From Jane Austen's original *Pride and Prejudice* to a graded reader for L2 learners: a computational study of the processes of text simplification

Emily Franzini
efranzini@etrap.eu
eTRAP Research Group
University of Goettingen, Germany

Marco Büchler
mbuechler@etrap.eu
eTRAP Research Group
University of Goettingen, Germany

Introduction

Authentic text and graded reader

One of the objectives of second language (L2) learning is to be able to read and understand a variety of texts, from novels to newspaper articles, written in the language of interest. These texts written with a native audience in mind are commonly referred to as authentic texts or ”real life texts, not written for pedagogic purposes” (Wallace, 1992). Authentic texts, however, can present too many obstacles for L2 learners with too low a level of knowledge. The complex language structures and advanced vocabulary of these 'real' texts can have the unwanted effect of demotivating the reader (Richard, 2001). The gap between the learner's limited L2 knowledge and the fluency of authentic texts creates an ideal space for graded readers. Graded readers are ”simplified books written at varying levels of difficulty for second language learners” (Waring, 2012). Through graded readers original classic works can be adapted to match the learner's level of knowledge, thus providing the ideal tool to tackle 'real' themes, narratives and dialogues.

From authentic text to graded reader

One such graded reader is a newly adapted version of Jane Austen's *Pride and Prejudice* (edition of 1813) that one of the authors of this paper wrote (Franzini, 2016) as part of a collection for learners of English as a foreign language (EFL).

For authors, the process of adaptation of a text for a learning audience is complex. In order to simplify the text the author will necessarily have to make grammatical changes and lexical substitutions following vocabulary lists, shorten the text by cutting out entire paragraphs and events, and in some cases eliminate entire chapters and characters. Together with these changes, which can be defined as 'structural' because they are dictated by hard requirements of length and standardised level of difficulty, the author will also make a series of judgment calls at a sentence and word level. These changes, which are here defined as 'cognitive', include processes that are more intangible and that are a consequence of a native author's 'feeling' that the original text is too difficult for learners. These include elaborating, clarifying, providing context and motivation for unfamiliar information and non-explicit connections (Beck et al., 1991).

Research Objective

The objective of this study is to computationally analyse the manual process behind the simplification of a historical authentic text aimed at producing a graded reader. More specifically, it aims to classify and understand the structural and cognitive processes of adaptation that a human author, more or less consciously, is able to perform manually. Do the applied changes follow strict rules? Can they be classified as forming a pattern? And if so, can they be reproduced computationally?

Related Research

Researchers have long been addressing the issue of text simplification for a variety of purposes. A similar study to this was made by Petersen who compared authentic newspaper articles with abridged versions (Petersen and Ostendorf, 1991). Similar studies have been made, for example, to create a reading aid for people with disabilities (Canning, 2000, Allen, 2009).

Data

This study considers two sets of data. The first is the entire original novel (ON) *Pride and Prejudice*. The second dataset the graded reader (GR) published by Liberty. The GR has been compressed from the 61 chapters of the ON to 10 chapters. When comparing word tokens, the GR has a size of 12.6% of the ON (Tab. 1). The language was simplified to match the upper intermediate level B2. To guide the choice of vocabulary, the author chose to follow the Lexitronics Syllabus (Lexitronics, 2009).

	Line count	Word tokens	Word types	Average sentence length
Original Novel	5,974	143,386	6,823	24.00
Graded Reader	1,115	18,086	1,813	16.22
% GR size in respect to ON	18.6%	12.6%	26.5%	67.5%

Table 1: Quantitative comparison between data sets

Methodology

Readability

As a first step towards analysing the differences and similarities between an authentic text and a graded reader, we decided to evaluate if what is published as a graded reader can computationally be considered a simplified version of the original. The method chosen to make this investigation was to conduct two different readability tests, namely the ARI test and the Dale-Chall Index test on the data. Both tests were designed to gauge the comprehension difficulty of a text by providing a numeric value, which corresponds to a particular school level of a native speaker of the language tested.

The results show that both tests yield similar scores and satisfy the hypothesis that this particular GR can be computationally proven to be, in terms of 'understandability', a simplification of the ON.

	ARI	Dale-Chall
Original Novel	14-15 year olds	14-16 year olds
Graded Reader	11-12 year olds	11-13 year olds

Table 2: Age level of text understandability

Difference Analysis

In order to analyse the process of adaptation, a difference analysis was conducted by considering both those elements that changed from the ON to the GR, and those that, by contrast, remained the same. The analysis is structured into chapters, sentences and words, so as to proceed in order from the largest unit of text to the smallest.

When adapting a text, whether it is for a graded reader, a play or a film, the rationale behind the selection of certain parts over others is normally content-based. Here the author selected the most dynamic parts of the novel, which included dialogues, moments of suspense, movements of the characters and revelations. The selection of some scenes of the plot over others is purely a 'cognitive' choice of the author because it is entirely subjective. However, by using a text reuse detection software (TRACER) on both texts it was possible to visualise where the majority of reuses occur. These concentrate in particular around the beginning and the end of the novel (dark green in Fig. 1).

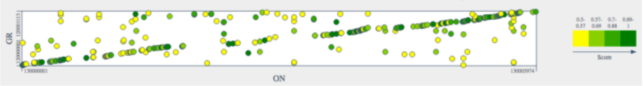


Figure 1: Dotplot visualisation of the reuses between the ON and the GR. The longer X-axis represents the larger original novel, the Y-axis the smaller GR. The darker the dot, the closer the similarities between the two datasets

'Structural' changes made at a sentence level present patterns that can be more systematically identified. For example, by comparing sentence length, it was noted that on average the ON contains longer sentences (24 words) than the GR (16.22 words) (Fig. 2). Though this might seem like an obvious result, it appears less so when one thinks that, in order to simplify a concept for a language learner, it is often necessary to use additional words to elaborate or clarify it.

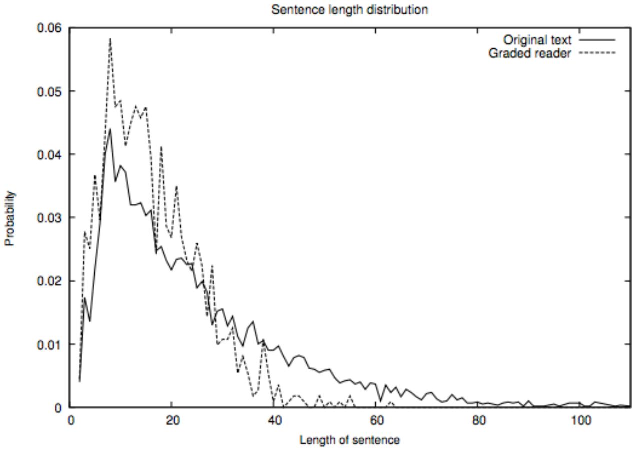
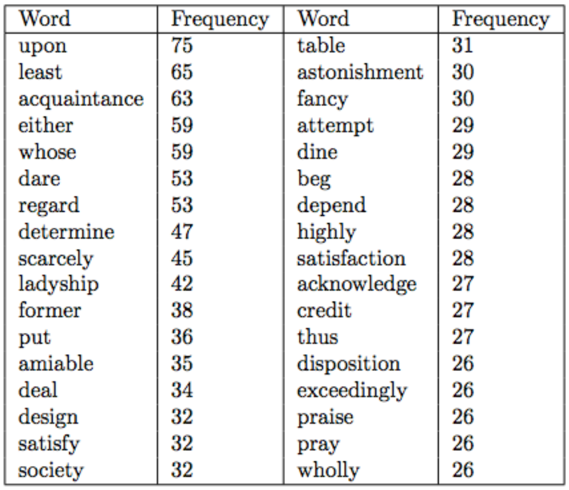


Figure 2: Sentence length distribution. The X-axis represents the number of words per sentence; the Y-axis is the probability of sentences of a specific length occurring in the texts

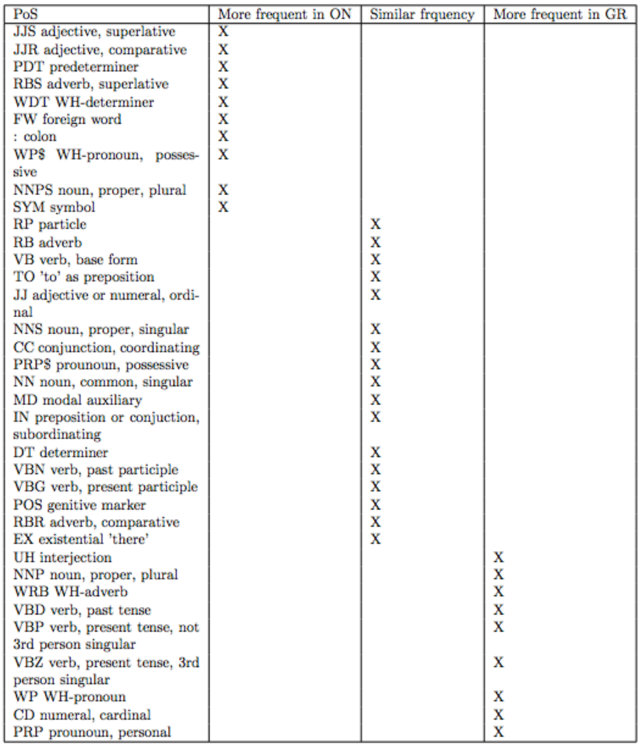
In order to conduct a difference analysis on the smallest unit of text - the word - we looked at all the words that appear frequently in the ON, but that never appear in the GR, in order to understand what kinds of words the author found necessary to drop.



Word	Frequency	Word	Frequency
upon	75	table	31
least	65	astonishment	30
acquaintance	63	fancy	30
either	59	attempt	29
whose	59	dine	29
dare	53	beg	28
regard	53	depend	28
determine	47	highly	28
scarcely	45	satisfaction	28
ladyship	42	acknowledge	27
former	38	credit	27
put	36	thus	27
amiable	35	disposition	26
deal	34	exceedingly	26
design	32	praise	26
satisfy	32	pray	26
society	32	wholly	26

Table 3: Words that appear only in the ON

Table (3) shows that 14 out of the 34 words listed (ca. 35%) are too advanced for level B2. Some of the other words, though accessible to B2 learners, were replaced with easier synonyms. We also conducted an analysis on parts of speech and how they differ in the two data sets (Tab. 4).



PoS	More frequent in ON	Similar frquency	More frequent in GR
JJS adjective, superlative	X		
JJR adjective, comparative	X		
PDT predeterminer	X		
RBS adverb, superlative	X		
WDT WH-determiner	X		
FW foreign word	X		
: colon	X		
WP$ WH-pronoun, possessive	X		
NNPS noun, proper, plural	X		
SYM symbol	X		
RP particle		X	
RB adverb		X	
VB verb, base form		X	
TO 'to' as preposition		X	
JJ adjective or numeral, ordinal		X	
NNS noun, proper, singular		X	
CC conjunction, coordinating		X	
PRP$ prounoun, possessive		X	
NN noun, common, singular		X	
MD modal auxiliary		X	
IN preposition or conjuction, subordinating		X	
DT determiner		X	
VBN verb, past participle		X	
VBG verb, present participle		X	
POS genitive marker		X	
RBR adverb, comparative		X	
EX existential 'there'		X	
UH interjection			X
NNP noun, proper, plural			X
WRB WH-adverb			X
VBD verb, past tense			X
VBP verb, present tense, not 3rd person singular			X
VBZ verb, present tense, 3rd person singular			X
WP WH-pronoun			X
CD numeral, cardinal			X
PRP prounoun, personal			X

Table 4: Parts of speech frequency in the ON vs. in the GR. Note the presence of comparative and superlative adjectives in the ON, which are totally absent from the GR

Conclusions and further research

This study is a first step into the realm of text simplification and adaptation regarding graded readers for L2 learners. By conducting a difference analysis between the two texts, it was observed that at plot level the selection of scenes has no impact on the difficulty of a text. The text reuse detection software used, however, identified which parts of the plot have been preserved and which have been eliminated for the sake of a consistent, yet shorter, story line. It was observed that the beginning and the end of the novel were the parts that were adapted most faithfully.

The identification of reuse over the whole novel was also a step towards pinpointing where sentences were reused verbatim and where they were not. Where the sentences have undergone heavy changes, we can observe to what extent they were modified, how and why. At a sentence level, we noted that reducing the length of the sentences is a successful simplification strategy. A further study would have to be conducted to best understand how sentences were split or reduced, and consequently how the syntax of a sentence was affected by its shortening.

At a word level, the simplification of the text appeared to be dictated by the elimination and replacement of difficult vocabulary and certain parts of speech, such as comparative and superlative adjectives. The word length does not appear to be an indicator of difficulty. While it was observed that both the readability tests were based on sentence length as a parameter, only the ARI test, however, considers word length as another parameter. A test on the word-length distribution of the ON versus the GR shows that, in this case, the word length bears no importance in assessing the difficulty of a text. Further research would have to be conducted in order to learn if it is easier for an L2 learner to remember a word not because of its length, but because of its repeated presence in a text. The insights gained from this study will be useful in future work on automating the simplification process.

Bibliography

Allen, D. (2009). "A study of the role of relative clauses in the simplification of news texts for learners of English." *System,* 37(4): 585–599.

Beck, I. L., McKeown, M. G., Sinatra, G. M., and Loxterman, J. A. (1991). "Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility." *Reading Research Quarterly*, Vol. 26(No. 3): 251–276.

Canning, Y. (2000). “Cohesive regeneration of syntactically simplified newspaper text.” *Proc. ROMAND*, pp. 3–16. 14.

Council of Europe (n.d.) European CEFR - Common Framework of References for Languages. Language Policy of the Council of Europe: http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

Franzini, E. (2016). *Adapted Edition of Jane Austen’s Pride and Prejudice.* Liberty Publishing.

Lexitronics (2009). Lexitronics Syllabus. https://tvo.wikispaces.com/file/view/20386024-Common-English-Lexical-Framework.pdf

Petersen, S. E. and Ostendorf, M. (1991). “Text simplification for language learners: A corpus analysis.” *Speech and Lanuguage Technology in Education SLaTE2007*.

Richard, J. C. (2001). *Curriculum development in language teaching.* Cambridge, C.U.P.

Wallace, C. (1992). *Reading.* Oxford, O.U.P.

Waring, R. (2012). “Writing graded readers.” Accessible at: www.robwaring.org/papers/Writing_graded_reader.doc