Classic Children's Literature – *Difficult to read?*

Dolf Trieschnigg
University of Twente, Enschede, The Netherlands
trieschn@ewi.utwente.nl

Claudia Hauff
TU Delft, The Netherlands
c.hauff@tudelft.nl

Introduction

Motivation: Classic children's literature is freely available thanks to initiatives such as Project Gutenberg. Due to diverging vocabularies and style, these texts are difficult to understand for children in the present day.

Goal: Aid children in the reading process of classic children's literature, by automatically identifying terms that result in low readability.

Vocabularies over time

	1825	1850	1875	1900	1925
1800	0.25	0.26	0.33	0.34	0.40
1825		0.25	0.34	0.35	0.42
1850			0.25	0.26	0.30
1875				0.25	0.29
1900					0.26

The Jensen-Shannon divergence between vocabularies (unigram word distributions) from different time periods.

An Example

"A gorilla, Esmeralda?" Mr. Philander, and the three men scarcely breathed as he voiced the horrible thought. "I thought it was the devil; but I guess it must have been one of them gorilephants. Oh, my poor baby, my poor little honey," and again Esmeralda broke into uncontrollable sobbing. Clayton immediately began to look about for tracks, but he could find nothing save a confusion of trampled grasses in the close vicinity, and his woodcraft was too meager for the translation of what he did see. All the balance of the day they sought through the jungle; but as night drew on they were forced to give up in despair and hopelessness, for they did not even know in what direction the thing had borne Jane. It was long after dark ere they reached the cabin, and a sad and grief-stricken party it was that sat silently within the little structure. Professor Porter finally broke the silence. His tones were no longer those of the erudite pedant theorizing upon the abstract and the unknowable; but those of the man of action--determined, but tinged also by a note of indescribable hopelessness and grief which wrung an answering pang from Clayton's heart."

Excerpt from *Tarzan of the Apes* by Edgar Rice Burroughs (1914). Underlined terms were flagged as difficult by one or more annotators; highlighted terms were detected automatically .

Detecting Difficult Terminology

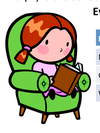
Approach:
1) Find dictionary terms in text.
2) Filter found terms based on ICF.

Dictionary: nouns, verbs, adverbs, adjectives found in Wiktionary

Inverse Collection Frequency (ICF):

$$ICF(t) = \min_{w \in t} \log \frac{n}{\max(cf(w),1)}$$

where w is a word in term t;
cf(w) is the collection frequency of w.



Evaluation results:

ICF-filtering	Precision	Recall	F-measure
No filtering	0.10	0.89	0.17
Gutenberg books	0.44	0.53	**0.46**
Wikipedia	0.37	0.66	0.24

User study

Task: Indicate difficult, rare or unusual terms in 3 or more paragraphs
Participants: 30 non-native English speakers, average age 32 (not children!).
Collection: 3 paragraphs from 10 books dating from 1719 to 1917 (with 134 to 268 words).
Results: 1) 3.9 difficult terms per paragraph;
2) relatively low agreement between participants: only 39% of indicated terms was agreed upon by at least one other participant.

Conclusion

The vocabulary gap is significant and needs to be addressed in a system that attempts to make classic children's books accessible to young readers.
There is relatively **low agreement** in the indication of difficult words; the basic system presented here yields reasonably good results.