Sentiment analysis
of OUP's Children's Corpus

123,436 stories. 54 million words. A unique window into the inner world of children and an opportunity to discover what they care about and what makes them happy, sad or anxious.

Earlier this year, children from all over the UK submitted stories for BBC Radio's '2,500 Words Competition'. In association with the Oxford University Press (OUP), department spin-out TheySay was invited to analyse the children's writing in order to provide valuable insights into the sentiment and emotions detected in the stories for the first time in the competition's history.

Using advanced computational linguistics and machine learning techniques TheySay was able to unearth fascinating information on the emotional signals detected in the stories and highlight how these change in different age groups and locations. The text from all submitted stories was analysed and data was collected around the positive, neutral, and negative sentiment as well as the emotional content of every story. TheySay also determined what entities, ideas, or opinions appeared most frequently in a positive or negative context in the entire set of submissions.

Overall, the stories submitted were complex tales that contained both negative and positive sentiment, with happiness and fear being the most common emotions. There was a significant drop of average positive sentiment with age. In fact, a 20% drop of average positive sentiment was detected from the youngest age group to the oldest one, showing that older children submitted stories that on average were darker, more complex, and multi-layered.

> **'There were more mentions of scary or unpleasant aunts in stories from Northern Ireland than any other region; aunts in other parts of the UK were presented as mostly harmless.'**

Happiness peaked in stories submitted by 7-year-old children, with a noticeable drop after that. The detected levels of fear and anger rose in stories submitted by children in the older age groups, perhaps a result of teenage angst.

There was also a small difference between the sentiment levels in stories submitted by girls and those submitted by boys. On average, girls' stories contained slightly higher levels of positive and neutral sentiment than those written by boys. Similarly, there was variation observed in the levels of related emotions: boys' stories expressed more fear and anger while girls' stories had higher levels of happiness and surprise.

Perhaps surprisingly, the words 'school' and 'teacher' were among those used in a positive context most frequently. Schools were often mentioned in association with happiness and excitement. The words 'adventure', 'heart', and 'chocolate' were also very popular words associated with positive sentiment and happiness. On the other end of the spectrum, the word 'door' was used most often in a highly negative context; many of the submitted stories talked about 'locker' or 'creaky doors', or doors behind which scary creatures like dragons or monsters were hiding.

Intriguing differences appeared between stories submitted from different parts of the country. There were more mentions of scary or unpleasant aunts in stories from Northern Ireland than any other region; aunts in other parts of the UK were presented as mostly harmless. The word 'maths' was used in a highly positive context much more frequently in stories submitted in Scotland compared to those submitted elsewhere. In stories written by English children, the words 'refugees' and 'Syria' were among those most frequently used in association with positive sentiment. Interestingly, these words appeared most often in stories that expressed high levels of hope and happiness, with the children's attitude towards refugees being largely positive and empathetic.

Finally, TheySay was able to provide a heat-map of happiness, showing how happiness in the children's stories varied by post-code. The highest average happiness levels were detected in stories submitted in Llandudno, Wales.

The insights provided by TheySay around the sentiment and emotions contained in the children's stories gave a new layer of understanding of children's language and a unique look at how age, gender, and location can affect children's writing.