INNOVATION IN LANGUAGE LEARNING International Conference

Introduction of Complex Vocabulary in Literature through Fine-Tuning: A Corpus-Based Study

Iglika Nikolova-Stoupak

Sorbonne University, France

Abstract

Methods of foreign language teaching that deliberately mimic the way children acquire their first language, such as Krashen's comprehensible input [5], have proven effective for a large number and variety of learners. As these methods rely heavily on knowledge of psychology and psycholinguistics, it comes naturally that new discoveries in the said fields should motivate their extension and, possibly, modification. Leung et al. claim that parents introduce complex vocabulary in a specific, fine-tuned manner, providing additional explanatory context to the child [6]. The present study adopts the framework of corpus linguistics to test the hypothesis that abridged literature for language learners and young readers also exhibits this trait. For the purpose, original and abridged versions of three classical literary works as well as their translations in several languages will be juxtaposed in relation to the context in which they introduce complex vocabulary. Furthermore, the presence and prominence of the examined textual characteristic will be analysed by target reader age as well as by language, thus shedding light on what Krashen refers to as the "natural order" of language acquisition.

Keywords: *comprehensible input, graded readers, second language acquisition*

1. Introduction and Theoretical Framework

The qualities of reading have been highly emphasised in recent years, notably as it is compared to popular pastime activities of more controversial nature, such as engagement in social media and other digital forms of entertainment. In particular, the reading of novels is proven to directly improve both academic performance and motivation [7]. Altered or "abridged" versions of literary works are specially created in order to aid children and foreign language learners in their language acquisition. The relative difficulty of a given text is traditionally measured through its "readability", a calculated and highly objective characteristic that has recently benefited greatly from advancements in the field of Natural Language Processing. Narrow research has sought optimal readability measures for texts in languages other than English (for example, Wilkens et al. provide a detailed readability assessment tool for French [10]) as well as for texts created for L2 learners (such as Xia et al.'s study [11]). In accordance with intuition, typical readability features, such as the length of a given text or the average number of words within a sentence, associate reduced length with reduced complexity. However, there are exceptions to this rule, one of which will be examined in detail in this paper: namely, the introduction of complex vocabulary with added auxiliary context.

The primary motivation for the current experiment is a recent study conducted by Leung et al., according to which parents tend to introduce advanced vocabulary in a specially fine-tuned manner, providing additional context to the child [6]. For instance, a leopard may be explicitly marked as "dotted" or said to be acting "like a cat". Could this tendency be extended to the domain of reading as well as to foreign language learners? According to Krashen's theory of language acquisition [5], such a link is anything but farfetched. With the idea of comprehensive input, the linguist emphasises exposure to altered and thus understandable by the learner spoken and written language and industriously establishes similarities between an infant's acquisition of their native language and foreign language learning. When it comes to the acquisition of discrete languages, Krashen notes that the so-called "natural order" may differ (i.e. pluralisation may be learned before or after verb conjugation or grammatical case endings), and that differences seem strikingly independent on the learner's L1 [5]. No extensive work has sought to establish the referenced natural order of different languages.

2. Experimental Setup

For the purpose of this study, three classical English-language novels of roughly the same time period are selected: *A Christmas Carol* (Charles Dickens), *Alice's Adventures in Wonderland* (Lewis Carroll), and *The Adventures of Tom Sawyer* (Mark Twain). A corpus is formed that consists of the original works, translations into French, Russian and Spanish and abridged versions of all full texts (up to two

different translations and three abridged versions per language are considered, based on availability). Some abridged versions are defined as targeting a specific reader audience, such as children of a certain age or foreign language learners of a specific level.

Firstly, all texts undergo basic preprocessing, which includes the removal of capitalisation and non-alphabetic symbols. Lemmas and part-of-speech tags are then derived from each word in a given text. A basic algorithm is utilised to find candidate complex words in the unabridged texts: a list of just the nouns is derived as exemplifying most strongly the examined readability feature; then, following Zipf's principle of least effort, the longest words are taken (following a process of trial and error, 300 words per text are opted for). The derived word lists are examined and manually cleaned of errors. Wrongly selected words include non-words derived from textual processing ("adventuresbeginning"), non-nouns ("aficionada", used as an adjective) and words that are objectively not complex despite their length ("Christmas," "cumpleaños"). Hyphenated collocations ("school-house") are purposely included in the list; whilst expressions in French and Russian that are hyphenated based on grammatical rules ("commença-t-elle", "как-нибудь") are removed.

All words from the finalised lists are sought within the respective book's abridged versions and, thus, shorter lists of words that appear in both works are derived for each full-abridged textual pair. The contexts of the word's introduction (i.e. first use) in the two texts are extracted and analysed. Instances of the examined characteristic; namely, introduction of a complex word within a more detailed context in an abridged work, are noted and studied. Words that appear in different contexts in the two works are disregarded.

3. Results and Observations

A total of 62 instances of complex vocabulary being introduced via additional context in an abridged version of a text were found out of a total of 377 complex words appearing in full-abridged textual pairs. Three main subtypes of the charactarestic were observed: 46.8 % of the instances included the addition of related vocabulary issuing from the same lexical field, at times bordering redundancy ("brouillard" vs "épais brouillard"; "гостеприимство" vs "щедрое гостеприимство"); 41.9 % consisted in transformation into simpler grammatical structure ("tejer una guirnalda de margaritas" vs "juntar margaritas para trenzar una guirnalda") and 11.3 % featured an explanation or definition ("в суде заседают, потому и называются 'присяжные заседатели'"). Occasionally, the distinctions between the three subtypes were not straightforward as, for instance, the addition of close but not fully synonymous vocabulary comes close to "explanation."

Language	Prominence of the Feature (proportion)	Prominence of the Feature (%)	Breakdown by Type of Additional Context
English	4/69	12.2%	50% gram. trans.
			50% related voc.
French	14/67	20.9%	28.6% gram. trans.
			42.9% related voc.
			28.6% explanation
Russian	15/85	17.6%	46.7% gram. trans.
			33.3% related voc.
			20% explanation
Spanish	29/156	18.6%	44.8% gram. trans.
			55.2% related voc.

Figure 1. Instances of complex vocabulary introduced with additional context in abridged literary works by language.

Fig. 1 shows the presence and distribution of the examined readability feature by language. Instances are fewest in English texts and highest in French, followed by Spanish texts; the proximity of scores for the latter two suggesting that their commonality as Romance languages might influence the examined feature. The distribution tends to be highly balanced between "grammatical transformation" and "related vocabulary", a limited number of explanations appearing only in French and Russian.

If one is to regard the works by intended audience (see Fig. 2), a few tendencies can be discerned. Very few complex words as defined in the full versions of the texts are present in the abridged versions for very young children, the only two instances of added context consisting in explanations. Similarly, none of the sought complex words are found in the abridged versions for lower-level foreign language learners (keeping in mind the limitation that only English works are examined). Inversely, the

characteristic is most prominent with higher-level foreign language learners (40% of considered words).

Audience	Prominence of the Feature (proportion)	Prominence of the Feature (%)	Breakdown by Type of Additional Context
General/Undefined	24/139	17.3%	50% gram. trans.
			41.7% related voc.
			8.3% explanation
FL Students	4/15	26.7%	50% gram. trans.
			50% related voc.
0-500 words	*N/A*	*N/A*	*N/A*
500-1000 words	*2/5*	*40%*	*50% gram. trans.*
			50% related voc.
Children	28/169	16.6%	35.7% gram. trans.
			44.4% related voc.
			17.9% explanation
Age 5-8	*2/10*	*20%*	*100% explanation*
Age 9-11	*14/95*	*14.7%*	*50% gram. trans.*
			50% related voc.

Figure 2. Instances of complex vocabulary introduced with additional context in abridged literary works by audience. “Age” is based on the lowest recommended age for a text. Works for children and foreign language learners that are not further specified are accounted for only in higher levels.

4. Conclusion and Future Directions

The presented experiment shows that in a non-negligible proportion of cases, complex vocabulary is introduced in abridged literary works with additional context as compared to their full counterparts. The said context mostly comes in the face of additional vocabulary of the same lexical field and what can be defined as grammatical transformations as accounted for by Harris, which render the text simpler whilst information remains constant [4]. The characteristic seems to be more present in the context of L2 learners as well as to be of increasing relevance as a reader’s level of proficiency (or age in the case of native speakers) increases. Variance by language is possible but inconclusive.

Whilst knowledge of the examined readability characteristic can be directly applied in the composition of texts as well as in FL classrooms in the context of introduction of vocabulary, additional research on and around the topic would be beneficial. As the corpus of examined works is limited, clear outliers can be pointed out (such as Sam’l Gabriel Sons and Company’s version of *Alice in Wonderland*, which features no instances of the feature in a large sample of 51 complex words). Also, attention should be accorded to similar yet distinct tendencies in abridged works that have been observed, which speak of possible general simplification of the context around newly introduced complex words rather than necessarily its expansion. The process of seeking the examined and related textual characteristics can be further automatised and refined; for instance, through reliance on relevant frequency lists rather than purely on a word’s length.

References

[1] Charyulu, G. M. (2018). Complications in Reading Abridged Texts: A Study on Cultural Destruction by ELLs in Meaning-Making Process, *Review of Research 7*, 9: 1-5.

[2] DuBay, W. H. (2007), *The Classic Readability Studies,* Clearinghouse.

[3] Gala, N., Todirascu, A., Bernhard, D., Wilkens, R. and Meyer, J.-P. (2020). Transformations syntaxiques pour une aide à l’apprentissage de la lecture : typologie, adéquation et corpus adaptés, *Congrès Mondial de Linguistique Française*, Montpellier, France.

[4] Harris, Z. (1988). *Language and Information,* Columbia University Press.

[5] Krashen, S. (1982). *Principles and Practices of Second Language Acquisition,* Pergamon.

[6] Leung, A., Tunkel, A. and Yurovsky, D. (2021). Parents Fine-Tune Their Speech to Children's Vocabulary Knoweldge, *Psychological Science*, 32: 975–984

[7] Moje, E. B., Overby, M., Tysvaer, N. and Morris, K. (2008). The Complex World of Adolescent Literacy: Myths, Motivations, and Mysteries, *Harvard Educational Review* 78, 1: 107–54.

[8] Reading for Pleasure: A Research Overview. (2006). *National Literacy Trust*. https://literacytrust.org.uk/research-services/research-reports/reading-pleasure-research-overview/.

[9] Rodriguez, A., Leonor, G. and Flórez, E. E. R. (2018). Using the Abridged Version of Some Novels as a Way to Encourage Students' Written and Oral Production, *GiST Education and Learning Research Journal,* 16: 6-32.
[10] Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022) FABRA: French Aggregator-Based Readability Assessment toolkit, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1217–1233.
[11] Xia, M., Kochmar, E. and Briscoe, T. (2016). Text Readability Assessment for Second Language Learners, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications,* 12-22.

Appendix: Corpus Content

Carroll, L. (1865). *Alice's Adventures in Wonderland*. Project Gutenberg.
Carroll, L. (1916). *Alice's Adventures in Wonderland* (Abr. ed.). Sam'l Gabriel Sons and Company. (Original work published 1865)
Carroll, L. (2000). *Alice's Adventures in Wonderland* (J. Bassett, Abr. ed.). Oxford University Press. (Original work published 1865)
Carroll, L. (1978). *Alisa v Strane chudes.* Nauka (N. M. Demurova, Trans.). (Original work published 1865)
Carroll, L. (1991). *Alisa v Strane chudes* (L. Yahnin, Abr. ed.). Eksmo. (Original work published 1865)
Carroll, L. (2000). *Alisa v Strane chudes.* Biblioteka Maksima Moshkova (Y. Nesterenko, Trans.). (Original work published 1865)
Carroll, L. (2018). *Alisa v Strane chudes* (Abr. ed.). Eksmo. (Original work published 1865)
Carroll, L. (1996). *Las aventuras de Alicia en el país de las maravillas* (L. Maristany, Trans.). Titivillus. (Original work published 1865)
Carroll, L. (2003). *Las aventuras de Alicia en el país de las maravillas* (M. Aguirre, Trans.). Ediciones del Sur. (Original work published 1865)
Carroll, L. (2017). *Las aventuras de Alicia en el país de las maravillas* (N. Schuff, Abr. ed.). Santa Fe. (Original work published 1865)
Carroll, L. (2018). *Las aventuras de Alicia en el país de las maravillas* (F. Díez de Miranda, Abr. ed.). Zig Zag. (Original work published 1865)
Carroll, L. (1908). *Les Aventures d'Alice au pays des merveilles* (H. Bué, Trans.). Hachette. (Original work published 1865)
Carroll, L. (1975). *Les Aventures d'Alice au pays des merveilles* (H. Parisot, Abr. ed.). Editions Corentin. (Original work published 1865)
Carroll, L. (1992). *Les Aventures d'Alice au pays des merveilles* (P. Rouard, Abr. ed.). Bayard Jeunesse. (Original work published 1865)
Carroll, L. (2001). *Les Aventures d'Alice au pays des merveilles* (J. Papy, Trans.). Gallimard Jeunesse. (Original work published 1865)
Carroll, L. (2012). *Les Aventures d'Alice au pays des merveilles* (P. Protet, Abr. ed.). Auzou. (Original work published 1865)
Dickens, C. (1905). *A Christmas Carol*. The Baker & Taylor Company.
Dickens, C. (2000). *A Christmas Carol* (C. West, Abr. ed.). Oxford University Press. (Original work published 1905)
Dickens, C. (2004). *A Christmas Carol* (P. Lagendijk, Abr. ed.). Mediasat Poland Bis. (Original work published 1905)
Dickens, C. (1986). *Canción de Navidad* (S. R. Santerbás, Abr. ed.). Anaya. (Original work published 1905)
Dickens, C. (2004). *Canción de Navidad* (Trans.). Ediciones del Sur. (Original work published 1905)
Dickens, C. *Canción de Navidad* (Abr. ed.). Ediciones la Cueva, https://www.argentina.gob.ar/sites/default/files/dickens_charles_-_una_cancion_de_navidad.pdf. (Original work published 1905)
Dickens, C. (1890). *Conte de Noël* (A. De Goy and De Saint-Romain, Trans.). La Bibliothèque électronique du Québec. (Original work published 1905)
Dickens, C. (2002). *Conte de Noël* (Trans.). Pitbook.com, https://www.pitbook.com/textes/htm/chant_noel.htm. (Original work published 1905)

Dickens, C. (2021). *Cuento de Navidad* (Abr. ed.). Blurb, https://www.studocu.com/es-mx/document/universidad-autonoma-agraria-antonio-narro/agricultura-sustentable/charles-dickens-cuento-de-navidad/28459149. (Original work published 1905)
Dickens, C. (1891). *Rozhdestvenskaya pesen v proze* (S. M. Dolgova, Trans.). Runivers. (Original work published 1905)
Dickens, C. (2021). *Rozhdestvenskaya pesen v proze* (T. Ozerskaya, Abr. ed.). ACT. (Original work published 1905)
Dickens, C. (2004). *Un Chant de Noël* (L. Papineau, Abr. ed.). Héritage. (Original work published 1905)
Twain, M. (1917). *Les Aventures de Tom Sawyer* (P. F. Caillé and Y. Dubois-Mauvais, Trans.). Ebooks libres et gratuits. (Original work published 1876)
Twain, M. (1996). *Les Aventures de Tom Sawyer* (F. De Gaïl, Trans.). Flammarion. (Original work published 1876)
Twain, M. (2003). *Las Aventuras de Tom Sawyer* (J. Torroba, Trans.). Biblioteca Virtual Universal. (Original work published 1876)
Twain, M. (2007). *Las Aventuras de Tom Sawyer* (L. I. Barrena, Abr. ed.). Anaya. (Original work published 1876)
Twain, M. (2010). *Las Aventuras de Tom Sawyer* (B. Palacios, Abr. ed.). Dirección General de Bibliotecas. (Original work published 1876)
Twain, M. (1917). *Les Aventures de Tom Sawyer* (P. F. Caillé and Y. Dubois-Mauvais, Trans.). Ebooks libres et gratuits. (Original work published 1876)
Twain, M. (1996). *Les Aventures de Tom Sawyer* (F. De Gaïl, Trans.). Flammarion. (Original work published 1876)
Twain, M. (2020). *Les Aventures de Tom Sawyer* (A. Culleton, Abr. ed.). Broché. (Original work published 1876)
Twain, M. (1972). *Priklyucheniya Toma Soyera* (K. Chukovskiy, Trans.). Kaliningradskoe Knizhnoe Izdatelstvo. (Original work published 1876)
Twain, M. (2014). *Priklyucheniya Toma Soyera* (I. O. Rodin, Abr. ed.). Biblioteka Shkolnika. (Original work published 1876)
Twain, M. (2014). *Priklyucheniya Toma Soyera* (N. L. Daruzes, Abr. ed.). Vita Nova. (Original work published 1876)
Twain, M. (1876), *The Adventures of Tom Sawyer*. Project Gutenberg.
Twain, M. *The Adventures of Tom Sawyer* (Abr. ed.). Global Publishing Solutions. https://americanenglish.state.gov/. (Original work published 1876)
Twain, M. (2000). *The Adventures of Tom Sawyer* (J. Kehl, Trans.) Pearson Education. (Original work published 1876)
Twain, M. (2000). *The Adventures of Tom Sawyer* (N. Bullard, Abr. ed.). Oxford University Press. (Original work published 1876)