Reading in a Foreign Language
Volume 17, No. 2, October 2005
ISSN 1539-0578

Simplification in graded readers: Measuring the authenticity of graded texts

Gillian Claridge
International Pacific College

Abstract

This study examines the characteristics and quality of simplification in graded readers as compared to those of 'normal' authentic English. Two passages from graded readers are compared with the original passages. The comparison uses a computer programme, RANGE (Nation and Heatley, 2003) to analyse the distribution of high and low frequency words in the passages. This is supported by a comparison of the texts in terms of Swaffar's (1985) characteristics of authentic message. The present study is in part a reanalysis and extension of Honeyfield's (1977) seminal study of simplification, but it reaches different conclusions. By not making the simplified versus original text comparison in absolute terms, but in terms of the respective readers, it finds that patterns of use of structure, discourse markers, redundancy, collocations, and high and low frequency vocabulary, are similar in both original and simplification. This suggests that the writing in well-written graded readers can be, for its audience, experienced as authentic and typical of 'normal' English.

keywords: graded readers, simplification, authenticity, high/low frequency words, random distribution, blandness, homogeneity, repetition, redundancy

Introduction

As language teachers of elementary and intermediate students of English, we are constantly enjoining our students to read, but to read what? With an understanding, possibly, of fewer than the first 1000 most common words in the English language, the choice is limited if the students' lack of vocabulary is not to interfere seriously with their comprehension. Graded readers, comes the cry. But there is an opinion, dating from the 1970s and still current among some EFL teachers today, which maintains that a graded reader cannot give the student an authentic reading experience.

Honeyfield (1977) says that the two principal aspects of text simplification, namely simplifying language and simplifying content, produce material which differs significantly from 'normal' English in the areas of information distribution and common structure. In a so-called normal, unsimplified text, there will be a random distribution of high and low frequency words, unexpected encounters with various low frequency words, and variations in sentence lengths and collocation, all of which give a text its unique, authentic character. Honeyfield claims that simplified graded readers, by removing low frequency vocabulary from a text, flatten and homogenize it, leaving a bland distortion of normal English in which the communicative structure is disrupted. This means that learners using graded readers develop "reading strategies that are inappropriate for reading unsimplified English" (Honeyfield, 1977). He concludes that these factors may limit the effectiveness of simplified material in training learners to read more advanced, or totally unsimplified English texts. If this were true it would have serious consequences for ESOL teaching.

Swaffar (1985) also takes issue with the effectiveness of graded readers for learners. She says that the primary intent of an original text is to communicate meaning, and in her view, authentic messages which perform this function have the following characteristics: authorial cues, repetition, redundancy, and discourse markers. She claims that simplified texts do not have these characteristics because they have a "pseudo-intent", which is to teach language, rather than to communicate.

But, pedagogical intent aside, if a simplified text aims primarily to communicate meaning no less than the original, might it not also have the characteristics of authenticity? When comparing simplified and original texts, Honeyfield (1977) and Swaffar (1985) both seem to ignore that fact that the point of view of native speaker readers is necessarily going to be different from that of language learners. For a comparison to be a true one, such text characteristics as random distribution of high and low frequency words, repetition and redundancy must be measured, not in absolute terms, but in terms of how they are likely to be perceived by their readers.

It is therefore the aim of this study to compare simplified and original texts using Honeyfield's criterion of variation in word frequency distribution and Swaffar's characteristics of authentic messages - and to make the comparison in audience-specific terms. The research question is whether such a comparison will reveal similarities. If is does, it suggests that, for language learners, the simplifications are experienced as neither homogeneous or distorted, but in ways similar to those of native or near-native speakers reading the originals.

Methodology

This study is based on two original texts, and the graded reader simplified versions of each. The first original was an extract from Edgar Allan Poe's *The Gold Bug,* the text which Honeyfield used in his exposé of simplifications. The graded reader simplification was the Longmans version of the same passage. These texts will be referred to in the study as *Poe Original* and *Poe GR (Graded Reader*). The second original text was Arthur Conan Doyle's *The Adventure of the Speckled Band,* and the simplified version was the one produced by the Oxford Bookworms Library, Level 2. These texts will be referred to as *SB Original* and *SB GR.* All four texts were analysed according to the characteristics discussed above, divided into the two broad categories of language and content. Under the heading of language, the word frequencies were analysed for the variation that Honeyfield (1977) specifies; and the length of words and sentences, the collocations, repetitions, redundancies and discourse markers – Swaffar's (1985) characteristics of authentic messages, were noted and tabulated. Under the heading of content, the environment or topic, the author's intent and the information given were all examined, these also being included in Swaffar's characteristics.

In order to determine the distribution of frequency of words at a given level of English, all the words in a text have to be categorized according to frequency, and to do this, a computer programme called RANGE was used.

The RANGE programme was developed by Paul Nation and Alex Heatley of Victoria University, Wellington. It can apply three distinct word lists, called Base Lists, to any text, and can sort the text vocabulary into three categories of headwords from each list, and a category of words outside all three lists, making four categories altogether. Headwords are defined here as the chief words in each word family, or group of words, coming from the same root, through not necessarily the same part of speech. RANGE can do this either by range across several texts, or by frequency within a text. It can also mark each word according to the category in which it falls. The Base Lists can be altered depending on requirements. The ones which come with the programme are the first and second thousand words from West's General Service List (West, 1953), referred to from now on as the GSL, and Averil Coxhead's Academic Word List, referred to as the AWL.

In this study all four texts are analysed using the GSL 1st and 2nd 1000 words, and the AWL, as the three Base Lists, to show the percentage and distribution of high and low frequency of words from the GSL and AWL in the texts. However, because the GSL was compiled using corpora of English written and read by native speakers, high frequency words are measured as frequent in the experience of a native speaker. It should also be noted that West (1953) used criteria other than frequency, such as coverage, learning burden, necessity, and stylistics. However, for a learner who routinely encounters words only within the first 1000 words of the GSL, the definition of high frequency is not going to be the same as West's. So, in order to get an idea of what a learner at the elementary or low intermediate level of English finds familiar, an analysis of the same texts was also made using three different Base Lists.

These Base Lists were the Oxford Bookworm Library (OBL) lists, which are based on the headwords used in their readers at Level 1 (400 words), Level 2 (400-700 words), and Level 3 (700-1000 words). The selection of headwords and their order of frequency does not correspond exactly with that in West's GSL, as presumably the only OBL source material is from books which are published in this series, although there is a degree of overlap.

This has produced an interesting discrepancy between the words learners might generally be expected to know at a certain level in their English learning, and the words that learners who are following the OBL graded reader course in particular, will have encountered at this level.

Roughly speaking, for learners embarking on Level 2, all the words outside the OBL first 400 words will probably be unfamiliar, although they may know a number of the 1st 1000 words of the GSL which do not come into this category. Any words over the 700 OBL level are likely to be unknown, but there will be several degrees of familiarity within their band of known vocabulary. The level of familiarity has been taken as being in inverse proportion to the intensity of impact a word will have upon a reader. This notion of intensity is presented schematically in Table 1.

Table 2 is included to show the percentages of words in the GSL and AWL that readers at a particular level might be expected to know. It is noticeable that the proportion of words in the 1st 1000 word list, for example, drops from 76.6% in Level 1 to 22.1% in Level 6, whereas the percentage in the AWL at Level 1 is a tiny 0.3%, which rises to 21.2% in Level 6. This may account for the absence of words from Base List Three in the analyses using the GSL and AWL Base Lists, which will be discussed in the results section.

It is also noticeable that the proportion of words outside the lists increases as the levels ascend, so apparently even for an excellent learner there will be a number of challenging, and even unknown words at Level 6. This can be contrasted with the experience of the native speaker, who will probably know all of these words even though he may not use them on a regular basis, and for whom any words which occur in the 1st 2000 GSL lists will be high frequency, not particularly unusual.

Table 1: Categories of 'Intensity' for Level 2 Readers Reading a Text With 700 Headwords, Ranging from 1: Familiar to 8: Very Unusual

1	**High frequency words from the 1st 1000 words in the GSL, known from Level 1**
2	**High frequency words from the 2nd 1000 words in the GSL, known from Level 1**
3	**High frequency words from the 1st 1000 words in the GSL, UNKNOWN from Level 1 but possibly encountered elsewhere.**
4	**High frequency words from the 2nd 1000 words in the GSL, UNKNOWN from Level 1 but possibly encountered elsewhere.**
5	**AWL words known from Level 1**
6	**Words from outside the GSL and AWL but KNOWN from Level 1**
7	**AWL words UNKNOWN from Level 1 and unlikely to have been encountered elsewhere.**
8	**Words from outside any of the lists and unlikely to have been encountered elsewhere.**

Table 2: Table Showing the Distribution of High and Low Frequency Words, Words in Coxhead's Academic Word List and Words Outside Those Lists Over the Six Levels of The Oxford Bookworms Library Graded Readers (Level 2 Is Highlighted)

Level in Oxford Bookworms	GSL:1st 1000 words	GSL:2nd 1000 words	Academic Word List	Words outside the lists
1: 400 headwords	76.6%	14.9%	0.3%	8.2%
2: 700 headwords	65.6%	26.2%	0.6%	7.6%
3: 1000 headwords	49.9%	39.4%	1.4%	9.4%
4: 1400 headwords	41.2%	45.3%	3.8%	9.7%
5: 1800 headwords	25.0%	44.7%	11.3%	19.0%
6: 2500 headwords	22.1%	35.2%	21.2%	21.4%

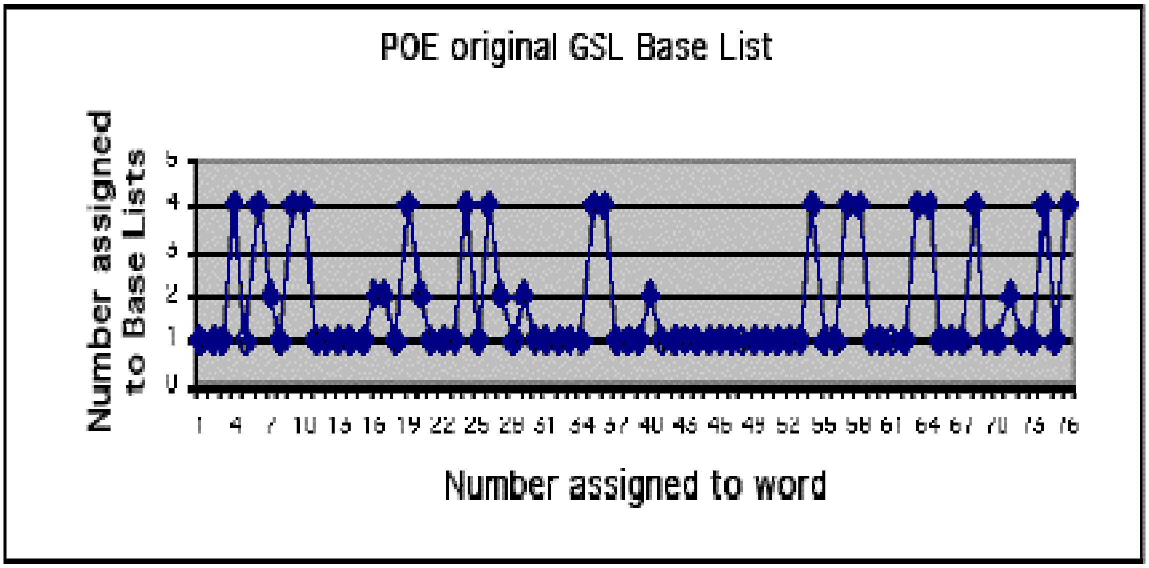
Note: each row in Table 2 gives the percentage of the total vocabulary known at each level, not just the particular level indicated.

Results: *The Gold Bug*

Language simplification

POE Original. The graph, (Figure 1), derived from the RANGE analysis of the text, shows a random distribution of high and low frequency words. It is notable that there are no results for Base List 3, which here represents the AWL, despite the fact that this is a piece of original, literary prose. But Coxhead (2000) finds that in general the AWL has a very low coverage of fiction, so this is not surprising.

Figure 1: POE Original



POE GR. Figure 2 represents the simplification of the Poe text analysed using the GSL/AWL Base Lists. This shows a much flatter profile, with most of the text within the 1st 1000 words, only a few in the 2nd 1000, none in the AWL (not surprisingly, as the original had none either), and only two in the list of unknown words. One of these is Charleston, a proper noun.

Figure 2: POE GR

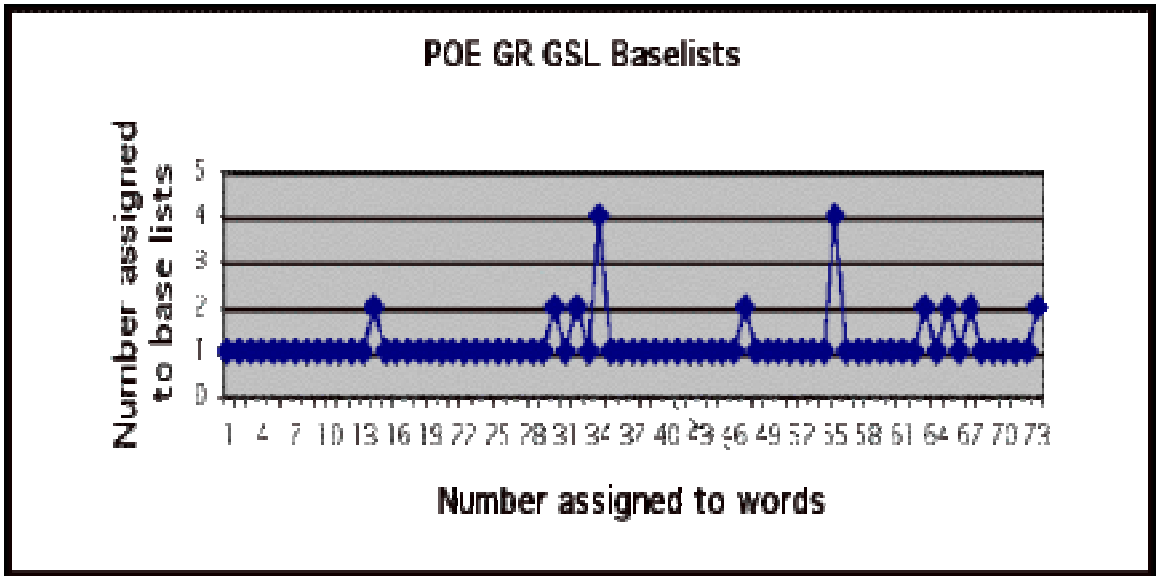
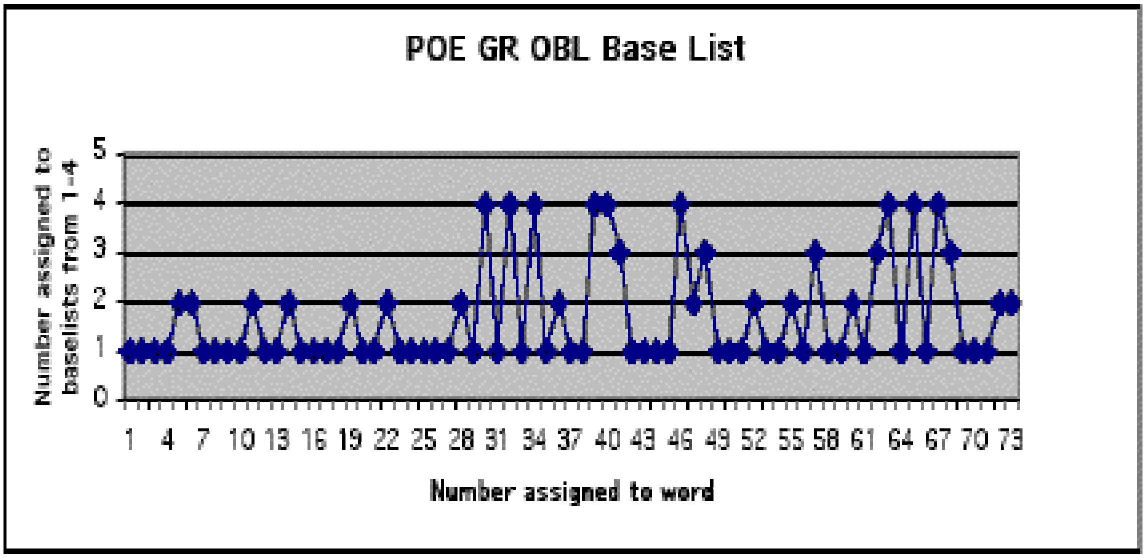


Figure 3 represents the same simplification analysed using the OBL Base Lists. This, by contrast with Figure 2, shows a profile which represents a much more random distribution of word frequencies, more similar to the analysis of the original text in Figure 1. Here, all the Base Lists are represented, showing the random distribution of words from Levels One, Two and Three of the OBL, and words which were outside the lists too.

Figure 3: OBL Base Lists



Other aspects of language simplification are represented in tabular form in Table 3. It is clear that the simplification is only a little shorter than the original, but some of the content is missing from it. The length must have been augmented by paraphrasing to simplify vocabulary, as for example, in *fugitives*, which become *those who wish to escape.* The entire text in the original is contained in one sentence, whereas it has been divided into five in the simplification. So, it is possible to see that in the simplified version, the structure has been altered to reduce the number of subordinate clauses and to make the main verbs easier to identify. It is true that in the original the main verb is very hard to identify, for the simplifier has missed it. In the original the first main clause is *Near the western extremity…may be found, indeed, the bristly palmetto.* The second clause is, *the whole island is covered with dense undergrowth….* Therefore it is a mistake in the simplification to give the primary importance to the few small buildings near the west end of the island. The point of the passage is that the island is covered with myrtle. The clue to the original meaning is provided by the discourse marker *but* in the fourth line. The same discourse marker is used twice in the simplification, in order to distinguish the passage about the buildings from that about the bushes. The *bushes* are repeated in the simplification, which is an example of redundancy, although there is more redundancy in the original.

The word length in the original averages 1.48 syllables, whereas in the simplification it is 1.23, showing that simpler, higher frequency words have been substituted for longer ones. Undergrowth has been traded for low bush, and miserable frame buildings have been left out in favour of small buildings of wood. Collocations, which Honeyfield and Swaffar regard as indications of authentic language, appear in the original, examples being *dense undergrowth, much prized, western* (or other points of the compass) *extremity*. These words are often found together, but there are examples of simpler collocations in the simplified version too, for instance, *west end.*

Content simplification

In Table 3 it is shown that the environment, the author's intent and the information have been significantly altered in the simplification. The myrtle has somehow acquired the epithet 'ugly' which was not in the original, and the author's intent, which was chiefly to describe the island as being covered with myrtle, has been altered in the simplification to describing the habitation as well as the vegetation.

Table 3: POE Original and GR

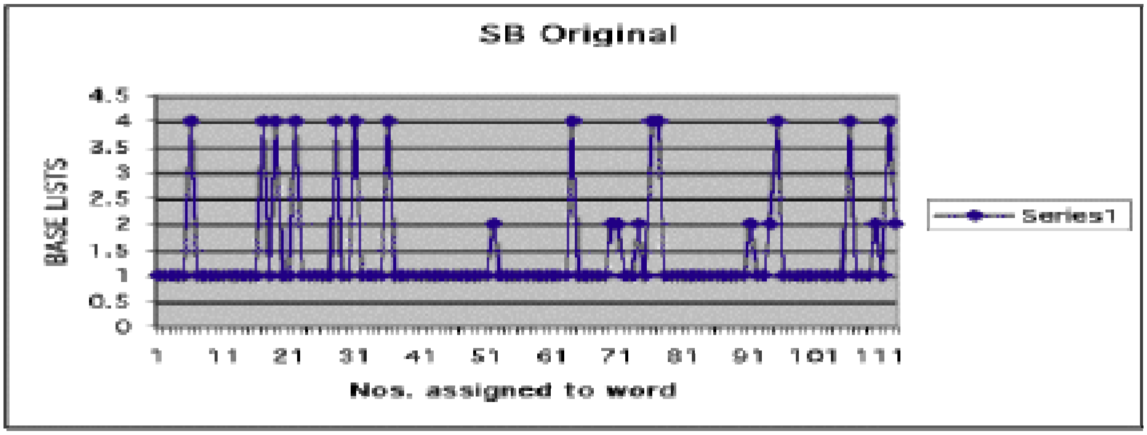
Linguistic Features	**Original Text**	**GR**
Words	76 WORDS 71.9% in 1st 2000 17 outside lists of which 4 proper nouns	72 WORDS 97.95% in 1st 2000 (close to AC1)
Structure	Sentence length 76 words in 1 sentence use of subordinate clauses and ; Av word length: 1.48 syllables	5 sentences averaging 14.4 words. Simpler sentences. Av word length 1.23 syllables
Discourse markers	But (the whole island) With the exception of	However, but, but
Redundancy	Where/where Bristly/palmetto Beach/on the seacoast	Small buildings/they Bushes/bushes
Collocations	Western extremity Dense undergrowth Much prized	West end
CONTENT Environment/topic	Island vegetation	Vegetation Habitations
Author's intent	To convey fact that island is covered in myrtle	To explain the habitation
Information	1. Island covered in myrtle 2. Some palmetto 3. Wooden habitations at west end of island	1. Ugly myrtle 2. Habitations

Results: *The Adventure of the Speckled Band*

Language

SB Original. Figure 4 shows an analysis of the original *Speckled Band* text, using the GSL/AWL Base Lists. This reveals a random distribution of word frequencies, with a number of leaps from high frequency words to very low frequency words. There are also a number of high frequency words in the 2nd 1000 word list, but, as in the Poe Original, there are no words from the AWL.

Figure 4: SB Original



SB GR. Figure 5 represents the simplified version of the *Speckled Band* analysed using the GSL/AWL Base Lists. It shows a much flatter profile than Figure 4 with, again, no results for Level 3.

Figure 5: SB GR

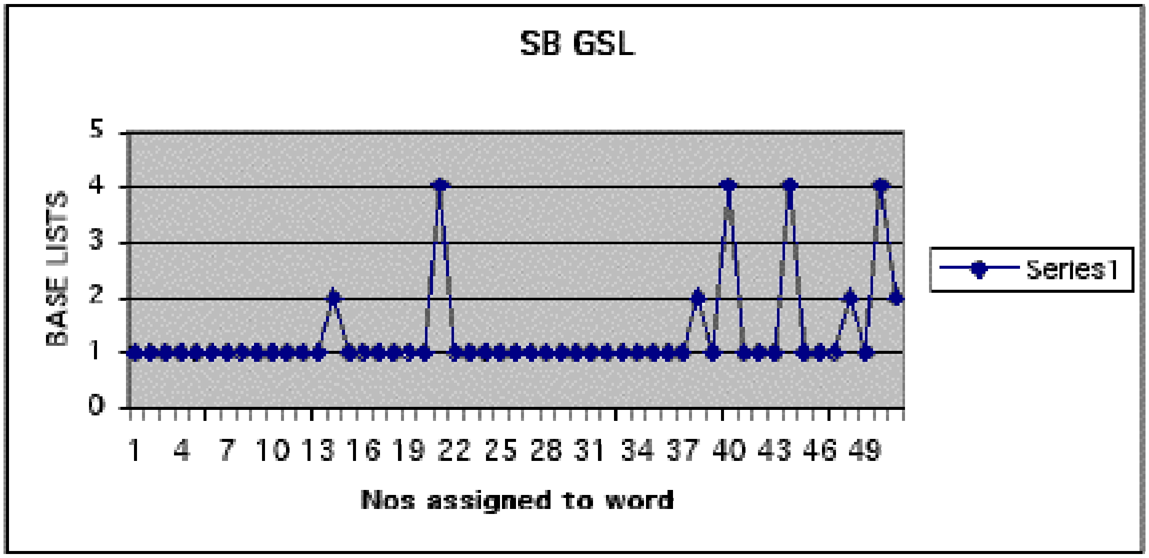
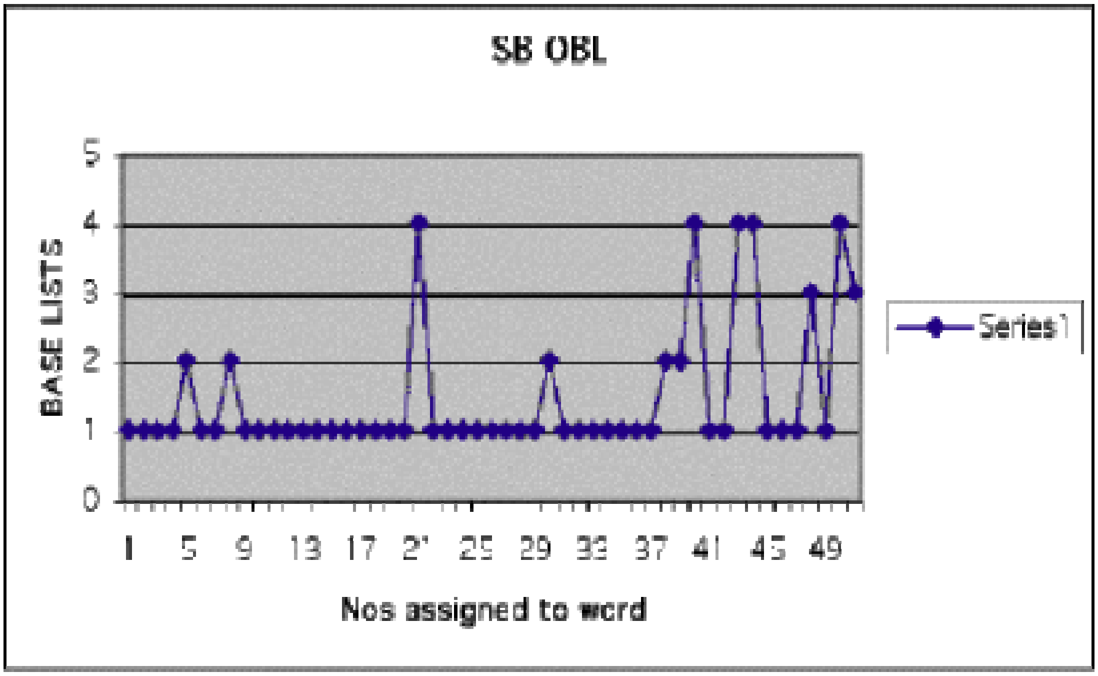


Figure 6 represents the same version analysed using the OBL lists. Here there is more variation than in Figure 5, though there is not as much as in the equivalent graph for the Poe text in Figure 3 and Base List 3 is represented in Figure 3. One reason for the difference between Figure 4 and both Figures 5 and 6, is that the simplification process has shortened the text considerably, which has had the effect of reducing the possibilities for frequency variation.

Figure 6: SB OBL



Other aspects of language simplification are represented in Table 4. The original text contains 9929 words and has been shortened in the OBL version to 2199 words. It therefore contains more variation. In the original only 66% of the text is composed of high frequency words, whereas in the simplified version the percentage is 94.3. Table 4 also reveals that an effect of the simplification has been to shorten the sentences by about 40%. Instead of subordinate clauses, new sentences have been formed, and instead of a description by the writer "*she writhed as one who is in terrible pain, and her limbs were dreadfully convulsed*", the simplified version has given direct speech to the victim, who says, "*Help me, help me Helen! I'm ill, I'm dying*." The collocations are considerably reduced in the simplified version. There is redundancy in both, and there are also discourse markers, *but* and *and* in the SB original and only *and* in SB GR.

Table 4: *Speckled Band* Analysis

LINGUISTIC Features	**TEXT SB: original**	**TEXT SB GR: Analysed using GSL and AWL**	**TEXT SB GR: Analysed using the OBL level word lists**
Words	95 WORDS (counted as types) 74.7% in the 1st 1000 GSL 8.9% in the 2nd 1000 GSL 16.5% not in lists Average word length 1.49 syllables (difference of 10%)	46 WORDS 85.7% in the 1st 1000 8.6% in the 2nd 1000 5.7% not in lists (but these are Helen, a proper noun, and Speckled, part of the title) Average word length: 1.39 syllables	Level 1: 74.3% Level 2: 14.3% Level 3: 2.9% Outside lists: 8.6% (but these are Helen, God and speckled)
Structure	4 sentences: average sentence length 23.75 words. First sentence = a main clause with 3 adjectival phrases	3 sentences: average sentence length 15.33 words. 1st sentence simple: 1 subject, 2 verbs. Description simplified by direct speech not in original.	
Collocations	• By the light of the lamp • Threw my arms around her • In a voice which I shall never forget Oh my God.	Oh my God	
Discourse markers	• But • At first …but	• And	
Redundancy	• her, her… • The band! The speckled band!	• Help me! • I'm ill/dying • The band! The speckled band!	
CONTENT:	The dying throes of Julia	Same	
Author's intent	A graphic picture of a horrible death. A clue embedded in the text as to the cause.	Same	

Information	Description of Julia: • Face blanched with terror • Groping for help • Swaying • Writhed • Convulsed	Description of Julia: • white and afraid • I'm ill! • I'm dying!	

Content

The information given in SB Original is more detailed than in SB GR, but is essentially the same. The content is the dying throes of Julia, the author's intent is to paint a graphic picture of a horrible death, with a clue as to the cause embedded in the description. The information does not differ over the two versions.

Discussion

The research question of this study is whether, when compared in terms of their respective audiences, original and simplified texts reveal similar patterns of variation of word frequency distribution, and of other characteristics of authentic messages. The results will be discussed below with reference first to the Poe texts, and then to the *Speckled Band* texts.

POE texts. The graphic representation of the Poe original text, Figure 1, using the GSL/AWL Base Lists, shows a random distribution of high and low frequency words, and this, if we accept that Poe is a master of the written language, can be taken as a model for frequency distribution in 'normal' texts. By contrast, the simplification shown in Figure 2 analysed according to the same GSL/AWL Base Lists, looks homogeneous. But the same simplification analysed using the OBL Base Lists in Figure 3 shows a good deal of variety, and can almost compete with the master's original in frequency distribution. By using RANGE, it has been possible to compare the word frequency distribution for each text.

The word count is almost the same in the original and the simplification, but difficult words have been paraphrased. This might have extended the passage far beyond the length of the original if the simplifier had not omitted some of the content. One of the criteria of a good simplification is that it communicates meaning. Even if a simplified version communicates meaning, if it does not communicate the same meaning as the original, it cannot be called a true simplification. Therefore if it is not faithful to the content of the original, it fails on that count. We can therefore agree with Honeyfield (1977) that it is not a good simplification of the original, but we cannot agree with his judgment that it is flat and homogeneous, because, set against his own criterion of word frequency distribution, and Swaffar's other characteristics of authentic messages, it passes every test. It contains authorial cues in its collocations, it has repetition and redundancy, and it has discourse markers.

SB texts. As the SB simplification is very much shorter than the original, the profile of word frequency distribution is likely to be less varied than that of the original, and this should be taken into account here. This is a reminder that in shortening to simplify we do sometimes lose the opportunity, as in this case, to build up a climate of suspense over several paragraphs, which culminates in a dramatic event. But an unintelligible dramatic build up which loses the learner is worse than a shorter, understandable version. In SB original, only 66% of the text is composed of high frequency words, whereas in the simplified version the percentage is 94.3% This comes close to Paul Nation's prescription (Nation, 2001: 150) that "extensive reading texts should contain no more than 5% unknown tokens (excluding proper nouns) and preferably no more than 2.5% to ensure that comprehension and guessing can occur, and no less than 1-2%, to learn." The effect of the simplification has been to reduce the proportion of unknown words (to our 1000 GSL headwords learner) to known, from 25.4% to 14.3%. The relative 'unusualness' of the words for the learner in the simplified version can be roughly assessed against the chart shown in Table 1. The new or relatively new words can be divided into the following categories of intensity:

2:	afraid
3:	voice, fell ground
4:	band, terrible
8:	Helen, speckled

This, plus the visual evidence of the graph in Figure 6, suggests that even at Level 2 of the OBL, it is possible to write simplified prose that displays a considerable level of variation to the reader, retaining not only the characteristics of 'normal' English, but also the content. By using RANGE, it has been possible to show the variations in the word frequency levels. Even the GSL analysis of the simplified version (Figure 5) has some variation, and interestingly, one word, *afraid*, appears in the 2nd 1000 words of the GSL, but not in the first 400 of the OBL lists, evidence that the overlapping of the lists provides a number of variations in familiarity with words at given levels, as shown in Table 1. The collocations in the original are not particularly unusual. However the clause, *she shrieked out in a voice, which I shall never forget,* contains the word *shrieked* , which is outside the lists in the GSL. This is rendered in the simplified version *by she cried out in a terrible voice.* Here it is the word *terrible*, which is more unusual than the others, coming from the 2nd OBL Base Lists, rather than the 1st. *Oh my god* appears in both texts, perhaps reflecting the fact that the intensity of this phrase has diminished since the original text was written. The discourse markers in the simplified version are both "*and*", compared with the original which contains *but*, and "*at first*". However, the simplicity of the structure is compensated for by the repetition and redundancy in the simplified version.

The final phrase in which the dying woman attempts to tell her sister the cause of her death has not been altered in the simplified version. The word *band* is in the 2nd 1000 word list of the GSL, but *speckled* does not occur in the lists. However, perhaps because the name of the deadly snake is germane to the story, the writers of the simplified version have decided to include it unedited. Their critics cannot say that in this case they have rendered the text unauthentic by failing to give the readers the opportunity to play "psycho-linguistic guessing games with unknown words" (Swaffar, 1985: 18).

The information given in the original text is certainly more detailed than in the simplified version. The phrases *her face blanched with terror, her hand groping for help, her whole figure swaying to and fro like that of a drunkard*, together with *her knees seemed to give way and she fell to the ground* are combined into two sentences: *She… fell to the ground. Her face was white and afraid.* But despite the reduction in detail the main points are clearly conveyed, even with a certain amount of redundancy, a mark of an authentic message, and the atmosphere of panic is conveyed by the use of direct speech.

Conclusion

The sample text from the OBL version of *The Adventure of the Speckled Band* differs from the one chosen by Honeyfield (1977), in that it has been skillfully simplified, preserving the variety in word frequency distribution without prejudicing the content. Honeyfield objected to graded readers on the basis of their blandness and homogeneity, without taking into account that what was bland for him might not strike the learner in the same way. However, the RANGE analysis of the simplified version is able to indicate that, for a learner at level 2 of the OBL graded readers, the word frequency distribution is nearly as varied, at that level, as that in the original. The preservation of the essential features of a 'normal' English text in the simplification has been shown in the areas of language and content. It seems fair to conclude that well-written graded readers can offer an authentic reading experience for learners, which will help prepare them for reading unsimplified texts.

References

Conan Doyle, A. (Ed.). (1961). *The complete Sherlock Holmes*. Edinburgh: Murray.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-239.

Honeyfield, J. (1977). Simplification. *TESOL Quarterly*, 11(4), 431-440.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. & Heatley, A. (2003). The RANGE computer programme. Available at http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P. & Wang Ming-tsu, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, 12(2), 355-380.

Poe, E. (1967). *The gold bug: The selected writings of Edgar Allan Poe.* New York: Penguin Books.

Swaffar, J. (1985). Reading authentic texts in a foreign language. *The Modern Language Journal*, 69, 115-134.

West, M. (1953). *A general service list of English words*. London: Longman.

Widdowson, H. (1976). The authenticity of language data. J. F. Fanselow & R. H. Crymes (Eds.), *ON TESOL* '76. Washington: TESOL.

About the Author

Gillian Claridge taught ESOL for many years in England before coming to New Zealand to teach English on the Foundation Programme at the International Pacific College. Her first degree is in Russian language and literature and she has an MA in Applied Linguistics from Victoria University, Wellington. She is also a qualified French teacher.