# Rethinking The Corpus: Moving towards Dynamic Linguistic Resources

*Andrew Rosenberg*[1]

[1]Department of Computer Science, Queens College (CUNY), Flushing, NY, USA

andrew@cs.qc.cuny.edu

The corpus is an invaluable resource in Spoken and Natural Language Processing. Consistent data sets have allowed for empirical evaluation of competing algorithms. The sharing of high-quality annotated linguistic data has enabled participation and experimentation by a wide range of researchers. However, despite dubbing these annotations as "gold-standard", many corpora contain labeling errors and idiosyncrasies. The current view of the corpus as a static resource makes correction of errors and other modifications prohibitively difficult. In this paper, a perspective of the corpus as dynamically changing is advanced. We highlight the problems of the static view of the corpus through case studies of the Penn Treebank, Switchboard, Hub-4 and Boston University Radio News Corpus. We propose the use of version control software as a mechanism to facilitate this dynamic view.

## Abstract

**Index Terms**: Linguistic Resources, Opinion paper.

## 1. Introduction

Linguistically annotated corpora are the lifeblood of empirical linguistic research. The availability of high-quality human annotations of linguistic phenomena has been a boon to computational linguistics and speech research.

Publicly available data allows for a wide and diverse group of researchers to participate in investigations that they would not otherwise be able to perform. For example, the Penn Treebank has allowed hundreds of researchers to test hypotheses about American English syntax without being required to generate a single parse tree by hand. Due to the investment of time required to generate trustworthy linguistic annotations, the corpus is a tremendous resource-saving device. It allows the annotation effort performed by one organization to have multiplicative impact by being used by an unconstrained pool of researchers. The amount of resources (time, money and linguistic knowledge) required for performing large scale linguistic annotation limit the groups that are able to and interested in undertaking this activity, though the whole community benefits.

The corpus enables reproducibility across analyses and experimentation. Reproducibility is a foundational principle of the scientific method; without it, claims cannot be verified by independent and disinterested parties (cf. [1]). By representing consistent sets of data, corpora enable researchers to publish an analyses, or experimental results, with the confidence that the work is consistent with previous work, and will serve as comparison to future work.

We typically refer to linguistic annotations that are part of a corpus as "gold-standard". The expectation is that these represent some TRUE quality of the annotated material. The aforementioned benefits of corpus-based investigations, in fact, depend on this. When an underlying corpus has annotation errors, the findings of analyses are biased, and evaluations of experiments are undermined. The repeatability of studies requires that corpora should remain unchanged across research labs; if errors are corrected, this is undermined. If each site has its own version of experimental data, expectations of consistency and repeatability are damaged.

The current view of the corpus relies on the fact that the contained linguistic material and associated annotations are both unchanging, and without error. These two qualities put a significant burden on the constructors of corpora. Specifically, the annotation must be completely correct when the corpus is released. This is an unrealistic expectation. Through four case studies, we point out that errors in linguistic annotation are common, even in the community's most thoroughly investigated corpora, the Penn Treebank, Switchboard, Hub-4 and the Boston University Directions Corpus. Since errors are a natural and unavoidable part of linguistic annotation, we propose a vision of the corpus as a dynamic resource that changes and is improved through use over time. A changing corpus, however, poses difficulty in ensuring reproducibility across sites and publications. There are parallels between software development and maintenance and the development and maintenance of a linguistically annotated corpus. Therefore, version control systems that have been developed for software engineering can provide a clear answer to the reproducibility problem. By using version control, legacy versions of a corpus can be retrieved for comparison with previous studies, while the most current version represents the latest revisions and highest quality annotations.

The rest of this paper is structured as follows. Sections 2.1, 2.2 and 2.3 include case studies of errors that have been reported in the Penn Treebank, Switchboard Corpus, Hub-4 and Boston University Radio News Corpus. Section 3 describes how version control can be applied to the process of corpus construction and maintenance. Conclusions and future directions are described in Section 4.

## 2. Errors in Linguistic Annotation

In this section, we explore four heavily investigated, linguistically annotated corpora, The Penn Treebank, Switchboard, Hub-4, and the Boston University Radio News Corpus. We highlight that errors and inconsistency in the linguistic annotation in each of these are common. This investigation is not meant to depress. By acknowledging errors as the rule, not the exception, we can entertain a new vision of the corpus. Rather than viewing corpora as "gold-standard" monoliths, we can treat them as dynamic resources that require improvement and maintenance. By first acknowledging that errors are common, we motivate the adoption of techniques to facilitate the correction of linguistic annotations and to enable their rapid and consistent dissemination to the community at large.

### 2.1. Errors in the Penn Treebank

The Penn Treebank [2] is a corpus of naturally occurring text annotated for linguistic structure. The Penn Treebank consists of over 4.5 million words of American English that have been annotated with part of speech tags and syntactic structure. The material annotated under the Penn Treebank Project includes Wall Street Journal text, the Brown Corpus, Switchboard (discussed in more detail in Section 2.2), ATIS and other material.

In the last twenty years, the Penn Treebank has enabled remarkable progress in the development of automatic parsers, and part-of-speech taggers. Google Scholar has identified 3,929 citations of Marcus et al.'s 1993 paper as of the writing of this paper. There is little doubt that The Penn Treebank is one of the most significant corpora available to computational linguistics.

Despite the longevity, and impact of the Penn Treebank, and the number of researchers who have investigated its contents, errors and inconsistencies still remain in the annotation. Manning [3] points out that there are a number of clear errors in the part of speech tagging in the Penn Treebank. Manning points out the following errors, from section 02 of WSJ, marked with an exclamation point, with the correct tag in parentheses.

- Time , the/DT largest/JJS newsweekly/RB!(NN) , had average circulation of
- below the $2.29 billion value United Illuminating places/NNS!(VBZ) on its bid
- Rowe also noted that political concerns also worried/VBN!(VBD) New England Electric .
- Commonwealth Edison now faces an additional court-ordered refund on its summer/winter rate differential collections that/IN!(WDT) the Illinois Appellate Court has estimated at $140 million .
- Joseph/NNP M./NNP Blanchard/NNP , 37 , vice president , engineering ; Malcolm/NNP A./NN!(NNP) Hammerton/NNP

Yuret [4], in a blog post with the rather hyperbolic title "Why you should not use the Penn Treebank to Train a Parser", observes that 15% of examined constituent strings contain multiple ambiguities. These include inconsistencies in POS tagging like those observed by Manning, constituent labels, and brackets. Some of these inconsistencies may represent correct parses, reflecting a human interpretation of syntactic ambiguity, but some are correctly ascribed to be errors.

Blaheta [5], and Dickenson and Meurers [6, 7] have all done research on the automatic identification of errors in the Penn Treebank. Blaheta specifically highlights the frustration of detecting annotation errors without a mechanism to incorporate them into the corpus or share them with the community.

While many errors have been observed, and documented by members of the research community, without an easy-to-use mechanism to submit, approve and disseminate corrections to the corpus, these errors remain uncorrected.

### 2.2. Transcription Errors in Switchboard and Hub-4

The Switchboard Corpus [8] is the most widely studied corpus of conversational, telephone speech. The corpus includes over an hour of speech from fifty speakers and several minutes from additional hundreds. The Hub-4 corpus [9] has been a significant corpus in the development of speech recognition systems since its publication in 1997. Hub-4 contains 97 hours of manually transcribed broadcast news speech.

The initial dissemination of the Switchboard corpus revealed a significant rate of transcription and segmentation errors. These errors were considered significant enough to promote the Switchboard Resegmentation Project. Deshmukh et al. [10] reports that the original corpus had a transcription error rate of approximately 10%. After the resegmentation project, the transcription error was reduced to 2%. This is a significant and worthwhile improvement to the corpus, but the effort necessary to make a revision of a corpus is significant.

Pitz and Molau [11] developed a method to automatically verify speech transcription. They manually inspected 3.5 hours of speech of the Hub-4 corpus, corresponding to 1,352 segments. In this analysis they found that 36% of segments contain a transcription error and that 22% contain errors that they considered to at least "moderately" severe. "Minor errors" were defined to be those containing untranscribed noise, or single phoneme errors.

These observations show that speech transcription errors are exceedingly common. Moreover, despite the fact that these corpora have been heavily used, errors remain. Speech transcription is considered to be a task that any native speaker without hearing problems is capable of performing. It would stand to reason that if there were a simple way to share corrections of these errors that the integrity of the transcriptions would significantly increase.

### 2.3. Errors in the Boston University Radio News Corpus

The Boston University Radio News Corpus (BURNC) [12] is the most heavily investigated prosodically annotated corpus. This corpus includes 4 hours of professionally read broadcast news speech. The corpus includes manual and force-aligned orthographic transcription, and full ToBI [13] labeling. Despite being used in over a two hundred published works on automatic prosodic analysis, the BURNC corpus distributed by the Linguistic Data Consortium contains a variety of errors. The literature does not include a study of possible transcription errors in this data, though the expectation is that the professional readers correctly read the material.

Some of the errors are idiosyncratic. For example, the term "school-based" and the abbreviation "WBUR" occur in the read material. In the orthographic transcription these are considered to be single words, while prosodic annotation include multiple break indices, which should only occur at word boundaries. This forces experimenters to make an alignment decision between the two annotations. Without manual intervention, 62 out of 416 force-aligned transcripts contain an unequal number of break indices and words.

There are other errors in the prosodic annotation. The ToBI standard includes a number of requirements of a valid annotation. The number of times each requirement is violated in the BURNC appears in parentheses. Each intermediate phrase must contain at least one pitch accented word (946). Each intermediate phrase boundary must be marked with a phrase accent (68). Each intonational phrase boundary must be marked with a boundary tone (68). Many of these errors are fixable with relatively trivial manual intervention, while others require listening to the source material and adding a missing annotation.

From personal communication, some researchers opt to perform no corrections to the corpus and instead omit files that contain detectable errors. Others make some modifications to the transcription or to the break-index timing to facilitate alignment but will not include a new prosodic annotation. These independent decisions lead to slightly different corpora at each research site, undermining the repeatability and comparability of reported experiments.

Most publications do not report the pre-processing performed to clean errors in the distributed corpus. However, many include the number of words or syllables that are analyzed. We find at least three different word counts reported in the literature: Chen et al. [14] report 14,844, Sridhar et al. [15] 12,608, and Yuan et al. [16] 11,203. While each of these papers purport to be working on the same subset of the same corpus, the material they describe is substantially different. In the extreme, the Chen et al. paper describes the corpus as having nearly 32% more data that the material described in the Yuan, et al. paper. It is impossible to reconstruct the corpus modification that led to this disparity.

A mechanism to share information about a corpus, including the correction of errors and manual pre-processing, would enable all researchers to work on identical data.

## 3. Version Control as a Solution

Blaheta in describing a technique for identifying transcription errors describes the problems with individual research sites generating transcription corrections [5]. In his discussion of "Practical Considerations", he raises specific concerns over multiple researchers "imposing their own corrections". First, even if everyone published their own corrections, and comparisons to previous corrections, "there is some danger that a variety of different correction sets will exist concurrently". Second, "there will be dispute as to what is correct".

Both of these are valid concerns, but version control systems provide solutions to both of them.

Version control was developed by the software engineering community to provide a mechanism for multiple developers to simultaneously work on the same code base. In its traditional usage, it allows for a team of people (users) to make asynchronous changes to a common set of files. To use a resource under version control a users CHECKS OUT the current version or HEAD. Any changes can be made locally. When the user is satisfied with the changes that have been made, the user CHECKS IN or COMMITS their current version. This CHECK IN merges the changes with the HEAD and assigns a new version number to the current state of the resource. Previous versions can be RESTORED in the case of a bad or erroneous change being CHECKED IN. If the merge can be performed automatically, it is. In some cases, however, a change conflicts with the current state of the HEAD, in the conflict must be corrected manually. Note that there is nothing specific to software engineering or programming in this paradigm. Version control can be easily generalized to provide asynchronous access to shared resources. In preparing grants, technical reports and writing papers with multiple authors, groups have used version control systems to facilitate collaboration.

Version control systems can provide a mechanism to allow linguistic corpora to become dynamic resources. By distributing and maintaining linguistic corpora under a version control system, users will easily be able to correct errors and disseminate the corrections to the broader research community. Since every revision committed to the system receives a unique version number, researchers can document precisely which version of the corpus that their work is using. Even if the current state of the corpus has incorporated a number of changes, previous versions can be retrieved, facilitating honest comparisons with previously published work.

The considerations raised by Blaheta are easily addressed by modern version control systems. First, it is acceptable and expected that multiple versions will exist concurrently. There is nothing wrong with this *per se*. The issues arise when two algorithms or analyses need to be compared on the basis of consistent data. Using an unpublished revision of a corpus in a research report should be discouraged, just as using any other non-repeatable technique is discouraged. Version control, however, enables the publication of revisions through a single command. Another option that version control provides is to construct a new branch of the corpus. A new branch results in a second HEAD being constructed based on a previous revision of the corpus. This would likely create some degree of confusion, but there is nothing to prevent the citation and maintenance of multiple branches of a corpus.

Second, most version control systems allow for a moderator to be required to approve a proposed change before it is merged with the HEAD. This moderator would have the responsibility of determining if a change is, in fact, a correction of an error, or if there is some reasonable ambiguity or subjectivity in the annotation. The owners of a corpus would be responsible for either 1) moderating the proposed changes, 2) assigning one or more moderators or 3) allowing the community of users to self-regulate the integrity of the corpus.

There are a variety of publicly available version control systems including cvs, Subversion (svn), and git. GitHub has emerged as a popular web interface to git repositories with over a million users, and an clean, easy-to-use graphical interface. While there are differences between each of these systems, each provide the necessary functionality to enable the use of version control to maintain and update linguistic corpora.

There are (at least) two potential limitations to using version control to maintain linguistic corpora. First, there is an additional burden on corpus creators to moderate proposed changes. This may prove to be a significant limiting factor. However, as the community as a whole benefits from improved data integrity, it should be possible to find a moderator or group of moderators willing to perform this service. If the corpus has a licensing fee associated with it, it seems reasonable to expect this fee to cover corpus maintenance.

A second limitation is more directly related to the task of linguistic annotation. There are cases where the annotation output is not consistent with the annotation process. For example, annotated data may be stored as XML, or Praat TextGrids. However, it is rare for an annotator to manipulate these files directly. More commonly, annotators use specially crafted tools like AGTK [17], or audio browsers like Praat, wavesurfer or Xwaves. A version control system will describe a proposed change in the annotation format, without any consideration of the semantics of this change. This is not a problem in the software engineering setting, since users are expected to be able to interpret the format of the file that they are submitting. However, most users of Praat are unable to read a TextGrid file. This makes the problem of moderation more difficult.

As part of Reciprosody, we are developing a wrapper around a git repository that allows for a version control system to describe changes to a TextGrid file in terms that an annotator can clearly understand. Reciprosody is a new repository of prosodically annotated material. Prosodic annotation is commonly performed as a set of tiers, including a time-aligned orthographic transcript, and prosodic markers. For example, ToBI annotation [13] includes four tiers: words, tones, breaks, and miscellaneous.

The output of a simple `diff` or version control DELTA may include an indication that the two lines preceded by left brackets have been added:

```
10.00
H*
>12.00
>L+H*
15.00
H*
```

However, the Reciprosody git wrapper describes this example change as “Point added to tier ‘tones’ at time 12.00s with label L+H*”. A description of each Checked In change is included in a comment, allowing users to quickly browse previous changes. If the system knew that the file represented a ToBI annotation this description could be further augmented to say “L+H* pitch accent added to word ‘Example’ at time 12.00s”. This description would allow for a user or moderator to quickly navigate to the modified position in Praat and inspect the surrounding material. A further improvement that we are considering is the ability for the repository to open two Praat windows containing the existing file and proposed change along with this text description.

## 4. Conclusion and Future Work

Annotated corpora are an indispensable resource for the study of linguistic phenomena and the engineering of speech and language processing tools. However, large scale annotation efforts invariably lead to annotation errors and labeler idiosyncrasies. The static view of “gold-standard” annotation leads to a rigid definition of the corpus as something unchangeable. By moving to a more dynamic definition of a corpus, the community is empowered to fix and share errors, and to discuss and address labeling inconsistencies. Version control software provides much of the functionality required to make this transition. Updates, conflict resolution and automatic assignment of version numbers are all standard fixtures of subversion and git. By adopting a dynamic view of the corpus, the quality of the annotation will improve. Errors can be addressed as soon as they are identified by any member of the community. This perspective will also allow a faster dissemination of linguistic annotation. If some degree of errors and incomplete annotations are an expected part of a corpus, creators of linguistic resources will be more likely to disseminate their product at an earlier phase of the process.

There is considerable work necessary for this vision to be realized. Primarily, such a transition requires support from the community of corpus users and creators. It is clear that a more dynamic view of the corpus allows for a higher quality product. However, there will be additional costs on those who maintain data sets, including approving changes to the corpus, and maintaining landmark version numbers. This increased cost must be made bearable through the development of effective tools for corpus maintenance and community support for maintaining data integrity.

We are currently, developing a web-interface to a version controlled backend as part of Reciprosody, a shared repository of prosodically annotated speech. Demonstrating the efficacy of this approach on smaller corpora with a smaller community of users will serve as a valuable scenario to realize the unforeseen consequences and issues of scalability that come with this new view of the corpus. This project will serve as an incubator of new ideas, tools and resources to enable data owners and users to engage with a dynamic view of the corpus.

## 5. Acknowledgements

The author would like to thank Julia Hirschberg and Stefanie Shattuck-Hufnagel for their roles in developing Reciprosody and Mark Hasegawa-Johnson, Karen Livescu and Margaret Fleck for a very inspiring discussion on this topic.

## 6. References

[1] J. R. Huizenga, *Cold Fusion: The Scientific Fiasco of the Century*. New York, NY, USA: Oxford University Press, 1993.

[2] M. Marcus, M. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[3] C. D. Manning, “Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?” *Lecture Notes in Computer Science 6608*, pp. 171–189, 2011.

[4] D. Yuret. Why you should not use the penn treebank to train a parser. [Online]. Available: http://denizyuret.blogspot.com/2006/10/why-you-should-not-use-penn-treebank.html

[5] D. Blaheta, “Handling noisy training and testing data,” in *EMNLP*, 2002.

[6] M. Dickinson and W. D. Meurers, “Detecting inconsistencies in treebanks,” in *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden, 2003, pp. 45–56.

[7] ——, “Detecting errors in part-of-speech annotation,” in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, Budapest, Hungary, 2003, pp. 107–114.

[8] J. Godfrey, E. Holliman, and J. McDaniel, “Switchboard: telephone speech corpus for research and development,” *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 517–520 vol.1, Mar 1992.

[9] D. Graff, Z. Wu, R. MacIntyre, and M. Liberman, “The 1996 broadcast news speech and language-model corpus,” in *Proceedings of the 1997 DARPA Speech Recognition Workshop*, 1997.

[10] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, “Resegmentation of switchboard,” in *ICSLP*, 1998.

[11] M. Pitz and S. Molau, “Automatic verification of broadcast news transcriptions,” in *Eurospeech*, 1999.

[12] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, “The boston university radio news corpus,” Boston University, Tech. Rep. ECS-95-001, March 1995.

[13] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “Tobi: A standard for labeling english prosody,” in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[14] K. Chen, M. Hasegawa-Johnson, and A. Cohen, “An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic-prosodic model,” in *ICASSP*, 2004.

[15] V. R. Sridhar, S. Narayanan, and S. Bangalore, “Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework,” in *HLT-NAACL*, 2007.

[16] J. Yuan, J. Brenier, and D. Jurafsky, “Pitch accent prediction: Effects of genre and speaker,” in *Interspeech*, 2005.

[17] K. Maeda, S. Bird, X. Ma, and H. Lee, “Creating annotation tools with the annotation graph toolkit,” in *Proceedings of the Third International Conference on Language Resources and Evaluation*, vol. cs.CL/0204005, 2002.