Internet Delivery of Time-Synchronised Multimedia
The SCOTS Project

Dr Wendy J Anderson and Mr Dave Beavan
Scottish Corpus of Texts and Speech Project
Department of English Language
University of Glasgow
scots@arts.gla.ac.uk

1 Introduction to the SCOTS Project

The Scottish Corpus of Texts and Speech (SCOTS) Project[1] at the University of Glasgow aims to deliver over the Internet (www.scottishcorpus.ac.uk) a 4 million word multimedia corpus of texts in the languages of Scotland, concentrating in the first instance on Scottish Standard English and varieties of Scots from 1940 onwards.[2] Twenty percent of the final corpus will comprise spoken language, made available as orthographic transcriptions accompanied by the source audio or video material. It is the aim of this paper to outline the approach taken by the SCOTS Project to the Internet delivery of these spoken texts.

The online SCOTS corpus was first launched on St Andrew's Day 2004, and regular additions are made to the corpus as new texts are acquired and processed, and as we are able to offer more sophisticated functionality. The current release of the corpus (Search System Version 2.1, Dataset 4, released in June 2005) holds a total of 527 documents, comprising close to 785,000 running words of text. SCOTS is not currently tagged for part-of-speech or any other linguistic level.

1.1 Text Sources and Corpus Design

The question of representation is always present in corpus design, and influences the SCOTS corpus at every stage. However, balance is no simple matter for a corpus which includes a non-standardised variety such as Scots. The range of text types available in Scots is much more restricted than that available for Scottish Standard English. Thus, while prose fiction, poetry and conversation in Scots are commonly found, journalistic texts and official writing are much rarer. In addition, with the exception of some of the language produced by educated enthusiasts, Scots tends to be used in more informal situations, a restriction not present in Scottish Standard English. It is appropriate in this context to take a more opportunistic approach to corpus building (see Rundell 1996), while continuing to monitor content and design throughout the project, and targeting specific genres to achieve balance. (See also Douglas 2003 for a discussion of issues of representation in the SCOTS corpus.) It is intended that in future versions of the corpus, users will be able to create their own sub-corpus from the total number of texts available, and analyse this in the same way as they would the complete collection, therefore designing to their own specification.

---

[1] The Scottish Corpus of Texts and Speech (SCOTS) Project is a venture by the Department of English Language and STELLA project at the University of Glasgow. The first stage of the project was grant-funded by the Engineering and Physical Sciences Research Council (EPSRC Grant no. GR/R32772/01): the current three-year stage is funded by the Arts and Humanities Research Council (AHRC Grant no. B/RE/AN9984/APN17387). The project website can be found at www.scottishcorpus.ac.uk.

[2] Scottish English is a localised variety of British English, characterised by Scottish features of lexis, grammar and pronunciation. Scots (or Broad Scots) is generally used as a cover term for a range of regional and social varieties such as the Doric and Ayrshire dialects, Lallans and the Glaswegian urban dialect.

1.2 Metadata

In addition to texts themselves, the SCOTS Project also gathers considerable quantities of metadata which is made available alongside each text. There are two categories of information: sociolinguistic details about the author of the text or the participants in an audio or video text (e.g. birthplace, occupation, decade of birth); and information about the context of production of the text, whether it is written or spoken (e.g. year produced, text type, audience, publication details where applicable, etc.). This information greatly enhances the value of the corpus for sociolinguistic researchers, lexicographers and general users. Especially with a minority language like Scots, for which there is no standard written form, likely factors behind variation, such as geographical factors, age of speaker, occupation, level of formality and spontaneity, may be suggested.

1.3 Delivery

Unlike many other corpus projects, SCOTS is freely available on the Internet, does not require registration or passwords, and the complete corpus may be searched and analysed without the need to download files or software (although a researcher can choose to download text files in order to use free-standing analysis tools). This ease of use and self-contained nature are essential for a project which seeks to involve the general community as well as academic researchers. It does, however, have implications for the corpus infrastructure, and this is nowhere more relevant than with multimedia texts. The remainder of this paper will consider the method adopted for making spoken texts available over the Internet in a form which is both user-friendly and a valuable resource for linguistic researchers.

2 Corpora and speech data

Spoken language, if not always the prime focus of a corpus, certainly forms an important subset of many corpora. Although not exclusively, most speech data is transcribed at least orthographically, and as such the examples in this paper will be based upon orthographic transcriptions. However, the principles and methods described here are just as applicable to transcriptions of phonetic form or gesture and facial expressions.

2.1 Traditional transcriptions

We are used to seeing orthographic transcriptions in a variety of formats, typically containing a number of speaker turns, with the speakers labelled by name or some other form of identifier. Below is a typical excerpt from a transcribed conversation:

```
M642: Is that high: I'm there on the bike. I am waving like this,
with one hand still on the throttle. Really r- low revs going over
the hump. I g- open the throttle again, and the bike dies on me.
M608: mm
M642: Right?
M608: mmhm
M642: Suddenly I'm
M642: //going this way//
M608: //oh no.//
```

Figure 1. Extract from a SCOTS transcription
(http://www.scottishcorpus.ac.uk/corpus/search/document.php?documentid=353)

Interruptions and overlap may be included in orthographic transcriptions (marked above by a double slash - //), with varying degrees of legibility for a non-specialist user depending on the conventions adopted. Additionally other events such as background noises or incidental events may be contained in the transcription.

While transcriptions are a very useful resource in their own right, a project such as SCOTS must bear in mind that many researchers will want to gain access to the original recording for a number of reasons. The ability to listen to passages identified by the transcriber as inaudible or open to interpretation may provide extra information to supplement the transcription. Finer analysis (of gesture mark-up, phonetic form, etc.) beyond the scope of orthographic transcriptions may be performed at a later stage. Greater contextual information can also be gained from studying intonation patterns and turn-taking in dialogue.

Unfortunately, in many cases access to the original recording is not available. Although such issues as legal permissions and speaker anonymity may lead to genuine restrictions, modern corpora should seek to address these issues where possible, and we would encourage corpus builders to aim towards opening up both the recording and the transcription to users. Historically, the technical requirements and physical storage of such recordings may have had a role to play in holding back such access, but with the shift to Internet delivered corpora, like the Scottish Corpus of Texts and Speech, these issues can be addressed with an appropriate technical approach. Indeed, with an appropriate methodology a rich resource can be made of transcribed speech recordings, by themselves or as part of a more general corpus.

It is worth noting that some projects choose to distribute their audio footage without any form of transcription; these are, of course, far less prevalent in the world of linguistic corpora. Transcription opens these resources up to alternative methods of access such as indexing and textual searching, as well as allowing computational processing and analysis. The issues discussed in this paper apply also if an existing footage-only data set is to be augmented by transcription.

2.2 Internet accessible corpora

Internet delivery has opened up the world of corpora, allowing for a much quicker and simpler distribution of corpus data. This greater accessibility has also extended the user base; no longer are corpora solely the tools of the linguistic community, they are valuable also in other disciplines such as history and social sciences. Depending on the aims of the corpus project, one particular group of users not to be ignored is the general public; their use of the corpus and indeed active contributions to it are fundamental for SCOTS.

Key to opening up such participation in a resource is ease of access to the data; this is something which must be addressed for casual users fully to appreciate the resource. Allowing playback of recordings is a big step forward in enabling more varied analysis and providing a greater feel for the data, something which can be appreciated by scholars and the wider public alike.

While the tools and methods used to search the SCOTS corpus are beyond the scope of this paper, the display and usability of a particular document once it has been found are of interest with respect to transcriptions. Where web corpora have chosen to distribute recordings, this is often in a very basic fashion. It cannot be denied that a link to a text file containing the transcription and a secondary link to a recording clip is a step in the right direction; however, a golden opportunity to add value and make the most of the resource is being overlooked. To do this we need to make an explicit link between the footage and the transcription.

3 Synchronisation

A link between two data sets can only successfully be made if it is across a common domain. It goes without saying that speaker-turns in a transcription occur successively later into the

passage. Similarly, a recording has defined start and end points which span time. Therefore, time is the common domain which can be used to tie together a recording and transcription. All that needs to be done is to synchronise these two pieces of data; simply put, to mark common points in time in both data sets.

Recordings are linear with respect to time, and the length of the recording is already known, therefore a point 25% into the recording file is 25% of the total time. Transcriptions are not so predictable; speakers do not talk at a constant rate of x words per minute, so we need to embed additional data, namely time information, into the transcription.

Our goal is to create the structure outlined below:

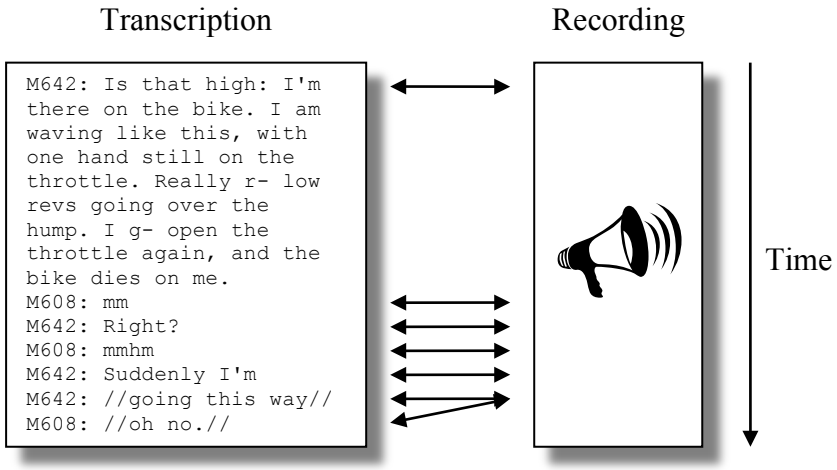

Figure 2. Alignment of transcription and recording

The example above features overlap between speakers M642 and M608, with the latter's final utterance occurring mid-way through the former's final utterance. Overlap is of course not unusual in speech but can be cumbersome to transcribe, particularly if there are many speakers involved. Therefore it is advantageous to separate the transcription for each individual speaker; this allows for greater flexibility of the output, such as enabling events or other data to be transcribed alongside speech. We then have a structure capable of handling any number of speakers, or tiers of other information (phonetic transcription, events etc.). Each transcription is time-aligned to the source footage, paving the way for enhanced access to the resource. Our model now looks like this:

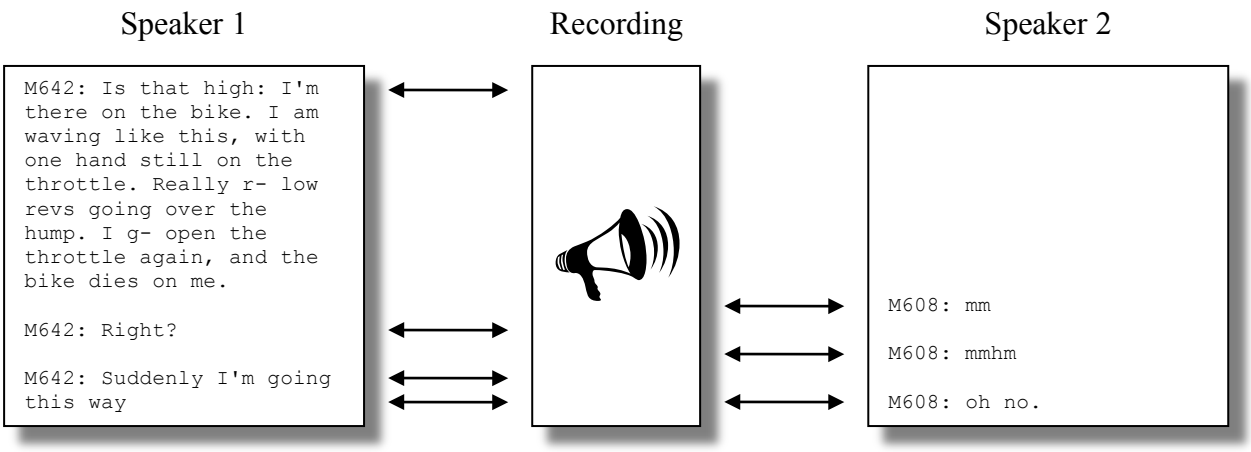

Figure 3. Alignment of transcription and recording – speakers separated

The main advantage of synchronised or time-aligned transcriptions is ease of access from transcription to recording and vice-versa. When viewing a transcription it becomes possible to jump to that particular section of the recording. It is also possible for this to work in reverse, to highlight the part of transcription a user is listening to, or to provide captions or annotations to the recording.

3.1 Transcription software

The task of creating data corresponding to the model above is not trivial, and realistically requires the use of software to automate the task. There are a number of software packages able to produce data in the correct structure, including (see References for web addresses):

- Praat
- TASX-Annotator
- Clan
- Multitool
- Anvil

Following trials of the above software we elected to use Praat for the Scottish Corpus of Texts and Speech. Praat is very user-friendly, has low system requirements, is an actively maintained Open Source project and has a large user base. One drawback, unlike TASX-Annotator or Multitool for example, is its lack of support for video. Therefore, in order to facilitate the transcription process, we strip off the audio from video footage and make the transcription as we would do audio recordings. While all the examples in this paper relate to Praat, the techniques should be applicable to other software packages offering similar functions.

All the software packages above provide an interface through which the transcriber or researcher can browse and play the media, often with associated wave form, along with a series of tiers, each associated with a speaker or other track of information. Tiers consist of boundaries, into which the transcription is typed or pasted. Each boundary has a defined start and end point, which are specified by time. The size of boundaries is flexible, and the granularity of the data is a free choice. While it is entirely possible to place a boundary around each word, given the nature of authentic speech, boundaries more or less at utterance level were found for our purposes to be more appropriate.

Using such software tools also speeds up the process of transcribing. It is very easy to search or navigate through the file to locate a particular passage or point in time. The ability to play back a chosen section in quick succession is extremely useful when transcribing rapid speech or audio which is difficult to hear. Where a transcription already exists, in say, a Word document, the process of moving to a time-aligned transcription can be tedious as it is often an exercise in copying and pasting. But especially when working from scratch, we have found the software-based approach takes less time and is less error-prone than transcribing into a text editor whilst listening back through the recording equipment.

Although the Scottish Corpus of Texts and Speech is not linguistically annotated, it includes minimal tagging; for speech data there is a small set of tags, expressed in XML:

- False starts
- Truncation
- Censorship (personal names etc.; this tag is also used in written documents where necessary)
- Unclear portions of speech
- Gaps in the recording or transcription (e.g. inaudible sections)
- Semi-lexical utterances e.g. “ah”
- Non-lexical utterances e.g. “laugh”
- Events e.g. “door bell rings”

Praat does not offer any native XML features and production of well-formed XML is therefore not guaranteed. It is possible to extend Praat to include, at the very least, a selection of pre-set tags which can be inserted into the transcription at a given point. XML mark-up is validated at a later processing stage, although it would be advantageous if this feature were incorporated into Praat.

The end result, below, is the waveform combined with the individual transcription of each speaker:

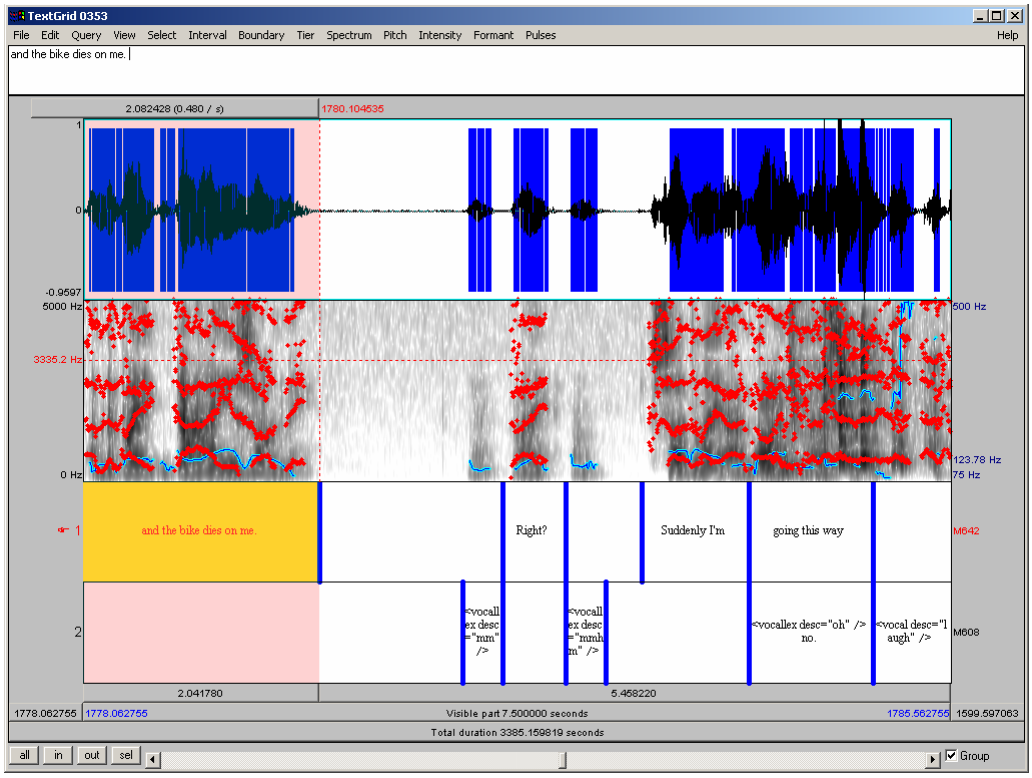

Figure 4. Praat screenshot

3.2 Web delivery of transcriptions

While we have achieved our goals in producing a synchronised transcription of the recording, what is lacking above is ease of use. Praat and similar tools are excellent for producing time alignment but they unfortunately cannot be readily used for the casual access which is so important to SCOTS. For web delivered corpora reliance on external tools should be kept to a minimum, and even when such tools are freely available this is an off-putting additional hurdle for many users. We should, therefore, aim towards delivering content to users in a way which embraces Internet and web standards. Typically this involves HTML pages (in our case XHTML) for the transcription. For the audio/video multimedia, we have adopted Apple QuickTime for its cross-platform support, although other options are readily available and may be just as successful.

Our goal is to produce an easy to read orthographic transcription tied to the media footage. This is where we hit our stumbling block; because the time alignment involved creating separate tiers for each speaker, it makes production of a sequential transcription challenging.

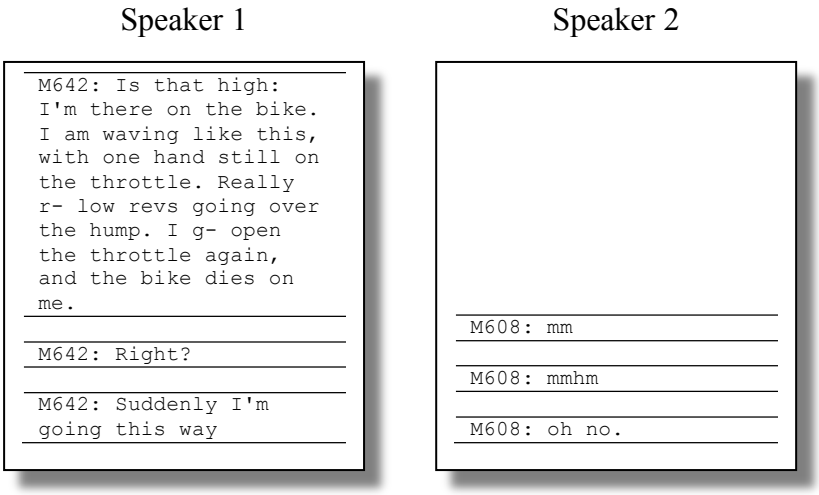

Figure 5. Boundaries, pre-subdivision

Our issue is with the overlap present at the end of this example, the very thing that multi-tier transcription makes it trivial to deal with. With boundaries placed at utterance level we know when each utterance occurs, and when the overlap begins, but what we do not know is at which exact point in terms of words the overlap begins and ends. It would be foolish to assume that because a second speaker starts talking half way through the first speaker's boundary that it is also half way through the number of words; this is not a reliable guide. This prevents us from directly creating an orthographic transcription with overlap marked in this way.

The word on which the overlap begins must be explicitly stated and we can do this by subdividing the overlapped utterance, and creating two boundaries from the original one:

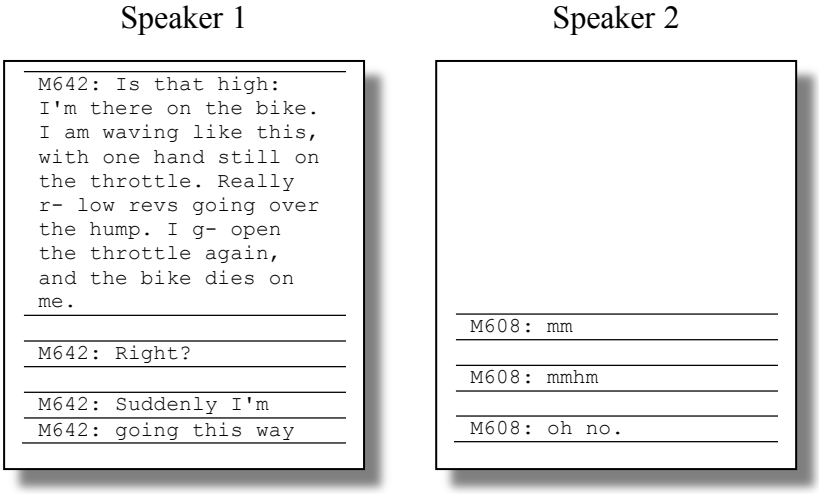

Figure 6. Boundaries, post-subdivision

Now we can ascertain exactly at which point the second speaker begins talking. In the example above, speaker M608 begins his final utterance just as speaker M642 reaches the

word “going” in his utterance. We now have all the information required; when speakers talk, and exactly when they overlap in terms of both time and words uttered. We can now go ahead and produce a sequential transcription, showing the individual speaker-turns in addition to marking overlap.

4 Generation of a user-friendly time-aligned orthographic transcription

The production of the textual transcription is automated by software created specifically for the Scottish Corpus of Texts and Speech (by D. Beavan). It validates the transcription, including the XML mark-up present, and the processed transcription is automatically entered into the corpus data structure. In parallel, the audio/video recording is also compressed for web delivery.

Logically, the process involves sequentially travelling through the parallel transcriptions by time, identifying overlap and producing output accordingly. Where possible speaker continuity is preserved; that is, if there is only one speaker for a number of boundaries, these are compressed into one line in the transcription, while preserving the boundary and time stamp information within. The text from each boundary is wrapped in tags specifying at which point in time it occurs. This is also true for overlapped content.

While it may appear we have now gone full circle, it is important to review the advantages to this workflow. The time taken to produce the transcription using this software is far less than by traditional methods. Additional tracks of information, such as expression or narrow phonetic analysis can be held side by side and optionally excluded from the output. Easily readable, sequential transcriptions with overlap and XML tags can be created in a straightforward manner.

4.1 Implementation in SCOTS

This process has been used for the Scottish Corpus of Texts and Speech for all our multimedia documents. A typical view of a transcription tailored for the project is shown below:

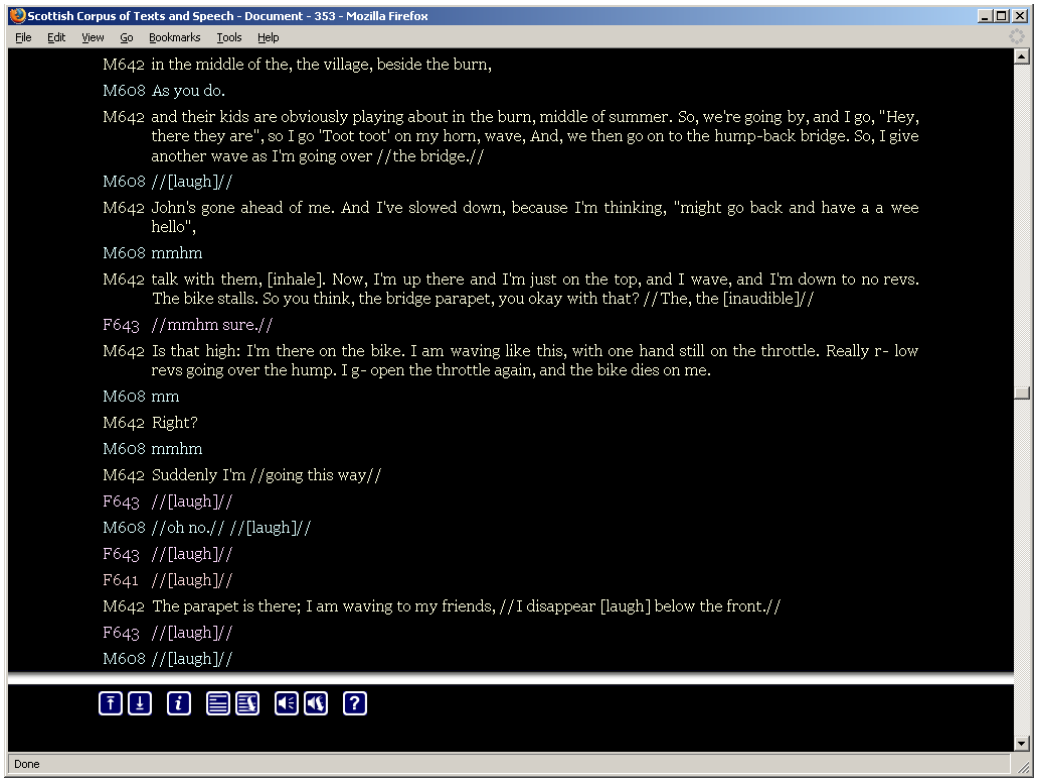

Figure 7. SCOTS screenshot - transcription

There are a number of features present in this view:

- Speakers are identified by number and optionally colour-coded
- False starts and truncation are present (marked with "-", e.g. "r-")
- Inaudible or unclear sections are marked in line
- Non-vocal utterances such as laughter are clearly presented
- Overlap when it occurs is indicated (marked by "//")
- Metadata covering the document and participants is accessible (from the Information button, third from the left along the bottom of the screen)
- Document (text and multimedia) download options are available (fifth and seventh from the left, respectively)
- Playback can be initiated (sixth button from the left along the bottom of the screen)

The benefits really come alive when the multimedia is played:

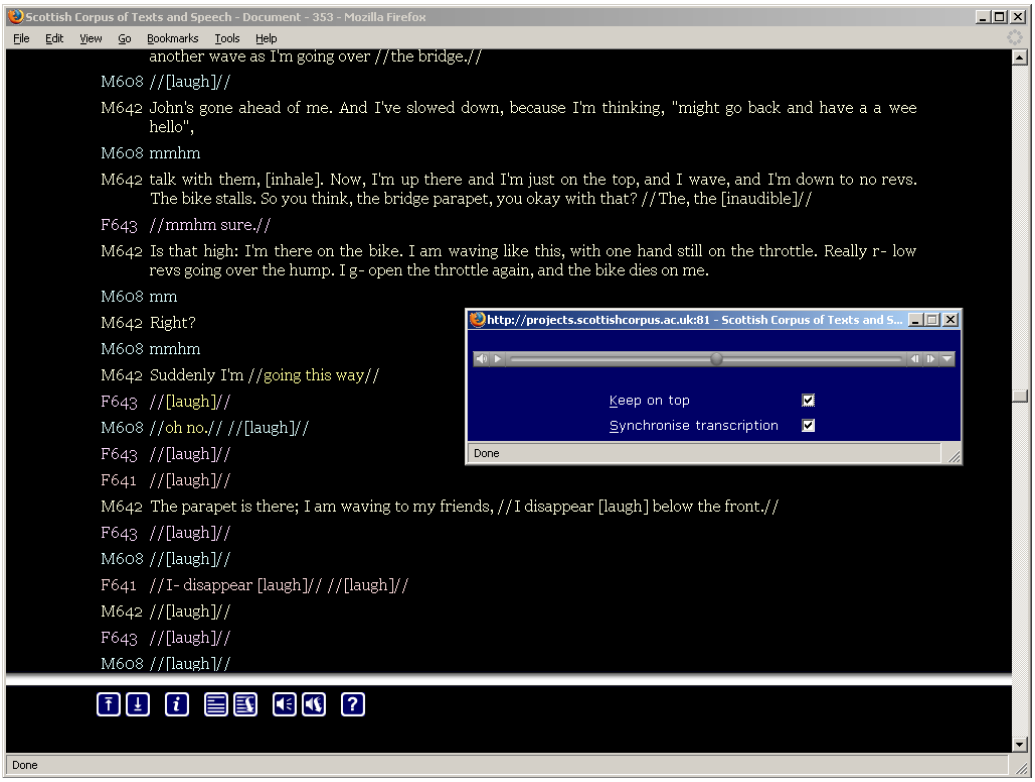

Figure 8. SCOTS screenshot – transcription and media window

The transcription can be viewed independently or synchronised to the playback, highlighting the passage the user is listening to in addition to scrolling the text. This can also work in reverse; any section of the transcription can be clicked on to reposition the (audio or video) playback accordingly.

5 Conclusion

The techniques described here will not be appropriate to all corpus projects, or even all corpora with a spoken component. However, where it is crucial to the success of a project that an audio resource can be easily accessed, such as by a non-specialist audience exploiting the corpus for a wide range of purposes, we believe that this procedure for synchronising audio and transcription data offers several benefits. The techniques used by SCOTS have been restricted by the requirement to deliver the corpus exclusively over the Internet; as such the choice of technologies has been limited. However, the choice of software and methods should make the implementation platform neutral and open the resource up to as many users as possible.

The different stages of the procedure have brought various advantages. For the transcriber, Praat has enabled a faster and more accurate result than more traditional methods. The choice of this particular software has also enabled separation of tiers of data (e.g. individual speakers' utterances, gesture) which makes the transcription process clearer and more intuitive. These can subsequently either remain separate (with, for example, speakers' words aligned vertically on screen), or be re-merged for a more traditional appearance.

The subsequent time-alignment allows the end-user to jump around freely in the recording with the transcription scrolling apace, and, vice versa, for the user to search for a word, locate

it in an audio transcription, and jump immediately to the corresponding place in the audio/video recording.

Perhaps most importantly for a project like SCOTS, the procedure lends itself well to the provision of a self-contained, integrated Web resource, where users are not required to download separate software, or battle with unfamiliar windows to carry out a straightforward search or analysis of written and spoken texts together. It remains therefore accessible both to a specialist researcher, and a member of the public with a casual interest in language.

5.1 The Next Step

This time-alignment may be further exploited to make additional improvements to usability, beyond the features which the Scottish Corpus of Texts and Speech currently offers. We intend in the next version of the corpus to provide an integrated concordance facility. Initially, this will simply return a KWIC concordance of the selected node word(s) in both written texts and transcriptions, and the results may be re-ordered according to the user's requirements. Subsequent versions of the concordancer, however, will be able to exploit more fully the time-synchronised audio material by offering hyperlinks to the immediate context of the word in the relevant audio or video file: an audio concordance. This will be of particular use to phonetics researchers, who will very quickly be able to compare variant pronunciations of an item even without a pre-prepared phonetic transcription.

It is only a short step to being able to add captions or annotations directly to the source recording interface, to providing an effect like subtitling. Owing to the underlying multi-tier approach, a single type of information (such as gesture mark-up) could be transferred to the audio window; alternatively, all tiers could move at once to reduce the number of active windows on screen.

Producing synchronised transcriptions is not only quicker and more accurate than traditional methods, it also adds value to the resource. This value can then be exploited in the many different forms of output generated by tying the source recording to the transcription. A user-friendly web interface can be achieved, allowing the resource to be used by a wide audience, general public included, with very few software prerequisites. This is ideal for the Scottish Corpus of Texts and Speech, and other projects which target a diverse user base.

References

Praat - Boersma, Paul and Weenink, David (2005). Praat: doing phonetics by computer (Version 4.3.12) [Computer program]. http://www.praat.org/

Anvil, video annotation research tool – http://www.dfki.de/~kipp/anvil/

Clan - http://childes.psy.cmu.edu/clan/

TASX-Annotator - http://medien.informatik.fh-fulda.de/tasxforce/TASX-annotator/

Multitool - http://www.ling.gu.se/projekt/tal/multitool/

Douglas, Fiona M. (2003) The Scottish Corpus of Texts and Speech: Problems of Corpus Design. *Literary and Linguistic Computing*, Vol. 18, No. 1, 23-37

Rundell, M. (1996) The corpus of the future, and the future of the corpus. Available online from http://web.archive.org/web/20020212093708/www.ruf.rice.edu/~barlow/futcrp.html (accessed 3 June 2005)