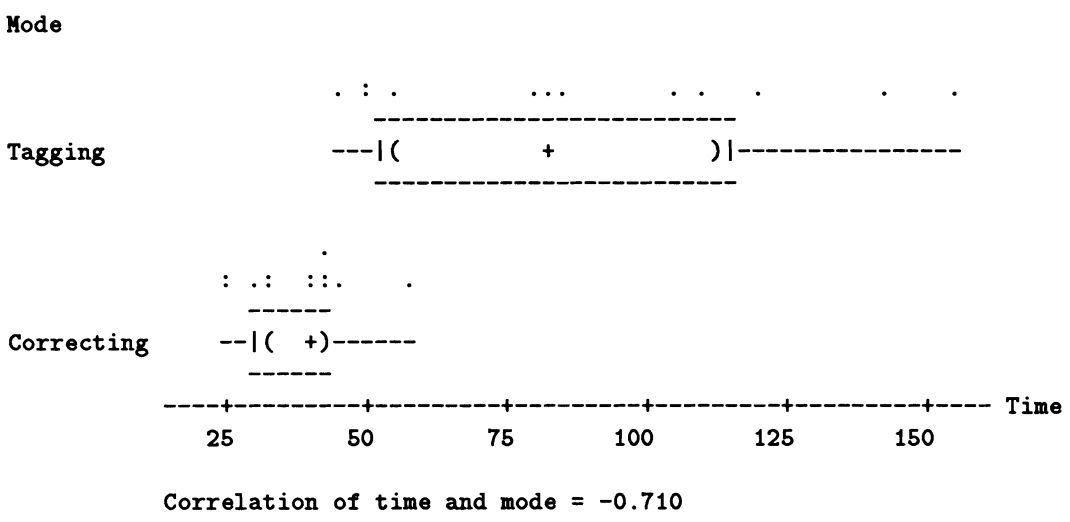
Table 5:
Annotation time by genre



Again, the median annotation times of all four genres are almost identical, the overlap among their 90% confidence intervals is considerable, and the correlation between time and genre is low.

Annotation time by annotation mode: Finally, Table 6 shows the effect of annotation mode on annotation time.

Table 6:
Annotation time by annotation mode



Three characteristics of this table are remarkable. First, the median annotation time for correcting (40 minutes) is half that for tagging (80 minutes). Second, the 90% confidence interval associated with correcting is much narrower than that associated with tagging. As a result, there is no overlap between the 90% confidence intervals associated with each of these modes. Finally, the correlation between time and annotation mode is approximately twice as high as that of time and any other factor. Annotation mode is thus the only factor of the four that we have been considering to have an important effect on annotation time.

2.3.2 Consistency

We turn now to the issue of inter-annotator consistency. We express this factor in terms of the rate at which annotators disagreed with one another over the tagging of a particular word.[2] The rate itself is calculated as the ratio of the raw number of such disagreements over the number of words in a given text sample. For each text, we thus obtain six disagreement ratios (one for each possible pair of annotators), and for each annotation mode, we obtain 24 such ratios (6 annotator pairs × 4 texts). The first two columns of Table 7 gives summary statistics for these figures.

A closer examination of the range of disagreement among annotators revealed that the highest disagreement rates occurred in connection with the correction of J04, a scientific text that contained many instances of a cover symbol for chemical and other formulas. In the absence of explicit tagging guidelines concerning this cover symbol, the annotators had made different decisions on what part of speech it represented. If we revise the figures in the second column of Table 7 in light of this exceptional circumstance by discarding the figures

[2] We counted each token of a disagreement and did not attempt to classify the disagreements by type.

for J04, we obtain the results in the third column.

Table 7:
Inter-annotator disagreement rates (Summary statistics)

Annotation mode:	Tagging	Correcting	
		Raw	Revised
Mean:	7.2%	4.1%	3.5%
Median:	7.2%	3.6%	3.6%
Absolute range:	4.2%–11.2%	2.7%–8.3%	2.7%–4.6%
Relative range:	7.0%	5.6%	1.9%

According to Table 7, the average disagreement rate among annotators, whether expressed as the mean or the median, is about half as high on the correction task as on the tagging task. Eliminating J04 has little effect on the mean, and none on the median, suggesting that our results are robust. It does, however, have the welcome effect of reducing the relative range of inter-annotator disagreement for correcting compared to that for tagging by a factor of three, from .80 (5.6%/7.0%) to .27 (1.9%/7.0%).

2.3.3 Accuracy

Consistency, while a desirable formal virtue, tells us nothing about the validity of the annotators' output. We therefore compared each annotator's output not only with the output of each of the others, but also to a benchmark version of the eight texts. We derived the benchmark version from the tagged Brown Corpus by (1) mapping the original Brown Corpus tags onto the Penn Treebank tagset and (2) hand-correcting the revised version in accordance with the tagging conventions in force at the time of the experiment. We then measured accuracy in terms of the rate at which annotators disagreed with the benchmark version, thus obtaining four error rates for each text (one for each annotator) and 16 error rates (4 annotators × 4 texts) for each of the two annotation modes. The summary statistics for these figures are given in the first and second columns of Table 8. Eliminating the text sample J04 for the reasons discussed in Section 2.3.2 gives the revised figures in the third column.

Table 8:
Annotators' error rates compared to revised Brown Corpus (Summary statistics)

Annotation mode:	Tagging	Correcting	
		Raw	Revised
Mean:	5.4%	4.0%	3.0%
Median:	5.7%	3.4%	3.3%
Absolute range:	2.0%–8.9%	1.8%–9.0%	1.8%–3.9%
Relative range:	6.9%	7.2%	2.1%

As in the case of consistency, the average error rate is appreciably lower for correcting than for tagging. Eliminating J04 brings down the mean error rate somewhat, has virtually no effect on the median and reduces the relative range of error rates for correcting as compared to that for tagging by a factor of more than three, from 1.04 (7.2%/6.9%) to .30 (2.1%/6.9%).

We obtained a further measure of the annotators' accuracy and productivity by comparing their error rates to the rates at which the output of Church's PARTS program—appropriately modified to conform to the Penn Treebank tagset—disagreed with our benchmark version. We show the relevant figures in Table 9.[3] The figures in the first and second rows are the means based on the series of texts that the annotators tagged and corrected, respectively; the figures in the second row include the exceptional text J04.

Table 9:
"Error" rates of modified output of PARTS vs. annotators' error rates

	PARTS	Annotators	Difference
H03-R01	9.6%	5.4%	4.2%
H04-R02	7.8%	4.0%	3.8%

The figures in Table 9 show that in the correction task, human annotators reduce the disagreement rate between the output of PARTS and the benchmark version by 3.8%, corresponding to a factor of .49

[3] We would like to emphasize that the percentages given for the modified output of PARTS in Table 9 cannot and should not be construed as error rates and that they do not reflect the accuracy of Church's algorithm. This is because certain syntactic distinctions are made in the benchmark version that are not made by PARTS. For instance, the Penn Treebank tagset distinguishes present-tense non-third-person verb forms (VBP) from non-tensed verb forms (VB), while PARTS follows the Brown Corpus in tagging both forms as VB. Similarly, the Penn Treebank tagset distinguishes between prepositions (IN) and particles (RP), while PARTS tags both categories indiscriminately as IN.

(3.8%/7.8%). In the texts that the annotators tagged by hand, their error rate is 4.2% lower than the disagreement rate between PARTS and the benchmark version; that is, had the annotators been correcting these texts instead of tagging them, they would have reduced the error rate by a factor of .44 (4.2%/9.6%). In sum, employing human annotators to correct automatically pretagged texts reduces error rates by almost half.

3 Conclusion

The results of the experiment described above show that hand-correcting automatically tagged text yields better results than hand-tagging in a number of different respects. First, correcting is about twice as fast as tagging. All other things being equal, this alone would lead one to prefer correcting to tagging, particularly where large quantities of text are concerned. Second, correcting yields inter-annotator disagreement rates and error rates that are about half as high as those for tagging. Finally, correcting eliminates almost half of the disagreements between automatically annotated text and a benchmark version.

References

Church, Kenneth W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, 136–143.

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 Cooparative Research Project No. E-007.* Providence: Brown University.