Can Statistical Tests Be Used for Feature Selection in Diachronic Text Classification?

Sanja Štajner and Richard Evans

Research Group in Computational Linguistics
University of Wolverhampton, UK
{sanjastajner,R.J.Evans}@wlv.ac.uk

Abstract. In spite of the great number of diachronic studies in various languages, the methodology for investigating language change has not evolved much in the last fifty years. Following the progressive trends in other fields, in this paper, we argue for the adoption of a machine learning approach in diachronic studies, which could offer a more efficient analysis of a large number of features and easier comparison of the results across different genres, languages and language varieties. We suggest the use of statistical tests as an initial step for feature selection in an approach which uses the F-measure of the classification algorithms as a measure of the extent of diachronic changes. Furthermore, we compare the performance of the classification task after the feature selection made by statistical tests and the CfsSubsetEval attribute selection algorithm. The experiments were conducted on the British part of the biggest existing diachronic corpora of 20th century written English language – the 'Brown family' of corpora, using 23 different stylistic features. The results demonstrated that the use of the statistical tests for feature selection can significantly increase the accuracy of the classification algorithms.

1 Introduction

Approaches to text classification continue to develop from those based on knowledge engineering techniques that prevailed in the 1980s, in which classifiers were defined manually by domain experts. In the 1990s, these methods were superseded by those relying on machine learning which provided high levels of efficacy, cost effectiveness, in terms of time and manual effort, and easy adaptation for use in different scenarios and domains [26]. Continuous methodological improvements in the field of text classification has led to the adoption of more effective and less labour intensive approaches in place of those requiring a large amount of human annotation. By contrast, approaches to the linguistic study of stylistic variation and change were more conservative and did not follow the progressive trends in other related fields.

Early work in the field of stylistic variation and change was based on historical and sociolinguistic approaches, e.g. [14,1,4]. The next generation of stylistic

A.-H. Dediu et al. (Eds.): SLSP 2013, LNAI 7978, pp. 273–283, 2013.

variation studies, e.g. [7,8] employed a corpus-based methodology and the multi-dimensional framework presented in [5,6]. The same methodology was used in a great number of subsequent diachronic studies, e.g. [32,33]. Another set of corpus-based diachronic studies was initiated by the emergence of the diachronic part of the 'Brown family' of corpora in the 1990s. These corpora offered a possibility for diachronic comparison of various lexical, grammatical, syntactic and stylistic features in two major English language varieties – British and American [21] in the period 1961–1991/2. Many diachronic studies of these corpora (e.g. [23,22,24]), shared the same methodology. The corpora were POS tagged, change was presented in terms of absolute and relative differences between the corpora and the statistical significance of that change was measured using the log likelihood function. The first attempt at completely automated feature extraction from the raw text version of these 'Brown family' of corpora in diachronic studies was reported in [30]. The corpora were parsed with Connexor's Machinese Syntax parser[1] and the features were automatically extracted from the parser's output. Statistical significance of the results was measured by the t-test.

In this paper, we adopt the hypothesis that diachronic language change could be seen as a classification problem and therefore addressed by machine learning techniques. To illustrate, if we wish to investigate the degree of change in certain features between the texts published in 1961 and 1991, we could train a classifier on a representative set of labeled texts (using the selected features as variables) and then classify a set of randomly selected unlabeled texts using this classifier. The performance of the classifier (in terms of the F-measure), would then represent the extent of diachronic change in the selected features. In the cases where diachronic changes were most pronounced, the F-measure obtained by the classification algorithm will be at its highest level. More importantly, by using the machine learning approach, we could also take advantage of existing attribute selection algorithms in order to single out from a large set of initial features, those features which underwent the most extensive changes over the observed period. In this paper, we wanted to investigate whether statistical tests and the CfsSubsetEval attribute selection algorithm [15] would improve the accuracy of diachronic classification and whether they would select the same subsets of features. In order to do so, we applied several well-known classification algorithms (Naïve Bayes and different versions of Logistic and Support Vector Machines functions) in Weka[2] on the texts from the British part of the 'Brown family' of corpora, using different subsets of the 23 initial features.

2 Related Work

Altmann et al. [3] and Kroch [19] proposed the logistic function as the underlying S-shaped curve of linguistic change. Although the correctness of this choice was not proved at the time, it was generally considered appropriate to use this function in statistical studies of changing percentages of alternating forms over

[1] http://www.connexor.eu
[2] http://www.cs.waikato.ac.nz/ml/weka/

time ([2,29] in [18]). Twenty years later, Geisler [13] used logistic regression in the study of relativisation variation in Ulster English [12]. Therefore, we decided to include the classifier based on the logistic function for our experiments.

A survey of previous diachronic studies of the 'Brown family' of corpora motivated the development of our initial feature set. Leech and Smith [22] reported a reduction in the use of passive voice between 1961 and 1991/2 in both British and American English. Stajner and Mitkov [30] investigated diachronic changes of four stylistic features: average sentence length, Automated Readability Index [27], lexical density and lexical richness [11]. The results revealed statistically significant changes in these features between 1961 and 1991/2 in both varieties of English and across all four main text categories (Press, Prose, Learned and Fiction). Although in both cases the authors differentiated only between texts across the four main text categories, it is reasonable to expect that some significant changes of these features would also be reported in a separate investigation of the sub-genres (A–R, see Table 1). Mair et al. [23] compared the frequency of occurrence of words with particular parts of speech in the British part of the corpora. They reported a significant increase in the number of nouns and adjectives and a decrease in the frequency of occurrence of pronouns in all four main text categories over the observed period (1961–1991). Usage of verbs underwent a significant increase in the Press and Science categories, and a significant decrease in the Prose and Fiction categories. In the study reported in the current paper, we investigated nine different POS tags. We differentiated between texts across sub-genres (A–R) and calculated two different types of tag frequencies – tag frequency as a percentage of the selected tag in the whole text and tag frequency as an average per sentence. Stajner and Mitkov [31] reported some significant changes in sentence complexity in the period 1961–1991 in three genres of the British part of the corpora.

Identifying the best set of features for a particular classification task is one of the central problems in machine learning. The CfsSubsetEval attribute selection algorithm uses a correlation based approach to the feature selection problem. It is based on the idea that "good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other" [15]. When compared with a wrapper, the CfsSubsetEval gave similar results to the wrapper and even outperformed the wrapper on small datasets [15].

3 Methodology

The corpora, features and experimental settings used in this study are presented in the following three subsections.

3.1 Corpora

We used only the British part of the aforementioned 'Brown family' of corpora [21]:

- the Lancaster-Oslo/Bergen Corpus of British English (LOB);
- the Freiburg-LOB Corpus of British English (F-LOB).

These two corpora are mutually comparable [21] and contain texts published in 1961 and 1991, respectively.[3] Each corpus consists 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 (A–C), Prose (D–H), Learned (J) and Fiction (K–R). The corpora were used in their untagged, raw text versions and parsed with

Table 1. Structure of the corpora

Category	Code	Genre	# texts
Press	A	Press: Reportage	44
	B	Press: Editorial	27
	C	Press: Review	17
General Prose	D	Religion	17
	E	Skills, Trades and Hobbies	38
	F	Popular Lore	44
	G	Belles Lettres, Biographies, Essays	77
	H	Miscellaneous	30
Learned	J	Science	80
Fiction	K	General Fiction	29
	L	Mystery and Detective Fiction	24
	M	Science Fiction	6
	N	Adventure and Western	29
	P	Romance and Love Story	29
	R	Humour	9

Connexor's Machinese Syntax parser in order to achieve consistent, highly accurate sentence splitting, tokenisation, lemmatisation and part-of-speech, syntactic and functional tagging.

3.2 Features

Twenty-three stylistic features (automatically extracted from the parser's output) were exploited (Table 2). Nine different POS tags were considered: N (noun)[4], A (adjective), PRON (pronoun), DET (determiner), ADV (adverb), V (verb)[5], CC (coordinative conjunction), CS (subordinate conjunction), PREP (preposition). Each POS tag was represented by two separate features: (1) the percentage of tokens tagged with that POS in each text; and (2) the average number of tokens tagged with that POS per sentence. Therefore, the last two rows in Table 2 account for 18 different features in total.

Connexor's Machinese Syntax parser was reported to achieve 99.3% accuracy in POS tagging on Standard Written English (benchmark from the Maastricht Treaty) [10]. Details of the parser's tokenisation and lemmatisation processes can be found in [30], while the details of passive and finite predicator marking procedures can be found in [31].

[3] Both corpora are publicly available as a part of the ICAME corpus collection at `http://www.hit.uib.no/icame`

[4] The ABBR morphological tag was counted as occurrence of a noun (N).

[5] The morphological tags ING (present participle) and EN (past participle) were counted as occurrences of a verb (V).

Table 2. Features (Key: c – total number of characters in a text; w – total number of words in a text; s – total number of sentences in a text; tokens – total number of tokens in a text; passive – total number of passive constructions in a text; active – total number of active constructions in a text; simple_s – total number of sentences in a text which have 1 finite predicator at the most; complex_s – total number of sentences in a text which have 2 or more finite predicators)

Feature	Code	Formula
Average sentence length	ASL	ASL = w/s
Coleman-Liau readability index	CLI	CLI = 5.89(c/w) - 29.5(s/w) - 15.8
Lexical richness	LR	LR = (unique lemmas)/(unique tokens)
Passive voice (%)	PASS	PASS = passive/(passive+active)
Sentence complexity	COMPL	COMPL = (simple_s)/(complex_s)
Part-of-Speech (%)	POS_per	POS_per = POS/tokens
Part-of-Speech (on average per sentence)	POS_av	POS_av = POS/s

3.3 Experimental Settings

First, we wanted to explore whether it is reasonable to expect that these 23 stylistic features would differ between the texts published in 1961 and those published in 1991, if we investigate them in each sub-genre (A–R) separately. Therefore, we conducted two sets of preliminary experiments. The Shapiro-Wilk's W test (offered by SPSS) was applied in order to determine whether the features follow the normal distribution across all thirteen genres in the two observed years. Additionally, the skewness and the existence of outliers was examined by using the box-plot. As the results demonstrated that the distribution of certain features in certain genres was significantly different from the normal distribution, we were not able to apply the t-test as a measure of statistical significance of the changes in all cases. In the cases where the distribution of the features was not approximately normal in both samples, we applied a non-parametric statistical test (Kolmogorov-Smirnov test).[6] The results of these statistical tests revealed significant differences in all 23 features, though in different subsets across the thirteen analysed genres. After these two preliminary experiments, which justified the use of the 23 initial features, we applied several Machine Learning algorithms in Weka Experimenter [34]: Support Vector Machines [25,17], Naïve Bayes [16], Logistic [9] and Simple Logistic [20,28] to classify the texts according to the year of publication – 1961 or 1991, using 5-fold cross-validation with 10 repetitions. The experiments were conducted separately for each text genre (A–P, excluding M)[7], thus enabling a comparison of diachronic changes in the period 1961–1991 across these thirteen text genres. We conducted three sets of experiments which differed in the subset of features they used:

- Experiment I: Using all 23 features;
- Experiment II: Using only the features marked as significant (at a 0.05 level of significance) by the statistical tests;

[6] We followed the same method for deciding on the appropriate statistical test as described in [31].

[7] Genres M and R were excluded from our analysis as they contain less than 10 texts in each corpus which is insufficient for the Machine Learning approach.

– Experiment III: Using only the features selected by the CfsSubsetEval attribute selection algorithm [15].

The comparison of the results obtained from these three experiments allowed us to further explore the potential of such a machine learning approach in diachronic studies. The goal was to answer the following questions:

1. Could the use of the statistical tests as a preprocessing (feature selecting) step improve the classification accuracy? (Comparison of the results of the first and second experiment).
2. Would the classification accuracy be improved if only the features selected by the CfsSubsetEval attribute selection algorithm were used? (Comparison of the results of the first and third experiment).
3. Would the CfsSubsetEval attribute selection algorithm be consistent with the results of the statistical tests? (Comparison of the results of the second and third experiment).

4 Results and Discussion

The results of the classification experiments are presented in Table 3. Column 'Exp.' contains the label of the experiment (I, II, III or III$^+$). While running the CfsSubsetEval attribute selection algorithm in the third experiment, it was noted that in the cases when it actually cannot find the best subset of features, the algorithm returns the first feature in the given list of all features as the best one. In those cases, the value of 'the merit of best subset found' is zero, while in the case of successful feature selection 'the merit of best subset found' has a value greater than zero. Therefore, in the first of these cases, an additional classification experiment was carried out – Exp. III$^+$– on the features selected by the CfsSubsetEval algorithm applied only on the subset of the initial set of features (those features reported as significant by the statistical tests). Columns 'NB', 'Log.', 'SLog.', 'SMO(s)', and 'SMO(n)' contain the F-measures of the five following classification algorithms: Naïve Bayes, Logistic, Simple Logistic, Support Vector Machines (with previous normalisation of the data), Support Vector Machines (with previous standardisation of the data) used in 5-fold cross-validation with 10 repetitions. Column '#feat.' contains the number of features used in each experiment. The highest obtained F-measure in each genre is shown in bold. As each genre contains the same number of texts published in 1961 and those published in 1991, the baseline accuracy in all genres could be considered to be 0.5. All comparisons between the results of experiment I and any other experiment were done pairwise using the paired t-test at a 0.05 level of significance. The statistically significant differences are shown in bold, with significantly lower results presented with an '*', and significantly higher results presented with a 'v'.

From the results presented in Table 3 it can be noted that in all cases where a statistically significant difference between the results of the first and second experiments was reported (genres B and N), the F-measure was lower in experiment I which uses all features. This indicates that the use of statistical tests

Table 3. Results of the classification experiments

Code	Genre	Exp.	NB	Log.	SLog.	SMO(n)	SMO(s)	#feat.
A	Press: Reportage	I	0.69	0.74	0.81	0.79	**0.84**	23
		II	0.76	0.81	0.78	0.78	0.77	8
		III	0.76	0.74	0.72	0.74	**0.73***	3
B	Press: Editorial	I	0.62	0.68	0.74	0.75	0.66	23
		II	**0.80v**	**0.81v**	0.79	0.77	0.78	4
		III	0.72	0.73	0.73	0.72	0.73	1
C	Press: Review	I	0.61	0.72	0.73	0.74	0.76	23
		II	0.69	0.71	0.72	0.73	0.71	2
		III	0.75	0.72	0.74	**0.78v**	0.75	1
D	Religion	I	0.71	0.72	0.76	0.74	0.74	23
		II	0.78	0.63	0.79	0.79	0.77	5
		III	**0.84**	0.80	0.81	0.81	0.82	1
E	Skills, Trades and Hobbies	I	0.56	0.62	**0.66**	0.64	0.62	23
		II	0.62	0.62	0.62	0.54	0.61	2
		III	0.61	0.61	0.61	0.59	0.61	1
		III$^+$	0.61	0.61	0.61	0.59	0.61	1
F	Popular Lore	I	0.45	0.54	0.65	0.61	0.61	23
		II	0.53	0.62	0.64	0.56	0.61	3
		III	0.59	**0.67**	**0.67**	0.50	0.66	1
G	Belles Lettres, Biographies...	I	0.57	0.70	**0.72**	0.68	0.71	23
		II	0.60	0.67	0.67	0.63	0.66	6
		III	0.63	0.62	0.61	0.62	0.64	2
H	Miscellaneous	I	0.55	0.60	**0.65**	0.59	0.61	23
		II	0.55	0.57	0.62	0.55	0.58	3
		III	0.63	0.62	0.62	0.64	0.61	1
		III$^+$	0.50	0.62	0.62	0.35	0.55	1
J	Science	I	0.65	**0.74**	0.71	0.69	**0.74**	23
		II	0.70	0.72	0.72	0.69	0.71	6
		III	0.71	0.73	0.72	0.69	0.73	3
K	General Fiction	I	0.52	0.47	0.48	0.55	0.50	23
		II	0.63	**0.65**	0.64	**0.65**	0.64	3
		III	0.54	0.55	0.56	0.43	0.51	1
		III$^+$	0.59	0.61	0.60	0.50	0.55	1
L	Mystery and Detective Fiction	I	0.35	0.57	**0.58**	0.54	0.56	23
		III	0.50	0.46	**0.58**	0.37	0.42	1
N	Adventure and Western	I	0.69	0.57	0.55	0.58	0.45	23
		II	0.70	0.71	0.68	0.69	**0.69v**	2
		III	0.69	0.67	0.68	**0.72**	**0.70v**	1
P	Romance and Love Story	I	0.60	0.56	0.54	0.56	0.51	23
		II	0.65	0.64	0.63	0.66	0.63	2
		III	0.63	0.63	0.62	0.58	0.59	1
		III$^+$	0.66	**0.68**	**0.68**	0.67	0.67	1

as a preprocessing step could enhance the diachronic classification of texts. In comparison with the results of the first experiment (Exp. I), the use of the CfsSubsetEval attribute selection algorithm (Exp. III) significantly increased the classification performance in two cases (genres C and N), while it significantly decreased the classification accuracy in genre A.

The use of the classification algorithms based on the logistic function (columns 'Log.' and 'SLog.') led to the highest F-measure in 9 genres (B, C, E–L, and P), while the classification algorithms based on Support Vector Machines (columns 'SMO(n)' and 'SMO(s)') led to the highest results only in 5 genres (A, C, J, K, and N). This might be interpreted as support for the idea that the diachronic change is best presented by the logistic function [3,19].

The results presented in Table 3 also indicate that the stylistic changes (in terms of these 23 initial features) were most pronounced in the Press category (genres A–C). Genres belonging to the Prose category underwent less extensive stylistic changes than those in the Press genre, as the F-measures are significantly lower in Prose than in the Press category. It can also be noted that within the Prose category, genre D (Religion) stands out in with the highest classification accuracy which leads to the conclusion that the stylistic changes were more pronounced in that genre than in the other four, thus making this genre an outlier in its category.

A more detailed analysis of features marked as significant by statistical tests and those returned by the CfsSubsetEval attribute selection algorithm as part of the best subset of features (Table 4) revealed that the features selected by the CfsSubsetEval algorithm are a subset of features marked as significant by statistical tests, in all cases where the CfsSubsetEval algorithm was successful ('the merit of the best subset found' above zero). In the cases when the CfsSubsetEval algorithm is unable to find the best subset of features, the algorithm selects the first feature in the given list of features, with 'the merit of best subset found' equal to zero.

Table 4. Selected features in experiments II, III and III$^+$

Genre	Exp. II	Exp. III	Exp. III$^+$
A	ASL, LR, PASS, COMPL, V_per, V_av, N_av, det_per	LR, COMPL, v_av	/
B	LR, det_per, prep_per, sc_av	LR	/
C	LR, COMPL	LR	/
D	n_av, prep_av, adj_per, adj_av, CLI	CLI	/
E	LR, CLI	CLI*	CLI*
F	pron_per, pron_av, CLI	CLI	/
G	LR v_per, n_per, sc_per, sc_av, CLI	CLI, n_per	/
H	ASL, det_av, prep_av	CLI*	ASL*
J	PASS, det_per, det_av, prep_per, prep_av, CLI	CLI, prep_per, det_per	/
K	COMPL, adv_av, cc_per	CLI*	adv_av
L	/	CLI*	/
N	n_per, CLI	CLI*	CLI*
P	LR, adv_per	CLI*	LR*

The results of experiment III$^+$ were found to be significantly better than those of experiment III when CfsSubsetEval:

- fails to find the best subset of the initial features (selected feature in column 'Exp. III' in Table 4 is marked by an '*'),
- succeeds in finding the best subset of those features reported as significant by statistical tests (selected feature in column 'Exp. III$^+$' in Table 4 is not marked by an '*').

Although in our data set we found only one such case (genre K), we could still say that the safest way to use the CfsSubsetEval attribute selection algorithm would be to apply it only to a subset of initial features (only those features which were marked as significant by the statistical tests).

5 Conclusions

The results presented in this study indicated that the stylistic diachronic changes of written British English in the period 1961–1991 were significantly more pronounced in the Press category than in three other text categories (Table 3). They also demonstrated that the genres within the same broad text category are very heterogeneous. In each of them, different groups of features underwent a significant diachronic change (Table 4) and the extent of those changes differed significantly across them (Table 3). The results also indicated that lexical richness (LR) and the Coleman-Liau readability index were the features which significantly changed in most of the investigated genres (Table 4).

On the basis of the comparison of the results of different experiments, we can conclude that the use of the statistical tests as a preprocessing (feature selection) step, significantly increases the classification accuracy in several cases, while in others it does not have any significant influence. Therefore, we suggest the use of the statistical tests as a preprocessing step in other diachronic text classification tasks. When compared with the CfsSubsetEval attribute selection algorithm, the statistical test achieved significantly better or equal performance (with the only exception in genre D, for the Naïve Bayes classification algorithm). In most cases, this was due to the fact that the CfsSubsetEval algorithm selects the first feature in the given list of features in the cases when it is not able to find a subset with 'the merit of best subset found' greater than zero. The use of the CfsSubsetEval attribute selection algorithm on the subset of features previously selected by the statistical tests, significantly improves the classification accuracy (genre K) or it leaves it unchanged. The statistical tests when used in the preprocessing step on their own, either significantly improve the classification accuracy (genres B and N) or they do not lead to any significant differences. Therefore, we suggest either the use of the statistical tests on their own or the combination of the CfsSubsetEval attribute selection algorithm with them in the preprocessing step of diachronic text classification.

Most importantly, the presented study demonstrated various possibilities that the machine learning approach can offer to the investigation of language change. By partially automating the process, it can speed up and facilitate the initial phases of language change studies, by providing a broad overview of possible changes and selecting the most important features from a potentially large initial set, which would be the subject of closer investigation. A machine learning approach could also offer an easier comparison of diachronic changes across different genres, languages and language varieties.

References

1. Adolph, R.: The Rise of Modern Prose Style. M.I.T. Press, Cambridge (1966)
2. Aldrich, J., Nelson, F.: Linear probability, logit, and probit models. Quantitative applications in the social sciences. Sage, London (1984)
3. Altmann, G., von Buttlar, H., Rott, W., Strau, U.: A law of change in language. In: Brainerd, B. (ed.) Historical Linguistics, pp. 104–115. Brockmeye, Bochum (1983)

4. Bennett, J.R.: Prose Style: A Historical Approach through Studies. Chandler, San Francisco (1971)
5. Biber, D.: Investigating Macroscopic Textual Variation through Multifeature/Multidimensional Analyses. Linguistics 23, 337–360 (1985)
6. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge (1988)
7. Biber, D., Finegan, E.: An Initial Typology of English Text Types. In: Aarts, J., Meijs, W. (eds.) Corpus Linguistics H: New Studies in the Analysis and Exploitation of Computer Corpora, pp. 19–46. Rodopi, Amsterdam (1986)
8. Biber, D., Finegan, E.: Drift and the evolution of English style: A history of three genres. Language 65, 487–517 (1989)
9. le Cessie, S., van Houwelingen, J.: Ridge Estimators in Logistic Regression. Applied Statistics 41(1), 191–201 (1992)
10. Connexor: Machinese language analysers (2006)
11. Corpas Pastor, G., Mitkov, R., Afzal, N., Pekar, V.: Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification. In: Proceedings of the AMTA, Waikiki, Hawaii (2008)
12. Geisler, C.: Relativization in Ulster English. In: Poussa, P. (ed.) Relativisation on the North Sea Littoral (LINCOM Studies in Language Typology 07), pp. 135–146. Lincom Europa, München (2002)
13. Geisler, C.: Statistical reanalysis of corpus data. ICAME Journal 32, 35–46 (2008)
14. Gordon, I.A.: The Movement of English Prose. Indiana University Press, Bloomington (1966)
15. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. In: McDonald, C. (ed.) Computer Science 1998 Proceedings of the 21st Australasian Computer Science Conference, ACSC 1998, pp. 181–191. Springer, Berlin (1998)
16. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)
17. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3), 637–649 (2001)
18. Kroch, A.: Function and grammar in the history of English: Periphrastic "do". In: Fasold, R. (ed.) Language Change and Variation, pp. 133–172. Benjamins, Amsterdam (1989)
19. Kroch, A.: Reflexes of grammar in patterns of language change. In: Language Variation and Change, vol. 1, pp. 199–244 (1989)
20. Landwehr, N., Hall, M., Frank, E.: Logistic Model Trees. Machine Learning 59, 161–205 (2005)
21. Leech, G., Smith, N.: Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB. ICAME Journal 29, 83–98 (2005)
22. Leech, G., Smith, N.: Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. In: Renouf, A., Kehoe, A. (eds.) The Changing Face of Corpus Linguistics, pp. 186–204. Rodopi, Amsterdam (2006)
23. Mair, C., Hundt, M., Leech, G., Smith, N.: 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 (2002)
24. Mair, C., Leech, G.: Current change in English syntax. In: Aarts, B., MacMahon, A. (eds.) The Handbook of English Linguistics, ch. 14. Blackwell, Oxford (2006)

25. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods – Support Vector Learning. The MIT Press, London (1998)
26. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
27. Senter, R.J., Smith, E.A.: Automated readability index. Tech. rep., University of Cincinnati. Ohio, Cincinnati (1967)
28. Sumner, M., Frank, E., Hall, M.: Speeding up Logistic Model Tree Induction. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 675–683. Springer, Heidelberg (2005)
29. Tukey, J.: Exploratory data analysis. Addison-Wesley, Reading (1977)
30. Štajner, S., Mitkov, R.: Diachronic Stylistic Changes in British and American Varieties of 20th Century Written English Language. In: Proceedings of the RANLP 2011 Workshop "Language Technologies for Digital Humanities and Cultural Heritage", pp. 78–85 (2011)
31. Štajner, S., Mitkov, R.: Diachronic Changes in Text Complexity in 20th Century English Language: An NLP Approach. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey (May 2012)
32. Westin, I.: Language Change in English Newspaper Editorials. Rodopi, Amsterdam (2002)
33. Westin, I., Geisler, C.: A multi-dimensional study of diachronic variation in British newspaper editorials. ICAME Journal 26, 133–152 (2002)
34. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers (2005)