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 narratives. These measures will also be analysed in relation to measures of fluency (text generation measures).

Generating the content: We will consider meaning-based dimensions of language (vocabulary and the development of ideas) and rule-based dimensions (sentence structure) within each modality. Complexity and diversity will be examined.

Transcribing ideas into written language: Spelling ability and writing conventions will be analysed. Children can also appear to be good or poor spellers by their choice of easy or hard words to spell in their written compositions, so spelling success will also be measurable in terms of written word frequency and length.

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 hardcopy material more efficient. Misspellings 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: 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[XX]

2. Child writes "there" instead of "they're"

there[they're]

3. It is not clear if the letter is 'o' or 'a' but the word in the script should be the contracted form of 'they are'

th<oa>re[they're]

4. Script has no space between two words:

playgroundand[playground and]

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

M_y[My]

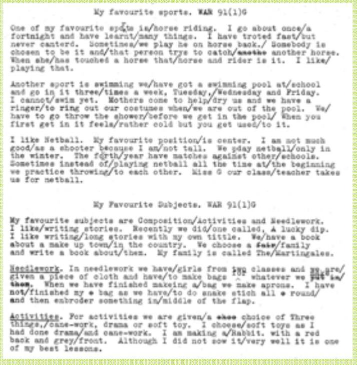
Abstract

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.

Digitisation of original transcription and audio recordings

1960s corpus



My favourite sports. WAR 91(1)G

One of my favourite sports is/horse riding. I go about once/a fortnight and have learnt/many things. I have troted fast/but never canterd. Sometimes/we play he on horse back./ Somebody is chosen to be it and/that person trys to catch/~~anothe~~ another horse. When she/has touched a horse that/horse and rider is it. I like/ playing that.

Another sport is swimming we/have got a swimming pool at/school and go in it three/times a week, Tuesday,/Wednesday and Friday. I cannot/swim yet. Mothers come to help/dry us and we have a ringer/to ring out our costumes when/we are out of the pool. We/ have to go throw the shower/before we get in the pool/ When you first get in it feels/rather cold but you get used/to it.

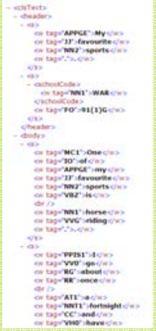
I like Netball. My favourite position/is center. I am not much good/as a shooter because I am/not tall. We pday netball/only in the winter. The forth/year have matches against other/schools. Sometimes instead of/playing netball all the time at/the beginning we practice throwing/to each other. Miss G our class/teacher takes us for netball.

My Favourite Subjects. WAR 91(1)G

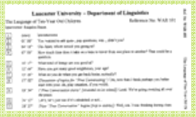
My favourite subjects are Composition/Activities and Needlework. I like/writing stories. Recently we did/one called, A lucky dip. I like writing/long stories with my own tittle. We/have a book about a make up town/in the country. We choose a ~~fake~~/family and write a book about/them. My family is called The/Martingales.

Needlework. In needlework we have/girls from two classes and we are/ given a piece of cloth and have/to make bags for whatever we put in/ them. When we have finished makeing a/bag we make aprons. I have not/finished my bag as we have/to do snake stich all round/ and then embroder something in/middle of the flap.

Activities. For activities we are given/a ~~choo~~ choice of Three things,/cane-work, drama or soft toy. I choose/soft toys as I had done drama/and cane-work. I am making a/Rabbit. with a red back and grey/front. Although I did not sow it/very well it is one of my best lessons.



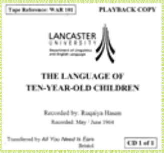
```
- <clsText>
  - <header>
    - <s>
        <w tag="APPGE">My</w>
        <w tag="JJ">favourite</w>
        <w tag="NN2">sports</w>
        <w tag=".">.</w>
      </s>
    - <s>
      - <schoolCode>
          <w tag="NN1">WAR</w>
        </schoolCode>
        <w tag="FO">91(1)G</w>
      </s>
    </header>
  - <body>
    - <s>
        <w tag="MC1">One</w>
        <w tag="IO">of</w>
        <w tag="APPGE">my</w>
        <w tag="JJ">favourite</w>
        <w tag="NN2">sports</w>
        <w tag="VBZ">is</w>
        <br />
        <w tag="NN1">horse</w>
        <w tag="VVG">riding</w>
        <w tag=".">.</w>
      </s>
    - <s>
        <w tag="PPIS1">I</w>
        <w tag="VV0">go</w>
        <w tag="RG">about</w>
        <w tag="RR">once</w>
        <br />
        <w tag="AT1">a</w>
        <w tag="NNT1">fortnight</w>
        <w tag="CC">and</w>
        <w tag="VH0">have</w>
```



Lancaster University – Department of Linguistics

The Language of Ten-Year Old Children

Reference No: WAR 101



Tape Reference: WAR 101

PLAYBACK COPY

LANCASTER UNIVERSITY

Department of Linguistics and English Language

THE LANGUAGE OF TEN-YEAR-OLD CHILDREN

Recorded by: Ruqaiya Hasan

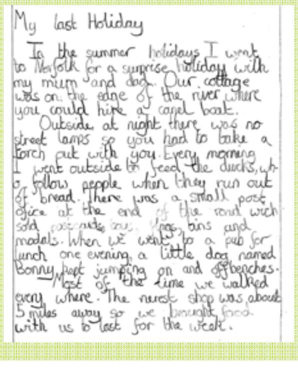
Recorded: May / June 1964

Transferred by All You Need is Ears Bristol

CD 1 of 1

Transcription of new written and spoken material

2008 corpus

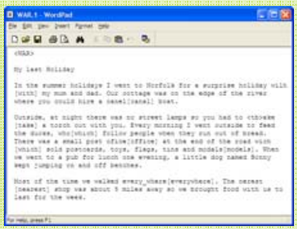


My last Holiday

In the summer holidays I went to Norfolk for a surprise holiday with my mum and dad. Our cottage was on the edge of the river where you could hire a canel boat.

Outside at night there was no street lamps so you had to take a torch out with you. Every morning I went outside to feed the ducks, who follow people when they run out of bread. There was a small post ofice at the end of the road wich sold postcards, toys, flags, tins and modals. When we went to a pub for lunch one evening a little dog named Bonny kept jumping on and off benches.

Most of the time we walked every where. The nerest shop was about 5 miles away so we brought food with us to last for the week.



<WAR>

My last Holiday

In the summer holidays I went to Norfolk for a surprise holiday with [with] my mum and dad. Our cottage was on the edge of the river where you could hire a canel[canal] boat.

Outside, at night there was no street lamps so you had to <th>ake [take] a torch out with you. Every morning I went outside to feed the ducks, who[which] follow people when they run out of bread. There was a small post ofice[office] at the end of the road wich [which] sold postcards, toys, flags, tins and modals[models]. When we went to a pub for lunch one evening, a little dog named Bonny kept jumping on and off benches.

Most of the time we walked every_where[everywhere]. The nerest [nearest] shop was about 5 miles away so we brought food with us to last for the week.

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