TOWARDS A FORMAL FRAMEWORK FOR LINGUISTIC ANNOTATIONS

Steven Bird *Mark Liberman*

Linguistic Data Consortium
University of Pennsylvania, Philadelphia, USA

ABSTRACT

‘Linguistic annotation’ is a term covering any transcription, translation or annotation of textual data or recorded linguistic signals. While there are several ongoing efforts to provide formats and tools for such annotations and to publish annotated linguistic databases, the lack of widely accepted standards is becoming a critical problem. Proposed standards, to the extent they exist, have focussed on file formats. This paper focuses instead on the logical structure of linguistic annotations. We survey a wide variety of annotation formats and demonstrate a common conceptual core. This provides the foundation for an algebraic framework which encompasses the representation, archiving and query of linguistic annotations, while remaining consistent with many alternative file formats.

1. INTRODUCTION

‘Linguistic annotation’ is a cover term for any orthographic, phonetic or prosodic transcription; any speech, part-of-speech, disfluency or gestural annotation; and any free or word-level translation. Linguistic annotations may describe texts or recorded signals; our focus will be on the latter, broadly construed to include any kind of audio, video or physiological signal, or any combination thereof.

To date there have been several independent efforts to provide tools for annotating linguistic signals, to provide general formats for annotations, and to provide tools for searching databases of annotations. Additionally, hundreds of annotated linguistic databases have been published, where each database typically contains several different tiers of annotation. While the utility of such tools, formats and databases is unquestionable, the lack of standards is becoming a critical problem. Attempts to standardise practice in this area have focussed on file formats (e.g. [3]). We contend that file formats, though important, are secondary.

In this presentation we report on a study of the logical structure of linguistic annotations. We demonstrate that, while the different annotation formats vary greatly in their form, their logical structure is remarkably consistent. In order to help us think about the form and meaning of annotations, we describe a simple mathematical framework endowed with a practically useful formal structure. This opens up an interesting range of new possibilities for creation, maintenance and search. We claim that essentially all existing annotations can be expressed in the framework.

The present paper gives an extended abstract for the presentation, itself to be made available in full at `[http://www.ldc.upenn.edu/sb/icslp98.html]`.

2. DESIDERATA FOR A LINGUISTIC ANNOTATION FRAMEWORK

We will focus on three evaluation criteria for a linguistic annotation framework:

Generality
The framework should be sufficiently expressive to encompass all commonly used kinds of linguistic annotation, including sensible variants and extensions. It should be capable of managing a variety of (partial) information about labels, temporal information and hierarchical structure.

Searchability
There should be an efficient algebraic query formalism, whereby complex queries are composed out of well-defined combinations of simple queries, and where the result of querying a set of annotations is just another set of annotations. Annotations, however incomplete, should still be searchable. There should be an efficient indexing scheme providing near constant-time access into arbitrarily large annotation databases. The framework should also support the projection of natural sub-parts of annotations. For example, we may wish to project out just the prosodic content of annotations, or just the orthographic content.

Maintainability
Annotation databases should be durable, remaining coherent and usable in the presence of corrections or the addition of new layers of annotation. Queries on prior versions should remain valid, and references into superseded annotations should persist whenever possible. Layers and versions of annotations should be modular so that revisions to one part do not entail global modification. For example, changing the spelling of a word should not entail changes to an annotation of phrase-level discourse function which covers the same text.

In addition to these desiderata, we shall be concerned to provide realisations of annotations and queries in the finite-state realm, in the graphical domain, and as XML markup.

3. EXISTING ANNOTATION SYSTEMS

Prior to presenting our proposed framework, a selection of examples drawn from a variety of existing annotation systems will be presented. Here, we just give one example taken from the TIMIT database [2]. The file `timit/train/dr1/fjsp0/sa1.wrd` contains:

```
2360 5200 she
5200 9680 had
9680 11077 your
11077 16626 dark
16626 22179 suit
22179 24400 in
24400 30161 greasy
30161 36150 wash
36720 41839 water
41839 44680 all
44680 49066 year
```

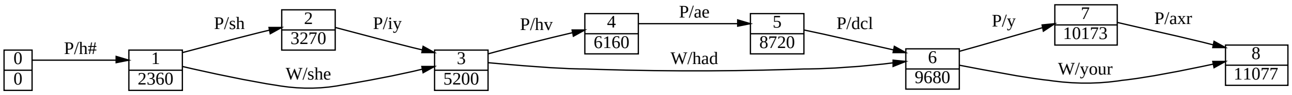


Figure 1: Graph Structure for TIMIT Example

This file combines an ordinary string of orthographic words with information about the starting and ending time of each word (measured in audio samples at a sampling rate of 16 kHz). The path name `timit/train/dr1/fjsp0/sa1.wrd` tells us that this is training data, from ‘dialect region 1’, from female speaker ‘jsp0’, and that it contains words and audio sample numbers.

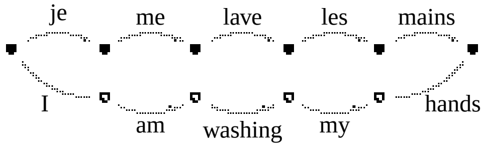
The file `timit/train/dr1/fjsp0/sa1.phn` contains a corresponding broad phonetic transcription, which begins as follows:

```
0 2360 h#
2360 3720 sh
3720 5200 iy
5200 6160 hv
6160 8720 ae
8720 9680 dcl
9680 10173 y
10173 11077 axr
11077 12019 dcl
12019 12257 d
```

We can interpret each line `<time1> <time2> <label>` as an edge in a directed acyclic graph, where the two times are attributes of nodes and the label is a property of an edge which connects those nodes. The resulting ‘annotation graph’ for the above fragment is shown in Figure 1.

Observe that edge labels have the form `<type>/<content>` where the `<type>` here tells us what kind of label it is. We have used `P` for the (phonetic transcription) contents of the .phn file, and `W` for the (orthographic word) contents of the .wrd file.

The top number for each node is an arbitrary node identifier, for ease of reference, while the bottom number is the time reference. We distinguish node identifiers from time references since nodes may lack time references. This may be because times were not measured, as in typical annotations of extended recordings where time references might only be given at sentence boundaries. Or it may be because time measurements are not applicable in principle, as may arise when an annotation is a phrasal translation. This last point is illustrated below, where time-marked vertices are represented as bullets and non-time-marked vertices are represented as hollow circles.



Observe that there is no meaningful way of assigning time references to word boundaries in the phrasal translation.

The presentation will cover comparable examples from other annotation models, in order to demonstrate the existence of a common conceptual core of linguistic annotations. The survey will include Emu [1], BAS Partitur [4], the NIST ‘Universal Transcription Format’ [3], and the speech concordance facility of LDC Online [5]. Full details and updated pointers will be available at `[http://www.ldc.upenn.edu/sb/icslp98.html]`.

4. AN ALGEBRAIC FRAMEWORK

We maintain that most, if not all, existing annotation formats can naturally be treated as directed acyclic graphs having typed labels on (some of) the edges and time-marks on (some of) the vertices. We call these ‘annotation graphs’.

On our algebraic approach, queries are nothing other than expressions in a calculus defined over annotation graphs. This calculus is built up recursively from elementary graphs by combining them in various ways, including conjunction, disjunction, concatenation and Kleene closure, in the analogous fashion to the way regular expressions are built up in an RE calculus. Coreference of arbitrary edges between conjuncts is accomplished using operations analogous to the reference operators available in extended regular expression formalisms (such as that of Perl).

In concert with this, we propose an indexing method based on the elementary annotation graphs from which queries are constructed. Indices will specify where particular elementary annotation graphs are to be found, and so a complex search expression can be limited to those regions for which these graphs are necessary parts.

After presenting the model we will show how it satisfies our desiderata for a linguistic annotation framework and demonstrate a prototype implementation.

5. REFERENCES

1. Steve Cassidy and Jonathan Harrington. Emu: An enhanced hierarchical speech data management system. In *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, 1996. [http://www.shlrc.mq.edu.au/emu/].

2. John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. NIST, 1986. [http://www.ldc.upenn.edu/lol/docs/TIMIT.html].

3. NIST. A universal transcription format (UTF) annotation specification for evaluation of spoken language technology corpora. [http://www.nist.gov/speech/hub4_98/utf-1.0-v2.ps], 1998.

4. Florian Schiel, Susanne Burger, Anja Geumann, and Karl Weilhammer. The Partitur format at BAS. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 1998. [http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html].

5. Zhibiao Wu and Mark Liberman. LDC Online: A digital library for linguistic research and development. In *Proceedings of the Second ACM Conference on Digital Libraries*. New York: ACM, 1997. [http://www.ldc.upenn.edu/lol].