Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness

Sanja Štajner and Ruslan Mitkov

Research Group in Computational Linguistics, RIILP
University of Wolverhampton, UK
S.Stajner@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract

This study from the area of language variation and change is based on exploitation of the comparable diachronic and synchronic corpora of 20th century British and American English language (the 'Brown family' of corpora). We investigate recent changes of lexical density and lexical richness in two consecutive thirty-year time gaps in British English (1931–1961 and 1961–1991) and in 1961–1992 in American English. Furthermore, we compare the diachronic changes between these two language varieties and discuss the results of the synchronic comparison of these two features between British and American parts of the corpora (in 1961 and in 1991/2). Additionally, we explore the possibilities of these comparable corpora by using two different approaches to their exploitation: using the fifteen fine-grained text genres, and using only the four main text categories. Finally, we discuss the impact of the chosen approaches in making hypotheses about the way language changes.

Keywords: corpus analysis, language change, lexical richness

1. Introduction

Kroch (2008) defines language change as "a failure in the transmission across time of linguistic features" and states that "over historical time languages change at every level of the language structure: vocabulary, phonology, morphology and syntax". He states that in principle, language change can occur within groups of adult native speakers of language as the result of the substitution of one feature with another as in the case of the substitution of old words with new ones, though he raises a doubt in the validity of this hypothesis in the case of syntactic and grammatical changes.

1.1. Lexical density and lexical richness

In this study, our focus was only at the vocabulary level of the language change. We wanted to investigate how the lexical density and lexical richness were changing during the 20th century. Lexical density is one of the most commonly used features for describing diversity of a vocabulary (Stamatatos et al., 2000). Smith and Kelly (2002), for instance, used this feature for dating works. Lexical density is calculated as the ratio between the number of unique word types and the total number of tokens in the given text. Therefore, a higher lexical density would indicate a wider range of used vocabulary. However, as lexical density counts morphological variants of the same word as different word types (tokens), Corpas Pastor et al. (2008) suggested the use of another measure – lexical richness, instead. The lexical richness is computed as the ratio between the number of unique lemmas and the total number of tokens in the given text. This second measure does not take into account different morphological counts of the same word as different word types. Therefore, Corpas Pastor et al. (2008) believed that it would be a more appropriate indicator of the vocabulary variety of an author.

1.2. Diachronic corpora of 20th century English language

There are several corpora of English language consisting of the texts published in the 20th century, compiled principally for purposes of grammatical researches, but they are usually not publicly available or they cover only a specific genre. The ARCHER corpus (Biber et al., 1994), for instance, belongs to the first of the mentioned groups. It covers a wide range of genres - drama, medical, historical and news reportage texts, from 1650 to 1990 divided into fifty-year blocks, but is not available to the research community (Leech and Smith, 2005). The Corpus of Late Modern English Prose (Denison, 1994), a collection of informal private letters written in British English between 1861 and 1919 is, on the other hand, available to the research community, but it covers only one genre and belongs more to the 19th than to the 20th century. The Corpus of English Newspaper Editorials – CENE (Westin, 2002; Westin and Geisler, 2002), which consists of institutional editorials of three 'broadsheet' British newspapers - The Times, The Guardian and The Daily Telegraph, sampled at ten-year intervals across the 20th century (Leech and Smith, 2005) and the Bauer's corpus of The Times (Bauer, 1994), also consisting of editorials sampled at decade intervals (Leech and Smith, 2005), both belong to the intersection of the above two types as they cover only a specific genre and they are not publicly available.

1.3. The 'Brown family' of corpora

The 'Brown family' of corpora is comprised of five mutually comparable corpora. The American part consists of two corpora:

- The Brown University corpus of written American English – Brown (Francis, 1965)
- The Freiburg - Brown Corpus of American English –

Main category	Code	Genre	Number of texts		
			(F/B)LOB	Brown	Frown
PRESS	A	Press: Reportage	44	44	44
	B	Press: Editorial	27	27	27
	C	Press: Review	17	17	17
PROSE	D	Religion	17	17	17
	E	Skills, Trades and Hobbies	38	36	36
	F	Popular Lore	44	48	48
	G	Belles Lettres, Biographies, Essays	77	75	75
	H	Miscellaneous	30	35	30
LEARNED	J	Science	80	80	80
FICTION	K	General Fiction	29	29	29
	L	Mystery and Detective Fiction	24	24	24
	M	Science Fiction	6	6	6
	N	Adventure and Western	29	30	29
	P	Romance and Love Story	29	29	29
	R	Humour	9	9	9

Table 1: Structure of the corpora

Frown (Hundt et al., 1998).

The British part consists of three corpora:

- The Lancaster1931 – BLOB (Leech and Smith, 2005)
- The Lancaster-Oslo/Bergen Corpus – LOB (Johansson et al., 1978)
- The Freiburg-LOB Corpus of British English – FLOB (Sand and Siemund, 1992).

The corpora contain texts published in years 1931±3 (Lancaster1931), 1961 (LOB and Brown), 1991 (FLOB) and 1992 (Frown) divided into 15 different genres (Table 1). These five corpora comply with the formal criteria of comparability as the texts have been compiled on the basis of the same sampling frame and with similar balance and representativeness. In particular, the texts have been selected to match the same domain and topics, and are of comparable size. Therefore, they fulfill all the necessary conditions for being widely used throughout the linguistic community – they are a diachronic corpora of 20th century written English texts, which cover a wide range of genres and are publicly available as part of the ICAME Corpus Collection[1].
The Brown corpus was published first, back in 1964. One of the ideas of compiling the Brown corpus was to help "to have a common body of material on which studies of various sorts can be based" (Leech and Smith, 2005) and in that way to provide some kind of 'standard' for the following parallel corpora of British English or for English of other periods to be matched (Francis, 1965 in Leech and Smith, 2005). It was a one-million-word corpus, consisting of 500 texts of about 2000 running words each, selected at random points from the original source and the texts covered fifteen different text genres. Following that idea, the LOB corpus (Johansson at al. 1978) of written British English was compiled as the first corpus to match the Brown corpus, respecting the year of sampling (1961) and its sampling frame and representation of different text types (Leech and Smith, 2005). The release of the LOB corpus enabled synchronic comparison between two major English language varieties across a wide range of text genres. In the 1990s, the FLOB and Frown corpora were compiled at Freiburg University representing, respectively, written British English in 1991 and American English in 1992. As their design matched closely to the design of the LOB and Brown corpora, this provided the opportunity to investigate and compare diachronic changes between two major varieties of the written English language. The exact procedure for diachronic matching applied during the compilation of the FLOB and Frown corpora could be found in (Leech and Smith, 2005, p.8). Later on, the research to extend the Brown model backwards in time, undertaken at the Lancaster University, led to the compilation of the Lancaster1931 corpus to match the design of the LOB and FLOB corpora. The target sampling year in this case was 1931 (± three years), in order to maintain the thirty-year gap already established between LOB and FLOB corpora, as well as between Brown and Frown corpora.
Being all mutually comparable, these five corpora (BLOB, LOB, FLOB, Brown and Frown) create the possibility for several different types of investigation:

- Synchronic comparison between British and American English in 1961 and in 1991/2
- Diachronic comparison among the texts published in 1931, 1961 and 1991 in British English
- Diachronic comparison among the texts published in 1961 and 1992 in American English
- Comparison of diachronic changes in 1961–1991/2 between British and American English

1.4. Structure of the corpora

Each of the corpora (BLOB, LOB, FLOB, Brown and Frown) consist of approximately 1,000,000 words – 500 texts of about 2000 running words each. The texts cover fifteen different text genres (Table 1), which could be further grouped into four more generalised categories: Press

[1] http://icame.uib.no/newcd.htm

(A–C), Prose (D–H), Learned (J) and Fiction (K–R). This structure of the corpora allows three different approaches to the exploitation of the corpora in diachronic studies:

1. Differentiating between texts only across two different language varieties or two different years of publication (without differentiating between texts across the text genres/categories).

2. Differentiating between texts across the four main text categories (Press, Prose, Learned and Fiction), thus exploring diachronic changes separately in each of the four main text categories.

3. Differentiating between texts across all fifteen fine-grained text genres (A–R), thus exploring diachronic changes separately in each of the fifteen fine-grained text genres.

2. Related work

The 'Brown family' of corpora has already been used in many diachronic studies of various lexical, grammatical, stylistic and syntactic features, e.g. (Mair and Hundt, 1995; Mair, 1997; Mair et al., 2002; Smith, 2002; Smith, 2003b; Smith, 2003a; Leech, 2003; Leech, 2004; Leech and Smith, 2006; Mair and Leech, 2006; Leech and Smith, 2009; Leech et al., 2009; Štajner and Mitkov, 2011). A large set of these studies shared the same methodology. The corpora were part-of-speech tagged, the change was presented in terms of the absolute and relative differences and the statistical significance was measured by the log likelihood function. The first attempt for a completely automated feature extraction from the raw text version of the 'Brown family' of corpora in diachronic studies was reported by Štajner and Mitkov (2011). The corpora were parsed with the state-of-the-art Connexor's Machinese Syntax parser[2] and the features were automatically extracted from the parser's output. Statistical significance of the results was measured by the t-test.

However, all of these previous studies used the aforementioned second approach, differentiating only between texts across the four main categories (Press, Prose, Learned and Fiction). Following the discussion in (Stajner, 2011) about the impact of the chosen genre granularity (aforementioned approaches 1–3), we decided to use the third approach and differentiate between texts across all fifteen fine-grained text genres (A–R), in order to obtain a better understanding of how lexical density and richness change. To the best of our knowledge, this is the first diachronic study conducted on the 'Brown family' of corpora using this approach.

Of the most relevance for this work was the study conducted by Štajner and Mitkov (2011), where the authors investigated diachronic changes of lexical density (LD) and lexical richness (LR) in the period 1961–1991/2 and used the same methodology for feature extraction. However, they only differentiated between texts across the four main text categories (Press, Prose, Learned and Fiction). In this study, we went one step further, by differentiating between texts across all fifteen fine-grained text genres (A–R). This approach allowed us to obtain a better insight into the way language changes. It also gave us the opportunity to compare the results obtained by these two different approaches and draw attention to the possible pitfalls in making hypotheses by differentiating between texts only across the four main text categories. In that sense, the results presented in this study could also be taken as an additional support for the claims made in (Štajner, 2011).

In this study, we also extended the time span in British English by using the Lancaster1931 corpus. Therefore, we were able to compare the trends of change in two consecutive thirty-year time gaps (1931–1961 and 1961–1991) in British English and examine whether the trend of change was stable during the whole sixty-year period.

3. Methodology

In this study, we followed the methodology for feature extraction proposed by Štajner and Mitkov (2011). All five corpora were used in their initial raw text format and then parsed with the state-of-the-art Connexor's Machinese Syntax parser for the purposes of tokenisation and lemmatisation. The main reason for using the same parser and the same methodology, although the tokenisation and lemmatisation could have been done by some lighter tools, was to be able to compare our results obtained for all fifteen text genres (the aforementioned third approach) with those results reported by Štajner and Mitkov (2011) when the authors were differentiating only between the texts across the four main text categories (the aforementioned second approach). As the performance of the parser in this task and its specificities regarding the tokenisation and lemmatisation processes were already discussed in details in (Štajner and Mitkov, 2011), here we will just highlight the most important ones in order to facilitate a better understanding of the presented results.

The lexicon of the Connexor's Machinese Syntax parser was built using various large corpora of different text genres – news, bureaucratic documents, literature etc. (Connexor, 2006) and contains hundreds of thousands of base forms. The words which are not found in the lexicon are assigned their word class and base form by using the heuristic methods (Connexor, 2006). The software which was used as a base for the current version of the parser reported an excellent accuracy (Samuelsson and Voutilainen, 1998) and the parser itself reported the POS accuracy of 99.3% on Standard Written English (benchmark from the Maastricht Treaty) with no ambiguity (Connexor, 2006).

3.1. Tokenisation

The Connexor's Machinese parser treats the contracted negative form (*n't*) and its antecedent verb as two separate tokens. E.g. *aren't* would be separated into two tokens *are* and *not* and assigned two separate base forms – *be* and *not*. The *'s* is treated in two different ways, depending on the role it has in the sentence. When it represents a genitive form, e.g. *"... Isaac's illness..."* (FLOB: K02), it is treated as one token and is assigned the corresponding lemma *isaac*. In other cases where *'s* represents the contraction of the verb *to be (is)* or *to have (has)*, e.g. *" He's at a table over there."* (FLOB: K01), the personal pronoun

[2] http://www.connexor.eu

and verb contraction are treated as two separate tokens *he* and *is* and assigned two separate base forms *he* and *be*, accordingly.

3.2. Lemmatisation

The output of the lemmatisation process done by the Connexor's Machinese parser expresses certain differences between the earlier versions and the current version of the parser. The main difference is in the way that possessive pronouns, derived adverbs, and EN and ING forms are treated.

While the earliest versions of the parser would assign the corresponding personal pronoun as the lemma of the given possessive pronoun (e.g. the word *theirs* would be assigned *their* as its lemma), the current version of the parser assigns their own base forms to possessive pronouns (e.g. the word *theirs* is assigned *theirs* as its lemma).

A similar rule applies to derived adverbs. In the previous versions of the parser, derived adverbs, such as *absolutely* or *directly* would be assigned *absolute* and *direct* as their lemmas, while in the current version of the parser, these same derived adverbs are assigned their own base forms – *absolutely* and *directly*.

The EN and ING forms, which can represent either present and past participles or corresponding nouns and adjectives, are assigned a POS tag (EN, ING, N or A) and different base forms in the current version of the parser, according to their usage in that particular case. For example, if the word *meeting* is recognised as a noun by the parser, it will be assigned *meeting* as the corresponding lemma. In case that the same word is recognised as a present participle of the verb *to meet*, it will be assigned *meet* as its corresponding lemma. The results would be similar in the case of an EN form. For example, if the word *selected* represents an adjective in the given context, it will be assigned *selected* as its lemma. In another case, if it represents a past participle, it will be assigned *select* as the corresponding lemma.

These differences between previous and current versions of the parser in lemmatising certain word forms is reflected in the differences between the lexical richness and lexical density. It is reasonable to expect that the calculated LD and LR will be much closer if we use the current version than if we use an earlier version of the parser.

3.3. Feature extraction

The lexical density (LD) and lexical richness (LR) were calculated for each text separately in order to enable later applied statistical tests. Lexical density was calculated as the total number of unique word forms (tokens) divided by the total number of tokens in the given text (eq.1).

$$LD = \frac{number\_of\_unique\_tokens}{total\_number\_of\_tokens} \quad (1)$$

Lexical richness was calculated similarly, this time using the total number of unique lemmas divided by the total number of tokens (eq.2).

$$LR = \frac{number\_of\_unique\_lemmas}{total\_number\_of\_tokens} \quad (2)$$

4. Experimental settings

The purpose of this study was two-fold: (1) to investigate diachronic changes of lexical density and lexical richness in 20th century English language in each of the fifteen fine-grained text genres, and (2) to compare the results of two different approaches to the exploitation of these comparable corpora. Therefore, we had two different sets of experiments. The first set of experiments consisted of investigating the following five changes using the third approach (differentiating between the texts across the all fifteen fine-grained text genres):

- Diachronic changes in British English in the period 1931–1961
- Diachronic changes in British English in the period 1961–1991
- Diachronic changes in American English in the period 1961–1992
- Synchronic differences between British and American English in 1961
- Synchronic differences between British and American English in 1991/2.

The second set of experiments consisted of the same five experiments, but this time using the second approach (differentiating between the texts only across the four main text categories).

4.1. Statistical significance testing

For each of the aforementioned five experiments we calculated the statistical significance of the mean differences between the two corresponding groups of texts. Statistical significance tests are divided into two main groups: parametric (which assume that the samples are normally distributed) and non-parametric (which do not make any assumptions about the sample distribution). In the cases where the samples follow the normal distribution, it is recommended to use parametric tests as they have greater power than non-parametric tests (Garson, 2012a). Therefore, we first applied the the Shapiro-Wilk's W test (Garson, 2012b) offered by SPSS EXAMINE module in order to examine in which cases/genres/categories the features were normally distributed. This test is a standard test for normality, recommended for small samples. It shows the correlation between the given data and their expected normal scores. If the result of the W test is 1, it means that the distribution of the data is perfectly normal. Significantly lower values of W (≤ 0.05) indicate that the assumption of normality is not met. Those cases are shown in bold (Table 2).

Following the discussion in (Garson, 2012c), for both approaches we used the following strategy: if the two data sets we wanted to compare were both normally distributed we used the t-test for the comparison of their means; if at least one of the two data sets was not normally distributed ($W \leq 0.05$ in Table 2), we used the Kolmogorov-Smirnov Z test (a non-parametric test) for two independent samples to calculate the statistical significance of the differences between their means.

Approach	Genre	LD					LR				
		British			American		British			American	
		1931	1961	1991	1961	1992	1931	1961	1991	1961	1992
III	A	.320	**.003**	.807	.448	.737	.221	**.015**	.963	.345	.575
	B	.935	.905	.326	.263	.958	.776	.644	.322	.256	.371
	C	.399	.716	.428	**.002**	.369	.370	.786	.692	**.002**	.574
	D	.777	.679	.643	.711	.089	.706	.409	.178	.816	**.047**
	E	.115	**.026**	**.011**	.238	.725	.289	**.047**	.093	.664	.353
	F	.818	.639	.319	.338	**.000**	.883	.652	.401	.383	**.000**
	G	**.013**	.065	.170	.054	.072	**.017**	**.018**	.285	.236	.240
	H	.202	.892	.952	.119	.303	.261	.992	.970	.109	.266
	J	.051	.883	.252	**.002**	.470	.127	.803	.158	**.003**	.826
	K	.403	.835	.511	.283	.523	.304	.722	.916	.353	.630
	L	.333	.599	.291	.230	.529	.365	.457	.359	.141	.277
	M	.528	.290	.940	.179	.812	.601	.55	.792	.107	.835
	N	.966	.127	.287	.990	.314	.886	.087	.183	.789	.572
	P	.587	.084	.322	.279	.362	.300	.068	.316	.379	.386
	R	.291	.913	.580	.555	.962	.182	.873	.683	.421	.805
II	Press	.834	.068	**.012**	.112	.490	.856	.230	**.014**	**.044**	.660
	Prose	**.000**	**.001**	**.000**	**.000**	**.000**	**.002**	**.002**	**.011**	**.001**	**.000**
	Learned	.051	.883	.252	**.002**	.470	.127	.803	.158	**.003**	.826
	Fiction	.756	.116	.850	.087	.169	.591	.101	.645	**.011**	.181

Table 2: Normal distribution testing (Shapiro-Wilk's W test results)

It is interesting to note that in some cases, even if the data in fine-grained text genres follow the normal distribution (e.g. genres A–C in columns LD and LR of British English in 1991), the data in that whole text category (Press in columns LD and LR of British English in 1991) do not follow the same distribution. Also, we can find examples of the opposite situation when some of the data in the fine-grained text genres (e.g. genre A in columns LD and LR of British English in 1961) do not follow the normal distribution, but the data in the corresponding broader text category (Press in columns LD and LR of British English in 1961) are normally distributed. This second case is intuitively more expected as we know that the bigger the data set, the more chance there is that the data would be normally distributed. However, both the cases force us to use different statistical significance tests for the second and for the third approach.

5. Results and discussion

Our study basically has two main parts: diachronic comparison (1931–1961 and 1961–1991 in British English; 1961–1992 in American English) and synchronic comparison of British and American English (in 1961 and in 1991/2). Therefore, we will present the results separately for diachronic (separately for LD and LR) and synchronic comparisons (together for LD and LR) in the next three subsections. In each of these subsections, together with our main results obtained by using the third approach (differentiating across fifteen fine-grained text genres) we will also present the results of the alternative second approach (differentiating across only four main text categories), in order to be able to compare the differences in the conclusions drawn from these two approaches. Statistically significant changes at a 0.05 level significance (sign. ≤ 0.05) are shown in bold.

5.1. Diachronic changes of lexical density (LD)

The results of the investigation of diachronic changes of lexical density (LD) in British and American English are presented in Table 3 (using the third approach) and Table 4 (using the second approach). In both cases we followed the same pattern of representing the results. Columns '1931', '1961' and '1991' under 'British English', and columns '1961' and '1992' under 'American English' represent the calculated average LD in those years for the corresponding language variety. Columns '1931–1961', '1961–1991' and '1961–1992' contain the information about the changes of LD in those periods for the corresponding language varieties. Their subcolumn 'sign.' represent the calculated two-tailed statistical significance of the differences between the corresponding means, by using t-test or Kolmogorov-Smirnov Z test, according to Table 2 and the discussion in Subsection 4.1. The subcolumn 'change' contains the relative change in the observed period, calculated as a percentage of the starting value. The sign '+' stands for an increase and the sign '−' for a decrease over the time.

5.1.1. British English

The results presented in Table 3 indicate several interesting phenomena. First, we can notice that diachronic changes in British English were generally not stable in the two subsequent periods 1931–1961 and 1961–1991. Most of the genres demonstrated significant changes only in one of the two observed periods. In genres G (Belles Lettres, Biographies, Essays) and R (Humour), LD had changed (increased) only in the first period 1931–1961, while in genres A (Press: Reportage), B (Press: Editorial), C (Press: Review), D (Religion) and P (Romance and Love Story) it had changed (increased) only in the second period 1961–1991. Genre E (Skills, Trades and Hobbies) was the only genre that showed a stable increase of LD throughout both periods

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	0.352	0.316	+0.64%	**0.355**	**0.000**	**+6.90%**	**0.379**	0.369	0.940	+0.13%	0.368
B	0.354	0.202	+1.95%	**0.361**	**0.000**	**+7.94%**	**0.389**	**0.378**	**0.031**	**+3.64%**	**0.392**
C	0.382	0.158	+2.83%	**0.392**	**0.001**	**+7.95%**	**0.424**	**0.395**	**0.006**	**+4.18%**	**0.411**
D	0.312	0.427	−2.79%	**0.304**	**0.027**	**+8.47%**	**0.329**	0.323	0.381	+3.26%	0.334
E	**0.327**	**0.045**	**+4.66%**	**0.342**	**0.002**	**+6.99%**	**0.366**	**0.331**	**0.014**	**+7.72%**	**0.357**
F	0.342	0.916	+0.23%	0.342	0.421	+1.91%	0.349	**0.342**	**0.027**	**+5.84%**	**0.362**
G	**0.341**	**0.047**	**+2.79%**	0.350	0.065	+2.66%	0.359	0.347	0.279	+1.42%	0.351
H	0.286	0.593	+1.79%	0.292	0.792	+1.01%	0.295	0.294	0.688	+1.74%	0.299
J	0.295	0.600	+1.34%	0.299	0.236	+2.84%	0.307	0.298	0.329	+4.69%	0.312
K	0.315	0.295	−2.81%	0.307	0.118	+4.51%	0.320	0.327	0.370	−2.99%	0.317
L	0.299	0.458	+1.93%	0.304	0.434	−2.31%	0.297	0.299	0.493	+2.17%	0.306
M	0.328	0.810	+1.46%	0.333	0.574	+4.48%	0.348	0.323	0.779	−1.55%	0.318
N	**0.314**	**0.048**	**−4.90%**	**0.299**	**0.020**	**+7.06%**	**0.320**	0.315	0.768	−0.92%	0.313
P	0.298	0.089	−4.46%	**0.285**	**0.010**	**+7.66%**	**0.307**	0.302	0.528	−1.86%	0.297
R	**0.311**	**0.000**	**+14.16%**	**0.355**	0.545	−1.96%	0.348	**0.359**	**0.011**	**−18.39%**	**0.293**

Table 3: Diachronic changes of lexical density (LD) – third approach

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
Press	0.358	0.168	+1.48%	**0.364**	**0.000**	**+7.43%**	**0.391**	0.376	0.084	+2.03%	0.384
Prose	**0.328**	**0.007**	**+2.02%**	**0.335**	**0.000**	**+3.54%**	**0.347**	**0.333**	**0.007**	**+3.76%**	**0.346**
Learned	0.295	0.600	+1.34%	0.299	0.236	+2.84%	0.307	0.298	0.329	+4.69%	0.312
Fiction	0.308	0.297	−1.35%	**0.304**	**0.009**	**+3.92%**	**0.316**	0.315	0.105	−2.51%	0.307

Table 4: Diachronic changes of lexical density (LD) – second approach

1931–1961 and 1961–1991. The most interesting might be the case of genre N (Adventure and Western) which demonstrated a significant change of LD in both periods although these changes had opposite directions. While in the first period (1931–1961) LD had decreased, in the second period (1961–1991) it had increased. At the same time, the decrease of LD in this genre is the only observed significant decrease of LD in British English in this study.

5.1.2. American English

In American English, the results (Table 3) indicated a significant increase of LD in four genres: B (Press: Editorial), C (Press: Review), E (Skills, Trades and Hobbies) and F (Popular Lore), and a significant decrease of LD in genre R (Humour). At the same time, this change of LD in genre R was of a significantly higher intensity than the changes reported in other genres.

5.1.3. British vs. American English

The comparison of diachronic changes of LD between British and American English in the period 1961–1991/2 indicates that the most of the genres did not undergo the same changes at the same time. For instance, genres A (Press: Reportage), D (Religion), N (Adventure and Western) and P (Romance and Love Story) demonstrated a change only in British English, while genres F (Popular Lore) and R (Humour) demonstrated a change of LD only in American English during the same period 1961–1991/2. The only genres which reported a significant increase of LD in both language varieties during that period were genres B (Press: Editorial), C (Press: Review) and E (Skills, Trades and Hobbies).

5.1.4. Second vs. third approach

The first obvious difference in conclusions drawn from the results of the second approach (Table 4) and those of the third approach (Table 3) is that by using solely the results of the second approach we would conclude that whenever there was a change, LD has increased. By closer examination of the corpora (Table 3), we notice that in fact a significant decrease of LD is also likely to happen, as in the case of genre N (Adventure and Western) in British English (1931–1961) and genre R (Humour) in American English (1961–1992).

The other differences between the conclusions drawn from these two approaches are more subtle but maybe even more important to mention. The most drastic difference can be noticed in Fiction category of British English (1931–1961), and Fiction and Press categories in American English (1961–1992). While the results of the second approach (Table 4) reported no changes of LD in these particular cases, the results of the third approach (Table 3) revealed some interesting phenomena in the corresponding genres. In American English, a very intensive decrease of LD in genre R (present in the results of the third approach), was probably masked in the second approach by the constancy of LD in other genres of this category (genres K–P) which have a greater number of texts than genre R (Table1

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	0.317	0.808	+0.09%	**0.317**	**0.001**	**+5.95%**	**0.336**	0.331	0.334	−1.84%	0.325
B	0.316	0.263	+1.81%	**0.321**	**0.000**	**+8.27%**	**0.348**	0.337	0.117	+2.93%	0.347
C	0.345	0.141	+3.21%	**0.356**	**0.002**	**+8.42%**	**0.386**	**0.358**	**0.017**	**+3.49%**	**0.371**
D	0.278	0.362	−3.59%	**0.268**	**0.030**	**+9.40%**	**0.293**	0.286	0.240	+3.92%	0.297
E	**0.290**	**0.012**	**+4.74%**	**0.303**	**0.005**	**+6.99%**	**0.324**	**0.292**	**0.024**	**+8.00%**	**0.316**
F	0.303	0.993	+0.02%	0.303	0.389	+2.36%	0.310	0.304	0.249	+5.55%	0.320
G	**0.304**	**0.018**	**+3.21%**	**0.313**	**0.004**	**+3.35%**	**0.324**	0.310	0.550	+0.89%	0.312
H	0.254	0.649	+1.75%	0.258	0.840	+0.78%	0.260	0.261	0.772	+1.28%	0.265
J	0.262	0.550	+1.61%	0.267	0.413	+2.11%	0.272	0.265	0.436	+4.87%	0.278
K	0.277	0.246	−3.54%	0.268	0.168	+4.56%	0.280	0.287	0.411	−3.17%	0.278
L	0.261	0.521	+1.88%	0.265	0.427	−2.63%	0.259	0.260	0.490	+2.49%	0.267
M	0.290	0.879	+1.06%	0.293	0.562	+5.40%	0.309	0.285	0.699	−2.47%	0.277
N	**0.276**	**0.030**	**−6.18%**	**0.259**	**0.020**	**+8.26%**	**0.281**	0.275	0.826	−0.79%	0.273
P	0.259	0.066	−5.44%	**0.245**	**0.014**	**+8.24%**	**0.266**	0.264	0.359	−3.05%	0.256
R	**0.271**	**0.000**	**+16.78%**	**0.317**	0.555	−2.06%	0.310	**0.320**	**0.012**	**−21.14%**	**0.253**

Table 5: Diachronic changes of lexical density (LR) – third approach

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
Press	0.322	0.279	+1.25%	**0.326**	**0.000**	**+7.17%**	**0.349**	0.338	0.387	+0.71%	0.341
Prose	**0.291**	**0.001**	**+2.06%**	**0.297**	**0.000**	**+3.95%**	**0.309**	0.296	0.096	+3.52%	0.307
Learned	0.262	0.550	+1.61%	0.267	0.413	+2.11%	0.272	0.265	0.436	+4.87%	0.278
Fiction	0.270	0.201	−1.89%	**0.265**	**0.012**	**+4.28%**	**0.276**	0.276	0.202	−3.04%	0.268

Table 6: Diachronic changes of lexical density (LR) – second approach

in Section 1.4). The differences in the Prose category of American English could be explained in the same way. In British English, however, the situation was even more complex. The results of the second approach did not only mask the changes of LD in certain genres (N and R), but they also hid the fact that the changes in these two genres went in opposite directions (an increase of LD in genre R and a decrease of LD in genre N).

Less pronounced, but still worth mentioning, were the differences between the results of the second and third approaches in Prose (1931–1961, 1961–1991) and Fiction (1961–1991) categories of British English, and Prose category of American English. In these cases, the results of the second approach reported significant changes of LD in these categories (Table 4), while the more detailed analysis used in the third approach (Table 3) actually demonstrated that these changes were present only in certain genres of the mentioned categories and not in all of them.

5.2. Diachronic changes of lexical richness (LR)

Diachronic changes of lexical richness (LR) in British and American English are presented in the same manner as in the case of lexical density. Table 5 contains the results of the third approach and Table 6 the results of the second approach.

5.2.1. British English

Similar to the case of LD, LR did not show the same trends of changes in both observed periods 1931–1961 and 1961–1991 in most of the genres. In genre R (Humour) a change was present only in the first period (1931–1961), while in genres A (Press: Reportage), B (Press: Editorial), C (Press: Review), D (Religion) and P (Romance and Love Story) a change was present only in the second period (1961–1991). In genres E (Skills, Trades and Hobbies) and G (Belles Lettres, Biographies, Essays), LR had increased in both periods, while in genre N (Adventure and Western) it first had decreased (in period 1931–1961) and then increased (in the period 1961–1991).

If we compare these results for LR with those obtained for LD (Table 3), we can notice that in most genres, LD and LR demonstrated similar diachronic changes. The only exception to this was reported in genre G (Belles Lettres, Biographies, Essays) in the period 1961–1991 , where LD did not show any statistically significant change, while LR reported an increase of +3.35%.

5.2.2. American English

The results of the investigation of diachronic changes of LR in American English (Table 5) reported a higher lexical richness in 1992 than in 1961 in genres C (Press: Review) and E (Skills, Trades and Hobbies). In genre R (Humour) the situation was the opposite. In this genre, LR was reported to be higher in 1961 than in 1992.

The comparison of diachronic changes between LD and LR (Table 3 and Table 5) indicate similar behaviour of these two features in all three genres in which a significant change of LR was reported. Additionally, LD demon-

Year	Genre	LD				LR			
		Br.	sign.	change	Am.	Br.	sign.	change	Am.
1961	A	0.355	0.043	+3.92%	0.369	0.317	0.012	+4.39%	0.331
	B	0.361	0.012	+4.79%	0.378	0.321	0.020	+4.91%	0.337
	K	0.307	0.035	+6.66%	0.327	0.268	0.041	+7.41%	0.287
	P	0.285	0.037	+6.04%	0.302	0.245	0.023	+7.45%	0.264
1991/2	G					0.324	0.031	−3.53%	0.312
	R	0.348	0.004	−15.91%	0.293	0.310	0.002	−18.63%	0.253

Table 7: Synchronic comparison of LD and LR in 1961 and 1991/2 (British vs. American English)

strated a change in genres B (Press: Editorial) and F (Popular Lore), in which LR did not report any changes.

5.2.3. British vs. American English

The results of the comparison of diachronic changes of LR between British and American English indicates that this feature underwent similar changes in both language varieties (in the period 1961–1991/2) in only two genres – C (Press: Review) and E (Skills, Trades and Hobbies). The number of genres in which LR reported a change in only one of the two language varieties was significantly higher, thus indicating different trends of change between these two varieties in general. On one side we have genres A (Press: Reportage), B (Press: Editorial), D (Religion), G (Belles Lettres, Biographies, Essays), N (Adventure and Western) and P (Romance and Love Story) for which the results (Table 5) indicate a significant increase of LR in the period 1961–1991 only in the British part of the corpora. On the other side we have genre R (Humour) in which a significant change (in this case a decrease) of LR was reported only in American English.

5.2.4. Second vs. third approach

The investigation of diachronic changes of LR revealed the same possible pitfalls in making conclusions solely based on the results of the second approach (Table 6) as in the case of LD (Section 5.1.4). For example, these results (Table 6) did not show any significant differences of LR between 1931 and 1961 in Fiction category, while the results of the third approach (Table 5) indicated a significant decrease of LR in genre N (Adventure and Western) and a significant increase in genre R (Humour). In this case not only did the results of the second approach fail to report significant changes in some genres of the Fiction category, but even more importantly, they failed to report that different genres which belong to the same broad category, exhibit different trends of change – increase and decrease, in the same period of time.

In American English, the results of the second approach (Table 6) did not indicate any changes of LR in the observed period 1961–1992, while the results of the third approach (Table 5) reported significant changes in one of the genres in each of the Press, Prose and Fiction categories – genres C (Press: Review), E (Skills, Trades and Hobbies) and R (Humour). In the case of Prose category (in both periods, 1931–1961 and 1961–1991) and Fiction category (in the period 1961–1991) in British English, the results of the second approach (Table 6) which reported a significant increase of LR were less misleading than in the previous case, though still hiding the fact that these changes were present only in certain genres of this category and not in all of them (Table 5).

5.3. Synchronic comparison

The results of synchronic comparison of LD and LR between British and American English are presented in Table 7. As LD and LR were already presented for both of these language varieties in the previous two sections (5.1 and 5.2), here we presented only the genres in which a statistically significant difference between British and American English was reported for at least one feature and one year.

It is interesting to note that the results (Table 7) did not report any genre in which a significant difference of LD or LR between these two language varieties was present in both years – 1961 and 1991/2. Actually, in 1961, a significant difference in LD and LR between British and American was reported in only four genres – A (Press: Reportage), B (Press: Editorial), K (General Fiction) and P (Romance and Love Story). In all these genres, the texts written in American English used a wider vocabulary than those written in British English. In 1991/2, a significant difference of LD between British and American English was reported in only one genre – genre R (Humour). In this genre, texts written in British English had a greater vocabulary variety than those written in American English. In the same year (1991/2), LR was reported to be significantly higher in British than in American English for two genres – genre G (Belles Lettres, Biographies, Essays) and R (Humour).

It is also interesting to notice that all reported differences in 1961 went in favour of a larger vocabulary used in American English, while all those differences reported in 1991/2 went in favour of a larger vocabulary used in British English.

6. Conclusions

The results of the experiments presented in this paper enabled us to make two different types of relevant conclusions. The first type of conclusions would be those regarding the investigated diachronic changes of lexical density and lexical richness and their behaviour in British and American English. The second type would be those regarding the influence of the chosen approach (chosen way of exploitation of the comparable corpora) – using only four main broad text categories (second approach) or using all fifteen fine-grained text genres (third approach), on making hypotheses about the way English language changes.

On the basis of the results of the third approach to the investigation of diachronic changes of LD and LR (Tables 3 and 5), we can conclude that the changes of these two stylistic features were very heterogeneous in various ways – across the genres (A–R), language varieties (British and American) and periods observed (1931–1961, 1961–1991/2). Most importantly, these results indicated different trends of change even among the genres which belong to the same broad text category, e.g. genres N and P in Fiction category reported a decrease and an increase of LD and LR in the same period 1931–1961. Furthermore, the investigated genres did not report many constant ongoing changes during the two consecutive periods 1931–1961 and 1961–1991. Genre N (Adventure and Western) reported a significant decrease in the first period 1931–1961 and then a significant increase of both features (LD and LR) in the second period (1961–1991) in British English. In other genres, a significant change was usually reported in only one of the two observed periods. The only exceptions were noticed in genre E (Skills, Trades and Hobbies), where LD and LR had increased in both periods, and in genre G (Belles Lettres, Biographies, Essays), where LR reported a significant increase during both periods.

Genre R (Humour) reported different behaviour between the two language varieties (no change in British English and a significant decrease in American English) for the same period 1961–1991/2, and different behaviour in two consecutive time periods in British English (an increase in 1931–1961 and no significant change in 1961–1991). Even more interestingly, the reported changes in British and American English (although not for the same period, but for 1931–1961 in British and for 1961–1992 in American English) did not follow the same direction, i.e. in British English, LD and LR had increased (in the period 1931–1961), while in American English, both of these features had decreased (in the period 1961–1992). Therefore, we cannot even say that the changes reported in British and American English were shifted in time (for thirty years). The results presented in this study actually indicate that the changes of LD and LR in British and American English were not mutually influenced.

All these findings lead to the conclusion that the time gap in diachronic studies of lexical density and lexical richness should ideally be smaller if we wish to gain a better insight into the way they change. They also indicate that different language varieties should be investigated separately as they generally do not follow the same patterns of change. Similarly, the presented results emphasise the necessity for separate investigation of the genres which belong to the same broad text category as they demonstrate different trends of changes among themselves.

The comparison between the results obtained by using the second approach (differentiating only across the four main broad categories) and those obtained by using the third approach (differentiating across all fifteen fine-grained text genres) clearly stated some of the potential pitfalls in making hypotheses about the way language changes solely on the basis of the results of the second approach. It pointed out two possible problems in using the second approach. The first problem would be the case in which the results of the second approach do not report any changes in the relevant text category, while a closer examination of the same category (using the third approach) clearly indicates significant changes in some of the genres belonging to that category. The second problem would be the case in which the results of the second approach again do not report any changes, while the results of the third approach not only indicate significant changes in some of the genres of that category, but also indicate different trends of changes among them (increase, decrease and no change). In the second approach these changes are probably masked by unbalanced distribution of texts or by a high heterogeneity of changes across different genres of that category.

Finally, this study presented various possibilities of the comparable 'Brown family' of corpora and different approaches to their exploitation in diachronic and synchronic language studies. Most of these ideas and the methodology used could also be applied to other existing comparable corpora in order to enable their better exploitation in various tasks.

7. References

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

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges. compiling and exploring a representative corpus of historical English registers. In G. Tottie U. Fries and P. Schneider, editors, *Creating and using English language corpora*, pages 1–14. Amsterdam: Rodopi.

Connexor. 2006. Machinese language analysers. *Connexor Manual*.

Gloria Corpas Pastor, Ruslan Mitkov, Afzal Naveed, and Pekar Victor. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA*.

David Denison. 1994. A Corpus of Late Modern English Prose. In *Corpora Across the Centuries. Proceedings of the First International Colloquium on English Diachronic Corpora*, pages 7–16. Amsterdam: Rodopi.

Nelson W. Francis. 1965. A Standard Corpus of Edited Present-Day American English. *College English*, 26(4):267–273.

David G. Garson. 2012a. Significance. Statnotes: Topics in Multivariate Analysis. [http://faculty.chass.ncsu.edu/garson/PA765/signif.htm].

David G. Garson. 2012b. Testing of assumptions: Normality. Statnotes: Topics in Multivariate Analysis.

David G. Garson. 2012c. Tests for two independent samples: Mann-Whitney U, Wald-Wolfowitz runs, Kolmogorov-Smirnov Z, & Moses extreme reactions tests. Statnotes: Topics in Multivariate Analysis. [http://faculty.chass.ncsu.edu/garson/PA765/mann.htm].

Marianne Hundt, Andrea Sand, and Rainer Siemund, 1998. *Manual of Information to Accompany the Freiburg-LOB Corpus of British English*. Freiburg.

Stig Johansson, Geoffrey Leech, and Helen Goodluck, 1978. *Manual of Information to Accompany the*

Lancester-Oslo/Bergen corpus of British English. Department of English, University of Oslo.

Anthony S. Kroch, 2008. *Syntactic Change*, pages 698–729. Blackwell Publishers Ltd.

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. *Language and Computers*, 55(1):185–204.

Geoffrey Leech and Nicholas Smith. 2009. Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. *Language and Computers*, 69(1):173–200.

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

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

Geoffrey Leech. 2004. Recent grammatical change in English: data, description, theory. *Language and Computers*, 49(1):61–81.

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 f´ur Anglistik und Amerikanistik*, 43:111–122.

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

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. 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, editors, *Language history and linguistic modelling: a festschrift for Jacek Fisiak on his 60th birthday*, pages 1537–1543. Berlin: Mouton de Gruyter.

Christer Samuelsson and Atro Voutilainen. 1998. Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eight Conference of the European Chapter of the Association for Computational Linguistics (ACL '98)*, pages 246–253. Association for Computational Linguistics.

Andrea Sand and Rainer Siemund. 1992. LOB-30 years on ... *ICAME Journal*, 16:119–122.

Joseph A. Smith and Colleen Kelly. 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36:411–430.

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

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

Nicholas Smith. 2003b. 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, editors, *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 714–723. Lancaster University: UCREL Technical Papers.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis T. 2000. Automatic text categorization in terms of genre and author.

Sanja Štajner and Ruslan Mitkov. 2011. Diachronic stylistic changes in British and American varieties of 20th century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP 2011*, pages 78–85.

Sanja Štajner. 2011. Towards a better exploitation of the Brown 'family' corpora in diachronic studies of British and American English language varieties. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 17–24.

Ingrid Westin and Christer Geisler. 2002. A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME Journal*, 26:133–152.

Ingrid Westin. 2002. *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.