*Gerry Knowles**

Annotating large speech corpora: building on the experience of Marsec

Abstract

This paper discusses a methodology for the processing of large amounts of speech data using database techniques and applying the lessons learned in the compilation of the Marsec database. The methodology is offered as an alternative to the conventional method of processing the orthographic transcription using only techniques designed for written texts. It is argued that while according to past practice it might appear that the first step in processing spoken texts is to make phonemic and prosodic transcriptions, these are not in reality necessary. Given the appropriate organisation of the data, much of the information in conventional transcriptions is predictable, and human expertise is required only to add unpredictable supplementary annotations.

1. Introduction

There are currently underway several initiatives - including the BNC and the ICE - which involve the compilation of large corpora containing a significant proportion of natural speech. Following past practice, the methodology for dealing with the spoken sections consists of first transcribing them orthographically, and then treating them in the same way as written texts. Most corpus linguists will probably agree that this is not the ideal way of handling speech data, but see no practical alternative, since it would be impossible to make phonetic or prosodic transcriptions of corpora running to millions of words.

This paper reviews the experience of developing that Machine Readable Spoken English Corpus ("MARSEC"), and identifies techniques that are already available and which can be used now in the processing of large amounts of speech data. It is argued that by using database techniques it is possible to process large amounts of data with the minimum of manual input and with an acceptable degree of accuracy.

* *Gerry Knowles*
Department of Linguistics & Modern English Language
Lancaster University
LA1 4YT, UK

Marsec is a relational database which was developed from the Lancaster/IBM Spoken English Corpus (Knowles, 1993). The original corpus consisted of different versions of the text including orthographic transcription, prosodic transcription and a grammatically tagged text, and a set of cassette recordings. The database contains digitised waveform files stored on CD-ROM and related tables derived from the different versions of the corpus texts.

2. Manipulating speech data

Before dealing with the technical problems themselves, it is necessary to consider two problems which essentially concern perception, and the way linguists tend to regard their data. The two major problems are (a) the nature of the primary data, and (b) the organisation and presentation of the data.

2.1. The nature of the primary data

A distinction needs to be made between *primary data* and *annotations*. The primary data is the data in its totally raw state, while an annotation is a label which interprets the data or identifies the different categories to which the data items belong. In practice it may be impossible to capture the primary data itself, and so some appropriate substitute is treated as though it were the primary data. The primary data of the present paper, for instance, consists of a computer file containing the orthographic text - words in standard spelling, spaces and punctuation marks - and formatting commands. For the reader it consists of a visual image. No great harm is done, however, if the orthographic text is treated as the primary data, irrespective of whether it is an interpretation of the visual image, or a subset of the computer file. This is because it is difficult to imagine a kind of linguistic research for which these differences would be relevant.

The problem arises as soon as we depart from modern standard printed texts. An ASCII version of a handwritten text is clearly not the same object as the original. Conventional edited versions of historical texts - where there may be conflicting originals, the text may be modernised, and the editor may have interpreted a manuscript - is even further from primary data. In this case the differences themselves are the legitimate object of linguistic research. It might be necessary, for instance, to use a specially prepared version of the text for standard processing, such as tagging and concordancing, alongside a computer-readable representation of the primary data.

In the case of spoken texts, the primary data consists of soundwaves, to which the nearest approximation which can be handled by computer is the digitised waveform. In the Marsec database, the waveforms are sampled at the rate of 16KHz, which means that there are 16000 samples or records for every second of speech. These digitised waveforms can be processed automatically, e.g. to identify pauses or to extract F0, or they can be annotated manually, e.g. with a phonemic, prosodic or orthographic transcription. When the digitised waveform is played back, the beginnings and ends of words marked in the orthographic transcription can be identified, and all the samples in between treated collectively as an instance of the word, e.g. samples 12345 to 12567 consitute an instance of the word *the*.

The orthography is self-evidently an annotation, and yet there is a general tendency among corpus linguists to treat it as the primary data. This is justifiable to a limited extent on practical grounds, first because it is possible to make and analyse orthographic transcriptions of large bodies of data, and second because the orthography enables many other kinds of useful annotations to be generated automatically. In this way the orthography captures a useful subset of the data. However, the orthographic annotation is ultimately not acceptable as a substitute for the primary data for several reasons:

* It is not fixed but subject to revision. Transcribers make mistakes and the text has to be amended as it is checked and errors come to light.

* Word forms are not necessarily fixed and may need to be represented differently for some purposes. For example, a number such as 509 can be represented in digits or in words *five-oh-nine* as opposed to *five hundred and nine*.

* Unskilled transcribers are likely to edit out performance errors made by the speaker.

* There are no fixed rules for punctuation, and different editors will vary widely in their practice, and some will simply be incompetent.

While there is no upper limit on the number of possible annotations, it simply does not make sense to have competing versions of the primary data.

Aspects of the primary data which are not represented in the orthography cannot subsequently be taken properly into account, and this necessarily distorts - and to an extent which it is impossible to assess - any investigations based upon it. This affects several areas including lexis, syntax, semantics and pragmatics. It is far from self-evident, for example, that the word *right* when it occurs in interactive discourse with a fall in pitch should count as a token of the same lexical item as *right* with a rise in pitch. Prosodic patterns are essential components of spoken relative clauses, cohesion in spoken texts, and of speech acts. If they are disregarded, spoken and written texts are compared only in respect of their common features, which necessarily gives a misleading view of their similarities and differences.

The lack of rigour in identifying the primary data leads to the treatment of spoken corpora as though they were written corpora with optional add-on sound. This leads to some bizarre methodology. In the development of Marsec, for instance, one of the most difficult tasks was to align the orthographic text with the waveforms, and to identify the points at which words and phonemes begin. This seems a perfectly sensible thing to do until one realises that this alignment was actually thrown away in the making of the orthographic transcription. This point is returned to below.

2.2. Data organisation and presentation

The orthographic transcription also governs the way in which the linguist subsequently thinks about the data. The spoken data exists in the form of sound waves, but as soon as the sound waves are transcribed, the text is assumed to be in *book format* (Knowles, in press) which means that it runs in linear fashion from left to right across the page and from top to bottom. Textfiles are not stored in this way on the computer, of course, but they are presented in this way when sent to the screen or the printer. Annotations - such as *grammatical_JJ tags_NNS* - may be forced to conform to this format.

But a text can also be organised in vertical format, with one word per line, and annotations arranged in columns, in the manner of a simple flatfile database. This is, for example, the format in which texts are presented for the CLAWS tagging programs. The length of the line varies according to the number of columns to be accommodated, and although this is less familiar than book format, it is actually much more convenient for handling linguistic data. If it is necessary to reproduce the text in the conventional form - e.g. to use in conjucntion with a conventional concordance package - it is easy enough to generate it. Furthermore if each line of data contains an identifier recording the original line number and position in the line, the original format can be recreated almost exactly. The point is that the organisation of data is independent of its presentation.

There are some kinds of information which will not fit easily into the flatfile format, and require the flexibility of a relational database. Much of the information relevant to the classification of the data is best held in the form of tables: tables of phonemes, tonetic stress marks, grammatical tags, lexical items, and so on. The advantage of a table is that information can be stored once as a property of the piece of data as a type, not of the individual token, which means that it does not have to be repeated each time the item occurs in the text. A table can also hold several pieces of information all of which are relevant to every token, and which could not practice be repeated each time in a flatfile or book format, e.g. a lexical entry might contain spelling, pronunciation and a range of syntactic and phonological information. The data should be stored according to the nature of the data itself, and not according to the requirements of some preconceived (and irrelevant) format. The relational database may be less familiar than book format, but it is the most appropriate structure to use for storing speech data (see further Knowles, in press).

2.3. Manipulating data

Database formats open up a variety of techniques for manipulating the data. To an extent this is achieved simply by freeing the analyst from the constraints of book format. A text in book format might be thought of as a sequence of words in a fixed order, but other orderings are possible, e.g. a text can be arranged in alphabetical order, according to grammatical tags, or according to morphological structure. It is also useful to examine a subset of words, which is easily done using a database filter.

When a text is in book format, it may also appear that the way to process it is to start at the beginning and work through to the end. However, there is no reason why it should not be processed in some other order or even randomly. Let us suppose the analyst is manually annotating a text. It would be possible to go through the text one word at a time, adding the annotation. This is clearly inefficient when the same instruction has to be repeated, e.g. if all occurrences of *the* have to be tagged as an article. It would be possible to automate the process with a find-and-replace command, so that *the* was replaced by *the_art*, but this sort of task is more easily performed in a database. It is much easier to make a change in a table or add information to it, retrieving the information for every token by using the normal facilities of the database management system. This is also true for the correction of errors. For example, in the development of the Marsec database errors occurred in the original look-up procedure to find the dictionary pronunciations of words. The pronunciation found for *job* actually belonged to the name of the Biblical character Job, so that *job* was transcribed /dʒəʊb̆/. The entry had to be corrected once in the lexicon, and then all tokens linked to it automatically retrieved the correct form. Another example is the word *accent* which only occurs as a verb in the corpus, but for which the original entry was for the noun; in this case /’æksnt/ had to be corrected to /æk’sent/.

The use of tables makes it possible to automate the process of annotation. This is trivial in the case of annotations which are totally predictable. Many typical linguistic annotations, however, can be predicted most of the time, but not with an accuracy of 100%. As a general principle there is no point in using human expertise to provide information which is totally or highly predictable. The human transcriber should provide information which is otherwise unreliable or unobtainable, for example by correcting or post-editing automatically generated annotations. This can take the form either of corrections, or of some other annotation that forces the calculation to work correctly. Suppose for instance we wish to investigate the phonological process of assimilation in a corpus. We could in theory make and store a phonetic transcription of the corpus with assimilations marked. Alternatively we can devise an algorithm to identify places where assimilation can occur, and keep a record of cases where it can take place but does not. The correct output can then be generated either by running the algorithm followed by the corrections, or by redesigning the algorithm to take account of exceptions.

Apart from enabling things to happen in practice - for nobody is really going to transcribe a corpus with assimilations - automatic annotations are an important means of saving storage space. If a value can be calculated it does not need to be stored permanently, and can be generated whenever it is needed or at most stored temporarily. Even if the output has to be corrected, a few lines of code and a small table can replace a very large file.

3. Automating transcription

A traditional phonemic or prosodic transcription carried out by a phonetician is a time consuming task which can only be carried out in reality on a small scale. Fortunately much of the output is predictable, which means that the task can be automated to a considerable extent. Instead of spending a high proportion of the time on low level work, the human expert can concentrate on problems that can only be solved by human intervention. This not only makes the work more interesting, but vastly increases the amount of data that can be usefully processed. Because it is necessary to store only the results of human work, and not the mass of predictable results of machine work, the output takes up relatively little storage space. In the discussion below I shall assume that the text is spoken in RP, but it would be a relatively trivial matter to revise the methodology to handle another variety of English.

3.1. Phonemic transcription

Phonemic transcription can be carried out simply by linking every word in the text to the appropriate entry in a pronouncing dictionary. In the development of Marsec words were looked up in a commercially available dictionary, and separate tables were compiled of the different orthographic forms and pronunciations. These have now been brought together in a phonological lexicon, and each entry contains information on its grammatical class and accentual behaviour (which is particularly important in the case of function words which have weak-forms).

One of the Marsec tables, called simply *Text*, contains a unique identifier for each word token (*“phon_id”*) and a reference number (*“lex_id”*) pointing to an entry in the lexicon. The file begins as follows:

Phon_id	Lex_id
10	6421
20	2357

The reference number 6421 recovers the lexical entry with the corresponding number, which is linked to several tables which together retrieve the information that it is the adjective *good* pronounced /gʊd/, and a word that is normally accented. Similarly 2357 retrieves the noun *morning* pronounced /mːnɪŋ/, which is also normally accented.

Given this kind of facility, there is no longer any point in expecting a phonetician to make a phonemic transcription of the words of a text. A new text has to be transcribed orthographically and grammatically tagged, and then the pronunciation of words can be found by matching the orthography and tag to entries in the lexicon. Any adjective spelt *good*, for example, will be identified as entry 6421, and this will give access to all the linked information, including the pronunciation. Some of the words of a new text can of course be expected not to be in the lexicon, but these only have to be added once. In practice transcription will be a two-pass operation, the first to update the lexicon, and the second to link the words of the text to lexicon entries. Note that this method gives systematic access to information which a human transcriber would not provide.

A problem could arise - although none has yet come to light in practice - in the case of words which have two or more pronunciations for the same spelling and tag, e.g. the adjective *lamentable,* which can be stressed on either of the first two syllables. Only one form will be retrieved arbitrarily from the dictionary. This will lower the accuracy rate slightly but it is not a major problem. All that is required is an additional lexical entry:

1234	lamentable	JJ	‘læməntəbl
9999	lamentable	JJ	lə’mentəbl

Tokens with stress on the second syllable will have to be located manually and the reference changed from 1234 to 9999. It is still much more economical to get a skilled phonetician to do this minor editing task than to do the whole transcription.

The lexical entries do not necessarily match entries in a conventional dictionary. Performance errors of various kinds will be identified in the text, and these must be included in the lexicon along with conventional items. *Ers* and *ums* can be included as words, as can word fragments and false starts. A mispronunciation can be treated in the same way as an alternative pronunciation.

3.2. Prosodic transcription

Some aspects of prosody are highly predictable. This includes accentuation and the grouping of words into tone groups or tone units. Work at Lancaster has achieved over 80% accuracy in these areas for the kind of text found in the SEC, and the probable upper limit of accuracy is probably over 90%. This means that much of the work can be done automatically, leaving the human expert to edit the predicted version. This involves supplementing the lexical information with the manual annotation of information which cannot otherwise be made available. For example, we would expect the phrase *the rector of the university*, when produced out of context, to be stressed *the RECtor of the uniVERsity*. In the context in which this phrase actually occurs in the corpus, *the university* has been mentioned before, and the speaker actually stresses the phrase *the RECtor of the university*. If *the university* is marked as ‘given’ information, the rules can be adapted to predict the correct pattern. Other things which will need to be marked include hesitation pauses, changes of tempo, contrastive stress, and rhetorical flourishes. These are the sort of things that the analyst wishes to discover in the prosody, and so they will be of interest for their own sake.

The unpredictable information can be kept in tables . Among the tables currently being developed for Marsec are the following:

(i) punctuation errors, where the punctuators of the orthographic transcript simply made mistakes, e.g. by putting a full stop before a time phrase where the prosody clearly indicates that the sentence ends immediately after it.

(ii) compound nouns, e.g. *seCUrity officer* which is accented only on the first element but which cannot automatically be distinguished from cases like *TOWN LIbrary* which is accented on both elements.

(iii) unexpected accentuation patterns including the failure of normal accentuation rules, such as accented prepositions or even *seCUrity OFficer* instead of the expected *seCUrity officer*.

(iv) given information.

(v) rhetorical flourishes, such as end-of-sentence cadences.

3.3. Contextual phonetic transcription

Given stress and tone groups, it is possible to predict the contextual form of words with some considerable accuracy. The dictionary forms discussed above do not of course represent the actual sounds of the text, and they are subject to processes of assimilation and elision, and words with weak-forms are subject to reduction. The contextual forms are derived by rule from the information in the lexical entries and from prosodic boundaries. In practice the predicted boundaries are likely to be insufficiently detailed. Consequently, processes which are generally blocked by a boundary will be found to apply in certain instances even though a boundary is marked, or to be blocked where there is no boundary indicated. Manual annotations will therefore take the form of additional boundary markers, indicating the failure of rules - thus ‘don’t assimilate / elide the final consonant / reduce the vowel here’ - or their unexpected operation.

In addition to the practical advantages of this approach, there is a logical gain. In the database, the dictionary pronunciation and the contextual pronunication are clearly distinguished as logically different annotations of the same speech events. When phoneticians make a transcription, they have to choose between them. In transcribing a phrase like *good morning*, should the transcription represent the individual words before the assimilation, i.e. /gʊd mɔːnɪŋ̆/, or what is actually heard as a result of it, i.e. /gʊb mɔːnɪŋ̆/? In principle they are different annotations and so both are required, but in practice a phonetician is unlikely to mark both. In the database there is no problem, as they stand side by side as parallel annotations.

4. Future work: transcribing spoken data

In compiling the original corpus a number of things were done which seemed perfectly sensible at the time, but which with the benefit of hindsight were perhaps not well advised. Two of the most important problems are:

* creating the orthographic version of the text
* defining the canonical version of the text

The simplest way to make an orthographic version of the text is for the transcriber to play the recording back on a cassette recorder and type the text into a wordprocessor. This is, however, to treat the recording and the transcription as unrelated objects. Later on it will be desirable to locate and replay the sound relating to any portion of the orthographic text, and for this it will be necessary to align them. It would be much better not to throw away the alignment in the first place.

This will require planning the organisation of the data before any transcription is carried out. The first step is to digitise the sound files and provide every point with a fixed time address. The second step is to find some kind of marker in the digitised file which can be represented in a database file containing the orthographic text. Possible candidate for this role are the beginning or end of pause.

Let us suppose that every pause in the corpus can be located automatically and the time of its end logged. By transcribing the data in the intervals between pauses it is possible to set up a time-aligned orthographic file. The beginning of the SEC, for example, would look something like this:

0000 Good morning.
1234 More news about the reverend Sun Myung Moon,
2345 founder of the Unification Church,
3456 who’s currently in jail for tax evasion.

In the case of interactive discourse it is possible to include information on the speaker or speakers, and to encode overlapping speech.

Once the orthographic version is completed, any variations and amendments to it must be very strictly controlled. It is necessary to decide on a canonical form, and record variants where appropriate and under specific conditions. For example *14th July* is a normal presentation of a date in orthographic texts, but for phonological purposes it might be better to write it out as *the fourteenth of July*. This information must be stored systematically so that either version is recoverable as required. Phonetic investigations may reveal further detail, for example a fragment of sound immediately after a pause which is ignored by the first transcriber may be identified as an instance of the word *and*. Errors of various kinds will come to light as work on the text progresses.

The difficulty is that an orthographic text very quickly acquires a life and legitimacy of its own, especially when it is passed on to other research groups and subjected to further processing. The only solution is to release occasional updates with new version numbers. Amendments have to be carried through systematically into all related files, so that if the orthography is changed, any necessary changes must be made to the grammatical tags and to references to the phonological lexicon e.g. if *Cosby* is amended to *Crosby* then the corresponding dictionary pronunciation must be amended to /krʊzbɪ/.

5. Conclusion

In this paper some doubt has been cast on the conventional practice of writing spoken texts down in conventional orthography before subjecting them to standard processing designed for written texts. Spoken texts need to be treated as spoken texts, and subjected to kinds of processing appropriate for spoken texts. For this purpose several bodies of data in addition to the orthography need to be assembled and related to each other in a systematic way.

Using the techniques traditionally employed by phoneticians to process texts - including phonemic and prosodic transcription - the task of processing large spoken corpora would be so enormous that it could not be carried out in practice. However, if the data is organised as outline in this paper, the vast bulk of this work would not need to be done at all, and the task of the phonetician would be to provide supplementary annotation. Looking ahead, it is even possible that once a methodology of this kind is established, it will be possible to extract some of this supplementary annotation automatically from the waveform itself.

References

Knowles, G. (1993): ‘From text to waveform: converting the Lancaster/IBM Spoken English Corpus into a speech database’ in C.Souter & E.Atwell, eds, *Corpus-based Computational Linguistics: Proceedings of the 12th ICAME conference*. Amsterdam: Rodopi; pp 47-58.

Knowles, G. (in press): ‘Recycling an old corpus: converting the SEC into the MARSEC database.’ in G.N.Leech, G.Myers & J.A.Thomas, eds *Spoken English on Computer: transcription and mark-u*p. London: Longman.