TR-IT-0335

SPACES: A Semantically and Prosodically Annotated Corpus Of English Speech

Jeremy Bateman & Nick Campbell

February 16, 2000

ABSTRACT

ATR possesses a corpus of approximately one million words of English text, of which every word has been syntactically and semantically tagged, and every sentence assigned a correct parse using the ATR grammar of general English [1,2]. This paper describes our work towards a large speech corpus (SPACES) containing the same material, annotated using a reduced version of the ToBI prosodic annotation system. For the first time a large corpus of prosodically annotated read English speech will also be annotated syntactically and semantically. We describe how correlations in the parallel annotations may provide a 'gold standard' against which to continue training and evolving the CHATR speech synthesizer.

1 Introduction: The Written Corpus

In 1994-97 ATR-ITL assembled a corpus of approximately one million words of written English text. Each word has been assigned a syntactic tag to indicate its part of speech and, if a content word (noun, verb, adjective, adverb), a semantic tag also. Each sentence has been parsed according to a grammar of approximately 1100 rules (the ATR Grammar of General English [1]. All these annotations were chosen or checked individually by human analysts to ensure accuracy [2]. For a full discussion of the assembly and use of this corpus so far, see [3].

The corpus contains about 2300 texts, ranging in size approximately from 30 to 3600 words. These fall into two categories:

- 380,000 words of transcriptions of the ATR Travel Corpus, which consists of dialogues such as travel arrangements and hotel bookings.
- 700,000 words of written nontravel texts drawn from a wide variety of electronically available sources including the AP corpus, Wall Street Journal (WSJ) financial articles, and some scanned ephemera such as advertisements.

Time does not allow us to record the whole corpus, but we have recorded over 108,000 words in the first year and each sentence has been prosodically annotated. This facilitates our investigation of correlations between the syntactic-semantic features of the sentences and prosodic features of corresponding recordings. It has already been demonstrated that using the prosodic features of prominence and interword break index duration improves listeners' understanding of syntactically ambiguous sentences [9].

Our principle interest lies in identifying and labelling break indices **2**, **3** and **4**, determining patterns in their occurrence and using these patterns to determine whether they will occur in text when it is realised as speech. We describe the distinctions between these labels in section 2.2 below.

2 ToBI Annotation In The Spoken Corpus

We based our prosodic analysis of recordings on the ToBI system, originating from the Linguistics Department at Ohio State University [4,5,6].

2.1 Labelling of prominent words

The ToBI labelling system provides a standard set of labels with which to annotate the prosodic features of recordings of speech. As it has become widely used, we are using it to enhance commonality with similar or related projects. Our version currently uses a very reduced label set as we mark prominent words rather than accented syllables. While ToBI notes 'pitch events associated with accented syllables (pitch accents)', we locate stress in the word by using dictionaries in conjunction with data relating to fundamental frequency, duration and power.

By not marking accent types, we are speeding up annotation of data. This multi-layered corpus annotation facilitates future entry of further ToBI labelling, such as for phrase tones on the boundaries of prosodic units, which we do not at present label. We have plans for

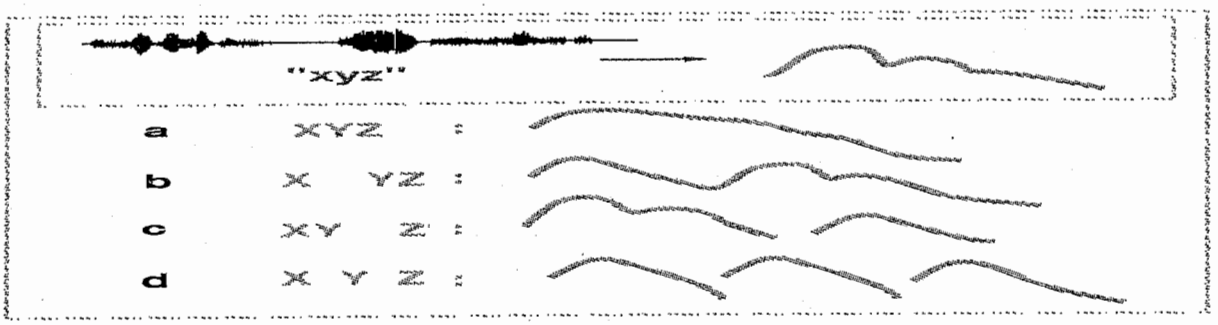

Figure 1: prosodic labelling by comparison with predicted alternatives

full ToBI-labelling of part of the corpus in order to provide reference material to check the automatically-derived annotations.

Research with the incomplete written corpus has already shown that incorporating lexical and syntactic information from the ATR General English Parser can enhance prediction of pitch events, pauses and phrasing [11].

Patterns of prominences indicate where, in a particular syntactic structure, CHATR should be giving words prominence. Most prominences occur in content words, but their precise distribution will vary according to factors such as the nature of the discourse and speaker attitudes. These features vary in the corpus as it is drawn from such a wide variety of texts[1]. It is therefore helpful for studying how such variables affect utterance production[2]

Figure 1 shows the set of pitch contour patterns predicted as possible for a given input text, and the one that was realised on the spoken utterance. This shows that information derived from the physical speech waveform can provide important details about its bracketing.

Figure 2 illustrates prominence detection. The prediction exhibits a close correlation to that of the actual pattern [12]. When the actual pitch contour (dotted line) differs significantly from the predicted (solid line) one, prominence can be inferred (especially when this is supported by evidence from segmental durations and power).

2.2 Labelling of break indices

We retain ToBI break index labelling more fully than tone labelling. These labels indicate 'the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance'. As **The ToBI Annotation Conventions** [5] describe these indices:

[1]Texts include: New York City Opera ticket (ocr76); Flier from bank about loan rates (ocr44); Judge Refuses To Deport Latvian Accused In Nazi War Crimes (b13); Futures Trader Has Heavyweights Behind Him as He Faces Charges (a884)

[2]Note that the analysts of the written corpus assigned to all files 'header information' such as TONE, STYLE, LINGUISTIC-LEVEL, POINT-OF-VIEW, TYPE-OF-DOCUMENT, DISCOURSE-MODE, and MODE-OF-PRESENTATION. Samples of these classifications include 'scientific'; 'literary'; 'educational'; 'technical'; 'formal'. These annotations enable us to easily categorise texts according to these discourse characteristics.

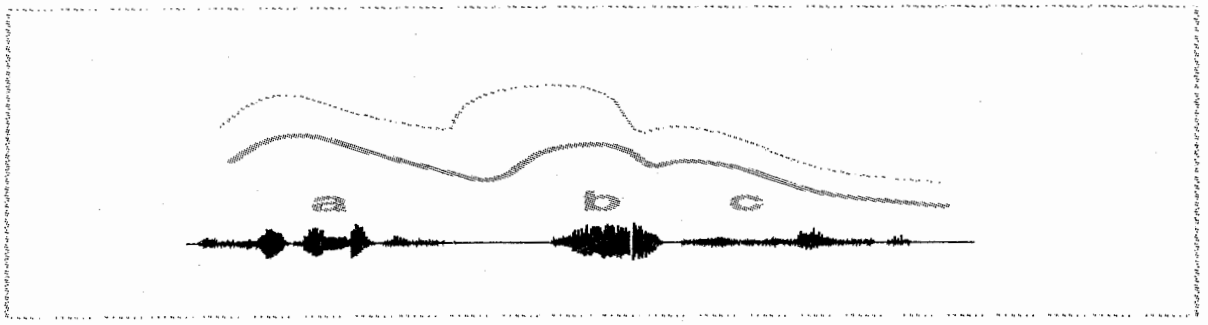

Figure 2: patterns of focus and segmentation

- **0** clear phonetic marks of clitic groups
- **1** most phrase-medial word boundaries
- **2** a strong disjuncture marked by a pause or virtual pause, but with no tonal marks, or a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary
- **3** intermediate intonation phrase boundary; i.e. marked by a single phrase tone affecting the region from the last pitch accent to the boundary
- **4** full intonation phrase boundary; i.e. marked by a final boundary tone after the last phrase tone

An interesting challenge lies in determining patterns in the occurrence of break indices **2** and **3**, and finding the syntactic and semantic cues which trigger them. This knowledge should also enhance CHATR's prosody by enabling the synthesizer to predict where the interphrase and interclause breaks will occur in a given sentence.

We take **2** to indicate tone group boundaries, or minor pauses for breath in the course of a (perhaps long) phrase or clause.

3 tends to indicate the boundary of a clause, marked by a noticeable pause at a major break in the syntactic structure of the sentence.

Although Hirschberg & Beckmann [5] state that 'there is no default juncture type', **1** is such a default interword boundary in our annotation scheme, and is therefore not explicitly marked. **4** conversely almost always correlates with sentence-final breaks, particularly in SPACES as we are recording written sentences (traveldata, the texts in the written corpus which are transcribed from speech, can differ from this, as we discuss below).

3 Recording And Analysis Procedure

The analyst chose sentences ranging across the text types in the corpus to reproduce its variegated nature. The corpus does contain some (tagged) sentences which the ATR grammar

had proven unable to parse, due to their complexity or particularly unusual features, and these sentences were correspondingly not recorded, as it was more ergonomical to record only those sentences which could contribute to our attempt at multilevel annotation[3].

This leads to the anomaly that directory b242 may contain files b242.01, b242.02, b242.04 but not b242.03. In these cases, the directory will contain a file b242.03.qrs noting that the sentence has not been parsed in the written treebank.

The sentence divisions of the written corpus were retained absolutely. Because the analysts annotating the written corpus made sentence splitting decisions based on the ATR General English Grammar, designed to handle analysis of complex sentences such as are found in written English texts, they frequently grouped the transcripts of utterances (Traveldata) into sentences which span more than one utterance. Hence some recordings may seem to take the form of two actual original utterances. What would be classed as separate utterances, therefore, were sometimes combined into a parse suitable for a multiclause sentence, and the recordings are marked as two intonation phrases, with a midsentence and a final **4**:

Excuse me, I was supposed to meet my friend here, but I can't find him

UF430011.01

0.510000 122 #SIL#
0.970000 122 Excuse
0.981220 121 **n**
1.030000 122 #SIL#
1.250000 122 me
1.253520 120 **4**
2.200000 122 #SIL#
2.240000 122 I
2.370000 122 was
2.710000 122 supposed
2.713613 121 **n**
2.770000 122 to
2.960000 122 meet
3.070000 122 my
3.160000 122 #SIL#
3.520000 122 friend
3.525818 121 **n**
3.760000 122 here
3.840372 120 **3**
4.090000 122 #SIL#
4.270000 122 but
4.330000 122 I
4.640000 122 can't
4.950000 122 find

[3]As these sentences are among those with the most complex grammar, they may be of great interest for prosodic analysis also. However, they are also those less likely to be encountered in either written or spoken language

4.960522 121 **n**
5.030000 122 #SIL#
5.240000 122 him
5.960000 122 #SIL#
5.256297 120 **4**

(UF430011.01)

In this example, it was appropriate to split the 'sentence' with another intonation phrase boundary after 'excuse me'.

The subject matter of the written texts in the corpus often needed to be expressed in long and complex sentences, which required particularly careful recording. In many texts from the financial and business domains, sentences averaged over 30 words.

The sentences were recorded in the same room in sessions of approximately one hour, using a Sony TCD-D10 Digital Audio Tape (DAT) recorder, set at recording level 7, and a Sennheiser head-mounted microphone.

The DAT recordings were then transferred to digital storage. A script was written which downsampled each file into speech data (.sd) and fundamental frequency (.f0) files and made a directory to store each recording with all its associated files (see appendix 2). This enabled the analyst to view each utterance's speech wave using the **xwaves** speech recording display and analysis system. A text file (.txt) was written containing the words of the utterance. These files were those required for analysis of data by Aligner, a tool which aligns text to its position in a sound file and displays the positions of words in the file using xwaves [7]. Aligner's BAlign function then aligned the words in the .txt file to their positions in the .sd file (see figure 3 below).

The initial stage of BAlign process required analyst involvement to enter phonetic transcriptions of words which were not in his personal Aligner dictionary, a process which has continued as we have continued to analyse texts from new subject areas with new vocabulary. The transcription software uses American English vowels, but we are adapting Aligner to recognise British English vowels. At present, for example, the vowels in **cot** and **cart** are both transcribed **aa** even though they are clearly different in British English. As a trace of this stage of the process, Aligner creates a .phones file recording its phonetic transcription of the utterance, including notes of stressed syllables and syllable boundaries.

Using Aligner's Align function, the analyst then uses xwaves to display the aligned .sd and .txt files (see figure 1) . He then marks the other two horizontal windows on screen to indicate the presence of accents and breaks. The annotations are saved as the sentence's .tones and .breaks files. The .words, .tones, and .breaks files record the time in the utterance at which these features occur. This time alignment enables researchers to match the prosodic annotation to the text, and therefore to examine relationships between prosody and the syntactic and semantic annotations.

As an example of this consider sentence 20 of file b228 (Figure 3 overleaf).

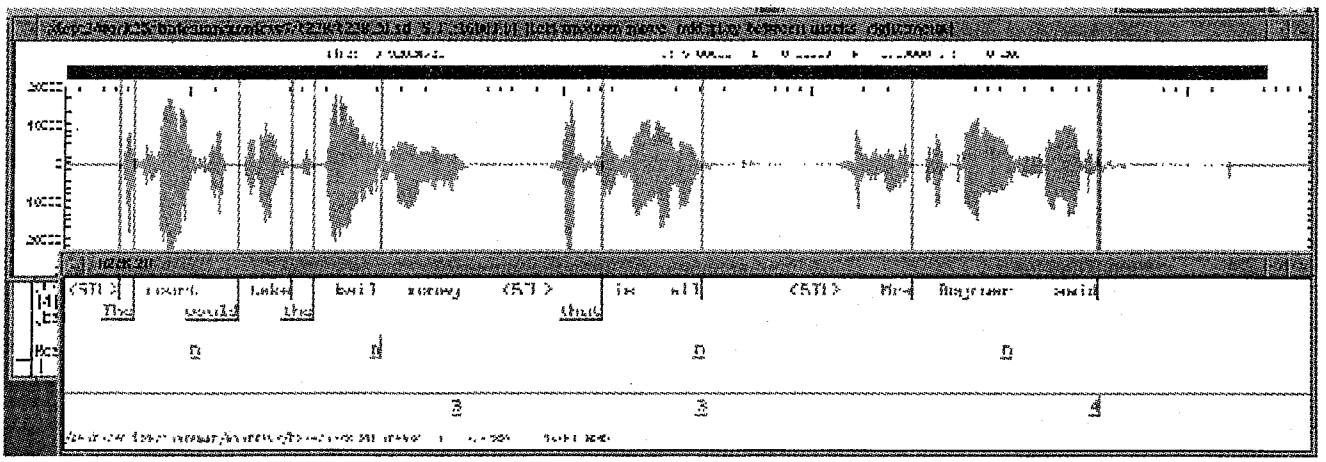

Figure 3: Aligner display of b228.20 .sd, .words, .tones and .breaks files

The input .txt file of b228.20 is

'The court would take the bail money, that is all,' Mrs. Dogruer said.

Figure 3 above shows the output Aligner display of this.

In the example (sentence b228.20), words given prominence in the recording are marked by **n**: *court*, *bail*, *all*, and *Dogruer*. A **2** break is marked after the first clause of the quoted sentence, a longer **3** after the second, and an utterance-final **4** at the end of the sentence[4].

Aligner labels the words of the .txt file with the times at which they occur to form the .words file:

0.220000 122 #SIL#
0.280000 122 The
0.540000 122 court
0.700000 122 would
0.910000 122 take
1.000000 122 the
1.260000 122 bail
1.590000 122 money
1.970000 122 #SIL#
2.160000 122 that
2.290000 122 is
2.560000 122 all
3.120000 122 #SIL#
3.400000 122 Mrs
3.800000 122 Dogruer
4.140000 122 said
5.160000 122 #SIL#

[4]**1** is the standard interword break and is assumed rather than explicitly marked

(b228.20.words)

(Aligner inserts #SIL# to indicate where it detects silences in the recording. These happen at the beginnings and ends of sentences and there is some correlation with breaks, although sometimes Aligner labels a pause where listeners find it hard to detect)

The timings on the .tones and .breaks files are generated by the point at which the words of the .txt files are labelled. As is conventional with ToBI, tone and break annotations are inserted to the right of the word and breaks to the right of the tone mark if there is one. The timings therefore link the annotated events to physical features that can be measured from the speech wavesform and used to determine automatically the location of breaks and accented words.

0.547687 121 **n**
1.274789 121 **n**
2.573186 121 **n**
3:800760 121 **n**

(b228.20.tones)

1.600569 120 **2**
2.582628 120 **3**
4.150147 120 **4**

(b228.20.breaks)

This shows how the xwaves display above is represented in the stored files written using the xwaves/Aligner display.

The corresponding sentence in the written corpus is:

> [start [sprpd4 "_" [coord9 [sprime1 [sd1 [nbar4 [d1 The_AT d1] [n1a court_NN1SYSTEM n1a] nbar4] [vbar2 [o8 would_VMPAST o8] [v2 take_VVIRECEIVE [nbar4 [d1 the_AT d1] [n4 [n1a bail_NN1FUNCTION n1a] [n1a money_NN1MONEY n1a] n4] nbar4] v2] vbar2] sd1] sprime1] [coord4 [cc9 ,_, cc9] [sprime1 [sd1 [nbar2 [d1 that_DD1 d1] nbar2] [vbar1 [v2 is_VBZ [nbar2 [d1 all_DB d1] nbar2] v2] vbar1] sd1] sprime1] coord4] coord9] ,_, "_" [si1 [nbar1 [n4 [n1a Mrs._NP1HON n1a] [n1a Dogruer_NP1LSTNM n1a] n4] nbar1] said_VSAYINGD si1] . _. sprpd4] start]

The syntactic-semantic tags follow each word, word_TAG, e.g. in court_NN1SYSTEM the tag is NN1SYSTEM. Labelled brackets indicate the ATR General English grammar rules which apply to each syntactic unit, for example noun rule **n1a** embraces [n1a money_NN1MONEY n1a].

The above sentence consists of three clauses, two of direct speech (those bracketed by [sprime1 sprime1] brackets and coordinated under the [coord9 coord9] brackets) and one indicating its nature as direct speech ('Mrs. Dogruer said', bracketed by [si1 si1]). Here the breaks clearly correspond to the three clauses of the sentence.

To indicate the prosodic annotations, the reproduction below indicates prominences and breaks by capitalising accented words and inserting the appropriate numerical indicator) to indicate breaks:

> [start [sprpd4 "_" [coord9 [sprime1 [sd1 [nbar4 [d1 The_AT d1] [n1a COURT_NN1SYSTEM n1a] nbar4] [vbar2 [o8 would_VMPAST o8] [v2 take_VVIRECEIVE [nbar4 [d1 the_AT d1] [n4 [n1a BAIL_NN1FUNCTION n1a] [n1a money_NN1MONEY n1a] n4] nbar4] v2] vbar2] sd1] sprime1] **2** [coord4 [cc9 ,_, cc9] [sprime1 [sd1 [nbar2 [d1 that_DD1 d1] nbar2] [vbar1 [v2 is_VBZ [nbar2 [d1 ALL_DB d1] nbar2] v2] vbar1] sd1] sprime1] coord4] coord9] ,_, "_" **3** [si1 [nbar1 [n4 [n1a Mrs._NP1HON n1a] [n1a DOGRUER_NP1LSTNM n1a] n4] nbar1] said_VSAYINGD si1] ._. sprpd4] start] **4**

The prosodic labelling is conducted without reference to the semantic and syntactic labelling already undertaken, to reduce the extent that the analyst's (extensive) knowledge of the ATR grammar interferes with his prosodic analysis of the utterances. This would corrupt the independence of the two analyses, which we hope will naturally coincide to reveal commonalities.

Such cooccurrence of different types of annotation facilitates investigation of consistent relations between the written corpus' grammatical units and part of speech (POS) tag sequences and occurrences in the spoken corpus of prominences and break indices. Some similar research has already been conducted using the annotations in the written corpus but using CHATR to generate their spoken renditions [11].

4 Text To Speech And Break Labelling Issues

4.1 Adapting Written Words To Speech

In recording the written corpus, we attempt to be as faithful as possible to the written texts in order to facilitate analysis of the links between the different types of annotation. However, the differences between written and spoken language necessitated some distinctions.

Consider the sentence in the written corpus:

> *Yes, it's Akira Okada*

> [start [sprpd23 [sprime2 [ibbar1 [i1f Yes_UH i1f] ,_, ibbar1] [sd1 [nbar6 it_PPH1 nbar6] [vbar1 [v2 's_VBZ [nbar1 [n4 [n1a Akira_NP1FRSTNM n1a] [n1a Okada_NP1LSTNM n1a] n4] nbar1] v2] vbar1] sd1] sprime2] sprpd23] start]

(U22002.07)

The genitive marker **'s**, and enclitics such as **'s** in **it's**, are considered separate text items in the written corpus due to their independent syntactic function. **'s** is therefore tagged exactly as is **is**. As such words show no phonetic distinction from the words they are attached to, they are generally not considered separate words in spoken corpora, and we adopted this convention as demonstrated in U22002.07 above.

Conversely, the various expressions of number, for times, dates, and prices, were each considered one long string of words which did not warrant tag differentiation in the written corpus. Although the last word in such a sequence is more likely than any other to be stressed, different contexts will require speakers to realize such strings with different prosody.

Because a high proportion of texts are either financial articles from the electronic Wall Street Journal, or traveldata files dealing with payment for goods or services, we have a considerable dataset with which to study these distinctions. The 'standard' arrangement described above occurs in the sentence below, for example:

0.910000 122 #SIL#
1.040000 122 so
1.240000 122 that
1.248283 121 **n**
1.252196 120 **2**
1.350000 122 will
1.480000 122 be
1.710000 122 one
2.010000 122 hundred
2.090000 122 and
2.420000 122 forty
2.820000 122 dollars
2.829180 121 **n**
2.856571 120 **2**
3.390000 122 altogether
3.408320 121 **n**
3.435712 120 **4**
4.120000 122 #SIL#

(12030121RT.36.words)

dollars is counted as being part of a price expression, which is given the 'multiword' MPRICEWORD tag (M indicates any numerical expression, whether in words or figures):

> [start [sprpd23 [sprime2 [ibbar2 [r2 so_RRCONCESSIVE r2] ibbar2] [sd1 [nbar2 [d1 that_DD1 d1] nbar2] [vbar2 [o8 will_VMPRES o8] [v2 be_VBI [nbarq12 [nbar1 [n1c [multiword4 one_MPRICEWORD51 hundred_MPRICEWORD52 and_MPRICEWORD53 forty_MPRICEWORD54 dollars_MPRICEWORD55 multiword4] n1c] nbar1] altogether_RRDEGREE nbarq12] v2] vbar2] sd1] sprime2] sprpd23] start]

so the rule that the final word of a number sequence will be stressed predicts this stress correctly. However, the following example sees a different distribution of prominence:

1.000000 122 #SIL#
1.130000 122 The
1.710000 122 inflation

1.723003 121 **n**
2.350000 122 adjustment
2.640000 122 also
3.420000 122 means
3.427227 121 **n**
3.431922 120 **2**
3.990000 122 #SIL#
4.130000 122 that
4.210000 122 the
4.760000 122 maximum
4.769951 121 **n**
5.120000 122 annual
5.600000 122 level
5.710000 122 of
6.390000 122 earnings
6.399058 121 **n**
6.399058 120 **2**
6.660000 122 #SIL#
7.130000 122 subject
7.170000 122 #SIL#
7.270000 122 to
7.370000 122 the
7.810000 122 wage
7.821595 121 **n**
8.370000 122 tax
8.930000 122 #SIL#
9.100000 122 that
9.660000 122 generates
10.110000 122 revenue
10.117651 121 **n**
10.240000 122 for
10.310000 122 the
10.720000 122 Social
11.300000 122 Security
11.660000 122 trust
11.681030 121 **n**
12.100000 122 fund
12.122344 120 **2**
12.810000 122 #SIL#
13.020000 122 will
13.230000 122 rise
13.330000 122 to
13.370000 122 #SIL#
13.670000 122 fifty
13.678290 121 **n**
14.240000 122 thousand

14.255755 120 **2**
14.300000 122 #SIL#
14.450000 122 four
14.840000 122 hundred
14.856693 121 **n**
14.930000 122 #SIL#
15.310000 122 dollars
15.460000 122 in
15.950000 122 nineteen
16.480000 122 ninety
16.495189 121 **n**
16.517558 120 **2**
16.970000 122 #SIL#
17.130000 122 from
17.490000 122 forty
17.680000 122 eight
17.695960 121 **n**
18.220000 122 thousand
18.650000 122 dollars
18.663096 120 **2**
19.020000 122 #SIL#
19.290000 122 this
19.310983 121 **n**
19.573893 122 year
19.578588 120 **4**
20.460000 122 #SIL#

(a1593.04)

In this example, 'FIFTY thousand' contrasts with 'forty EIGHT thousand', so the contrasting elements are accented. The phrase 'thousand dollars' is given information in the contrast, so need not be foregrounded prosodically. Accentuation of the last word of these sequences would be very marked.

4.2 Break Labelling Distinctions

One problem in annotation encountered in building the corpus is that discrimination between break indices **2** and **3** can be difficult. The annotation system is absolute, so cannot 'compare' prominences or break lengths.

However, in the data analysed so far, certain patterns are emerging, such as the likelihood of a **2** break between the subject and predicate of a declarative sentencek[5]:

No new talks **2** are scheduled between the UAW **3** and Caterpillar **4**

[5]To facilitate reading the text is presented horizontally with only the break indices **2 3 4** marked

(b243.06)

If a noun phrase subject is postmodified, increased length and complexity increase the probability of a **3** break before the predicate, with the probability of another between the head noun phrase and the postmodification:

> The three o three X line of IBM's biggest computers **3** introduced in nineteen seventy seven **3** included the three o three one [...]

(b242.03)

> an eighty percent postal ballot **3** of the one hundred and fifty two thousand member workforce **3** voted nearly seven to one **2** to back his pruning proposals.

(b241.02)

The sentence above also illustrates, in its second break, a common prosodic pattern in verb phrases. If the phrase contains several arguments, breaks seem more likely, from data analysed so far. This may occur after the first argument, in this case 'nearly seven to one', especially as the speaker has accented both **seven** and **one**. Such heavy accenting effort usually leads to a break soon afterwards, and the end of the prepositional phrase provides a suitable point for one in that a whole syntactic unit has been completed.

Also typical of breaks noted in the corpus recordings we have annotated to date is the occurrence of **3** at clause boundaries, as in the following sentence:

> The new and unique approach **2** is not only capable of finding much smaller tumors **3** he said Thursday **3** but can use computers so effectively **2** that you can plan your operation **2** before you even start **4**

(b246.04)

Interpolated attributive sentences such as 'he said Thursday' are typically marked by such major breaks to distinguish them from the rest of the utterance they interrupt.

The examples of indices below also illustrate the role of breaks in coordinating phrases and clauses:

> For several decades **3** Newark Bay has been the dumping ground for mercury **2** zinc **2** cadmium **2** oil **2** various toxic wastes **3** and almost anything else you can think of **4** Mrs Weis said **4**

(b247.04)

> how many tickets would you want **3** and on what day **2** please **4**

(TRS33002.05)

Another emerging pattern is for a **2** break index before the head noun in a noun phrase consisting of a determiner, adjective (or numeric or sometimes descriptive noun) and head noun:

> the dismal results of ordinary **2** conventional treatment **2** for brain tumors [...]

(b246.02)

> I found **3** something hard **3** in this vegetable **2** sandwich **4**

(12040336ZF.04)

Then **3** for these two books **3** the posted rate **3** will be fourteen **2** dollars **4** (12040151ZF.24)

These breaks may serve to draw attention to a contrast with another noun phrase with the same head noun, or to emphasize a foregrounded aspect of the phrase (as in 'fourteen' above, whose action of specifying an amount of money is the focus of the sentence).

4.3 Break Labelling Definitions

Although we have suggested what are essentially heuristics for distinction between **2** and **3**, there is evidently considerable overlap in labelling patterns. This is because even a single speaker may produce utterances differently on different occasions. This is especially true in long and complex sentences, such as AP and WSJ reportage, the bulk of the nontravel material in the corpus. The analyst could only choose one of these for inclusion in the corpus.

The syntactic-semantic analysis of the corpus paid a great deal of attention to consistent application of ATR grammar rules and semantic tagging, with the assumption that a single parse is the single acceptable one (or at least the most appropriate use of clearly defined grammar rules) even for complex sentences[6] However, such categorisation is easier to achieve in a written corpus than a spoken. The amount of data in the corpus will produce vlid consistent patterns and a range of acceptable alternative productions and prosodic patterns.

Another problem with data classification is that the analyst may assign a **2** break in a longer sentence which may actually be a stronger break than a **3** in another. This arises from the relative significance of the breaks. Consider the following sentence:

> *While the Tigreans are communists, like the Eritreans they are among the most anti-Soviet guerrillas in the world, having suffered more than a decade of aerial bombardment by the Soviet-supplied Mengistu air force.*

(a1424.32)

Combining the .words and .breaks files produces the following indication of where breaks are noted in this sentence:

[6]Although tokenisation and sentence splitting rules need close definition for this exercise, when teh text of a sentence has been determined there is notionally a single acceptable parse for it.

While the Tigreans are communists **3** like the Eritreans **2** they are among the most anti-Soviet guerrillas in the world **3** having suffered more than a decade of aerial bombardment **3** by the Soviet-supplied Mengistu air force **4**

One could argue for including more than one of the range of possible realisations for this sentence. The **3** after 'communists' and the **2** after the next sentence-introductory phrase 'like the Eritreans' could plausibly be reversed. It is also noteworthy that the breaks notated by **3** in this sentence can be distinguished by their durations, indicating a distinction which the single notation may be insufficient to discriminate.

In a shorter sentence, such as are common in the dialogues in traveldata, comparatively major breaks may actually be shorter than 'lesser' breaks in longer sentences such as a1424.32 above:

And **3** could you describe your friend's symptoms for me **3** please **4**

(U22005.14)

Here, the breaks were of somewhat shorter duration than those in the previous example, as might be expected in conversational speech as opposed to complex sentences of news reporting, which are written to be read silently.

From examining sentences such as the above, we can model the distribution of breaks before sentence-initial and sentence-final elements ('And' and 'please' above), with a record of break duration and resource for determining the tone type. Although this does not provide an exhaustive record of break patterns characteristic of a native British English speaker, the data generated from the corpus will be useful for enhancing this aspect of CHATR's production, especially when it is combined with information about the physical attributes of the speech.

Traveldata files typically consist of a series of requests for information on single points, each followed by answers to that point, so the units in the exchange of information (the sentences) are rapid and focussed. By contrast, the written texts required more consideration before recording in order to produce a hesitation-free recording true to the precise words of the text (which is essential to maintain consistency between the written and spoken corpora). Sentences (and, indeed, words) in nontravel texts are considerably longer than in traveldata, and the sentences use more complex syntax to carry such discourse elements as logical arguments and comments on themes reported within clauses of the sentence.

Native speakers may produce such longer utterances by using a range of acceptable prosodies. The complexity of sentences in the nontravel element of the corpus[7] occasionally allows even for some syntactic ambiguity with respect to the writer's intention. The following example shows the problems for both speaker and syntactic analyst:

Mr. Harris may still want a boy to mow his lawn;

(I5.84)

[7]A random selection of texts from the substantial financial element of the corpus showed a mean sentence length of 24.14 text items

The sentence is ambiguous as to whether the infinitival clause *to mow his lawn* postmodifies the noun phrase *a boy*. It is more likely that it does, but the clause may be a separate argument of the verb. So the syntactic analyst could legitimately parse the verb phrase

> [vbar2 [o8 may_VMPRES o8] [vr1 [r2 still_RRTIME r2] [v2 want_VVIOPTATIVE [nbarq4 [nbar4 [d1 a_AT1 d1] [n1a boy_NN1PERSON n1a] nbar4] [i1b [t1 [vibar1 to_TO [v2 mow_VVIPHYS-ACT [nbar4 [d1 his_APP$ d1] [n1a lawn_NN1PLACE n1a] nbar4] v2] vibar1] nbarq4] v2] vr1] vbar2]

where the v2 indicates that the verb has one argument - that is, the postmodified noun phrase (nbarq4) 'a boy to mow his lawn'; or

> [vbar2 [o8 may_VMPRES o8] [vr1 [r2 still_RRTIME r2] [v4 want_VVIOPTATIVE [nbar4 [d1 a_AT1 d1] [n1a boy_NN1PERSON n1a] nbar4] [vibar1 to_TO [v2 mow_VVIPHYS-ACT [nbar4 [d1 his_APP$ d1] [n1a lawn_NN1PLACE n1a] nbar4] v2] vibar1] v4] vr1] vbar2]

where the infinitival clause (vibar1) is another argument of the verb phrase labelled v4[8].

If the infinitival clause is a separate element, the speaker may be liable to insert a longer break before it to represent the constituents' separation in the sentence structure, and to suppress any break before *a boy*:

> Mr. Harris – may still want – a boy to mow his lawn;

(emphasizing the unity of the postmodified noun phrase)

> Mr. Harris – may still want a boy – to mow his lawn;

(breaking to indicate that the noun phrase is separate from the following infinitival clause)

Knowledge of the context can determine which of these interpretations is correct, but as independent sentences both are valid potential realisations[9]. If the focus of the sentence was on the fact that Mr. Harris may want a boy, rather than a girl, to mow his lawn, **boy** would also have much greater prominence, but this is not yet marked in the corpus.

In some contexts, the specialized subject matter of texts requires some knowledge of the domain to interpret, whether parsing or speaking. Those drawn from the electronic Wall Street Journal (WSJ) are prominent examples:

> *$150 million of 9 1/4% subordinated notes due Nov. 1, 2001, priced at 99.821 to yield 9.275%.*

[8]In the ATR grammar, v2 is a verb followed by any of a range of syntactic structures; v4 is a verb followed by a noun phrase and then any of the same range of syntactic structures

[9]This sentence is taken from a file of such independent sentences giving examples of English usage

(a1161.16)

Here the speaker, or the syntactic analyst, has a range of options in breaking the sentence into its constituents:

> [$150 million of [[9 1/4% subordinated notes due Nov. 1, 2001] — priced at 99.821 to yield 9.275%]]
>
> [[[$150 million of 9 1/4% subordinated notes] — due Nov. 1, 2001] — priced at 99.821 to yield 9.275%]
>
> [[$150 million — of [9 1/4% subordinated notes due Nov. 1, 2001]] — priced at 99.821 to yield 9.275%]
>
> [[$150 million of [9 1/4% subordinated notes due Nov. 1, 2001]] — priced at 99.821 to yield 9.275%]

Only the first of these bracketings is technically correct given the subject matter. However, there are various points at which speakers could insert longer breaks. As the choice of prosodic analysis is made without reference to the parse, the break pattern could differ from that implied by the parse without use of particular subject matter expertise. Ideally, the speaker should be a subject matter expert in all subjects represented in the corpus. This shows that it is impossible to convert text to speech precisely without much more work being done on annotation of input for synthesis.

Another example of potential variations in prosody arises below:

> *The shop deals primarily with beginners*

(V13002.15)

The treebank parse is

> [start [sprpd23 [sprime1 [sd1 [nbar4 [d1 The_AT d1] [n1a shop_NN1SYSTEM n1a] nbar4] [vbar1 [v2 deals_VVZINTER-ACT [pr1 [rmod1 [r2 primarily_RRDEGREE r2] rmod1] [p1 with_IIWITH [nbar1 [n1a beginners_NN2PERSON n1a] nbar1] p1] pr1] v2] vbar1] sd1] sprime1] sprpd23] start]

The rule **pr1** labels

> *primarily with beginners*

as a single syntactic group (a prepositional phrase modified by a preceding adverb). The syntactic analyst did not take the option of separating the adverb from the prepositional phrase, another possible and syntactically acceptable option which would show the adverb limiting the scope of the whole verb phrase rather than the prepositional phrase only.

The pr1 **primarily with beginners** can also be realised as a prosodic group, but in fact it is separated into two separate prosodic units in the spoken corpus:

1.300000 122 ¡SIL¿ 1.360000 122 The 1.620000 122 shop 1.910000 122 deals 1.939697 120 **2** 1.980000 122 ¡SIL¿ 2.500000 12 primarily 2.524458 120 **2** 2.710000 122 with 3.300000 122 beginners 3.319591 120 **4** 3.760000 122 ¡SIL¿

The utterance is emphasizing that it is primarily with beginners that the shop deals, so the production apparently misrepresents the meaning, or at least the structure, of the utterance by breaking **primarily with**. It should instead be treated as a single prosodic foot. However, the interpretation the reader gave it is possible from the written text. Both performances of the sentence would be acceptable responses to the immediately preceding speaker's utterance

I'm just a beginner and I'm wondering if the shop can give me some advice

(V13002.14)

which shows that stressing both **primarily** and **with** is justified, and a **2** break after **primarily** is optional for normal production of the utterance.

The recording fails to represent the treebank parse, for which the syntactic analyst assumed or chose a slightly different emphasis. This illustrates the impossiblity of attempting to predict prosody from the text alone. To quantify the frequency of such mismatches in the corpus would be difficult.

Exposing the reader to the parse before recording each sentence may eradicate such anomalies. Making several recordings to cover all possible realisations of each sentence seems equally valid, but would constrain the amount of output and would not guarantee coverage of every possible interpretation and pattern of tones and phrase accents.

5 Conclusion

We have recorded, analysed and stored corpus texts totalling 50302 text items of traveldata and 57919 of nontravel, a total of 108221 text items. We cannot give a precise figure of the number of words this entails, as the 'text items' figure in corpus texts includes many numerical (monetary, date and time) expressions which expand to a larger number of words in the text files, and conversely a small proportion of unparsed sentences which we have not recorded. These numbers are small compared to corpora of written language but already constitute a large spoken and prosodically annotated corpus.

Appendix A: Corpus Structure

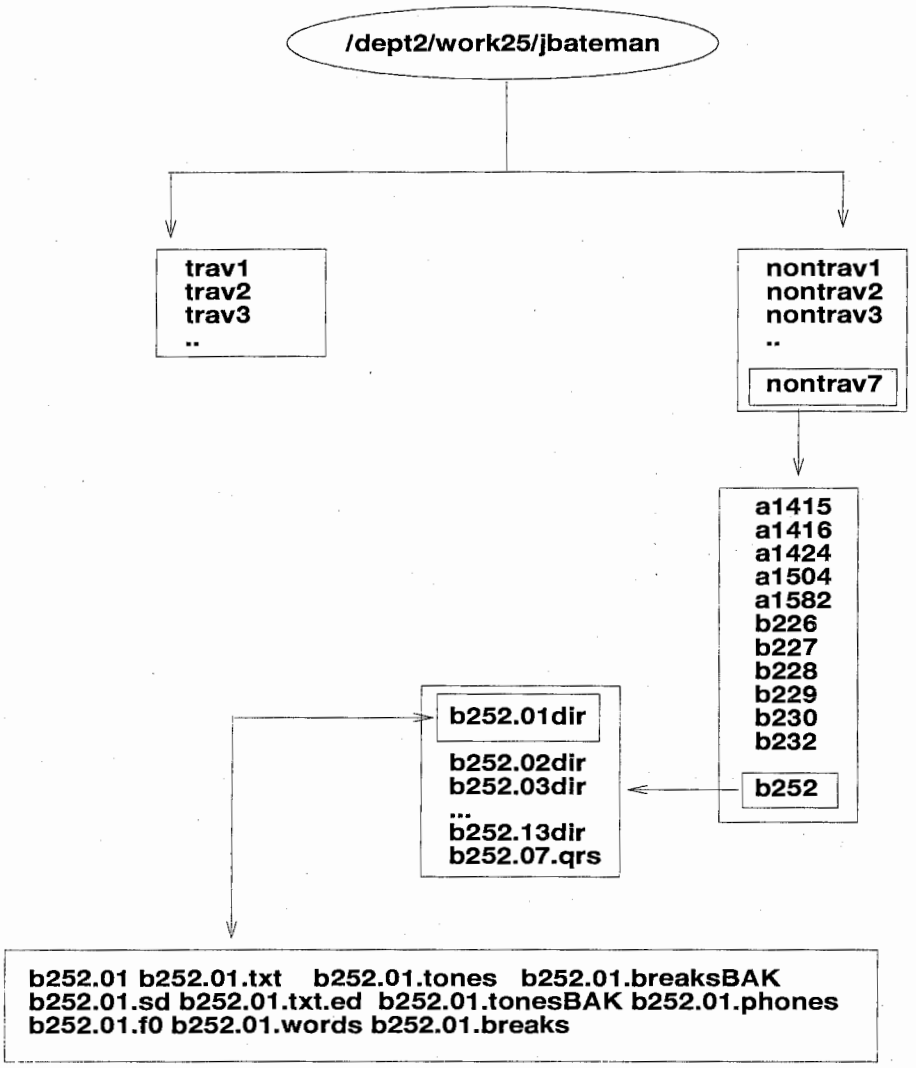

Figure 4: Structure of the Corpus

The numbered subdirectories (b252.01dir, b252.02dir...) and queries files (b252.07.qrs) directly under directory b252 correspond to the sentences as split for annotation in the written corpus.

If the analyst wishes to draw attention to any anomalies or interesting points in the data, a query file **.qrs** will also be found in the subdirectory for the sentence/utterance. If a comment was made for b252 sentence 01, the relevant file would be /b252/b252.01dir/b252.01.qrs.

Backup files are made automatically for .tones and .breaks files.

Appendix B: Locations

The written corpus files are stored at

/data1/atra08/itlusers/sbnlp/data

under the directories lanctb and lanctb[1-6]. /lanctb/orig and /lanctb/output contain the approximately 700,000 words of the first ATR-Lancaster contract, employing six analysts. The numbered subdirectories /lanctb[1-6]/OUT cover output from the subsequent contracts involving the three analysts whose output the supervising grammarian determined to be significantly most accurate and consistent.

The .orig files contain the text from which the spoken corpus' .orig files were copied. The .out files are the fully tagged and parsed texts.

(All Traveldata is in /lanctb[1-6])

The spoken corpus files are stored at

/dept2/work25/jbateman/

as shown above.

Appendix C: Textual Amendments

For direct comparison betwen a parsed file and its representation in the spoken corpus, it is important to remember that for most texts there is no one-to-one correspondence of ATR-defined 'text items' to 'words'.

The .txt files normally strip out punctuation. For Aligner to align text files to speech data, they also represent all figures as words.

Hyphenated numbers, such as **seventy-three**, are split, the hyphen is removed, and they are treated as two words, even though they are classed as a single text item in the written corpus and we use this 'text item' count for measuring the size of the spoken corpus. This enables us to distinguish more fully emphases and contrasts in utterances, e.g. in

> The inflation adjustment also means that the maximum annual level of earnings subject to the wage tax that generates revenue for the Social Security trust fund will rise to $50,400 in 1990 from 48{,}000$ this year.

(a1593.04)

the written 'forty-eight' is split to 'forty eight' so that the contrastive stress on EIGHT can be noted in the .tones file.

This can lead to dramatic differences, as in

> GPA is indeed talking about leasing Western planes to Aeroflot and even about buying Soviet-built Tupolev 204s

(a2151.12)

where the aircraft number has to be transcribed **two o four**, and needs to be reedited back to a single word if it is to be treated as an individual word. However, if a text on a similar theme is contrasting a **seven FOUR seven** with a **seven SEVEN seven**, there seems good reason for a tripartite split of the terms **747** and **777**.

Acronyms were usually represented as one word if they occurred frequently in the same formulation, e.g. in

> Both Shearson's Mr. Will and Stephen Reitman, European auto analyst at the London brokerage firm UBS-Phillips & Drew, recently switched their Jaguar recommendations to hold from buy.

(a2086.19)

Examples which occurred several or many times in the corpus include UBS, USA, and OTC (in 'OTC [over the counter] trading').

However, such formulations as **JAGRY** (a2086.44), a Stock Exchange symbol for a company, were realised as individual letters rather than as single unified verbal entities. JAGRY, for example, was split into five items in the .txt file.

Some unusual hyphenisations such as 'over-the-counter', 'most-actives' (both from a2086.13) and 'takeover-stock' (a2300.24) were also usually split, as the word groups from which they were formed could be given different prosodic patterns in speech.

The analyst has examined the earlier recordings and annotations, up to a1797 of directory nontrav3, to ensure consistency of tokenisation and labelling in the corpus.

Appendix D: Rerecordings of Spoken Texts

As explained above, the Traveldata texts of the corpus were transcripts of recorded scripted conversations. These texts included many hesitations and false starts, and other spoken-language characteristics (sometimes erroneously termed 'performance errors'), which were given a special tag and parsed using rules to indicate that they were 'random' interruptions. Initially, we attempted to reproduce these speech-performance characteristics in the recordings of the texts, but later they were omitted in order to maximise the number of utterances in the corpus which could be directly represented by well-formed English sentences. The sentences which were originally read to include these features have been replaced with readings from cleaned-up text and the original readings (with interruptions and hesitations) grouped in a separate subdirectory, trav.orig.

Appendix E: Duration of recordings in each directory

nontrav1 2093.18

nontrav2 2045.83

nontrav3 2350.87

nontrav4 2753.96

nontrav5 1404.46

nontrav6 4020.72

nontrav7 3527.94

nontrav8 1682.79

nontrav9 3100.76

nontrav10 1889.95

total nontrav = 24870.46 seconds

= 6 hours 54 minutes 30 seconds

trav1 4066.02

trav2 766.838

trav3 2336.61

trav4 2375.51

trav5 3382.28

trav6 3395.83

trav7 5.01244

total trav = 16328.10 seconds

= 4 hours 32 minutes 8 seconds

trav.orig 151.26

(traveldata with hesitations and restarts)

= 2 minutes 31 seconds

suki:

460 utterances 2395.12

= 39 minutes 55 seconds

Total duration of recorded utterances = 43744.94

= 12 hours 9 minutes 5 seconds

References

[1] J. Bateman, J. Forrest, T. Willis, The Use of Syntactic annotation tools: partial and full parsing, in R. Garside, G. Leech, A. McEnery, Corpus Annotation: Linguistic Information from Computer Text Corpora, Longman: Harlow 1997.

[2] E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, and D. Magerman, Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis,. In Proceedings of the 16th Annual Conference on Computational Linguistics, pages 107-112, Copenhagen, 1996.

[3] E. Black, The ATR General English Parser, ATR Technical Report TR-IT-0309, September 1999

[4] M. E. Beckman, G. M. Ayers, The ToBI Handbook, Technical Report, Ohio-State University, U.S.A. 1993.

[5] J. Hirschberg and M. E. Beckman, The ToBI Annotation Conventions, draft of 8 July 1993.

[6] M. E. Beckman and G. A. Elam, Guidelines for ToBI Labelling (version 3, March 1997), Ohio State University Research Foundation, http://www.ling.ohio-state.edu/phonetics/E_ToBI/.

[7] C. Wightman and D. Talkin, The Aligner: A system for automatic time alignment of English text and speech. Entropic Research Laboratory, Inc..

[8] ニックキャンベル、樋口宜男, マルチ言語マルチ話者音声合成システム CHATR. ATR Journal, 26, pp.8-9, Winter, 1997.

[9] W. N. Campbell and C. W. Wightman, Prosodic Encoding of Syntactic Structure for Speech Synthesis, in *Proceedings of International Conference on Spoken Language Processing*, Vol 2, pp167-70, 1992.

[10] W. N. Campbell, Prosodic Encoding of English Speech, in *Proceedings of International Conference on Spoken Language Processing*, Vol 1, pp663-666, 1992.

[11] W. N. Campbell, T. Hebert, E. Black, Parsers, Prominence, and Pauses, in Eurospeech '97: Proceedings of 5th European Conference on Speech communication and Technology, pages 979-982, Patras, 1997.

[12] S. Kitagawa, W. N. Campbell, Focus Detection by Comparison of Speech Waveforms, in Eurospeech '99: Proceedings of 6th European Conference on Speech communication and Technology, pages 1867-1870, Budapest, 1999.