Variability in Child Language

Nicola Pooley, Katie Alcock, Kate Cain (Psychology)
Andrew Hardie, Sebastian Hoffmann (Linguistics & English Language)
Paul Rayson (Computing)
Lancaster University

Background: The Child Language Survey

In the 1960s, the Nuffield Foundation funded the Child Language Survey (CLS), a project which gathered a vast collection of data on child language from the ages of about 8 to about 15. Consisting of transcripts of child language, both written and spoken, collected from a number of schools around the UK, this corpus was published in the late 1960s. Its extent has been estimated as a million words (of which 80% was spoken, 20% written).

While some university libraries possess copies of the transcript booklets, the CLS has long been unexploited, despite its potential value, because it is not in the digital format crucial to modern large-scale text analysis.

Pilot research programme

In this pilot project, we have:
• digitised a selection of the CLS data (both spoken and written, in the former case including audio recordings)
• created a comparable modern-day sample of data from the same or equivalent schools in London and Leeds
• investigated the use of these data in studying children’s linguistic variability

Our research will allow us to investigate the evidence in the data for the following three skills on the part of the children:
Planning the text: We will compare the coherence and cohesion of written and spoken language skills and to explore the interrelations between these skills. In addition, the developmental path of a range of written and spoken language skills and to explore the interrelations between these skills. In addition, the developmental path of a range of

Transcription examples: intermediate format (prior to automatic conversion to TEI-XML and CHAT standards)

1. Child writes “pogeket” but it is not clear what it should be normalised to (e.g. project or pocket). This is transcribed as: pogeket(){X}
2. Child writes “ther”) instead of “they’re” then “they’re”
3. It is not clear if the letter is ‘n’ or ‘y’ but the word in the script should be the contracted form of ‘they are’ then “they’re”
4. Script has no space between two words: playground and playground
5. Script has an extra space between two letters of a word. We transcribe this as: M_y{My}

Problems and solutions

In this pilot project funded by a Lancaster University small grant, we have carried out a proof of concept in order to estimate the effort required to digitise the whole corpus and creation of a new comparable corpus of modern data. We have digitised a small portion of the original 1960s transcription, previously only available in hardcopy. We have also digitised a number of the original audio recordings and transcribed one of these again in order to check the accuracy of the original transcription carried out in the 1960s.

We have employed Optical Character Recognition (OCR) technology to make the transcription of the 1960s hard copy material more efficient. Mosspelings and corrections have been preserved in the original material and this causes difficulties when automatically processing data in the OCR tool.

For digitisation of the audio recordings from the original half-track quarter-inch tape, we have employed digitisation specialists who are able to remove some of the distortion in the original recordings. Degradation of the recordings and the tape material itself is also a concern.

In the preparation of the 2008 corpus material, we needed to transcribe handwritten material, for example from 8 year old children. This has resulted in the need to develop new transcription guidelines for an intermediate format. There are three overriding principles observed here (a) preservation of the original (supplemented by retaining a full image scan) (b) efficiency of transcription and (c) consistency, in order to subsequently convert the transcription automatically to CHAT and XML/TEI formats.

We have used angled brackets to mark uncertain characters and square brackets to enclose a normalised equivalent.

Transcription examples:

5. Script has an extra space between two letters of a word. We transcribe this as: M_y{My}

Abstract

Our poster reports on the early stages of a large-scale inter-disciplinary project currently being carried out at Lancaster University. In the 1960s, the Nuffield Foundation funded the Child Language Survey (CLS), a project which gathered a large collection of written and spoken data on child language from the ages of about 8 to about 15. This data source has to date largely been unexploited, despite its potential value, because it is not in the digital format crucial to modern large-scale text analysis. We report on the digitisation of this data as well as the creation of a modern-day parallel corpus, which we have started compiling in cooperation with schools from the same areas – or in some cases, the same schools – as in the original data.

On the basis of some preliminary findings, the CLS and its modern counterpart provide a unique opportunity to compare the developmental path of a range of written and spoken language skills and to explore the interrelations between these skills. In addition, the availability of comparable data for school children from time periods that are almost two generations apart will of course also raise a number interesting questions relating to the educational policies that have been implemented in Britain over the past few decades and social change.

This research was supported by a Lancaster University small grant (2007). For more details see http://www.ling.lancs.ac.uk/activities/540/