Fairy Tales and ESL Texts: An Analysis of Linguistic Features Using the Gramulator

Rachel M. Rufenacht, Philip M. McCarthy, and Travis A. Lamkin

Department of English
University of Memphis
Memphis, TN 38152

rmrfncht@memphis.edu, pmmccrth@memphis.edu, talamkin@memphis.edu

Abstract

Using the Gramulator, we analyzed the linguistic features of ESL texts and fairy tales. Our goal was to determine if fairy tales had the potential to be used as reading material for English language learners. The results of our analyses suggest that there are significant similarities between fairy tales and ESL texts, but that differences lie in the content of the text types with fairy tales appearing significantly more narrative in style and ESL texts appearing more expository.

Introduction

This study investigates the linguistic features of traditional fairy tales. More specifically, we are interested in assessing the potential suitability of traditional fairy tales as reading material for English language learners (ELL). Fairy tales are essential stories that native-English speaking primary school students are familiar with early on in their language acquisition, whether through reading the actual stories or watching the Disney movies. As such, educators might consider using these texts when teaching English as a Second Language (ESL). Because ESL students could benefit from becoming familiar with and/or being able to reference fairy tales, it is necessary to discover whether ESL students are potentially capable of comprehending the language in fairy tales or using these texts as primary or supplementary reading material in their education. Therefore, we conducted the current study to address the following question: Are the linguistic features of traditional fairy tales sufficiently similar to standard ESL texts to have the potential to be used as material for language learners?

The purpose of the current study is to examine reading material ESL students are already receiving as compared to readings more often used with native-language students. This study is designed to facilitate ESL teachers' development of supplementary material that they can use to help increase student motivation in reading and identification of features specific to text types.

We know of no other study that specifically looks at fairy tales compared to existing ESL texts; however, there are a variety of studies that look at the positive effects of using fairy tales in the classroom. Davidheiser (2007), Fakharzadeh and Rasekh (2010), among others, advocate using fairy tales because students can identify with the characters and issues in the tales. Haulman (1985) supports the use of fairy tales because they provide language instruction through reading. They are also an important way for students to glean cultural information from reading material. Peltzman (1994) encourages educators to use storytelling to engage students in reading and learning, while Hoewisch (2001) uses fairy tales to help students improve their writing skills.

The role of schema is also an important part of this study. Schema helps students process information and construct understanding from previous knowledge. Carrell (1983), Al-Issa (2006) and Zhang (2008) investigate the ways schema affect reading comprehension, noting that background and culture play an important role in understanding texts. Carrell (1983) reviews multiple studies on both formal and content schema, noting that using background building exercises is important to help students improve comprehension. Al-Issa (2006) concludes that the closer passages relate to students' cultures, the

better their comprehension of the text, and that ESL students have difficulty activating appropriate schema. Zhang (2008) tests Chinese students' understanding of multiple types of schema through a Cloze test, determining that teachers need to be aware of formal schemata in the texts they use for their classes as it affects students' recall.

To address our research question, we formed two contrasting hypotheses: the language of fairy tales will be similar to the language of ESL texts because they are both for audiences with limited English skills; however, the content of fairy tales will be different from the content in ESL texts because fairy tales are focused on traditional narratives while the language in ESL texts is typically focused on helping students to learn the culture of the target language. Thus, ESL texts will contain content to help students learn the language associated with everyday life in America (e.g., *school*, *work*, *shopping*).

This study is of importance to current and future ESL teachers. Students need to read motivational and comprehensible material that will help them improve their English. Because narrative texts are the default form of reading, students of English may be well-served by exposure to this form of material (McCarthy et al, 2009). Also, content-based instruction is becoming an important part of the classroom setting (Grabe & Stoller 1997; Kasper, 2000). By using fairy tales in the ESL classroom, students are exposed to the language and culture simultaneously during their readings and discussions.

Methods

The Corpus

Our corpus comprises a total of 300 texts (see Table 1). This corpus is divided into three sub-corpora: The first, and main, sub-corpus is a collection of 50 fairy tales. The second sub-corpus is a collection of 100 ESL texts (equally divided into ESL upper and ESL lower corpora). The third sub-corpus is a collection of 150 baseline texts (equally divided into Narrative, History, and Science corpora).

Fairy tales are not easily distinguished from similar genres such as folk tales. As such, we used only authors that appear in the collection of *The Classic Fairy Tales* (Tatar, 1999). All the available fairy tales by these authors were downloaded from Project Gutenberg at www.gutenberg.org. A 50-text sample of the smallest text files from each author was taken to have a comparable corpus to the other sub-corpora in the study.

The ESL sub-corpus is taken from Healy et al. (2009). The texts were all gathered from internet sources and published textbooks used at the University of Memphis. These texts are a representative sample of texts that ESL students are currently receiving. The original ESL texts were categorized into four sections (*advanced*, *upper-intermediate*, *lower-intermediate*, and *beginner*) with 30 text files in each. To better compare the various corpora, we made all the sets of the total corpus as similar a size as possible. To do this, we merged the 25 smallest files each in beginner and lower-intermediate into the category *lower* and the 25 smallest files each from advanced and upper-intermediate into the category *upper*, resulting in 50 files altogether.

The third sub-corpus, previously published in Duran et al. (2007), contains 50 narrative, 50 science, and 50 history texts. These texts are randomly selected paragraphs from Science, History, and Narrative genre texts from textbooks published for junior high and high school students. Because the texts are generic and written for native English speakers, they form a useful comparison (or *baseline*) for our analyses.

Because the ESL texts and baseline genre texts were used in previous studies, the formatting of the files was not changed. The fairy tales were downloaded from a public website. Therefore, we cleaned the texts using the following procedure: All double hard returns were removed and replaced with a single hard return. Any added text (e.g., title, author name, date, printing information) before and after the beginning and end of the text was removed. Also, any instances where a picture caption was included in the middle of the text were removed.

Table 1: Descriptives of the three sets of corpora

Name	N	Avg. Words	Source	Date
Fairy Tales	50	1128.0	Gutenburg.org	---
ESL Upper	50	420.9	Healy et al.	2009
ESL Lower	50	344.4	Healy et al.	2009
Narrative	50	408.3	Duran et al.	2007
History	50	399.7	Duran et al.	2007
Science	50	411.8	Duran et al.	2007

Contrastive Corpus Analysis

Contrastive Corpus Analysis (CCA: McCarthy et al., in press) is a method through which the meaningfulness of lexical features is generated as a process of relativity. The principle of CCA is that any discourse unit (e.g., text-type, register, genre, variety, or section of text) is best understood, and perhaps only understandable, within the context of its contrast to some other similar discourse unit. Contrasting one corpus to another reveals the features indicative of each corpus, as relative to one another. In this study, the discourse units in question are the sister corpora (fairy tales, ESL, baseline), analyzed one relative to the other. Because we are interested in finding the relationships and similarities between these corpora (especially ESL texts and fairy tales), CCA is an appropriate approach for the current study.

The Gramulator

Recent developments in computational linguistics and discourse processing have made it possible for researchers to develop a wide range of sophisticated techniques that facilitate CCA. Some such tools (e.g., Coh-Metrix: Graesser et al. 2004 and LIWC: Pennebaker & King 1999) are useful in this respect, and have certainly contributed to ESL knowledge (e.g., Crossley et al. 2007). However, tools such as these only estimate pre-defined constructs (e.g., *cohesion*, *affect*) but they do not (and often cannot) specify where and to what degree the lexical features of that measure occur in text. For materials development and assessment, we need to know which specific linguistic features make an ESL text potentially readable and motivating. Thus, we are less interested in aggregated measures of words, and more interested in individual (strings of) words that identify linguistic features that make texts suitable (or not) for English language learners.

The Gramulator, like its forerunners (Coh-Metrix and LIWC), is a textual analysis tool (McCarthy, Watanabe & Lamkin in press). It allows users to combine quantitative and qualitative assessments of two or more sets of corpora to identify differential linguistic features. The Gramulator is typically used to produce *n-grams*: any string of adjacent lexical features. By processing n-grams across two sets of corpora, users can determine similarities and differences between the lexical features of the contrasting corpora. Various features of the Gramulator allow for analysis of the linguistic features: Thus, *typicals* are characteristic features found in the texts of any one corpus, and *differentials* are features indicative of a particular corpus (i.e. those features that distinguish a corpus relative to a contrasting corpus).

Differential n-grams are those n-grams that are among the most commonly occurring in one corpus (i.e., among the 50% most frequent n-grams) but are uncommon to the contrasting corpus (i.e., *not* among the 50% most frequent n-grams). The differentials are derived from the typicals by following the principals of *machine differential diagnostics* (Garg et al. 2005; Rahati & Kabanza 2010): namely, all typicals that are common to both corpora (called *shared n-grams*) are diagnostic of neither corpus, and therefore they are removed. The remaining n-grams are the most frequently occurring n-grams that are present in just one of the corpora (McCarthy, Watanabe & Lamkin in press). The Gramulator is the appropriate tool for the current study because we are as interested in the linguistic features (identified through differential n-grams) that define suitable ESL texts as we are interested in the amounts that those features are present.

The Gramulator has been used in many recent studies to analyze differentials in various corpora. For example, Min and McCarthy (2010) used the Gramulator to distinguish between American and Korean scientific writing styles; and Lamkin and McCarthy (2011) looked at differentials that distinguish two types of detective fiction.

The Gramulator includes eight modules: two pre-processing and six post-processing. In this study, we used the Cleanser module of the Gramulator to clean the texts. We also used the Evaluator, Viewer, and Concordancer modules to produce and analyze the results.

Results

Using the Gramulator, we processed both main sets of corpora (i.e., ESL and fairy tales) relative to the baseline corpus (i.e., NHS). This processing resulted in two differential indices: FT(NHS) and ESL(NHS), where FT refers to fairy tales, ESL refers to English as a Second Language texts (where the original upper and lower texts were combined), and NHS refers to the Narrative, History, and Science baseline texts. To determine the degree of similarity between ESL texts and fairy tales, we conducted a Pearson's Correlation using the *typicals* of each sub-corpora (see Table 2). The results suggest that there is a high correlation between fairy tales and ESL texts, indicating that the texts are more similar than they are different ($r^2 = .567$).

Table 2: Correlations between ESL and FT typicals

	ESL lower	ESL upper	FT
ESL all	0.911	0.806	0.753
ESL lower		0.708	0.703
ESL upper			0.511

Because the correlation showed that ESL texts and fairy tales are more similar than different, we conducted a further series of t-tests to determine the direction of the difference between the two corpora (i.e., is ESL an example of fairy tales or is fairy tales an example of ESL?). We used the previously described FT(NHS) and ESL(NHS) indices; that is, the arrays of n-gram differentials for ESL and FT relative to the NHS baseline. Using the Gramulator's Evaluator module, we then processed the FT (NHS) index against the ESL corpus and the ESL (NHS) index against the FT corpus.

We conducted a between texts t-test (see Table 3) to assess the difference between the language features of fairy tales and ESL texts (both relative to NHS). Thus, in Gramulator nomenclature, we write: FT → ESL (NHS) and ESL → FT (NHS). We predicted that there would be more ESL features in fairy tales than visa-versa because fairy tales is a more narrowly defined genre and ESL, by its nature, is seeking a general audience. The results confirmed our prediction: FT → ESL (NHS): $M = 0.083$, $SD = 0.037$; ESL → FT (NHS): $M = 0.055$, $SD = 0.038$), and reached a level of significance: t (1,148) = 4.331, $p < 0.001$, $d = 0.75$. The result suggests that FT contains more of the features of ESL (NHS) than ESL contains features of

FT (NHS). Therefore, the vocabulary used in fairy tales includes a significant amount of the language structures used in ESL texts. This result might be attributable to the broader content matter of ESL texts (see n-gram examples in the section *Viewer and Concordancer*). As such, we performed further analyses to assess whether fairy tales had more in common with the texts of upper level ESL or lower level ESL. We analyzed the FT corpus using the differentials (D) and typicals (T) of the ESL texts relative to the baseline texts (NHS). Thus, the indices were ESL_U (NHS)_D; ESL_L (NHS)_D; ESL_U_T; ESL_L_T.

For the differentials indices, we predicted that the fairy tales would have more in common with upper ESL texts because, presumably, fairy tales contain a higher frequency of complex language structures than beginning level ESL texts. The result supported out hypothesis: ESL_U (NHS)_D: $M = 0.090$, $SD = 0.073$; ESL_L (NHS)_D: $M = 0.063$, $SD = 0.029$; $t\ (1, 49) = 2.786$, $p = 0.008$, $d = 0.394$.

For the typicals index, we again predicted that the fairy tales would have more in common with upper ESL texts; the results once more indicated there was a significant difference: ESL_U_T: $M = 0.459$, $SD = 0.137$; ESL_L_T: $M = 0.283$, $SD = 0.068$; $t\ (1, 49) = 12.989$, $p < 0.001$, $d = 1.837$. As such, the two tests provide evidence that fairy tales texts contain more of the features of upper ESL.

Having shown that fairy tales contain more ESL upper material, we next turned our attention to the genre that best identified this difference. Thus, we ran four t-tests between the fairy tales corpus and the combined ESL corpora using the typicals of the baseline texts as the indices: Narrative, History, Science and all three combined (NHS): These analyses indicated that fairy tales have a higher level of Narrative features than ESL texts: (FT: $M = 0.452$, $SD = 0.102$; ESL: $M = 0.329$, $SD = 0.127$; $t\ (1,148) = 5.986$, $p < 0.001$, $d = 1.037$). In contrast, Science was in the direction of ESL texts: ESL: $M = 0.293$, $SD = 0.096$; FT: $M = 0.262$, $SD = 0.091$; $t(1,148) = 1.92$, $p = 0.057$, $d = 0.333$. And the History index was not significant: ESL: $M = 0.375$, $SD = 0.125$; FT: $M = 0.382$, $SD = 0.120$; $t\ (1,148) = -0.298$, $p = 0.766$, $d = 0.052$.

For the combined NHS index, the result was in the predicted direction: (FT: $M = 0.292$, $SD = 0.065$; ESL: $M = 0.254$, $SD = 0.057$) and reached a level of significance: $t\ (1,148) = 3.684$, $p < 0.001$, $d = .638$. The result suggests that fairy tales contain more narrative features than ESL and that the narrative features out-weight the history and science results. Taken as a whole, the results indicate that 1) there are more linguistic features of ESL texts in fairy tales than features of fairy tales in ESL texts 2) the features of upper ESL texts are more common in fairy tales than lower ESL texts, and 3) Fairy tales contain a significantly higher amount of both narrative and baseline features than ESL texts, whereas ESL texts might contain more expository elements.

The Viewer and Concordancer Modules

To analyze the differences in content between ESL texts and fairy tales, we used the Gramulator's *Viewer* and *Concordancer* modules, together with Fisher's Exact Test. The highest ranked differentials in fairy tales (FT) were *the king* ($p < .001$) and *said to* ($p < .001$). These bigrams point to the narrative quality of the texts. *The king* represents a character in a story, while *said to* represents the dialogue of the narrative. Assessing both bigrams using the *Concordancer*, the results confirm the narrative aspects of the differentials (see Table 3). Other high ranking differentials that encourage the theme of a narrative text type are the flexigrams "the + character": (*the prince, the queen, the princess;* $p < .001$); the dialogue indicators (*said the, I am, I will;* $p < .001$), and chronology structures (*at last, and when, and so;* $p < .001$). There are also significantly more instances of articles followed by concrete nouns in FT than ESL (FT: *the ground, the fire, the wood, the ogre, the palace*, $p < .001$; ESL: *the u.s., the way, the idea* $p < .001$).

The king as an n-gram also appears in ESL texts (2 texts out of 100). However, this bi-gram is not being used as a character in a narrative (see Table 4); instead, in three out of the four instances, it is referencing Elvis Presley. From

Table 3. Examples of *the king* in Fairy Tale corpus

his princess and wife, sending to invite *the king* of wood-valley to come to the feast.
was once upon a time in the service of *the king* of wide-river an excellent youth named corvetto,
bursting with envy at the kindness which *the king* showed to corvetto; so that all day long, in
desire to marry again...." at these words *the king* broke into piteous cries, took his wife's hands
the world, and so felt assured that *the king* would never marry again. be this as it may

Table 4. Examples of *the king* in ESL corpus

they sought revenge on *the king* who had killed their mother and driven them
for the 30th anniversary of the death of *the king o*f rock and roll. but there were devoted elvis
sales. and the sudden, tragic death of *the king* of rock and roll did sell a lot of records.
there are a lot of elvis presleys around *the king* lives on through 85,000 official impersonators

these examples, we can confirm our second hypotheses: the subject matter of ESL texts and fairy tales is different. This bigram demonstrates that FT is mainly concerned with stories and characters. ESL texts, on the other hand, are providing students with information about the United States political and cultural history.

The top two ESL differentials, *the united* (p <.001) and *united states* (p <.001), (which obviously combine as *the united states* (p < .001) demonstrate the subject matter provided for ESL students (Table 5): ESL texts are focused on helping students learn more about the United States. Although it may be assumed that students are being given texts that show them how to be good citizens, the *Concordancer* does not support this. Instead, the phrases using *the united* and *united states* contain information about the *melting pot*, rich culture of America, and amount of immigrants present in the country. This observation suggests that the texts ESL students are receiving are providing examples of people like themselves, living in the United States. For example, the flexigram *immigra*, *custom(s)*, *new home(s)*, and *melting pot* is more common to ESL texts (p = .030). Many of the top differentials in ESL texts also provide evidence of more expository features than narrative: *there are, for example, are the,* and *such as* (p = .002).

Table 5: Top 20 differentials from fairy tales and ESL

Fairy Tales		ESL	
1	the king	1	the united
2	said to	2	united states
3	who was	3	there are
4	as he	4	part of
5	who had	5	a lot
6	said the	6	for example
7	at last	7	want to
8	to his	8	such as
9	and said	9	the new
10	i am	10	have to
11	to him	11	new york
12	of her	12	according to
13	the poor	13	the way
14	could not	14	to help
15	and that	15	lot of
16	and when	16	the right
17	i will	17	he is
18	upon the	18	are the
19	went to	19	can be
20	him to	20	people who

Discussion

This study examined the linguistic features of ESL texts and fairy tales. The research question addressed whether fairy tales are sufficiently similar to ESL texts to be used as reading material for ESL students. We hypothesized that the language would be similar because they are written for similar audiences; however, the subject matter would be different because the texts are used for different purposes. Previous studies have encouraged the use of fairy tales in the classroom. This study began looking at whether or not fairy tales could also be used as reading material for ELLs by comparing the differentials of ESL texts and fairy tales.

The combined results indicate 1) there is a high correlation between fairy tales and ESL texts, 2) that fairy tales contain a significant amount of ESL language structures, and 3) that fairy tales also contain a high amount of baseline text features. Therefore, the linguistic features of fairy tales (as identified through n-gram analysis) can be considered similar to ESL texts, and we argue that they have the potential to be used as material for second language learners.

Limitations for this study include the discrepancy of file lengths between the fairy tale texts and those of ESL and the baseline texts. Clearly, the original lengths of fairy tales may not be suitable for many sections of ESL text books and further studies need to define an appropriate length of text. Such studies will also need to include a wider array of fairy tale texts as well as studies on the parts of texts (beginnings, middles, ends) to ensure consistency. Finally, the fairy tales are translations and some might view such texts as non-authentic. However, because we are investigating the use of fairy tales in the ESL classroom, we have to use translations because that is what native-English speakers would be accustomed to reading.

More analysis is also needed to better assess the level of similarity present between the two main types of texts. Such analyses will allow even further studies to be conducted that may lead to analyzing the usefulness of fairy tales as reading material in ESL classrooms. Finally, future research needs to assess the degree of text suitability for ESL students. Thus, while much remains to do be done, the results of this initial study contribute to the field of applied linguistics by providing teachers and materials designers with an analysis of non-traditional classroom texts and an initial evaluation of the potential suitability of such texts in ESL instruction.

Acknowledgments

The authors would like to acknowledge the following people for their help in making this paper possible: Lucille Booker, Hyunsoon C. Min, and Nick Duran.

References

Al-Issa, A. 2006. Schema theory and L2 reading comprehension: Implications for teaching. *Journal of College Teaching & Learning, 3* (7): 41-48.

Carrell, P.L. 1983. Some Issues in Studying the Role of Schemata, or Background Knowledge, in Second Language Comprehension. *Reading in a Foreign Language, 1,* p.81-92.

Crossley, S.A., Louwerse, M., McCarthy, P.M., and McNamara, D.S. 2007. A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91,* 15-30.

Duran, N.D., McCarthy, P.M., Graesser, A.C., and McNamara, D.S. 2007. Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 39,* 212-223.

Davidheiser, J.C. 2007. Fairy tales and foreign languages: Ever the twain shall meet. *Foreign Language Annals:* 40 (2). p. 215-225.

Fakharzadeh, M. and A. Rasekh. 2010. On the applicability or non-applicability of the gricean maxims to nursery rhymes. *Journal of Linguistics and Language Teaching*: 1.1, pp. 37-73.

Garge. A.X., Adhikari, N.K, McDonald, H., Rosas-Arellano, M.P., Deveraux, P.J., Beyene, J., Sam, J., and Haynes, R.B. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 293:1223-1238.

Grabe, W., and Stoller, F. 1997. Content-based instruction: Research foundations. In M. Snow, & D. Brinton (Eds.), *The content-based classroom*. White Plains, NY: Addison-Wesley, Longman, p. 5-21.

Graesser, A.C., McNamara, D.S., Louwerse, M., & Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202.

Haulman, A. 1985. Fairy tales in the ESL classroom. *Presented at the International Conference on Second/Foreign Language Acquisition by Children*. p. 3-12.

Healy, S.L., Weintraub, J.D., McCarthy, P.M., Hall, C.E., and McNamara, D.S. 2009. Assessment of LDAT as a Grammatical Diversity Assessment Tool. *Proceedings of the Twenty-Second International FLAIRS Conference*, 249-253.

Hoewisch, A. 2001. Do I have to have a princess in my story?: Supporting children's writing of fairy tales. *Reading & Writing Quarterly*, 17: 249-277.

Kasper, L. (Ed.). 2000. Content-based college ESL instruction. Mahwah, NJ: Lawrence Erlbaum Associates.

Lamkin, T.A., and McCarthy, P.M. 2011. The Hierarchy of Detective Fiction. In C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference.* Menlo Park, CA: The AAAI Press.

McCarthy, P.M., Myers, J.C., Briner, S.W., Graesser, A.C. and McNamara, D.C. 2009. A Psychological and computational study of sub-sentential genre recognition. *Journal for Language Technology and Computational Linguistics* 24: 23-55.

McCarthy, P.M., Watanabe S., and Lamkin, T.A. (in press), The Gramulator: A tool for the identification of indicative linguistic features. In P.M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

Min, H.C. and McCarthy, P.M. 2010. Identifying varietals in the discourse of American and Korean scientists: A contrastive corpus analysis using the gramulator. In H.W. Guesgen & C. Murray (Eds.), *Proceedings of the 2rd International Florida Artifical Intelligence Research Society Conference*. Menlo Park, CA: the AAAI Press, p. 247-252.

Peltzman, B. 1994. The art of storytelling teaches listening, speaking and comprehension skills. *The Reading Instruction Journal*: 37 (2). p. 7-11.

Pennebaker J.W., King L.A. 1999. Linguistic styles: language use as an individual difference. *J. Personal. Soc. Psychol.* 77:1296–312.

Rahati, A., and Kabanza, F. 2010. Persuasive dialogues in an intelligent tutoring system for medical diagnosis. *Proceedings for the 10th Annual Intelligent Tutoring Systmes International Conference*. Berlin, Germany: Springer, p. 51-61.

Tatar, M. (Ed.). 1999. *The Classic Fairy Tales*. New York: W. W. Norton & Company, Inc.

Zhang, X. 2008. The effects of formal schema on reading comprehension—An experiment with Chinese EFL readers. *Computational Linguistics and Chinese Language Processing*, 13(2), p. 197-214.