Corpus-driven Analysis on the Language of Children’s Literature

Evynurul Laily Zen

Universitas Negeri Malang

Keywords: Corpus-driven; linguistic analysis, children’s literature.

Abstract: This current paper examined distinctive patterns of language that characterize children’s literature using a corpus-driven approach. I built a limited corpus—CoCL (Corpus of Children’s Literature)—of 28 novels and short stories that were published in the late 19th to the early 20th available on Project Gutenberg and written by four prominent writers; Carlo Collodi, Lewis Carrol, Beatrix Potter and Hugh Lofting. With the utilization of WMaRtix and AntConc as the corpus tools, the 319,968 tokens of CoCL were further analyzed and compared to the BNC Written Imaginative. The findings demonstrated several features distinguishing the language of this particular genre to adult fictions including significant uses of noun and subjective pronoun, explicit articulations of smallness, animals, and food, as well as cultivation of positive vibes, joyful tones, and optimism. The paper attempted to enrich evidence on the effectiveness of corpora in both linguistic and literary analysis that was, at the same time, seen to mark the advancement of digital world in language research.

1 INTRODUCTION

A tradition of story writing for children and probably also by children has dated back 250 years ago. This explains that children’s literature has truly taken thousand miles of development following the changes of people and their cultures. Kennedy (2017) points out that it was in the seventeenth century where children’s literature emerged as an independent genre stimulated by an increasing awareness of repositioning children as the center of agency as well as the point of interest. This was clearly manifested in the emerging moral values relevant for children through the portrayal of adventures together with a massive growth of picture books in the nineteenth century (Kennedy, 2017).

In the twentieth-century, children stories become progressively diverse yet remain didactic as they are written in an age-specific language (Coghlan, 2017; Leland et al., 2013). In this way, scholars agreed that children’s literature should be distinctive in a sense that it should talk about children and use ‘child-oriented’ language. Taking a child-centeredness as a point of departure, I put forward a corpus-driven analysis toward children stories with the aim of figuring out distinctive features of this specific genre in comparison to the adult’s literature.

Corpus analysis becomes a primary backdrop of this paper as I refer to (Llaurado, et al., 2012) argument stating that corpus linguistics enables researchers to obtain samples of authentic language uses in different contexts for various analytical purposes ranging from capturing developmental shifts in language use to encapsulating genre specific features. The employment of corpus is also seen to be able to build a connecting link between linguistic analysis and literary interpretation (Hardstaff, 2015; Cogo, and Dewey, 2012; Forceville, 2006). Through a corpus-based study, Hardstaff herself carefully examines child agency and character development embedded within grammatical patterning in Roll of Thunder.

In a specific context of children literature analysis, Hardstaff’s study is influential as it not only approaches a literary analysis from a different angle, but also draws a bridging line between two sub-disciplines. It, at the same time, is able to fill the gap of previous studies that have repeatedly investigated children’s literature from a very specific literary issue in a single story, such as style shifting in Peter Rabbit (Mackey, 1998; Rudman, 1995), boundaries of properties in Beatrix Potter’s tales (Blomley,
2004), animals’ right in Doctor Dolittle (Elick, 2007), and translating animals’ language in Doctor Dolittle (Hague, 2007; Heine, Narrog, & Biber, 2015).

Following Hardstaff’s line of research, I specifically work to find prominent linguistic patterns and literary elements that characterize children’s literature by making use of corpus data and corpus tools. More explicitly, my study replicates Thompson and Sealey’s (2007) comparative analysis of language patterns used in three corpora: CLLIP (Corpus-based Learning about Language in the Primary-school), adult fiction, and newspapers to identify specific features of writing aimed for child audiences compared to adult audiences. Their findings demonstrate that the language of CLLIP and adult fiction was much similar than those in newspapers. However, in terms of methodology, Thompson and Sealey (2007) limit their corpus exploration only on POS (Part of Speech) tagging analysis. It is therefore necessary to broaden their investigation on semantic tagging and concordance analysis to get a closer look at different picture of linguistic patterns of children’s literature. In addition, instead of using their corpus data that they obtain from BNC (British National Corpus) Imaginative, I build my own corpus that I will elaborate in the following sub section.

2 METHOD

I employed a corpus-driven approach where the corpus files were created before being investigated. I primarily utilized WMatrix and AntConc to locate keyness/keyword lists, POS tags, Semantic tags, and concordance analysis by using BNC Written Imaginative as a reference corpus to reveal regular patterns of language within the frame of literary works. The corpus—I termed it CoCL (Corpus of Children’s Literature)—was compiled from samples of novels and short stories published during the late 19th to the early 20th century. Stories written by Carlo Collodi and Lewis Carrol were taken to represent the late 19th century, whereas Beatrix Potter and Hugh Lofting were to represent the early 20th. This was the Golden Age of Children’s Literature in Britain. Avoiding the use of random sampling, I took those with everlasting international popularity having been adapted into screen plays as my corpus data which all are listed in Table 1.
3 FINDINGS AND DISCUSSION

3.1 Linguistic Patterns

To find the most preferred linguistic items used in children stories, I look at three features: keyness, POS tags, and Semantic Tags. The keyness analysis presented in Figure 1 shows that character names (Alice, Pinocchio, doctor (Doctor Dolittle), Polynesia) and grammatical bins (at, the, and, im) appeared most prominently. This is unsurprising as these terms should generally appear in all fictional texts. The word little (988 times), however, can be claimed as a distinctive feature of CoCL due to its high frequency compared to the reference corpus (208 times). This is in line with the result of semantic tag analysis partly shown in Figure 2 in which the concept of small occurred more regularly than big. In this context, little is used to portray children’s ability in viewing the world and their surroundings where everything they could see, hear, taste, and touch must be in an equivalent ‘size’ to them.

![Figure 1: Keyness Analysis](image1)

Further observation is carried out to find group of words that collocates with the word little. The result in Figure 3 indicates that they commonly denotes (1) human and animals (e.g. man, donkey, boy, girl, fairy, pig, dog, etc), (2) places (e.g. house, boat), and (3) objects (e.g. way, door). This particular empirical evidence has strengthened a claim that children stories should be child-oriented in the way that they must contain more concrete objects (either animate or inanimate) rather than the abstract ones (McDowell, 2006).

POS tagging analysis provides another interesting point especially when this CoCL is compared to Thompson and Searley’s (2007) CoAL (Corpus of Adult’s Literature). The result in Chart 1 illustrates that the frequency of noun, article, preposition, pronoun, and conjunction are higher in CoCL than in CoAL. This significant use of noun and pronoun in children’s literature has suggested an emphasis of ‘subject’ and ‘object’ in child’s point of view.

![Chart 1: Comparative POS Tags of CoCL and CoAL](image2)

A distinctive pattern of language is further maintained by the overuse items of semantic category in Figure 4 below.

![Figure 4: Overuse Items of Semantic Category](image3)

The category of living creatures: animals in Figure 4 appears as the key concept in CoCL. It suggests that children’s stories were constantly capturing animals. More importantly, the stories featured animals as talking creatures that personify human with their life experiences (e.g. said the crow, said the four Rabbits, the judge was a monkey, etc.) as seen in Figure 5.
A semantic field of food is fascinating in particular association to animal. It is to say that animals’ life is commonly centered around food finding. A variety of food (e.g. blackberries, cherries, beans) in Figure 6 is seen to depict a close connection of animals to their habitat and environment. The significant use of food in CoCL, therefore, sustains the importance of it both in animal and children’s life serving as the basic need of all living creatures.

2.3 Literary Elements

Literary elements commonly include settings, characters, plots, meaning, point of view, and style (Temple, et al., 2002). In analyzing the outstanding concepts appeared in children’s literature literary elements, I look at characters, style and meaning.

To assess the first element, I analyze the use of pronoun as a relevant POS to describe characters. Chart 2 below illustrates the comparative use of pronoun in CoCL and CoAL to pinpoint the findings that all types of pronoun were exploited more frequently in CoCL rather than in CoAL. This finding articulates a critical role played by the ‘agent’ or ‘doer’ in children stories which, to a large extent, uncovers children’s distinctive point of view for it seems easier for them to understand who do things before what things are done. Furthermore, Chart 2 indicates the greater use of ‘subjective’ pronoun (e.g. I, you, she, he, we, they) and the lesser use of ‘objective pronoun (e.g. me, us). It supports the previous claim on the importance of ‘self’ in childhood which I assume to be shifting in adulthood.

A closer look at all characters in the CoCL, I find that 16 out of 119 were human (e.g. Alice, Doctor Dolittle, Mr. Jackson, etc), whereas 103 out of 119 were animals (e.g. Petter the rabbit, Polynesia the parrot, Mr. John Dormouse the mouse, etc). This is predictable as the semantic tag of living creatures: animal is overused. What I consider interesting is that these animal characters were mostly portrayed as male (See Chart 3). This illustrates the focal point of gender representation as key issues in children world. With a more thorough examination on this phenomenon, I believe further study will be able to provide more elaborate explanation.

Figure 5: Concordance of Living Creatures Category

Figure 6: Concordance of Food Category

Figure 7: Semantic Category of Emotion
Words within this field include verbs (e.g., cried, laughed), adjectives (e.g., funny, glad), and adverbs (e.g., angrily, anxiously). There seems to be a tendency to use greater positive emotional expressions (e.g., gently, joy) and lesser negative emotional expressions (e.g., angry, frightened) in CoCL. It strongly implies that children are similar to adult in the way they experience both joy and sorrow, but they are different in that joy and all those enjoyable experiences were valued more. Also, Figure 7 clearly indicates that words expressing unpleasantness and sorrow were underused, whereas pleasantness and joy were overused. This corpus evidence is in particular support to the argument stating that children’s books tend to use language expressing optimism rather than depression with a major purpose to entertain children and provide moral values (McDowell, 2006; O’Sullivan and Whyte, 2017; Glynn, 2010; Guo, 2015).

The meaning of a story can be broadly treated as a certain theme or value to share embedded within that story. In this way, theme often defines the segmentation of the readers. Thompson and Sealey (2007) figured out that intimacy and sex was frequently used as a theme in adult fiction, whereas the CoCL demonstrates that nature-related issues are noteworthy in defining children’s state of mind. I refer my finding to the semantic categories of geographical terms (ranked 15th), plants (ranked 27th), and farming and horticulture (ranked 28th). It is then convincing to claim that nature is the ultimate necessity that children need to know. Geographical terms in Figure 8 below, for example, is seen to function not only as places where daily activities were conducted, but also things attached to their daily life.

In addition, words within the category of plants (e.g., bushes, trees, flowered) in Figure 9 function more than only as supplementary elements, but as a center of interests where stories are about. The semantic field of farming and horticulture in Figure 10 shows a similar tendency. These all come to prove that nature is one of the distinctive themes being valued as the key element of children’s stories. This supports Pike’s (2010) argument on the nature of spaces in children’s perspective where fairgrounds, amusement parks, and zoological gardens are commonly successful in creating an enjoyment and pleasure for children as they can interact with the natural world.

In conclusion, a corpus analysis I carried out has facilitated me in discovering distinctive language patterns as well as literary elements embodied within the CoCL (Corpus of Children’s Literature). Compared to the literary texts written for adults, the language used in children’s literature tends to be centered around the idea of smallness, animals, and food through the significant uses of noun and subjective pronoun. In terms of literary concepts, children’s literature tends to cultivate the idea of personifying animals as talking characters, elevating positive vibes and joyful tones, making nature and optimism the most preferable themes of the stories.

4 CONCLUSIONS

In conclusion, a corpus analysis I carried out has facilitated me in discovering distinctive language patterns as well as literary elements embodied within the CoCL (Corpus of Children’s Literature). Compared to the literary texts written for adults, the language used in children’s literature tends to be centered around the idea of smallness, animals, and food through the significant uses of noun and subjective pronoun. In terms of literary concepts, children’s literature tends to cultivate the idea of personifying animals as talking characters, elevating positive vibes and joyful tones, making nature and optimism the most preferable themes of the stories.
A corpus of this kind will impart a practical implication of cross-sectional studies mainly for pedagogical purposes where it provides teachers with big data of children stories as well as typical patterns of children’s language.

ACKNOWLEDGEMENTS

This paper was written upon a completion of a term-paper of Corpus Linguistics Class 2017 at the Department of English Language and Literature, National University of Singapore with a special reference and appreciation for Prof. Vincent Ooi for his constructive feedback toward the earlier draft of the paper.

REFERENCES


