Oxford Children's Corpus: a Corpus of Children's Writing, Reading, and Education

Nilanjana Banerji
Oxford University Press, Education Division
nilanjana.banerji@oup.com

Vineeta Gupta
Oxford University Press, Education Division
vineeta.gupta@oup.com

Adam Kilgarriff
Lexical Computing Ltd.
adam@lexmasterclass.com

David Tugwell
Lexical Computing Ltd.
dtugwell@gmail.com

1 Introduction

The Oxford Children's Corpus (OCC), as it was in 2011, is described in Wild et al (2011, 2012). This was a corpus of writing for children. Since then OUP has developed a 'children's writing' component of the corpus, primarily with data from the BBC Radio 2 '500 Words' short story writing competition. This is a competition that runs in the spring every year with children aged 4-13 submitting entries up to 500 words long, with winners announced at the Hay Literary Festival. All shortlisted items can be read online.[1]

Lexical Computing Ltd is working with the Children's Dictionary and Language team at Oxford University Press to analyse the language that the children use. The 74,000 entries received in 2012 (called Beebox below) form a large part of OCC-W, the Children's Writing component of the OCC. The OCC as it was when last reported on forms the hub of the Reading component (OCC-R) and we have also gathered curriculum materials to form the Education component (OCC-E).

Here we focus on Beebox, describing the data and presenting some first results from the analysis of the 2012 data. In April this will be joined by the 2013 data, and any conference presentation in July 2013 will talk about the new data too.

2 The Beebox data

There are a total of 73,875 stories, with distribution by age and gender as in Figure 1.

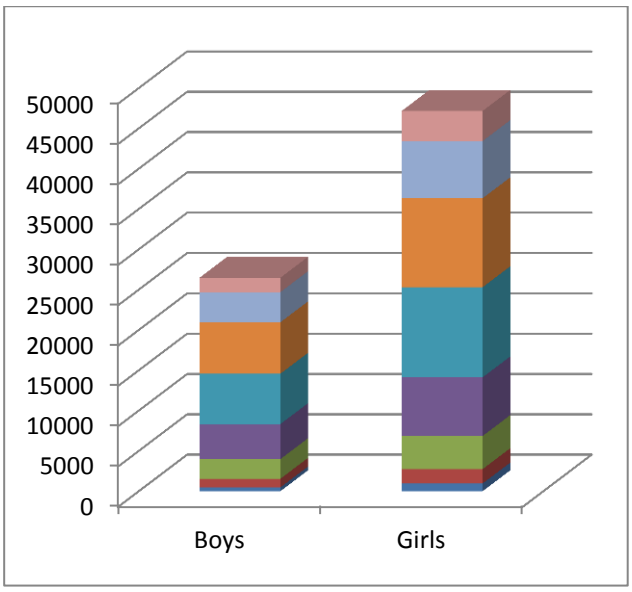

Figure 1. Age bands are, from bottom: up to 6, 7, 8, 9, 10, 11, 12, 13.

Most stories are close to 500 words. The total corpus size is 32.4 million words. From a statistical point of view this is a dream: very large numbers of same-size samples.

We also have the BBC region (in most cases, this is the same as the county) for each story. There are 54 of these regions, and for all but two, there are over 100 stories. For 37 of 54 regions there are over 1000 stories.

The stories have been delivered online, with no editing or correction by the BBC or OUP, so are complete with grammar, spellings and punctuation as provided. There are 48,000 hits for *friend* – and 311 for *freind.*

The data has all been lemmatised and part-of-speech-tagged, and then loaded into the Sketch Engine (Kilgarriff et al 2004).

3 Analyses

We have looked at contrasts with writing for children (OCC-R) and variation by age, gender and region.

3.1 Contrast with writing *for* children

We looked at the 200 keywords of Beebox in contrast to the 9 million words of 21st century fiction written for children that we had within OCC-R. These were examined by one of the authors and classified. [2] At the most general level, the classification was between writing problems, and

[1] http://www.bbc.co.uk/radio2/500words/2012/

[2] The keyword lists included only lowercase lemmas of at least three characters, with a simplemaths parameter of 100: for details of the statistic and method see Kilgarriff (2012).

themes. The writing problems included uncapitalised names, missing apostrophes (*cant, wont)*, hyphens (*hearted, haired, headed)* and inter-word spaces (*anymore, aswell, infront*) as well as spellings (*whent, thay, solder* for *soldier, minuet* for *minute, cheater* for *cheetah*).

More interesting were the themes that children wrote about notably more than adults writing for children:

- **Scary stories**
  - *creepy creaky croaky dreaded foggy ghost gloomy graveyard haunted mansion misty mysterious petrify scared scary spooky undead vampire zombie*
- **Traditional**
  - *pixie elf genie goblin leprechaun gnome*
  - *prank potion robber*
- **People**
  - *mum mummy mom dad daddy auntie grandpa grandma*
- **Space/war**
  - *alien asteroid astronaut galaxy portal rocket spaceship teleporter*
  - *ammo ninja sniper spaceship teleportal*
  - *airport*
- **Animals**
  - *cheetah dolphin hippo kitten ladybird panda penguin squirrel zebra*
  - *unicorn*
  - *bunny teddy*
  - *woof meow tweet* (what birds do)
  - *vet zoo*
- **Food**
  - *candy cupcake coke marshmallow*
- **Jewels**
  - *diamond emerald gem locket necklace*
- **Other nouns**
  - *clown diary bully snowman*
  - *gymnastics karate sleepover medal*
  - *foster orphanage*

These (but for the scary ones) were largely nouns. There were also:

- **Adjectives**
  - *adorable adventurous bouncy comfy fluffy ginormous horrific horrifying humongous magical sparkly stormy super wrinkly yummy*
- **Adverbs**
  - *extremely happily luckily speedily unfortunately worriedly*
- **Verbs**
  - *cuddle investigate sprint stroll stutter unpack wake*
- **Other:**
  - *(*ding*) dong phew*
  - *bye okay soo*

3.2 Gender

The gender analysis is somewhat painful.

Girls in contrast to boys

- **Romance**
  - *blush boyfriend cheek cuddle darling hug kiss snuggle sweetheart sweetie wedding xxx*
- **Horses**
  - *canter chestnut groom mane neigh pony riding stable unicorn*
- **Nature**
  - *butterfly cherry daisy flower kitten lilac lily petal poppy rainbow rose*
- **Dance**
  - *ballet chorus dance*
- **Adjectives**
  - *adorable beautiful cute dainty delicate flowery fluffy glittery gorgeous hazel pink silky sparkly rosy*
- **Traditional**
  - *diary fairy locket maid pixie mermaid*
- **Hard stuff**
  - *cancer comfort cope fault foster upset*
- **Textures/clothes**
  - *cardigan stroke (v) velvet ribbon silk silky skirt*
- **People**
  - *daddy daughter lady princess sibling sister twin*
- **Food**
  - *candyfloss bun(1)*
- **Hair and beauty**
  - *bun(2) glossy wavy blonde curly plait makeup necklace*
- **Pronouns**
  - *her hers herself she*
- **Other**
  - *doll giggle girl girlie pink soo sparkle sprinkle teddy skip sleepover shyly*

Boys in contrast to girls:

- **Fighting**
  - *aim ambush ammo armed armor armored army arrow assassin assassinate assault attack base battle blast bullet bunker cannon captain chopper civilian cockpit combat commander defend defender defense destroy device engine explosion explosive fighter fireball fuel general grenade guard gun gunfire helicopter helmet himself knight laser launch launcher leader league machine military missile mission nuclear opponent parachute patrol pilot pistol radar rifle robot scout sergeant shield shoot shot shotgun smash sniper soldier spear survivor sword system tank target teleport temple terrorist troop warrior weapon*
- **Sport**
  - *football footballer goal goalkeeper penalty player ref referee score squad stadium striker tackle team training*
- **Other**
  - *himself galaxy teleport zombi*e

3.3 Age

We divided the authors into three bands: up to eight, nines and tens, and 11+, and found the keywords of each age group in contrast to the remainder.

Up to 8

- **Fairy stories**
  - *once upon magic end happily castle fairy adventure magical king princess spell wand queen palace*
- **Other adjectives**
  - *naughty sunny sad sparkly lovely excited shiny friendly*
- **Food**
  - *cake party chocolate eat yummy tea*
- **Pirates**
  - *pirate cave dragon treasure*
- **Other**
  - *dinosaur swim play pet lot*

9 and 10

- **Reporting verbs**
  - *mumble moan yell stammer shout agree exclaim sneak boom*
- **-ly adverbs**
  - *suddenly excitedly sadly loudly angrily extremely luckily*
- **Scary adjectives**
  - *dusty gloomy spooky*
- **Other adjectives**
  - *gigantic exciting famous ugly colossal annoying enormous cute bore sunny super lovely brilliant*
- **Nouns**
  - *alien cage robot rope potion lightning ginger breakfast adventure mansion lady mum hamster sword ship portal*
- **Other**
  - *meanwhile later bye once hello yes zoom*

11+

- **Body parts**
  - *blood body cheek eye face fear hand heart shoulder throat*
- **Body/mind functions**
  - *breath feeling memory mind pain smile sweat tear thought*
- **Abstract nouns**
  - *darkness death echo force life murder silence soul word*
- **Atmospherics**
  - *alone cold dead pale silent slowly wind*
- C**onnectives**
  - *against almost since though within yet*
- **Verbs**
  - *die feel glance lie fill seem sense stand stare*
- **Romance**
  - *figure woman*
- **Pronouns**
  - *myself nothing*

The steps from childhood towards adolescence are vividly shown. The 11+ keywords (deeply indebted, we suspect, to the Twilight novels) scarcely need commentary, so loud do they sing of teenage concerns. The two pronouns which have made it into the list – *myself, nothing* – sum up all by themselves the agony of being a teenager.

Less obvious, and more intriguing, are the clusters of reporting verbs and –ly adverbs that the nines and tens use, and the adjectives, in the younger two age groups, switching to connecting words amongst the 11+s. They may relate to the National Curriculum, and story-telling techniques that children are taught at particular stages.

3.4 Region

The top keyword for Birmingham-and-the-Black-Country is *mom*. The top keyword for Tyne-and-Wear is *mam.* Children tend to write as they speak, and in the northeast the usual short name for a mother rhymes with 'Sam' and around Birmingham it rhymes with 'Tom'. For the rest of us it rhymes with 'plum'. The corpus is closer to a spoken data collection than most written corpora would be.

At a level of themes, the keywords for Norfolk have seven animals in the top twelve; the top three keywords for Wales are *sheep, bus, dragon;* for Scotland, *beside, wee, gran.*

Our explorations in this area are very preliminary, but we suspect the corpus offers a great deal to dialectologists.

References

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, David Tugwell 2004. *The Sketch Engine.* Proceedings of Euralex, Lorient, France.

Adam Kilgarriff 2012. *Getting to know your corpus.* in: *Proc. Text, Speech, Dialogue (TSD 2012)*, Lecture Notes in Computer Science. Sojka, P., Horak, A., Kopecek, I., Pala, K. (eds). Springer.

Kate Wild, Adam Kilgarriff, David Tugwell 2011. Oxford Children's Corpus: A corpus of writing for children. Poster at ICLIC (*International Corpus Linguistics Conference*), Birmingham, UK.

Kate Wild, Adam Kilgarriff, David Tugwell 2012. The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. International Journal of Lexicography, doi: 10.1093/ijl/ecs017. Oxford University Press.