Diachronic Changes in Text Complexity in 20th Century English Language: An NLP Approach

Sanja Štajner and Ruslan Mitkov

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

Abstract

A syntactically complex text may represent a problem for both comprehension by humans and various NLP tasks. A large number of studies in text simplification are concerned with this problem and their aim is to transform the given text into a simplified form in order to make it accessible to the wider audience. In this study, we were investigating what the natural tendency of texts is in 20th century English language. Are they becoming syntactically more complex over the years, requiring a higher literacy level and greater effort from the readers, or are they becoming simpler and easier to read? We examined several factors of text complexity (average sentence length, Automated Readability Index, sentence complexity and passive voice) in the 20th century for two main English language varieties – British and American, using the 'Brown family' of corpora. In British English, we compared the complexity of texts published in 1931, 1961 and 1991, while in American English we compared the complexity of texts published in 1961 and 1992. Furthermore, we demonstrated how the state-of-the-art NLP tools can be used for automatic extraction of some complex features from the raw text version of the corpora.

Keywords: language change, text complexity, corpus analysis

1. Introduction

Language change could be defined as "a failure in the transmission across time of linguistic features" (Kroch, 2008). This change occurs at various levels of the language structure – vocabulary, phonology, morphology and syntax (Kroch, 2008). The wide area of sociolinguisic and historical linguistic studies is concerned with how and why these changes occur. It is expected that syntactic changes require more time to be perceived. However, studies conducted on the 'Brown family' of corpora, e.g. (Mair and Leech, 2006), demonstrated that a 30-year time period is enough for many syntactic changes to be noticed.

In this study, we investigated diachronic changes of several features which could count for syntactic text complexity. Syntactically complex text could impede its comprehension by humans (those with some kind of language impairment or second language learners, for instance) and its processing by computer (e.g. parsing, machine translation, summarisation). Numerous studies in the area of text simplification have the aim of transforming the given text in a simplified form in order to facilitate its use either by humans or machines. We wanted to explore what the natural tendency of text is in 20th century English language. Do they tend to become simpler or more complex over the years?

We examined four factors of text complexity (average sentence length, Automated Readability Index, sentence complexity and passive voice) in the 20th century for two main English language varieties – British and American. In British English, we compared the complexity of texts published in 1931, 1961 and 1991, while in American English we compared the complexity of texts published in 1961 and 1992 (Figure 1). All experiments were conducted on the 'Brown family' of corpora – diachronic corpora of 20th century written English language. As the corpora of British and American English texts from 1961 and 1991/2 are mutually comparable, we were also able to compare the trends of change between the two language varieties.

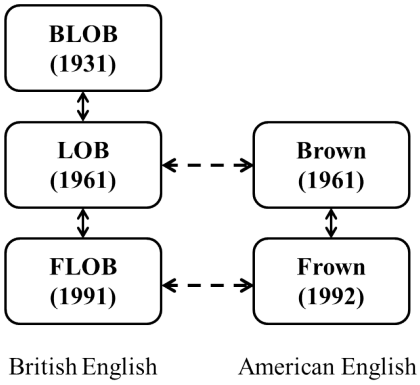

Figure 1: The 'Brown family' of coprora

Furthermore, we demonstrated how the state-of-the-art NLP tools can be used for automatic extraction of certain syntactic features from the raw text version of the corpora, thus enabling this type of study to be conducted without time-consuming and labour-intensive human annotation.

1.1. Corpora

The 'Brown family' of corpora is a group of comparable diachronic corpora of British and American English (Figure 1). The American part (Brown[1] and Frown[2]) contains texts published in 1961 and 1992, respectively. The British part (BLOB[3], LOB[4] and FLOB[5]) contains texts published

[1]The Brown University corpus of written American English
[2]The Freiburg - Brown Corpus of American English
[3]The Lancaster1931 Corpus
[4]The Lancaster-Oslo/Bergen Corpus
[5]The Freiburg-LOB Corpus of British 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

in 1931±3, 1961 and 1991, respectively. The four corpora (Brown, Frown, LOB and FLOB) are publicly available as a part of the ICAME corpus collection[6], while the fifth corpus (BLOB) is still not publicly available. As they are all mutually comparable (Leech and Smith, 2005), they provide the possibility for two types of investigations – diachronic (for each of the language varieties separately) and synchronic (between the two language varieties).
Each corpus is a million word corpus, consisting of 500 texts of about 2000 running words each, selected at a random point in the original source. The sampling range in all five corpora covers 15 text genres, further grouped into four more generalised categories (Table 1). This structure of the corpora allows three different approaches to the exploitation of the corpora in diachronic studies:

- Differentiating between texts only across two different language varieties.
- 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.
- 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.

The 'Brown family' of corpora were 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). All these studies used the second approach, differentiating only between texts across the four main categories (Press, Prose, Learned and Fiction). Following the discussion in (Štajner, 2011), 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 text complexity changes. To the best of our knowledge, this is the first diachronic study conducted on these corpora using this approach.

1.2. Features

We investigated diachronic changes of four factors that indicate text complexity:

- Average sentence length (ASL)
- Automated Readability Index (ARI)
- Sentence complexity (COMPLEX)
- Passive constructions (PASS)

Average sentence length is computed as the total number of words (*w*) divided by the total number of sentences (*s*) in the given text (Eq. 1).

$$ASL = \frac{w}{s} \quad (1)$$

It is known that longer sentences are more difficult to follow and require more effort to be understood (Graesser et al., 2001). Long sentences are especially difficult for language-impaired users (Siddharthan, 2002; Klebanov et al., 2004) and adult learners of English language (Siddharthan, 2002). Shorter sentences also demonstrate better performances in various NLP tasks, e.g. parsing, machine translation or text summarisation (Siddharthan, 2002).
Automated Readability Index (Senter and Smith, 1967; Kincaid and Delionbach, 1973) is one of the many readability measures which are used for assessing the necessary US grade level for understanding the given text. In the early eighties, it was listed among eleven most commonly used readability measures (McCallum and Peterson, 1982). Despite being one of the first readability indices, it is still widely in use, most probably due to the fact that it could be easily computed automatically and with a high precision. Unlike most of the other readability indices, which require

[6]http://www.hit.uib.no/icame

counting syllables in the text (a process which cannot be automatically done with a high precision), ARI requires only the number of characters (c), words (w) and sentences (s) in the texts (Eq. 2). These can be computed with very high precision using the standard NLP tools.

$$ARI = 4.71\frac{c}{w} + 0.5\frac{w}{s} - 21.43 \quad (2)$$

Sentence complexity (Eq. 3) was measured in terms of the number of verb chains (finite predicators) in the sentence, as the ratio between the number of sentences with one finite predicator at the most (*simple sentences*) and the number of sentences with two or more finite predicators (*complex sentences*). The number of finite predicators was calculated automatically using the state-of-the-art Connexor's Machinese Syntax parser[7] (Section 3.1.). The value of the feature COMPLEX is 1 for the text which contains equal number of simple and complex sentences, less than 1 for the text which contains more complex than simple sentences, and greater than 1 for the text which contains more simple than complex sentences (i.e. higher value of the feature COMPLEX indicates a simpler text)

$$COMPLEX = \frac{simple\_sentences}{complex\_sentences} \quad (3)$$

A high number of verb chains (finite predicators) in the sentence indicate presence of many clauses within the sentence, which can impede its comprehension by language-impaired readers (Siddharthan, 2002; Klebanov et al., 2004).

Passive constructions were extracted automatically using the information given by the parser. We counted all passive and active constructions recognised by the parser and then presented the feature as the ratio between the number of passive constructions and all recognised passive and active constructions in the text (Eq. 4).

$$PASS = \frac{passiv}{passiv + active} \quad (4)$$

The passive sentences were reported to be difficult for the language-impaired readers (Carroll et al., 1999; Klebanov et al., 2004).

2. Related work

Diachronic changes in the average sentence length (ASL) and Automated Readability Index (ARI) in the period 1961–1991/2 were already investigated in the same corpora by Štajner and Mitkov (2011), using similar 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. 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 30-year time gaps (1931–1961 and 1961–1991) in British English and examine whether the trend of change was stable during the whole 60-year period.

Diachronic changes in the use of passive voice were already investigated using the same corpora by Leech and Smith (2006) and Leech (2004) and the results indicated an overall decrease in the use of passive voice in both British and American English. However, the methodology used in those studies had many differences from the one presented here. They differentiated only between texts across the four main text categories (Press, Prose, Learned and Fiction), the methodology for extracting passive constructions was not specified and the log likelihood function was used for testing the statistical significance of the results.

To the best of our knowledge, there have not been any previous studies investigating diachronic changes in this type of sentence complexity.

3. Methodology

We conducted two sets of experiments:

- Diachronic changes in text complexity (ASL, ARI, COMPLEX, PASS) in British English in two periods: 1931–1961 and 1961–1991
- Diachronic changes in text complexity (ASL, ARI, COMPLEX, PASS) in American English in the period 1961–1992

The corpora was used in the raw text format and parsed with the state-of-the-art Connexor's Machinese Syntax parser. The XML output of the parser provided the information about the sentence and word boundaries, passive and active constructions and finite predicators. This information was used for the automatic feature extraction. For each language variety, year, category and genre, the value of the corresponding feature was calculated separately for each text. Details of detecting the passive and active constructions, and finite verbs from the parser's output are given in the following two subsections.

3.1. Finite predicators

The parser's output contains four functional tags: @+FMAINV (finite main predicator), @−FMAINV (nonfinite main predicator), @+FAUXV (finite auxiliary predicator) and @−FAUXV (nonfinite auxiliary predicator). For instance, in the sentence:

> *"All that has been said in the foregoing pages about what is meant by a lady, is true for all women and young girls."* (LOB:F08),

the used verbs have the following corresponding functional tags: *has* → @+FAUXV, *been* → @−FAUXV, *said* → @−FMAINV, *is* → @+FAUXV, *meant* → @−FMAINV and *is* → @+FMAINV. For the sentence complexity features used in this study, the number of finite predicators was counted as the number of tokens with the @+FMAINV or @+FAUXV functional tag. In the aforementioned sentence, we counted 3 finite predicators, which correspond to the following 3 verb chains: {*has been said*}, {*is meant*} and {*is*}.

[7] www.connexor.eu

3.2. Passive voice

The <syntax> tag of the parser's output is built from the surface syntactic and functional tags. We used this syntactic information to count the number of passive and active verb forms in the given text. More specifically, we were interested in two surface syntactic tags: %VP (main verb in a passive verb chain) and %VA (main verb in an active verb chain); and in two functional tags: @+FMAINV (finite main predicator) and @−FMAINV (nonfinite main predicator). We counted the number of passive and active verb forms by: (i) increasing the number of found active forms whenever a @+FMAINV %VA or @−FMAINV %VA <syntax> tag was found, (ii) increasing the number of found passive forms whenever a @−FMAINV %VP <syntax> tag was found.

Each main verb in the parser's output represents one verb form and has one of the three previously mentioned tags. The combination <syntax>@+FMAINV %VP</syntax> cannot possibly occur, as passive constructions always contain an auxiliary verb and therefore the functional tag %VP can stand only next to nonfinite main predicator (@−FMAINV).

3.3. Statistical significance

Statistical significance tests of differences of means are divided into two main groups: parametric (which assume that data is approximately normally distributed) and non-parametric (which do not assume any specific data distribution). In the case that the normality assumption is met, it is preferable to use parametric tests (e.g. t-test) as they have a greater power than the non-parametric ones (Garson, 2012a). Gardner (1975) calls for caution when using the t-test, as it can be unreliable when the samples come from two widely different shaped distributions (Garson, 2012a). Moore (1995) suggests that for sample sizes smaller than 15 (in our case genres M and R, Table 1) data for t-test should be normally distributed, while for samples between 15 and 40 (genres B, C, D, H, K, L, N and P in our case, Table 1) it should be approximately normal and without outliers. For sample sizes greater than 40 (genres A, F, G and J in our case, Table 1), Moore (1995) believes that data for t-test can even be markedly skewed (Garson, 2012a). Therefore, we first examined whether our data followed approximately normal distribution for each feature, genre and corpus. We used the standard test of normality (Shapiro-Wilk's W test) offered by SPSS EXAMINE module, recommended for small and medium sample sizes (n $<$ 2000). Additionally, we applied the Boxplot tests of the normality assumption in SPSS which detected the outliers in each data set. As an illustration, we present the output of this test for feature complexity in the BLOB corpus (Figure 2). The height of each rectangle inside the graph (Figure 2) indicates the spread of the values for the corresponding feature and genre, while the horizontal dark line indicates the mean. If this line is not in the middle of the rectangle it indicates that the distribution of the feature is not normal for that genre. The "whiskers" of the rectangles represent the smallest and largest observed values which are not outliers, while the outliers are marked as numbered cases beyond the whiskers. As it can be noted from Figure 2, in most of the genres in the BLOB corpus, the feature COMPLEX was either not normally distributed or it contained outliers. Similar results were obtained for other features and other corpora. Therefore, following the suggestions of Moore (1995) and the aforementioned discussion, we decided not to use the t-test for comparison of the means. Instead, we used the Kolmogorov-Smirnov Z test (a non-parametric test). This test does not assume any specific distribution of the data (Garson, 2012b).

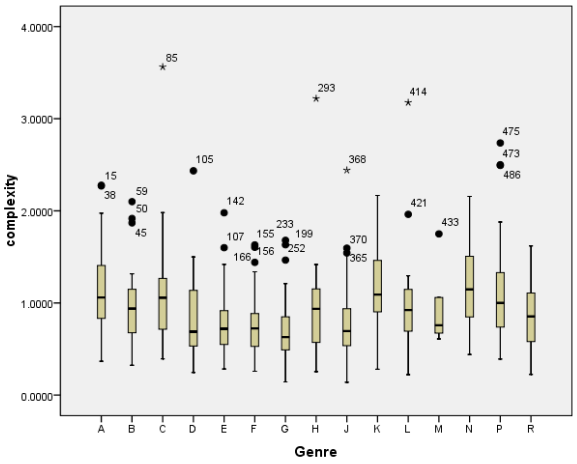

Figure 2: Sentence complexity in BLOB (1931)

4. Results

The results of the investigation of diachronic changes in the four features (ASL, ARI, PASS and COMPLEX) are given separately in the following four subsections. In all 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 value of the feature in those years for the corresponding language variety. Columns '1931–1961', '1961–1991' and '1961–1992' contain the information about the changes of the feature in those periods for the corresponding language varieties. Their subcolumn 'sign.' represents the calculated two-tailed statistical significance of the differences between the corresponding means, using the Kolmogorov-Smirnov Z test. Statistically significant changes at a 0.05 level of significance (sign. $<$ 0.05) are printed in bold. The subcolumn 'change' contains the relative change in the observed period, calculated as a percentage of the starting value. Sign '+' stands for an increase and sign '−' for a decrease in the observed period.

4.1. Average sentence length (ASL)

The results of the diachronic comparison of ASL in British and American English (Table 2) indicate that the average sentence length increased in two genres of British English – G (Belles Lettres, Biographies, Essays) and R (Humour) in the period 1931–1961. In neither of these two genres was a significant change in ASL reported in the next 30-year period (1961–1991). However, the results presented in Table 2 indicated a significant decrease of ASL in the pe-

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	18.90	0.634	+4.47%	19.75	0.076	+8.30%	21.39	**21.71**	**0.043**	**−5.93%**	**20.43**
B	20.94	0.744	−1.08%	20.72	0.324	−0.70%	20.57	20.42	0.187	−6.06%	19.18
C	19.54	0.240	+16.45%	22.76	0.734	+2.96%	23.43	22.68	0.954	−0.68%	22.52
D	22.91	0.240	−8.02%	21.07	0.112	+21.15%	25.53	24.99	0.112	−9.78%	22.54
E	22.14	0.897	+0.97%	22.36	0.897	−2.51%	21.79	20.86	0.211	−5.77%	19.66
F	21.13	0.316	+5.04%	22.20	0.634	+4.420%	23.18	22.47	0.847	−3.16%	21.76
G	**23.76**	**0.047**	**+7.47%**	**25.53**	0.908	−1.49%	25.15	24.09	0.787	−0.95%	23.86
H	25.95	0.799	+1.12%	**26.24**	**0.016**	**−12.85%**	**22.87**	**29.15**	**0.007**	**−20.73%**	**23.11**
J	25.46	0.054	+4.08%	26.50	0.560	−2.16%	25.93	25.09	0.120	−6.03%	23.57
K	14.88	0.564	+9.78%	16.33	0.367	−13.05%	14.20	15.89	0.367	−6.92%	14.79
L	14.55	0.893	+0.17%	14.58	0.893	−8.21%	13.38	13.42	0.139	−6.17%	12.59
M	15.86	0.893	−8.88%	14.46	0.893	−4.09%	13.86	13.98	0.893	−10.99%	12.45
N	14.85	0.220	−12.63%	12.97	0.782	+9.77%	14.24	13.74	0.564	+1.76%	13.98
P	13.78	0.367	+0.02%	13.78	0.782	−2.41%	13.45	15.15	0.220	−12.33%	13.28
R	**15.69**	**0.037**	**+18.21%**	**18.55**	0.124	−13.66%	16.02	19.74	0.336	−17.23%	16.34

Table 2: Diachronic changes of average sentence length (ASL)

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	10.70	0.316	+1.16%	10.83	0.076	+10.47%	11.96	12.20	0.634	−3.12%	11.82
B	11.18	0.187	+2.62%	11.47	0.928	−0.18%	11.45	11.46	0.324	−0.01%	11.46
C	10.56	0.112	+17.77%	12.44	0.734	+7.00%	13.31	12.78	0.954	+0.72%	12.87
D	12.01	0.240	−13.98%	**10.33**	**0.006**	**+37.72%**	**14.23**	13.40	0.734	+3.34%	13.85
E	11.21	0.731	+2.22%	11.46	0.539	+4.89%	12.02	11.49	0.699	−2.62%	11.18
F	**10.75**	**0.023**	**+9.48%**	**11.76**	0.206	+12.02%	13.18	12.32	0.161	+6.76%	13.16
G	12.52	0.157	+8.82%	13.62	0.307	+4.55%	14.24	13.22	0.292	+3.75%	13.72
H	14.52	0.998	+1.61%	14.75	0.799	−4.79%	14.05	17.34	0.388	−12.83%	15.12
J	13.86	0.329	+4.37%	14.47	0.172	+6.07%	15.35	14.54	0.329	+0.98%	14.68
K	6.45	0.782	+7.75%	6.95	0.367	−17.50%	5.73	6.98	0.122	−12.26%	6.12
L	6.13	0.441	−3.02%	5.95	0.893	−10.90%	5.30	5.24	0.441	−0.20%	5.23
M	7.68	0.893	−7.38%	7.11	0.893	−4.35%	6.80	6.88	0.893	−18.07%	5.64
N	**6.34**	**0.031**	**−22.92%**	**4.89**	0.064	+30.58%	6.38	5.58	0.945	+6.26%	5.93
P	5.49	0.782	−6.33%	5.14	0.564	+5.31%	5.42	6.17	0.220	−17.77%	5.07
R	**6.39**	**0.037**	**+41.72%**	**9.06**	0.124	−16.61%	7.56	9.87	0.336	−28.30%	7.08

Table 3: Diachronic changes of Automated Readability Index (ARI)

riod 1961–1991/2 in genre H (Miscellaneous) of both language varieties (British and American). They also reported a significant decrease of ASL in genre A (Press: Reportage) of American English in the same period. This change was much less pronounced than the one reported in genre H.

4.2. Automated Readability Index (ARI)

The investigation of ARI did not report any statistically significant changes of this feature in any of the text genres in American English (Table 3). In the Prose category of British English, the results reported a significant increase of ARI in genre D (Religion) in the period 1961–1991, and genre F (Popular Lore) in the period 1931–1961. These changes indicated a tendency of texts to become more complex in terms of the education and literacy level required for their comprehension. However, in neither of these two genres, the change was constant over the whole 60-year period (1931–1991). Two genres - N (Adventure and Western) and R (Humour) in British English, both belonging to the broader Fiction category, demonstrated a significant change of ARI in the period 1931–1961 (Table 3). It is interesting to observed that these two genres exhibited opposite directions of change (a decrease of ARI in genre N and an increase in genre R) over the same period (1931–1961), thus indicating that different genres inside the same broad text category do not necessarily follow the same trend of change.

4.3. Sentence complexity (COMPLEX)

The results presented in Table 4 indicate a tendency of texts in genre A (Press: Reportage) to become more complex[8] in terms of investigated sentence complexity (Section 1.2.) in both language varieties. In British English, this tendency was present during the whole 60-year period (1931–1991).

[8]Note that decrease of feature COMPLEX corresponds to a smaller ratio between the number of simple and complex sentences in the text, thus indicating a greater text complexity.

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	**1.16**	**0.043**	**−16.31%**	**0.97**	**0.000**	**−40.55%**	**0.58**	**0.91**	**0.001**	**−30.55%**	**0.63**
B	1.00	0.324	−8.29%	0.91	0.928	−11.16%	0.81	0.81	0.324	+8.85%	0.88
C	1.17	0.240	−21.92%	**0.91**	**0.017**	**−24.48%**	**0.69**	0.76	1.000	+1.82%	0.77
D	0.89	0.454	−18.57%	0.73	0.240	−13.03%	0.64	0.66	0.112	+26.32%	0.84
E	0.79	0.897	+3.59%	0.82	0.369	−10.81%	0.73	1.00	0.504	−13.84%	0.86
F	0.76	0.808	+4.21%	0.79	0.461	−2.74%	0.77	0.77	0.996	−1.35%	0.76
G	**0.68**	**0.018**	**−14.60%**	**0.58**	0.072	+12.70%	0.66	0.61	0.653	+5.50%	0.65
H	0.94	0.799	+1.76%	0.96	0.388	+17.44%	1.12	1.09	0.388	+28.69%	1.40
J	0.76	0.241	−0.93%	0.76	0.692	−3.58%	0.73	0.87	0.241	+5.08%	0.91
K	**1.21**	**0.002**	**−28.13%**	**0.87**	**0.031**	**+49.71%**	**1.30**	1.05	0.782	+3.70%	1.09
L	0.99	0.893	−10.43%	0.89	0.675	+16.11%	1.03	1.07	0.893	+11.69%	1.20
M	0.93	0.893	+37.70%	1.29	0.441	+14.84%	1.48	1.12	0.893	+22.48%	1.38
N	1.20	0.782	−8.03%	1.10	0.782	+1.70%	1.12	1.06	0.782	−0.53%	1.05
P	1.17	0.564	−23.49%	0.89	0.367	+4.91%	0.94	0.96	0.564	+7.03%	1.03
R	0.93	0.336	−4.84%	0.86	0.336	+39.61%	1.20	0.67	0.699	+23.85%	0.82

Table 4: Diachronic changes of sentence complexity (COMPLEX)

Genre	British English							American English			
	1931	1931–1961		1961	1961–1991		1991	1961	1961–1992		1992
		sign.	change		sign.	change			sign.	change	
A	**38.30**	**0.012**	**−18.60%**	**31.18**	**0.023**	**−12.95%**	**27.14**	**30.67**	**0.000**	**−28.64%**	**21.88**
B	**29.11**	**0.049**	**−11.82%**	**25.67**	0.744	−3.37%	24.81	24.57	0.100	−13.27%	21.31
C	32.68	0.734	+0.64%	32.89	0.112	−15.40%	27.82	30.73	0.454	−5.14%	29.15
D	35.19	0.454	+2.75%	36.16	0.240	−8.46%	33.10	29.02	0.734	+8.46%	31.48
E	41.09	0.144	−13.90%	35.38	0.237	−16.83%	29.42	32.55	0.124	−22.04%	25.38
F	34.04	0.993	−3.99%	32.69	0.634	−2.17%	31.98	31.16	0.687	−2.37%	30.42
G	31.06	0.800	−1.95%	**30.46**	**0.030**	**+9.52%**	**33.36**	28.71	0.970	−0.99%	28.42
H	**44.94**	**0.035**	**−11.40%**	**39.81**	0.388	−5.24%	37.73	41.38	0.134	−16.22%	34.67
J	45.06	0.560	+3.00%	**46.42**	**0.008**	**−13.00%**	**40.39**	43.25	0.120	−10.45%	38.73
K	16.79	0.782	−8.70%	15.33	0.782	+0.37%	15.38	17.46	0.945	−2.78%	16.97
L	17.64	0.441	+0.70%	17.77	0.259	−20.38%	14.15	15.19	0.893	+7.07%	16.26
M	24.45	0.441	−17.80%	20.10	0.893	+0.13%	20.13	20.06	0.441	−21.17%	15.81
N	18.41	0.122	−19.29%	14.86	0.122	+27.14%	18.89	16.05	0.367	+19.41%	19.17
P	**14.60**	**0.014**	**−21.72%**	**11.43**	0.220	+16.13%	13.27	12.67	0.367	−10.07%	11.40
R	20.05	0.124	+18.00%	23.66	0.336	−15.65%	19.96	**22.04**	**0.009**	**−44.58%**	**12.21**

Table 5: Diachronic changes in the use of passive voice (PASS)

In genre C (Press: Review) in British English, texts also demonstrated a tendency of becoming more complex in the period 1961–1991, as well as in genre G (Belles Lettres, Biographies, Essays) in the period 1931–1961. In genre K (General Fiction) in British English, texts first indicated a tendency to become more complex in the period 1931–1961 and then simpler in the period 1961–1991 (Table 4).

4.4. Passive voice (PASS)

The results of the investigation of diachronic changes in the use of passive voice (Table 5) indicated a decrease in most of the cases where the change was reported to be significant – genre A (Press: Reportage) in both language varieties, genres B (Press: Editorial), H (Miscellaneous) and R (Humour) in British English in the period 1931–1961, genre J (Science) in British English in the period 1961–1991 and genre R (Humour) in American English in the period 1961–1992. The only exception was observed in genre G (Belles Lettres, Biographies, Essays) in British English where the use of passive voice was reported to be increased in the period 1961–1991 (Table 5).

5. Conclusions

The presented study demonstrated possibilities of using the state-of-the-art NLP tools for automatic extraction of some complex features in diachronic studies. It also showed that a 30-year time gap is enough for some specific syntactic changes to be noticed. Furthermore, it indicated that these changes are usually not constant throughout two consecutive 30-year time intervals for most of the text genres (the exception being genre A – Press: Reportage in the case of COMPLEX and PASS features). More surprisingly, the results presented in this study indicated that the changes reported in these two consecutive time intervals could often follow opposite directions (first an increase and then a decrease or vice versa). They also showed that different genres which belong to the same broader text category (Press,

Prose, Learned or Fiction), do not necessarily follow the same trends of change.

Most importantly, the results presented in this study allowed us to make some preliminary conclusions about the natural tendencies of text complexity in 20th century English language. They indicated that in the period 1961–1991/2, average sentence length had decreased in all genres of British and American English where the change was reported to be statistically significant. However, this change was not necessarily followed by a decrease of sentence complexity (in terms of number of finite predicates). In most genres where the change in the use of passive voice was reported to be significant, the results indicated a decrease of passive constructions (only exception being genre G – Belles Lettres, Biographies and Essays in British English in the period 1961–1991). Text complexity in terms of readability index (ARI) did not indicate any significant changes in any genre of American English in the period 1961–1992. In the corresponding period (1961–1991) in British English, only one genre (D – Religion) reported a significant increase of the readability index, thus indicating that texts belonging to this genre became more difficult to read over the observed period, requiring a higher level of literacy and education for their comprehension.

6. Acknowledgements

We would like to thank Prof Geoffrey Leech for his advice and assistance with the resources.

7. References

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for langauge-impaired readers. In *Proceedings of EACL99.*

Paul Leslie Gardner. 1975. Scales and statistics. *Review of Educational Research*, 45:43–57.

David G. Garson. 2012a. Significance. Statnotes: Topics in Multivariate Analysis.

David G. Garson. 2012b. Tests for two independent samples: Mann-Whitney U, Wald-Wolfowitz runs, Kolmogorov-Smirnov Z, & Moses extreme reactions tests. Statnotes: Topics in Multivariate Analysis.

Arthur C. Graesser, Ashish B. Karnavat, Frances K. Daniel, Elisa Cooper, Shannon N. Whitten, and Max M. Louwerse. 2001. A computer tool to improve questionnaire design. In *Statistical Policy Working Paper 33, Federal Committee on Statistical Methodology*, pages 36–48. Washington, DC: Bureau of Labor Statistics.

Peter J. Kincaid and Leroy J. Delionbach. 1973. Validation of the Automated Readability Index: A follow-up. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 15(1):17–20.

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text Simplification for Information-Seeking Applications. On the Move to Meaningful Internet Systems. In *Lecture Notes in Computer Science*, volume 3290, pages 735–747. Springer-Verlag, Berlin Heidelberg New York.

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

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.

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

David S. Moore. 1995. *The basic practice of statistics*. NY: Freeman and Co.

R. J. Senter and E. A. Smith. 1967. Automated Readability Index. Technical report, Defense Technical Information Center. United States.

Advaith Siddharthan. 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (Hyderabad, India, December 13-15, 2002). IEEE Computer Society*, pages 64–71.

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.

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.