Low-level parameters reflecting the naturalness of texts

Oliver Mason[1], Ylva Berglund[2]

[1] Department of English – School of Humanities – University of Birmingham – Edgbaston – Birmingham – B15 2TT – United Kingdom – O.Mason@bham.ac.uk

[2]Oxford Text Archive – University of Oxford – 13 Banbury Road – Oxford – OX2 6NN – United Kingdom – ylva.berglund@oucs.ox.ac.uk

Abstract

The present study reports on work on automatic stylistic assessment of students' essays. Data from the Uppsala Student English Corpus, produced by advanced Swedish learners of English, is compared to native English data produced by both adult and school-age speakers of English. The method used is to classify texts according to a number of low-level 'surface' parameters to find clusters. It is shown that these parameters, which are not generally considered in traditional, manual assessment of text, allow near perfect distinction between different sets of data.

Keywords: automatic stylistic assessment, PCA, cluster analysis, second language acquisition, learner corpus, child language

1. Introduction

The work described in the present paper is part of a larger project on the automatic stylistic assessment of students' essays. The overall aim of the project is to identify low-level features that affect the naturalness of English text produced by non-native speakers, that is, features that make the L2 texts seem un-natural, 'foreign' or non-English. The project involves the development of suitable tools and methods as well as evaluation of the results obtained.

1.1. Low-level Parameters

We have previously shown that different low-level 'surface' parameters can be used to identify groupings of text with similar extra-linguistic features (Mason & Berglund, 2000), and that this method can be used to successfully distinguish text produced by advanced learner from that written by native speakers of English (Berglund & Mason, 2001). In the present paper, we will draw on the knowledge gained in our previous studies about variation between different kinds of text with regard to the distribution of low-level parameters. We will demonstrate how text produced by children with English as a native language compares to the data produced by adult native speakers and the advanced learners of English. We will also discuss how our results can be used to further refine and develop our method.

The basic idea behind our approach is to evaluate how far these low-level parameters can be used to assess style or naturalness of texts, even though they are not measuring something which is in itself meaningful to a human analyst. However, the perceived style of a text has to be mirrored in such features, as the choice of higher-level constructs is reflected in the length of words or the frequency distribution of letters. In previous (unpublished) work it was found that measuring character entropy on a set of individual sample texts was sufficient to group the texts according to authorship, and that stylistically closer texts (by intuitive human

judgment) were more similar in their respective entropy values than others by the same author. The background of our study is thus a further development of early experiments in quantitative linguistics, which we are trying to revive in the context of corpus linguistics.

By comparing reasonably large amounts of texts with known external features (eg text type and genre classification) we hope to be able to include features in our statistical analysis which were not in that form available to early studies.

1.2. Related Work

There are also parallels between our work and studies of language variation by Biber (for example Biber (1986), Biber (1987), and Biber (1988)). Biber, however, uses linguistic features, thus presupposing a framework of analysis which makes more assumptions on the data, and the methods he uses require quite substantial human intervention, thus allowing for a possible bias to be introduced by the analyst. We tried to keep interference to a minimum, relying mainly on fully automatic methods. These methods provide an objective assessment of the quality of the outcome, so that the results can be improved by adjusting a few parameters only. Doing this does in no way introduce any preconceived ideas about the 'desired' result, but could in principle be done fully automatically, using computational methods such as genetic algorithms.

The main 'intellectual' challenge of this approach lies in the interpretation of the results. Once we have identified the low-level features of a text, or a set of texts, which differentiates it or them from other texts, how can we then make the link between the individual feature and linguistically meaningful properties? Obviously, this is easier for some parameters than others, but it is a problem which we are still working on.

2. Method

The method we have used, and which we are continually refining, has been presented in Berglund & Mason (2000), and Mason & Berglund (2001). In short, the process involves the following steps (steps 1-6 are iteratively repeated):

1. selection of a number of low-level parameters that are to be examined
2. correlation analysis (to exclude features that correlate with each other)
3. Principal Components Analysis (to further reduce the dimensionality of the data set for visualisation)
4. cluster analysis (to identify groupings of text with similar values)
5. evaluation of results and method
6. refinement of method
7. formation of hypotheses about the linguistic reasons why certain texts are differently classified

2.1. Choice of Parameters

An initial set of nineteen parameters was chosen from a wide variety of well-known text measurements, such as type-token ratio, perplexity, vocabulary grade, average word length, and a number of readability scores. It was considered important at this point to restrict the

selection to parameters which can be computed automatically, not only because it is necessary to calculate the parameter values for a large number of texts but also to limit the influence of human intervention. An attempt was made to specifically include established measurements which in the past have been used for different purposes, in order to investigate the versatility of these parameters when applied in different contexts. Once they had been selected, a number of custom Java classes were implemented to perform the parameter extraction in batch processing mode.

Each of the texts in our sample data sets is thus characterised by a feature vector containing the values of the measured parameters. During the correlation analysis those parameters which strongly correlate with others are identified and then removed to reduce the initial dimensionality, as they would not contribute a lot of extra information to start with. This stage of the analysis is mostly trial-and-error, as we have not been able to find any regular patterns in the different parameter combinations. Again, human choice at this stage could in theory have been replaced by suitable automatic methods, which might even have found a suitable parameter combination in less time.

2.2. Clustering and Visualisation

Once an appropriate set of parameters has been selected, an agglomerative cluster analysis (PAM, see Kaufman & Rousseeuw (1990)) is used to find groups in the data based on parameter similarities, which can then be interpreted according to their composition with regards to the data sets under investigation. Unlike other agglomerative cluster algorithms, PAM (Partitioning Around Medoids) provides a quality assessment of the structure(s) it finds, which was used as a criterion for adjusting the procedure. An important feature of most agglomerative algorithms is that the number of clusters to be found has to be specified in advance; by iterating through a number of possible settings and comparing the quality assessments provided by PAM it was made sure that the optimal number of clusters was found.

At the same time, a principle components analysis (PCA) is used to visualise the configuration of the individual texts in parameter space. By extracting the two most important components we reduce the dimensionality from seven (the number of parameters retained after the correlation analysis) down to two, which allows us to plot the data set easily. This serves as an important aid when inspecting the clustering, as we would hope to identify the clusters visually on the component plot. It would also allow initial judgment on the coherence of the clusters found during the analysis.

2.3. Final Parameters

As mentioned above, the initial set of parameters considered for the analysis comprised 19 different measurements, from which 7 were ultimately chosen for the clustering. In this section we will give a brief overview of the kinds of parameters we were investigating.

The parameters can be divided into several more or less homogeneous groups. The first of these groups contains a number of readability scores. These scores have been devised by scholars over the past few decades in order to grade texts according to their perceived difficulty, usually based on the basic assumptions that long words and sentences are more difficult to comprehend than shorter ones. They all work according to the same principles: a set of text characteristics (such as word length in letters or syllables, or sentence length) are collected of the text in question and are combined with a set of weightings to reach a certain

target range with the result. The score is then interpreted on a scale which, for example, reflects the number of years of school the reader must have had before understanding the text. Other scales depend on the area a particular score is applied in. These scores are mainly used in education, for example to assess teaching material, but some have been developed for other purposes, such as training manual readability.

The different scores considered for our investigation are mostly implemented by the Unix utility *style*, which is available open source from the GNU project (see Stutz (2000)). They are:

- Kincaid (range between 5.5 (easy) and 16.3 (difficult)), developed for Navy training manuals
- Automated Readability Index (ARI)
- Coleman-Liau Formula
- Flesch Reading Easy Formula, developed in 1948, and still widely used, especially in the USA
- Fog Index, developed by Robert Gunning, supposed to yield a school grade. The 'ideal' score is 7 or 8, more than 12 would be too difficult for most people
- WSTF, Wiener Sachtextformel, developed for German texts, also giving a school grading for a text
- Wheeler-Smith, corresponds to school grades via a mapping table
- Lix, developed by Björnsson, also mapping onto school grades
- SMOG-Grading, developed by McLaughlin in 1969

Since they mostly measure the same textual features (word or sentence length) only three of them were used: SMOG, Coleman-Liau, and Wheeler-Smith. It is, in fact, quite surprising to note that these three scores are all necessary for the final clustering; leaving one of them out on the basis that they are too similar does not lead to a clear-cut result. This is one of the questions we are still trying to find an answer for.

Apart from readability scores, a selection of more 'traditional' measurements previously used in linguistics has been considered. These scores are:

- Average word length, measured in characters per word
- Average sentence length, measured in words per sentence
- Proportion of personal pronouns
- Yule's characteristic *K*, an early quantitative attempt of describing the diversity of vocabulary

- Type-token-ratio, the ratio between different word forms used and the length of the text (and thus sensitive to the text length)
- Vocabulary grade (between 1 and 3), calculated using three graded lists used in English language teaching, measured as a proportion of words of grade *n* in the text, with 1 being easy and 3 being hard (this is treated as three distinct parameters)
- Relative perplexity, based on unigram entropy (as described in Sekine (1997))
- Collocational richness, which measures how many typical collocations (derived from the Cobuild Bank of English) are used in the environment of a word

Of these additional parameters, four have been used: the proportion of pronouns, the average word length, and the proportions of grade 1 and grade 3 vocabulary.

3. Previous Work

3.1. Corpus data

We have previously used our method on three sets of data; a sub-set of the Freiburg-Brown Corpus (Frown), the Freiburg-LOB Corpus (FLOB) and data from the Uppsala Student English project (USE). The Frown corpus contains present-day American English data from 15 genres. The texts are from 1992, and the corpus set-up is mirrored on that of the earlier Brown Corpus, created by Francis and Kučera using data from 1961. The FLOB corpus is similar in composition to the Frown corpus, but comprising British English texts from 1991, as an updated analogue of the LOB corpus, the British English version of Brown, compiled by Johansson and Leech, again using data from 1961. Frown and FLOB have been compiled by Christian Mair at the University of Freiburg. The USE Corpus consists of five types of essays produced by advanced Swedish learners of English (see Axelsson (2000) for details). These corpora are homogeneous in the sense that the authors are all either (US or UK) native speakers (Frown and FLOB respectively) or non-native speakers of English resident in Sweden (USE). With respect to their internal composition, the corpora are varied and all comprise texts from different genres.

In order to assess the influence of genre on the procedure, we selected a number of genres from the American data which were supposedly similar to the style of essay as written by the Swedish students. We selected American data because it was judged to have a higher impact on learners in Sweden, due to cultural influences and students staying abroad there. We then repeated the analysis with British English data for comparison. In order to not bias the study through our pre-selection of texts we included the whole FLOB corpus, comprising a number of quite varied genres. As described below, similar results were obtained.

3.2. Results

The results from our previous studies are presented in Mason & Berglund (2000) and Berglund & Mason (2001). As reported there, we have found that our method can be used to distinguish with a high degree of accuracy texts produced by Swedish students learning English from those authored by native speakers (both British and American English). We have also got indications that the method can be used to identify genre-like groupings in the data, but only in terms of tendencies, not as obvious clusters.

The cluster algorithm allows assessment of the quality of the clustering. It reported that a 'strong' structure had been found when trying to group the texts into two clusters, the number of distinct data sets. For confirmation the procedure was repeated with a few other prescribed cluster numbers, but no other configuration achieved a better result.

The clusters that have been identified map almost exactly onto the data sets. Both with American English (AE) and British English (BE) the distinction is very accurate: less than 1% (3 out of 440 for AE and 5 out of 440 for BE) of learner texts are assigned the wrong cluster. From the AE corpus, 3 out of 215 were classified together with the Swedish data, while not a single one of the 500 BE texts was misclassified.

The plot of the PCA of the USE/Frown (AE) data is shown in figure 1. Here it can easily be seen that the data set neatly splits into two clusters.

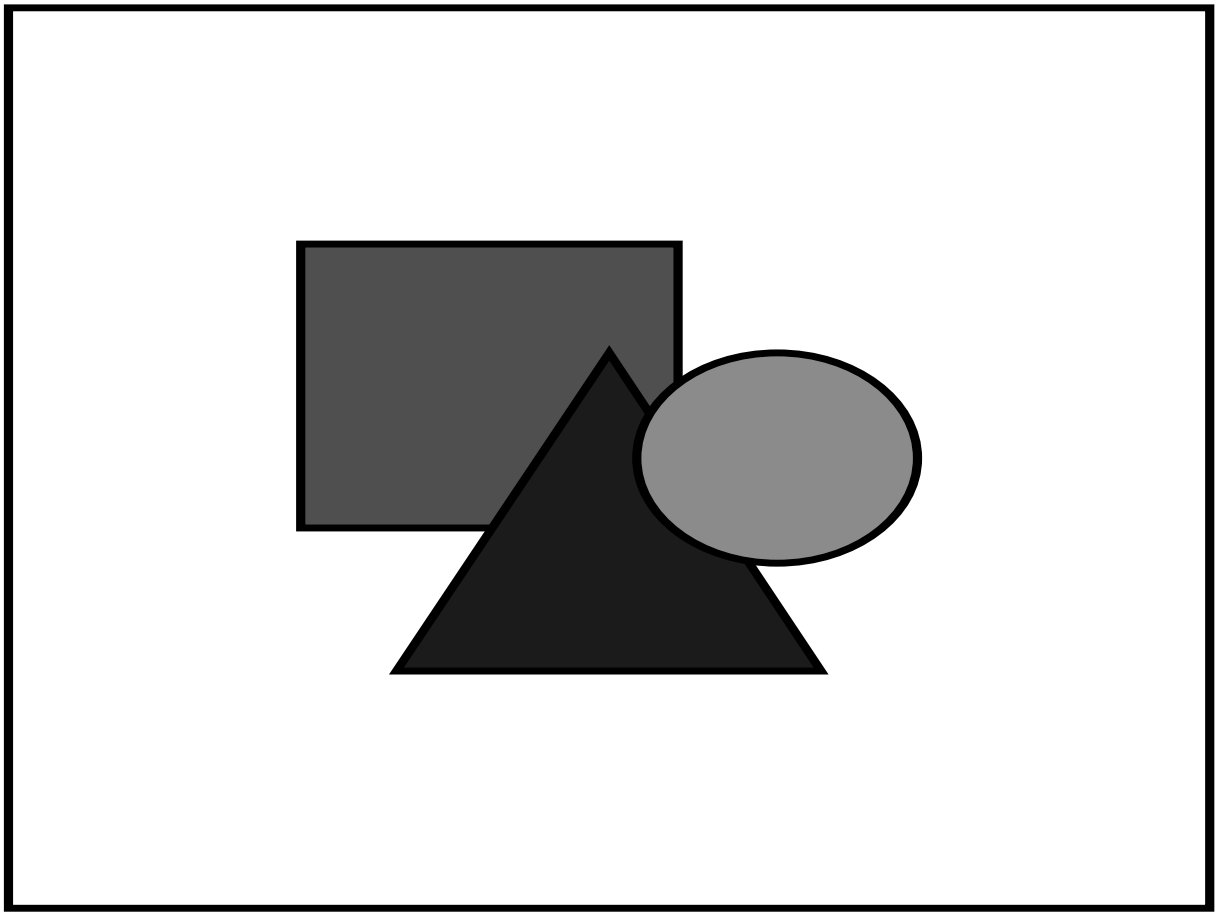

Figure 1 : USE and Frown

A manual analysis of the outliers was then employed to find out the reasons for the misclassifications, and in most cases it turned out that something was peculiar about the text in question. One learner text that had been clustered with the native speaker data was using copious amounts of quotes from a literary work under discussion. This provides an encouraging confirmation that the method is reliable and can distinguish between the learner and native corpus data.

3.3. Discussion

In our studies we have shown that a basic cluster analysis with small number of easily computable parameters in principle can be used to distinguish between native and non-native writing. We were pleasantly surprised to find such a high level of consistency where the grouping of the two kinds of texts is concerned. It could be mentioned that the non-native writing was produced by advanced learners with a high level of proficiency in an academic context. The students in question had access to, and were encouraged to use, dictionaries and spell-checkers, which would have reduced the amount of misspelled words and possibly increased the richness of vocabulary.

It needs to be stressed that the procedure we use only involves minimal human intervention (mainly choosing the parameters at the initial stage), and is thus free from any theoretical preconceptions which go beyond the definition of 'word' or 'sentence'. As all parameters can be extracted fully automatically, the procedure can be used on large amounts of data. However, as there was little linguistic input at the processing stage, further effort is required when interpreting the result. This is where more work is required in the following stages of this project.

Also, a further challenging avenue of research is to extend the scope from simply native *vs.* non-native data to other sources. That is done in the current stage of the project.

4. Present study

In the present study we draw on the results obtained in our previous work, expanding it in two directions. Firstly, the original method is applied to a new set of data, essays produced by British school children (the UCLES-Reading Corpus, held at Reading University, from now on referred to as URC). For a description of this corpus see Chipere *et al.* (2001). The data constitutes an interesting scope for comparison in more than one way. The child data is similar to the native speaker data previously investigated in that it has been produced by native speakers of English. It differs from the FLOB/Frown data, however, in that it is produced by school children, whose writing skills are expected to be different from those of mature users of the language, both in grammar and in lexis. The texts in the USE corpus, on the other hand, have been written by adult, non-native language users. By comparing the USE data to written language produced by young and adult native speakers, it may be possible to identify features that are shared between the different user groups, which in turn can add insights into the factors which make L2 language seem un-natural or less advanced than native, adult English.

Initially we conducted two two-way comparisons, in addition to the existing comparison USE *vs*. FLOB as reported above: the missing pairs URC-USE and URC-FLOB. This was then followed by a three-way comparison of all corpora involved.

One problem with the URC data was that it varied greatly in length. The texts are taken from different stages and levels of difficulty, and an arbitrary sample was taken, as there was a large number of texts available. The texts were sorted according to size (measured in characters), and the top 584 were chosen; these were at least comparable in length with the other corpus data used in the study. The total number of texts in each set was thus 500 FLOB, 440 USE, and 584 URC.

4.1. URC and USE

When the URC and USE corpora (child data and learner data) were analysed, PAM reports a 'reasonable' structure for two clusters, with the following compositions: 5 USE texts and 541 URC texts in cluster 1, and 435 USE and 43 URC in cluster 2. This is again a very good result, with 1% of cluster 1 being USE, and about 9% of cluster 2 being URC. This is not as clear-cut a separation as between USE and FLOB, but the result must still be regarded as satisfactory, with 99% of cluster 1 being URC texts, and about 91% of cluster 2 constituting USE texts.

4.2. URC and FLOB

Interestingly, the clustering algorithm applied to the native speaker samples, URC and FLOB, yields the best result. The quality is judged as 'strong', and cluster 1 consists exclusively of 582 texts from URC. The second cluster contains all 500 FLOB texts, and also the remaining 2 URC texts. The two URC texts are incidentally the two longest ones, so they have arguably more scope for elaborate exposition. Texts in the FLOB corpus are generally longer, but this should not have any influence on the procedure, as none of the parameters used is dependent on the text length.

4.3. URC, USE, and FLOB

When we analyse the data from all three corpora in our study, the best result is found for two clusters (rated 'strong'): cluster 1 then contains 418 USE texts and 582 URC texts, but not a single FLOB text, whereas cluster 2 comprises all 500 FLOB texts, 22 USE and 2 URC texts. A three-cluster-solution is only judged to be 'reasonable', and basically splits cluster 1 into two separate URC and USE clusters. The three clusters have the following compositions (USE/URC/FLOB): 430/42/0, 5/1/500, and 5/541/0.

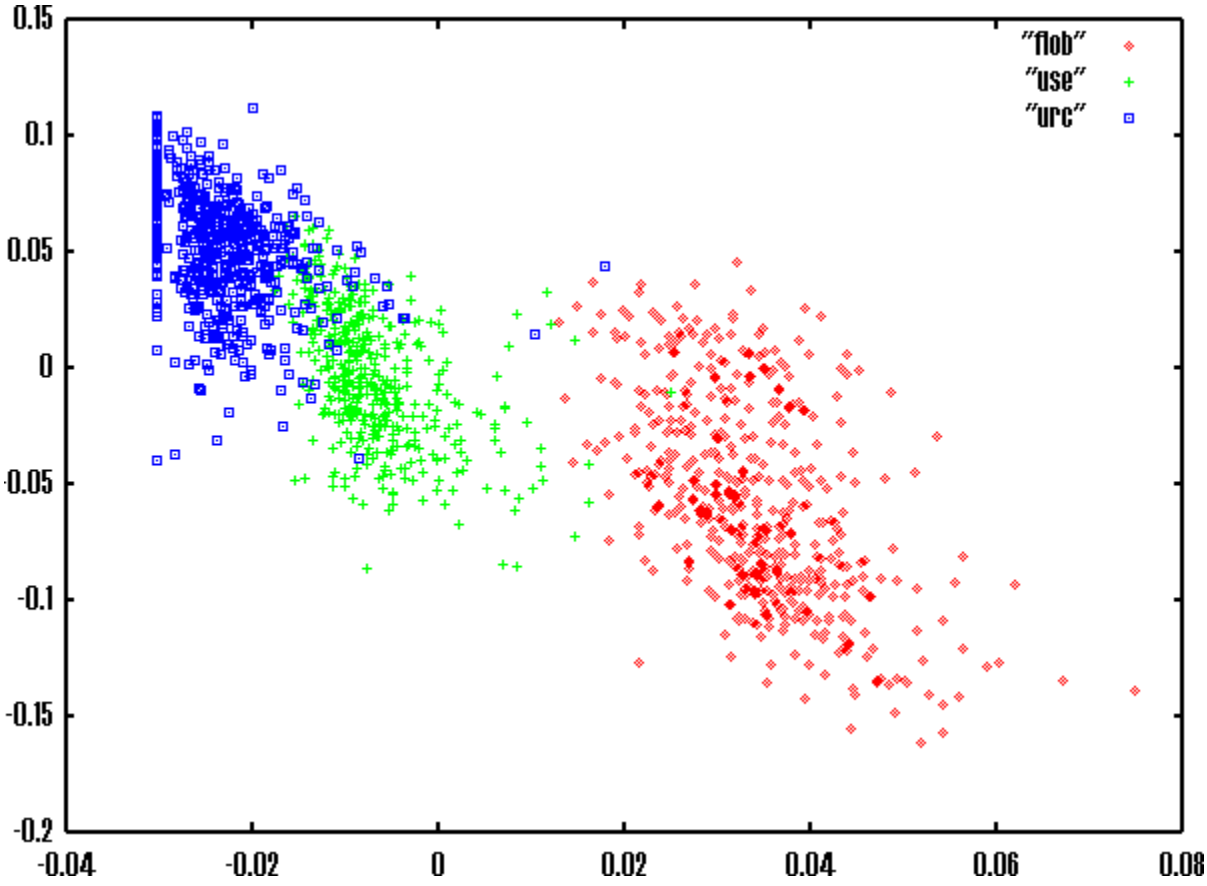

Figure 2 : URC, USE, and FLOB

From this we have to conclude that the school children's writing has more in common with the non-native learners than with native speaker adults. This becomes apparent when looking at the visualisation (the first two principal components from the PCA applied to the data set): while the cluster analysis can only clearly separate two clusters, one can clearly see a tendency for the school children's texts to be located in a different part of the reduced parameter space (see Figure 2).

While three samples are obviously not sufficient to draw any far-reaching conclusions, one can hypothesise that there is some relationship between the level of language proficiency exhibited in/by a certain text and its position in the seven-dimensional parameter space as defined by the low-level parameters described above. The next logical steps in investigating this further would be to add a number of other texts of differing standard in order to find out where they would be positioned, such as native speaker students' essays. If this correlation turns out to be valid, the results could, for example, be applied to computer-aided language assessment. In a co-operative environment a student could have an essay processed and would get feedback on its position in parameter space in relation to selected benchmarks. While this would not allow formative feedback it might still be useful for students to get a general idea of how their essay compares to other texts of known standard.

5. Conclusion

We have extended previous work on assessing the linguistic 'naturalness' of non-native speaker essays on University level by comparing British school children's essays with both

Swedish University students and British adult writing. The result of the analysis shows that seven low-level parameters can successfully be employed to separate off the British adults from the other two samples; there can also be found a separation between the Swedish students and the British school children, but it is not as clear-cut.

From their position in parameter space the University students seem to be closer to adult language than the school children, even though they are themselves closer to the school children than they are to the adults.

Further text samples need to be analysed now in order to corroborate the results, as three samples are clearly not enough to draw any conclusions yet. However, the procedure described shows consistency in that texts from the same source generally cluster closely together. This is an encouraging result, showing that the procedure has indeed descriptive power.

References

Axelsson, M.W. (2000) USE-The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal* 24: 155-157.

Berglund, Y. and Mason, O. (2001) *But this formula doesn't mean anything...!?": Some reflections on parameters of texts and their significance*. IN: edited collection to be published (by Peter Lang) in honour of Geoffrey Leech. (available on-line at `http://www.clg.bham.ac.uk/staff/oliver/publications/cl2001/berglund-mason.html`)

Biber, D. (1986) *Spoken and written textual dimensions in English: Resolving the contradictory findings*. Language 62, p. 384-414

Biber, D. (1987) *A textual comparison of British and American writing.* American Speech 62. p.99-119

Biber, D. (1988) *Variation across speech and writing.* Cambridge University Press

Chipere, N., Malvern, D., Richards, B and Duran, P. (2001) *Using a corpus of school children's writing to investigate the development of vocabulary diversity*. IN P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.) Proceedings of the Corpus Linguistics 2001 Conference. University of Lancaster.

Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data.* New York: John Wiley & Sons

Mason O. and Berglund Y. (2000) *Measuring the influence of external factors on learner performance*. Paper to appear in Proceedings of TaLC 2000, Graz.

Sekine, S. (1997) *A new direction for sublanguage NLP*. IN: Jones, D. and Somers, H. (eds) "New Methods in Language Processing", UCL Press, p.165-177

Stutz M. (2000) *Living Linux Column*, `http://www.oreillynet.com/pub/a/linux/2000/05/05/LivingLinux.html`, last visited on 30/10/2001