Towards a Better Exploitation of the Brown 'Family' Corpora in Diachronic Studies of British and American English Language Varieties

Sanja Štajner
University of Wolverhampton, UK
S.Stajner@wlv.ac.uk

Abstract

Since the 1990s, the Brown 'family' corpora have been widely used for various diachronic studies of 20th century English language. However, the existing methodologies failed to exploit its full potential as they only used the four main text categories. In this paper, we present the results of two experiments on diachronic changes of the Coleman-Liau readability Index (CLI) in British and American English in the period 1961–1991/2. The first experiment used all fifteen fine-grained text genres, while the second only used the four main text categories. The comparison of the results of these two experiments demonstrated the importance of using all fifteen fine-grained text genres for obtaining a better understanding of how language changes.

1 Introduction

The Brown University corpus of written American English[1] was published in 1964 with the aim of standardising the future parallel corpora of British English or American English of other periods (Francis, 1965 in Leech and Smith, 2005). Following this idea, the LOB corpus[2] of written British English was compiled as the first corpus to match the Brown corpus, in respect of the year of sampling (1961) and its representation of different text types (Leech and Smith, 2005). This provided the possibility for a synchronic comparison between two major English language varieties – British and American. In the 1990s, the emergence of the FLOB[3] and Frown[4] corpora, representing written British English in 1991 and American English in 1992, respectively, added a diachronic component. It created the opportunity to use the Brown 'family' corpora in diachronic studies of 20th century written English texts in these two regional language varieties. As they are publicly available as part of the ICAME Corpus Collection[5] and cover fifteen different text genres over the four main text categories (Press, Prose, Learned and Fiction), the Brown 'family' corpora have been widely used in various diachronic studies throughout the linguistic community.

Readability formulas provide assistance to a writer in producing comprehensible text and maintaining a consistent reading level throughout a document (McCallum and Peterson, 1982). They were initially designed for educational purposes with the aim of defining the appropriate reading levels for primary and secondary school text books (McCallum and Peterson, 1982). The most commonly used variables in a readability formula are the measures of sentence and word difficulty (Klare 1968; 1974 in McCallum and Peterson, 1982). In a survey on the most commonly used readability formulas of that period (McCallum and Peterson, 1982), special attention was given to the Coleman-Liau Index (Coleman and Liau, 1975) and Automated Readability Index (Smith and Kincaid, 1970), as these formulas are simpler to compute. Unlike most of readability formulas which use the number of syllables per word, these two formulas use the number of characters per word as a measure of word difficulty. Therefore, we decided to use one of these – Coleman-Liau Index, as a measure of text readability. The result of this formula is the U.S. grade level necessary to comprehend the given text.

The primary focus of this study was to highlight possibly misleading interpretations of the results in diachronic studies when using only the four main text categories, instead of all fifteen fine-grained text genres of the Brown 'family' corpora. In order to achieve this, we conducted two experi-

[1] http://khnt.aksis.uib.no/icame/manuals/brown
[2] http://khnt.aksis.uib.no/icame/manuals/lob
[3] http://khnt.aksis.uib.no/icame/manuals/flob
[4] http://khnt.aksis.uib.no/icame/manuals/frown
[5] http://www.hit.uib.no/icame

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 17–24,
Hissar, Bulgaria, 13 September 2011.

ments – first using all fifteen fine-grained text genres and then using only the four main text categories (Section 4). Both experiments were based on the investigation of diachronic changes of the Coleman-Liau Index in British and American English in the period 1961–1991/2. The results of those experiments are compared in Section 5. The main conclusions and suggestions for future exploitation of the Brown 'family' corpora in diachronic studies are given in Section 6.

2 Related Work

The four corpora – LOB, FLOB, Brown and Frown, were used for investigating the trends of change in various lexical, grammatical and syntactic features by Mair and Hundt (1995), Mair (1997; 2002), Mair, Hundt, Leech and Smith (2002), Smith (2002; 2003a; 2003b), Leech (2003; 2004), Leech and Smith (2006), Mair and Leech (2006). Mair, Hundt, Leech and Smith (2002) demonstrated the possibilities of these corpora in the investigation of diachronic changes of POS frequencies. Leech and Smith (2006) and Mair and Leech (2006) further exploited the corpora by investigating diachronic changes of core modals, semi-modals, passive, *wh-* and *that* relativisation, personal pronouns, nouns, *of-* and *s-* genitive constructions. More recent studies (Leech and Smith, 2009; Leech, Mair, Hundt and Smith, 2009) expanded the time-span for diachronic studies in British English by using the Lancaster1931 corpus together with the LOB and FLOB corpora.

All these studies shared the same methodology:

(1) They used the POS tagged versions of the Brown and Frown corpora and the manually post-edited versions of the LOB and FLOB corpora.

(2) The experiments were conducted first on the whole corpora and later separately on each of the four major subdivisions of the corpora: Press, General Prose, Learned and Fiction.

(3) The log likelihood test was applied as a measure of statistical significance of the results.

Although the Brown 'family' corpora provided an opportunity for separate investigation of diachronic trends across all fifteen fine-grained text genres, none of the above mentioned diachronic studies utilised this trait. They only differentiated between texts across the four main text categories and made the hypotheses about the trends of language change accordingly.

There are numerous readability studies of different texts genres and comparisons among them. However, to our best knowledge, there have been no diachronic studies of text readability. For this reason, we decided to use the Coleman-Liau Index as an initial experiment to trace the diachronic changes of text readability in 20th century English language.

3 Corpora

The Brown 'family' corpora is comprised of two corpora of American English:

- The Brown University corpus of written American English (**Brown**)
- The Freiburg - Brown Corpus of American English (**Frown**),

and two corpora of British English:

- The Lancaster-Oslo/Bergen Corpus (**LOB**)
- The Freiburg-LOB Corpus of British English (**FLOB**),

while the fifth corpus to join the 'family' – Lancaster1931 (BLOB) is still not publicly available.

All five corpora are mutually comparable (Leech and Smith, 2005) and contain texts published in the years 1931 ± 3 (Lancaster1931), 1961 (LOB and Brown), 1991 (FLOB) and 1992 (Frown). Each corpus consists of approximately one million words – 500 texts of about 2000 running words each, selected at a random point in the original source. The sampling range covers 15 text genres, which can be grouped into four more generalised categories:

- **Press**
  - Press: Reportage (A)
  - Press: Editorial (B)
  - Press: Review (C)
- **General Prose**
  - Religion (D)
  - Skills, Trades and Hobbies (E)
  - Popular Lore (F)
  - Belles Lettres, Biographies, Essays (G)
- **Learned**
  - Miscellaneous (H)
  - Science (J)

- **Fiction**
  - General Fiction (K)
  - Mystery and Detective Fiction (L)
  - Science Fiction (M)
  - Adventure and Western (N)
  - Romance and Love Story (P)
  - Humour (R)

The distribution of the texts for each genre and corpus is given in Table 1. As the LOB and FLOB corpora share exactly the same text distribution across all fifteen genres, they are presented in the same column – '(F)LOB'.

Genre	(F)LOB	Brown	Frown
A	44	44	44
B	27	27	27
C	17	17	17
D	17	17	17
E	38	36	36
F	44	48	48
G	77	75	75
H	30	35	30
J	80	80	80
K	29	29	29
L	24	24	24
M	6	6	6
N	29	30	29
P	29	29	29
R	9	9	9

Table 1: Text distribution in the corpora.

It can be noticed that the number of texts varies significantly among genres belonging to the same broad text categories. For example, in the Press category (A–C), the number of texts in each of the genres A, B and C is 44, 27 and 17, respectively. Therefore, it is reasonable to expect that the trend of change in genre A (Press: Reportage) will have the greatest impact on the overall trend of change in the whole Press category. This could lead to a failure to observe the changes present in some of the genres with a smaller number of texts. More importantly, different directions of changes (increase and decrease) in two genres of the same broad text category might lead to the overall perception of no change in that category. Neglecting the changes present in those genres could result in misleading conclusions and hypotheses regarding the way language changes.

4 Methodology

As the primary focus of this study was to compare the conclusions which can be drawn from the results obtained from two different approaches, we conducted two separate experiments:

- **Experiment I** – Investigation of diachronic changes of CLI in the period 1961–1991/2 across all fifteen text genres (A–R)
- **Experiment II** – Investigation of diachronic changes of CLI in the period 1961–1991/2 across the four main text categories (Press, Prose, Learned and Fiction)

Both experiments were conducted separately for each of the English language varieties (British and American), using the Brown 'family' corpora (LOB, FLOB, Brown and Frown).

4.1 Sentence Splitting and Tokenisation

The Brown and LOB corpora are available in their POS tagged and tokenised versions with sentence boundaries, while the Frown and FLOB corpora do not contain markers for sentence and word boundaries. In order to achieve a higher consistency for sentence splitting and tokenisation and offer a fairer comparison of the results among the corpora, we used the raw text versions of all four corpora and parsed them with the state-of-the-art Connexor's Machinese Syntax parser[6].

The parser tokenises contractions and hyphenated words in the following manner: the verb and its negation (e.g. *isn't*) are treated as two separate tokens (*is* and *not*), while *'s* is treated in two different ways, depending on its role in the sentence. In cases where *'s* represent a genitive form, *'s* and its antecedent noun are treated as one token. In other cases where *'s* represent a contracted form of the verb *be (is)* or *have (has)*, *'s* is treated as a separate token. E.g. In the sentence *"That's a Tory doctor's reaction to the new health charges..."* (LOB: A01), *That's* is treated as two separate tokens – *that* and *is*, while the *doctor's* is treated as one token *doctor's*. Hyphenated words, e.g. *30-year-old, built-in, type-recorder* (LOB: A10) are treated as one token. All punctuation marks are treated as separate tokens.

4.2 Feature Extraction

The Coleman-Liau Index (CLI) was calculated separately for each of the 500 texts in each of the

[6] www.connexor.eu

corpora, using the following formula:

$$CLI = 5.89\frac{c}{w} - 29.5\frac{s}{w} - 15.8 \qquad (1)$$

where *c, w* and *s* represent, respectively, the total number of characters, words and sentences in the text. The number of characters, words and sentences were calculated using the parser's output. Sentences were counted as the number of sentence tags (<s>) in the parser's output, words – as the number of word tags (<text>) excluding those which contained only punctuation marks, and characters – as the number of characters inside the word tags counted as words.

4.3 Statistical Significance

First we examined whether the data follow the normal distribution, using the Kolmogorov-Smirnov Z test. The results of this test demonstrated that the distribution of the CLI is not significantly different from the normal distribution (at a 0.05 level of significance), in each language variety, year, category and genre. Therefore, we used the two-tailed t-test as a measure of statistical significance of the change.

5 Results and Discussion

The results of the experiments on diachronic changes of CLI are given separately for British and American English in Sub-sections 5.1 and 5.2, respectively. Trends of change are compared between these two language varieties in Sub-section 5.3.

Table 2 (Sub-section 5.1) presents the results of the two experiments for British English, while Table 3 (Sub-section 5.2) presents the results of the same experiments for American English. The tables contain the results of both experiments in two consecutive columns – 'Exp. I' and 'Exp. II', thus enabling their direct comparison. For each of the experiments, results are presented in two columns – 'change' and 'p'.

Column 'change' presents the absolute change of CLI over the period 1961-1991/2. Both – starting (1961) and ending (1991/2) values were calculated as an arithmetic mean of the feature value in all texts of the relevant text genre/category and corpus. The direction of change is indicated by the sign '+' for increase and '−' for decrease.

Column 'p' represents the p-value of the two-tailed t-test. Statistically significant changes at a 0.05 level of significance ($p < 0.05$) are shown in bold.

5.1 Diachronic Changes of CLI in British English

The results of the experiments on diachronic changes of the Coleman-Liau Index (CLI) in British English are presented in Table 2.

Genre	Exp. I		Exp. II	
	change	p	change	p
A	+0.54	0.063		
B	+0.09	0.762	**+0.44**	**0.038**
C	+0.74	0.061		
D	**+2.35**	**0.001**		
E	**+1.04**	**0.002**	**+1.21**	**0.000**
F	**+1.26**	**0.002**		
G	**+1.01**	**0.000**		
H	**+1.10**	**0.009**	**+1.35**	**0.000**
J	**+1.44**	**0.000**		
K	−0.49	0.210		
L	−0.25	0.573		
M	+0.01	0.994	+0.19	0.143
N	**+1.17**	**0.006**		
P	+0.52	0.072		
R	−0.62	0.267		

Table 2: CLI in British English (1961–1991).

On the basis of the results of the second experiment (Exp. II, Table 2), it could be concluded that the change of CLI in the period 1961–1991 were significant in the Press, Prose and Learned text categories and not in the Fiction category. However, the results of the first experiment (Exp. I, Table 2) lead to different conclusions regarding the trend of change of CLI in the Press and Fiction text categories. In the Fiction category, the results of the first experiment (Exp. I, Table 2) indicate a statistically significant change of CLI in genre N (Adventure and Western). This change was not reflected in the second experiment (Exp. II, Table 2) probably due to the following two reasons: (1) a high heterogeneity of the results in the category, i.e. different directions of change among genres belonging to this text category (genres K–R, Exp. I, Table 2) and (2) unbalanced distribution of texts among the genres inside this text category (genres K–R, Table 1, Section 3). In the Press category, the results of the first experiment (Exp. I, Table 2) indicate that the changes of CLI in the period 1961–1991 were not statistically significant in any of the three genres (A–C) inside this category. It is interesting to notice that the p-value of the t-test in

genres A and C (0.063 and 0.061, respectively) is very close to the chosen critical value (0.05). Most probably, the results of the second experiment in the Press category (Exp. II, Table 2) reflect the cumulative effect of those changes in genres A and C, which were not reported as statistically significant in the first experiment (Exp. I, Table 2).

Furthermore, the results of the first experiment in the Prose and Fiction categories (Exp. I, Table 2) revealed two interesting phenomena, that the genres inside the same broad text category manifest: (1) different trends of change (genres K–R in the Fiction category) and (2) different intensities of change (genres D–G in the Prose category). In the Fiction category, CLI had a statistically significant increase in genre N (Adventure and Western), in genre M (Science Fiction) CLI stayed unchanged ($p > 0.99$), while in genres K (General Fiction) and R (Humour) the results indicated a possible decrease of CLI during the same period 1961–1991. The high heterogeneity of the results among different genres in the Fiction category raises a question: "is it possible to talk about a general trend of change in a text category if different genres inside that text category manifest different trends of change?" In the Prose category, CLI had a statistically significant increase over the observed period in all four genres (D–F), but the intensity of the increase was significantly higher in genre D (+2.35) than in the other three genres (+1.04, +1.26 and +1.01). These two phenomena, though important for obtaining a better understanding of the way text readability changes in British English, could be overlooked by using only the results of the second experiment (Exp. II, Table 2).

A general conclusion based on the results of the first experiment is that all genres which manifested a statistically significant change of CLI (genres D–J and N) had the same direction of change – an increase. This could be interpreted as a tendency in these genres to make texts more complex, using longer words and sentences.

5.2 Diachronic Changes of CLI in American English

The results of both experiments investigating diachronic changes of the Coleman-Liau Index (CLI) in American English are presented in Table 3.

Similarly as in the case of British English (Sub-section 5.1), the results of the second experiment in American English (Exp. II, Table 3) could lead to potentially incorrect conclusions regarding the change of CLI in the Press, Prose and Learned text categories. They indicate that in the Press category there had been no statistically significant changes of CLI in the observed period, while the results of the first experiment (Exp. I, Table 3) clearly demonstrate a statistically significant increase of CLI in one genre of this category – genre B (Press: Editorial). This result, important for obtaining a better understanding of the way text readability was changing in the Press category of American English, could be overlooked by using only the results of the second experiment (Exp. II, Table 3).

Genre	Exp. I		Exp. II	
	change	p	change	p
A	+0.22	0.501		
B	**+0.71**	**0.049**	+0.36	0.093
C	+0.19	0.506		
D	**+2.01**	**0.015**		
E	+0.31	0.558		
F	**+1.46**	**0.000**	**+0.99**	**0.000**
G	**+0.77**	**0.013**		
H	+0.80	0.152		
J	**+1.05**	**0.001**	**+0.98**	**0.037**
K	−0.56	0.209		
L	+0.27	0.445		
M	−0.96	0.412		
N	+0.20	0.606	−0.31	0.280
P	−0.44	0.248		
R	−1.80	0.069		

Table 3: CLI in American English (1961–1992).

The difference between the conclusions made about the changes of CLI in the Prose and Learned category, based on the results of the first and second experiment, is more subtle. If we assume that the trend of change for a broad text category should correspond to the trend which is the most common among its genres, the results of the first and second experiment in the Prose category (Table 3) are consistent. However, the phenomenon that the genres inside the same broad text category manifest different intensities of change (already discussed in Sub-section 5.1) could be overlooked if we relied solely upon the results of the second experiment (Exp. II, Table 3). The results of the first experiment (Exp. I, Table 3) demonstrated that all three genres (D, F and G), which manifested a statistically significant increase of CLI in the Prose category, exhibited significantly differ-

ent intensities of that change (+2.01, +1.46 and +0.77, respectively).

The result of the second experiment (Exp. II, Table 3) suggests that CLI had a statistically significant increase in the Learned category over the period 1961–1992. However, the results of the first experiment (Exp. I, Table 3) demonstrate that CLI had a statistically significant increase only in genre J (Science), while the results of the t-test in genre H (Miscellaneous) do not allow us to be certain about the behaviour of CLI in this genre. As the Learned category is comprised of only these two genres (J and H), the result of the second experiment misleadingly creates the impression that the increase of CLI was present in the whole category. This result is probably a reflection of the unequal distribution of texts between these two genres – 80 texts in J genre and 30 (35) texts in H genre (Table 1, Section 3).

5.3 Comparison of Diachronic Changes between British and American English

The fact that the British and American part of the Brown 'family' corpora are mutually comparable (Leech and Smith, 2005) allows us to compare the trends of change between these two English language varieties in both experiments.

The results of both experiments (Table 2 and Table 3) lead to a central conclusion that all statistically significant changes of CLI in the period 1961–1991/2 had the same trend of change – an increase, in both English language varieties. However, those changes were not present in the same genres and text categories across the two language varieties. The differences are noticeable even at the level of the four main text categories, where CLI manifested a statistically significant increase in the Press category only in British (Exp. II, Table 2) and not American English (Exp. II, Table 3). The results of the first experiment revealed some additional differences in the behaviour of CLI between British and American English. Genre B (Press: Editorial) had a statistically significant increase only in American English (Exp. I, Table 3), while genres E (Skills, Trades and Hobbies), H (Miscellaneous) and N (Adventure and Western) had a statistically significant increase only in British English (Exp. I, Table 2).

The comparison of the results between British and American English in the Press category emphasised the importance of carefully choosing the granularity of genres in diachronic studies. The results of the first and second experiment in the Press category led to the opposite conclusions. The results of the second experiment suggested an increase of CLI only in British English (Exp. II, Table 2), while the results of the first experiment demonstrated an increase of CLI only in genre B of American English (Exp. I, Table 3) and no statistically significant changes of CLI in any of the three genres of the Press category in British English (Exp. I, Table 2).

6 Conclusions

The results presented in this study indicated that in all genres of the Prose and Learned text categories and one genre (N – Adventure and Western) of the Fiction category in British English, a tendency existed to render texts more complex, using longer words and sentences. Furthermore, the results demonstrated that different genres inside the same broad text category do not follow the same trend of change. In the Fiction category of British English, genre N (Adventure and Western) manifested a statistically significant increase of CLI between 1961 and 1991, while in genre M (Science Fiction) texts from both years – 1961 and 1991 had approximately the same value of CLI, thus indicating a stable text complexity in terms of sentence and word length in the observed period. They also demonstrated that different genres inside the same text category, even if they follow the same trend of change, differ by the intensity of those changes. In the Prose category, genre D (Religion) exhibited a significantly higher intensity of increase than the other three genres of the same category – E (Skills, Trades and Hobbies), F (Popular Lore) and G (Belles Lettres, Biographies, Essays).

According to the results of the first experiment, several genres – B (Press: Editorial), D (Religion), F (Popular Lore), G (Belles Lettres, Biographies, Essays) and J (Science) of American English demonstrated a statistically significant increase of CLI between 1961 and 1992. Similarly as in the case of British English, all three genres of the Prose category in American English which manifested a statistically significant increase of CLI, differed by the intensity of those changes.

Most importantly, the comparison between the results of the two experiments – using all fifteen fine-grained text genres and then using only the

four broad text categories, revealed the potential pitfalls of hypothesising about the trends of diachronic change solely based on the results of the second approach. It also pointed out two types of misleading results. The first type would indicate that there were no significant changes in the observed broad text category, while after closer scrutiny some of the genres of that category did actually demonstrate significant changes. Those changes in the fine-grained text genres are probably masked by a high heterogeneity of changes or unbalanced distribution of texts among different genres in the relevant category. Therefore, they will not be reflected in the results of the examination of the whole broad text category. The second type of misleading result would indicate a specific trend/direction of change in the whole observed text category, while after closer examination, different genres in that category actually demonstrated different trends of change. We would expect that the general trend of change in the broad text category is determined by the trend which is most common among its genres. However, as the distribution of texts is unbalanced, what we actually see reflected is the trend of the genre(s) with the greatest amount of texts.

Acknowledgements

I would like to express my gratitude to my supervisor Prof. Ruslan Mitkov for his guidance and support. This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT.

References

Laurie Bauer. 1994. *Watching English change: and introduction to the study of linguistic change in standard English in the twentieth century*. London: Longman.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60 (2): 283–284.

David Denison. 1994. A Corpus of Late Modern English Prose. In: M. Kytö et al. eds. *Corpora Across the Centuries*: 7–16 Amsterdam: Rodopi.

Nelson W. Francis. 1965. A standard corpus of edited present-day American English. *College English*, 26: 267–273.

George R. Klare. 1968. The Role of Word Frequency in Readability. *Elementary English*, 45: 12–22.

George R. Klare. 1974. Assessing Readability. *Reading Research Quarterly*, 1: 62–102.

Geoffrey Leech. 2003. Modality on the move: the English modal auxiliaries 1961-1992. In: R. Facchinetti, M. Krug and F. Palmer, eds. *Modality in contemporary English*. Berlin/New York: Mouton de Gruyter, 223–240.

Geoffrey Leech. 2004. Recent grammatical change in English: data, description, theory. In: K. Aijmer and B. Altenberg, eds. *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. Amsterdam: Rodopi, 61–81.

Geoffrey Leech and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB. *ICAME Journal*, 29: 83–98.

Geoffrey Leech and Nicholas Smith. 2006. Recent grammatical change in written English 1961–1992: some preliminary findings of a comparison of American with British English. In: A. Renouf and A. Kehoe, eds. *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, 186–204.

Geoffrey Leech and Nicholas Smith. 2009. Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. In: A. Renouf and A. Kehoe, eds. *Corpus Linguistics: Refinements and Reassessments*. Amsterdam/New York, 173–200.

Geoffrey Leech, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Christian Mair and Marianne Hundt. 1995. Why is the progressive becoming more frequent in English? A corpus-based investigation of language change in progress. *Zeitschrift für Anglistik und Amerikanistik*, 43: 111–122.

Christian Mair. 1997. The spread of the going-to-future in written English: a corpus-based investigation into language change in progress. In: R. Hickey and St. Puppel, eds. *Language history and linguistic modelling: a festschrift for Jacek Fisiak on his 60th birthday*. Berlin: Mouton de Gruyter, 1536–1543.

Christian Mair. 2002. Three changing patterns of verb complementation in Late Modern English: a real-time study based on matching text corpora. *English Language and Linguistics*, 6: 105–131.

Christian Mair, Marianne Hundt, Geoffrey Leech and Nicholas Smith. 2002. Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7: 245–264.

Christian Mair and Geoffrey Leech. 2006. Current change in English syntax. In: B. Aarts and A. MacMahon, eds. *The Handbook of English Linguistics*. Oxford: Blackwell, Ch.14.

Douglas R. McCallum and James L. Peterson. 1982. Computer-based readability indexes. *In Proceedings of the ACM '82 Conference*: 44–48. New York, NY.

Edgar A. Smith and Peter J. Kincaid. 1970. Derivation and Validation of the Automated Readability Index for Use with Technical Materials. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 12(5): 457–464.

Nicholas Smith. 2002. Ever moving on? The progressive in recent British English. In: P. Peters, P. Collins and A. Smith, eds. *New frontiers of corpus research: papers from the twenty first International Conference on English Language Research on Computerized Corpora, Sydney 2000*. Amsterdam: Rodopi, 317–330.

Nicholas Smith. 2003a. A quirky progressive? A corpus-based exploration of the will + be + -ing construction in recent and present day British English. In: D. Archer, P. Rayson, A. Wilson and T. McEnery, eds. *Proceedings of the Corpus Linguistics 2003 Conference*: 714–723. Lancaster University: UCREL Technical Papers.

Nicholas Smith. 2003b. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In: R. Facchinetti, M. Krug and F. Palmer, eds. *Modality in contemporary English*. Berlin/New York: Mouton de Gruyter, 241–266.