Sharing Digital Dialect Data with Silver Surfers and the Google Generation

*Alexander Docherty, alexander.docherty@ncl.ac.uk, 091313191, Q100 Linguistics

<table>
<thead>
<tr>
<th>Age Patterns</th>
<th>Variability</th>
<th>Gender Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Although there was variability, as described above, speakers of all generations expressed positive views about the Geordie dialect. Is Geordie a lazy variety?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The Silver Surfer female found this view offensive. • The Google Generation male expressed considerable pride in his dialect: “I’m quite proud of where I’m from”.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

- Standard format: a standard digitised format enables comparative studies and interoperability with other similar corpora.
- Sustainability: unlike the original analogue tapes and written transcripts, the digitised data has a longer life and is more easily updated.

The sociolinguistic analysis revealed interesting insights into the awareness of how Geordies speak today. While there was some variability, all participants were proud of their regional variety.

The project was designed also to allow access to a wider audience by broadcasting the findings over the Culture Lab Radio. The Podcast will be made available on the DECTE website (http://research.ncl.ac.uk/decte).

Reference


Transformation

XML-encoded, it is in a suitable format for transformation into XHTML, through a process illustrated by the following diagram.

| XML Document | XSLT Processor | XHTML Document |

Results:
- The transcript can be displayed in a dynamic web format.
- The data can be manipulated without requiring major changes to the underlying code.
- The content is standardised – this makes it compatible with other software and corpora and therefore suitable for comparative studies.

Using the Data: Sociolinguistic Interviews

Six men and women from Tyneside and one non-local (control) speaker were interviewed about their perceptions of the Geordie accent and dialect.

**Ages: Google Generation = 16-30, Middle Aged = 30-60, Silver Surfers = 60+**

Extracts were chosen from DECTE interviews and played to the informants to elicit their views in relation to three major themes:

- **Attributes** In general, informants viewed the dialect favourably.
- **Stereotypes** There was agreement that *howay man* is typical Geordie.
- **Variability**

Stereotypes

There was agreement that *howay man* is typical Geordie.

Unsurprisingly, Middle Aged females, rather than Google Generation informants, rated media figures such as Cheryl Cole and the Byker Grove cast most highly as stereotypical Geordies.

Conclusion

- Standard format: a standard digitised format enables comparative studies and interoperability with other similar corpora.
- Sustainability: unlike the original analogue tapes and written transcripts, the digitised data has a longer life and is more easily updated.

The sociolinguistic analysis revealed interesting insights into the awareness of how Geordies speak today. While there was some variability, all participants were proud of their regional variety.

The project was designed also to allow access to a wider audience by broadcasting the findings over the Culture Lab Radio. The Podcast will be made available on the DECTE website (http://research.ncl.ac.uk/decte).

Reference


**Time Alignment**

Time-alignment is part of the process of preparing plain text versions of the DECTE transcripts for transformation into XML.

This is achieved by inserting ‘$$’ tags into the text at 20-second intervals (shown below) so the text and audio can be aligned using a computer algorithm.

The process:
- Open the audio file in Audacity and zoom in to view 20 second intervals.
- Locate the point in the transcript that corresponds to this time interval.
- Input the temporary ‘$$’ time tag and repeat this for every 20 seconds of the file.
- Tagged files are then transformed so that ‘$$’ tags become XML anchor-tags, e.g. <anchor xml:id=”decten2y10i001ortho0020”> (the tag for the first 20 second point)

This is a labour-intensive task. My dissertation will investigate the methods for automating the process.

The Diachronic Electronic Corpus of Tyneside English (DECTE) is an extensive corpus of sociolinguistic interviews AHRC funded. It is being developed for language teaching in primary, secondary, further and higher education, as well as for use by the wider public.

In this project, I contributed to the digitisation of the DECTE interviews, and then used them in a sociolinguistic study of attitudes, stereotypes and variability.

I thus had two key aims:

- Time-align DECTE transcripts with their corresponding audio files
- Use the DECTE data in a linguistic study and present findings via a radio podcast.

The current stage of development involves making the DECTE material available online, as an attractive and interactive multimedia resource, accessible to users of all ages, from “Silver Surfers” to the “Google Generation”.

In my dissertation I will investigate the methods for aligning DECTE transcripts for linguistic study and corpora and therefore suitable for comparative studies.

Once the transcript is XML-encoded, it is in a suitable format for transformation into XHTML, through a process illustrated by the following diagram.

**Results:**

- The transcript can be displayed in a dynamic web format.
- The data can be manipulated without requiring major changes to the underlying code.
- The content is standardised – this makes it compatible with other software and corpora and therefore suitable for comparative studies.