Linguistic profiling of texts for the purpose of language verification

Hans VAN HALTEREN
Dept. of Language and Speech, Univ.of Nijmegen, The Netherlands
P.O. Box 9103, 6500 HD Nijmegen
hvh@let.kun.nl

Nelleke OOSTDIJK
Dept. of Language and Speech, Univ. of Nijmegen, The Netherlands
P.O. Box 9103, 6500 HD Nijmegen
n.oostdijk@let.kun.nl

Abstract

In order to control the quality of internet-based language corpora, we developed a method to verify automatically that texts are of (near-) native quality. For the LOCNESS and ICLE corpora, the method is rather successful in separating native and non-native learner texts. The Equal Error Rate is about 10%. However, for other domains, such as internet texts, separate classifiers have to be trained on the basis of suitable seed corpora.

1 Introduction

Research in linguistics and language engineering thrives on the availability of data. Traditionally, corpora would be compiled with a specific purpose in mind. Such corpora characteristically were well-balanced collections of data. In the form of metadata, record was kept of the design criteria, sampling procedures, etc. Thus the researcher would have a fair idea of where his data originated from. Over past decades, data collection has been boosted by technological developments. More and more and increasingly large collections of data have been and are being compiled. It is tempting to think that the problem of data sparseness has been solved – at least for raw data or data without any annotation other than can be provided fully automatically – especially now that large amounts of data can be accessed through the internet. However, with data coming to us from all over the world, originating from all sorts of sources, we now possibly have a new problem on our hands: often the origins of the found data remain obscure.

It is not always clear what exactly the implications for our research are of employing data whose origin we do not know. Is it legal to use these data, ethical, appropriate, …? In this paper we will focus on the last point: the appropriateness of the data in the light of a specific application or research goal. More in particular, we will investigate to what extent we can devise a procedure that will enable us to identify texts produced by native speakers of the language (and thus by default those produced by non-native speakers). The present study is motivated by the fact that for many uses the (near-)nativeness of the data is a critical factor in the development of adequate resources and applications. Thus, for example, a style checker or some other writing assistant tool which has been based on erroneous materials or at least materials deviant from the language targeted, will not always respond appropriately.

1.1 Assessing (near-)nativeness

In the general absence of metadata which attest that texts have been produced by native speakers, there is one obvious approach that one may consider in order to assess the (near-)nativeness of texts of unknown origin and that is to exploit their specific linguistic characteristics.

Previous studies investigating language variation (eg Biber, 1995, 1998; Biber et al., 1998; Conrad and Biber, 2001; Granger, 1998, Granger et al., 2002) have shown that language use in different genres and by different (groups of) speakers displays characteristic use of specific linguistic features (lexical, morphological, syntactic, semantic, discoursal). These studies are all based on data of known origin. In the present study, we take a somewhat different approach as we aim to profile texts of unknown origin and identify native vs non-native language use, a task for which we coined the term *language verification*.

1.2 Non-native language use

Texts produced by non-native speakers will generally pass superficial inspection, i.e. they are deemed to be texts in the target language and will be treated as such. However, on closer inspection there is a wide range of features in the language use of non-natives which may have a disruptive effect on for instance derived language models. It is important to realize that non-native use is the complex result of different processes and conditions. First of all, there is the level of achievement. A non-native user gradually developes language skills in the target language. As he/she masters certain lexical items or morpho-syntactic structures and feels confident in using them, certain items and structures are bound to be

overused. At the same time, other items and structures remain underused as the user avoids them since he is not familiar with them or does not (yet) feel confident enough to employ them. Moreover, even for speakers who have attained a relatively high degree of proficiency, the influence of the native language remains. This may lead to transfer effects and interference (the effects of which are found, for example, in the use of false friends and word order deviations).

In the present paper, we report the results obtained in some experiments that were carried out and which aimed to assess whether texts are of (British English) native or non-native origin using the method of linguistic profiling. The structure of the paper is as follows: In section 2, we describe the method of linguistic profiling. Next, in section 3, its application in establishing the nativeness of texts is described, while in section 4 it is investigated whether the approach holds up when we shift from one domain to another. Finally, section 5 presents the conclusions.

2 Linguistic profiling

In linguistic profiling, the occurrences in a text are counted of a large number of linguistic features, either individual items or combinations of items. These counts are then normalized for text length and it is determined to what extent (calculated on the basis of the number of standard deviations) they differ from the mean observed in a profile reference corpus. For each text, the deviation scores are combined into a profile vector, on which a variety of distance measures can be used to position the text relative to any group of other texts.

2.1 Language verification

Linguistic profiling makes it possible to identify (groups of) texts which are similar, at least similar in terms of the profiled features (cf. van Halteren, 2004). We have found that the recognition process can be vastly improved by not only providing positive examples (in the present case native texts) but also negative examples (here the non-native texts). So we expect that, given a seed corpus containing both native and non-native texts, linguistic profiling should be able to distinguish between these two types of texts.

2.2 Features

As previous research has shown (see e.g Biber 1995), there are a great many linguistic features that contribute to marked structural differences between texts. These features mark 'basic grammatical, discourse, and communicative functions' (Biber, 1995: 104). They comprise features referring to vocabulary, lexical patterning, syntax, semantics, pragmatics, information content or item distribution through a text. Here we restrict ourselves to lexical features.

Sufficiently frequent tokens, i.e. those that were observed to occur with a certain frequency in some language reference corpus, are used as features by themselves. In the present case these are ítems that occur at least five times in the written texts from the BNC Sampler (BNC, 2002). For less frequent tokens, we determine a token pattern consisting of the sequence of character types. For example, the token *Uefa-cup* is represented by the pattern "#L#6+/CL-L", where the first "L" indicates low frequency, 6+ the size bracket, and the sequence "CL-L" a capital letter followed by one or more lower case letters followed by a hyphen and again one or more lower case letters. For lower case words, the final three letters of the word are also included in the pattern. For example, the token *altercation* is represented by the pattern "#L#6+/L/ion". These patterns were originally designed for English and Dutch and will probably have to be extended for use with other languages. Furthermore, for this specific task, we wanted to avoid recognizing text topics rather than nativeness, and decided to mask content words. Any high frequency word classified primarily as noun, verb or adjective (see below), which had a high document bias (cf. van Halteren, 2003) was replaced by the marker #HC# followed by the same type of pattern we use for low frequency words, but always without the final three letters. This occludes topical words like *brain* or *injury*, while leaving more functional words like *case* or *times* intact.

In addition to the form of the token, we also use the syntactic potential of the token as a feature. We apply the first few modules of a morphosyntactic tagger (in this case the tagger described by van Halteren, 2000) to the text, which determine which word class tags could apply to each token. For known words, the tags are taken from a lexicon; for unknown words, they are estimated on the basis of the word patterns described above. The most likely tags (with a maximum of three) are combined into a single feature. Thus *still* is associated with the feature "RR-JJ-NN1" and *forms* with the feature "NN2-VVZ". Note that the most likely tags are determined exclusively on the basis of the current token; the context in which the token occurs is not taken into account. The modules of the tagger which are normally used to obtain a context dependent disambiguation are not applied.

On top of the individual token and tag features we use all possible bi- and trigrams. For example, the token combination *an attractive option* is associated with the complex feature "wcw= #HF#an#HC#JJ#HC#6+/L". Since the number of features quickly grows too big to allow for efficient processing, we filter the set of features. This done by requiring that a feature occur in a set minimum number of texts in the profile reference corpus (in the present case a feature must occur in at least two texts). A feature which is filtered out contributes to a rest category feature. Thus, the complex feature above would contribute to "wcw=<OTHER>".

The lexical features currently also include features that relate to utterance length. For each utterance two such features are determined, viz. the exact length (e.g. "len=15") and the length bracket (e.g. "len=10-19").

2.3 Classification

When offered a list of positive and negative texts for training, and a list of test texts, the system first constructs a featurewise average of the profile vectors of all positive texts. It then determines a raw score for all text samples in the list. Rather than using the normal distance measure, we opted for a non-symmetric measure which is a weighted combination of two factors: a) the difference between text score and average profile score for each feature and b) the text score by itself. This makes it possible to assign more importance to features whose count deviates significantly from the norm. The following distance formula is used:

$$\Delta_T = (\Sigma\ |T_i - A_i|^{D}\ |T_i|^{S})^{1/(D+S)}$$

In this formula, T_i and A_i are the values for the i^{th} feature for the text sample profile and the positive average profile respectively, and D and S are the weighting factors that can be used to assign more or less importance to the two factors described. The distance measure is then transformed into a score by the formula

$$Score_T = (\Sigma\ |T_i|^{(D+S)})^{1/(D+S)} - \Delta_T$$

The score will grow with the similarity between text sample profile and positive average profile. The first component serves as a correction factor for the length of the text sample profile vector.

The order of magnitude of the score values varies with the setting of D and S, and with the text collection. In order to bring the values into a range which is suitable for subsequent calculations, we express them as the number of standard deviations they differ from the mean of the scores of the negative example texts.

3 Language verification

In order to test the feasibility of language verification by way of linguistic profiling, we need data which is guaranteed to be written by native and non-native speakers respectively. Moreover, the texts (native and non-native) should be as similar as possible with respect to the genre they represent. For the present study, therefore, we opted for the student essays in the Louvain Corpus of Native English Essays (LOCNESS) and the International Corpus of Learner English (ICLE; Granger et al., 2002).

3.1 LOCNESS and ICLE

ICLE is a collection of mostly argumentative essays written by advanced EFL students from various mother-tongue backgrounds. The essays each are some 500-1000 words long (unabridged) and although they 'cover a variety of topics, the content is similar in so far as the topics are all non-technical and argumentative (rather than narrative, for instance)' (cf. Granger, 1998:10). The size of the national sub-corpora is approx. 200,000 words per corpus. With the data metadata are available as they have been collected via a learner profile questionnaire.

The LOCNESS in various respects is comparable to ICLE. It is a 300,000-word corpus mainly of essays written by English and American university students. A small part of the corpus (60,000 odd words) is constituted by British English A-level essays. Topics include transport, the parliamentary system, fox hunting, boxing, the National Lottery, and genetic engineering.

3.2 Training and test texts

In order to be able to control for language variation between British and American English, we opted for only the British part of LOCNESS. Because this totalled only some 155,000 words, we decided to hold out about one third as test material and use the other two thirds for training. In order to have as little overlap as possible in essay type and topic between training and test material, we used sub-corpora 2, 3 and 8 of the A-level essays and sub-corpus 3 of the university student essays for testing.

For the ICLE texts, we chose to use each tenth text for training purposes. The remaining texts were used for testing.

3.3 General results

In the first step of training, we selected the features to be profiled. We used all features which occurred in more than one training text, i.e. about 470K features. In the second step, we selected the system parameters D and S for two classification

models: similarity to the native texts (D=1.0, S=0.0) and similarity to the non-native texts (D=1.2, S=0.2). The selection was based on the quality of classifying half of the training texts with the system having been trained on the other half.

The verification results for the test set of A-level texts are shown in Figure 1. The further the texts are plotted to the right, the more similar their profile is to the mean profile for the A-level training texts. The further the texts are plotted towards the top, the more similar their profile is to the mean profile for the ICLE training texts.

Most of the texts form a central cluster in the bottom right quadrant. A small gap separates them from a group of five near outliers, while there are two far outliers. We decided to use the limits of the central cluster as our classification separator, accepting that 10% of the LOCNESS texts would be rejected. We added the separation line to the plot. In order to create a reference frame linking this figure to the following ones, we add a second line, along the core of the cluster of the LOCNESS texts. Even though the core of the clusters in the successive figures may shift, this line remains constant, as does the plotting area.

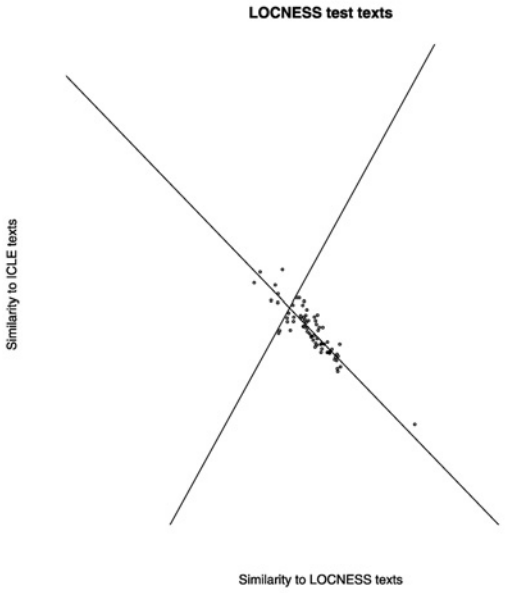

Figure 1. Text classification of the LOCNESS test texts in terms of similarity to native texts (horizontal axis) and similarity to non-native text (vertical axis). The separation line (top right to bottom left) divides the plot area in a native part (bottom right) and a non-native part (top left). The second line (top left to bottom right) is a reference line which allows comparison between this Figure and Figures 2-4.

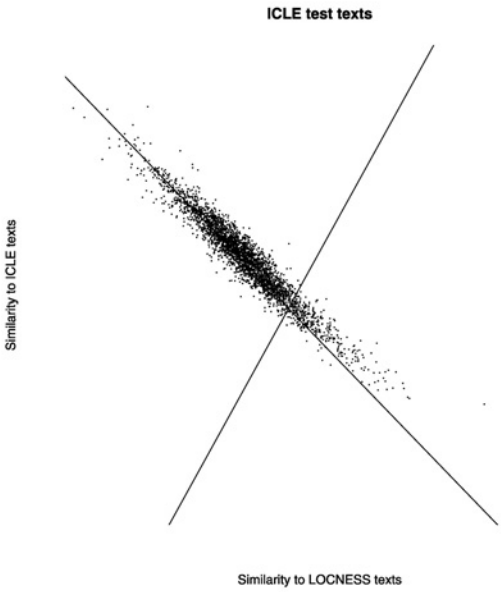

Figure 2. Text classification of the ICLE test texts

Figure 2 shows the results for the ICLE test texts. 89% of the texts are rejected. The verification results differ per nationality. A more detailed examination of such variation, however, is beyond the scope of the present paper.

The two dimensions, the degree of similarity to native texts and the degree of similarity to non-native texts, are strongly (negatively) correlated. Still, there are also clear differences, so that both dimensions contribute substantially to the quality of the separation.

3.4 Distinguishing features

When examining some of the features that emerge from studies reported in the literature as salient in describing different language varieties, we find that none of these dominates the classification. Table 1 shows the influence of each feature in terms of its contribution (expressed as millionths of the total influence, so e.g. 3173 corresponds to 0.3% of the total influence) to the decision to classify a text as native or non-native. The second and third column show the influence of the words (or word combinations) by themselves, which is extremely low. However, when examining all patterns containing these words, the fourth and fifth columns, their usefulness becomes visible.

Previous studies into the use of intensifying adverbs have shown an overuse of the token *very*. Thus it is a likely candidate to be considered as a marker of non-native language use. The second column in the Table confirms this, but the contribution is a mere 0.001%. The picture changes when we consider all patterns in which *very* occurs, it appears that there is indeed a

difference in use of the token by natives and non-natives. However, there are as many patterns that point to nativeness as there are that point to non-nativeness. Furthermore, the patterns provide a sizeable contribution in the classification either way.

Word(s)	Sep → ICLE	Sep → LOC	Patterns → ICLE	Patterns → LOCNESS
if *If* *if*	13	4	3931	4529
because	4	-	3230	2925
very	10	-	2860	3173
however	-	1	686	644
therefore	-	10	953	734
for instance	4	-	30	32
thus	2	-	411	287
yet	4	-	606	349

Table 1. Relative contribution to the overall classification of allegedly salient features

Although the expected features (or rather features related to expected word or word combinations) have a visible contribution, their influence is still only a small part of the total influence. In fact, all features have only very little influence. The most influential single feature is ccc=#HF#AT--#HF#NN1--#HF#CC—RRx13, one of the representations of *the*, followed by a single common noun, followed by *and*, a pattern unlikely to be spotted by humans. It contributes 0.06% of the influence classifying texts as non-native. Only 137 features in total contribute more than 0.01% either way. Classification by linguistic profiling is a matter of myriads of small hints rather than a few pieces of strong evidence. This is probably also what makes it robust against high text variability and sometimes small text sizes.

4 Domain Shifts

Now that we have seen that language verification is viable within the restricted domain of student essays, we may examine whether it survives the shift to a new domain. We tested this on two corpora: the FLOB corpus and (small) internet corpus that was especially collected for this purpose.

4.1 FLOB

The Freiburg LOB Corpus, informally known as FLOB (Hundt et al., 1998) is a modern counterpart to the much used Lancaster-Oslo/Bergen Corpus (LOB; Johansson, 1978) It is a one-million word corpus of written (educated) Modern British English. The composition of FLOB is essentially the same as that of LOB: it comprises 500 samples of 2,000 words each. In all, 15 text categories (A-R) are distinguished. These fall into four main classes: newspaper text (A-C), miscellaneous informative prose (D-H), learned and scientific English (J), and fiction (K-R).

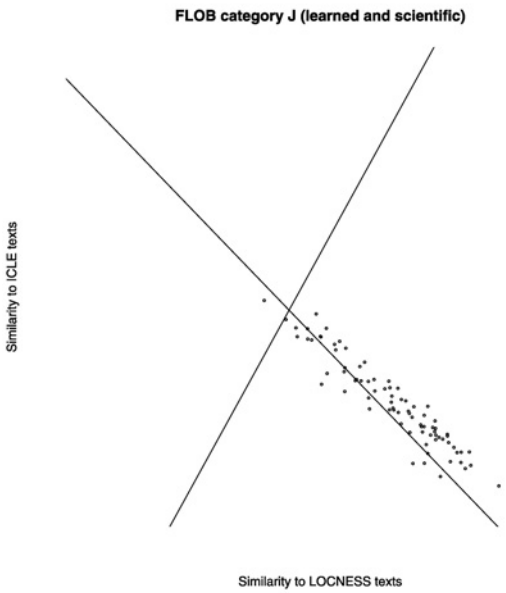

Figure 3. Text classification of the FLOB learned and scientific texts (category J)

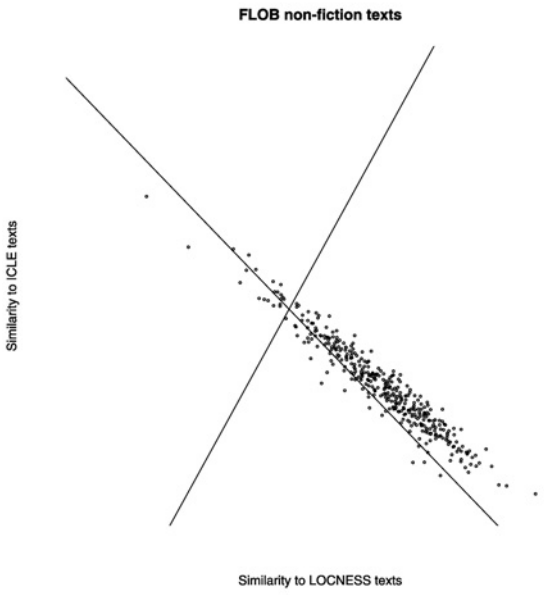

Figure 4. Text classification of the FLOB non-fiction texts (categories A-J)

Of these texts, the learned and scientific class (J) is closest to the ICLE and LOCNESS texts, and we should expect that the FLOB texts of this category are all accepted. This is indeed the case, as can be seen in Figure 3, which shows the classification of these texts. Only 1 text is rejected (1.25%). This

seems to confirm that we are indeed recognizing something like '(near-)native English'.

As soon as we shift the domain of the texts, however, the native texts are no longer distinguished as clearly. The larger the domain shift, the more texts are rejected. Within the non-fiction portion of FLOB, the system rejects 2.3% of the newspaper texts (categories A-C) and 8.7% of the miscellaneous and informative prose texts (D-H). This leads to an overall reject rate of 5.6% for the non-fiction texts (Figure 4), which is still reasonably acceptable. When shifting to fiction texts (K-R), the reject rate jumps to 39.2% (Figure 5), indicating that a new classifier would have to be trained for a proper handling of fiction texts.

Figure 5. Text classification of the FLOB fiction texts (categories K-R)

4.2 Capital-Born

Since our original goal was the filtering of internet texts, we compiled a small corpus of such texts. We chose texts which were present as HTML. These, we expected, were likely to be rather abundant, while they would have been subjected to a relatively low degree of editing. Thus they would constitute likely candidates for filtering. In order to be able to decide whether the texts were native-written or not, we searched autobiographical material, as indicated by the phrase *I was born in CITY*, with CITY replaced by a name of a capital city. The initial set of documents appeared to be of a reasonable size. However, after filtering out webpages by multiple authors (e.g. guest books), fictional autobiographies (e.g. a joke page about Al Gore), texts judged likely to be edited possibly with the help of a native speaker (e.g. a page advertising Russian brides), misclassified city names (e.g. authors from *Paris, Texas* should not be assumed to be French) and texts outside the desired length of 500-1500 words, we ended up with a mere 20 native British English texts and 18 non-native texts. We nicknamed the corpus "Capital-Born corpus".

When classifying these texts with the A-level versus ICLE classifier, we see that they cluster tightly, outside the area plotted so far, and showing no useful separation of native and non-native texts. This implies that if we want a filter for such texts, we have to train a new classifier.

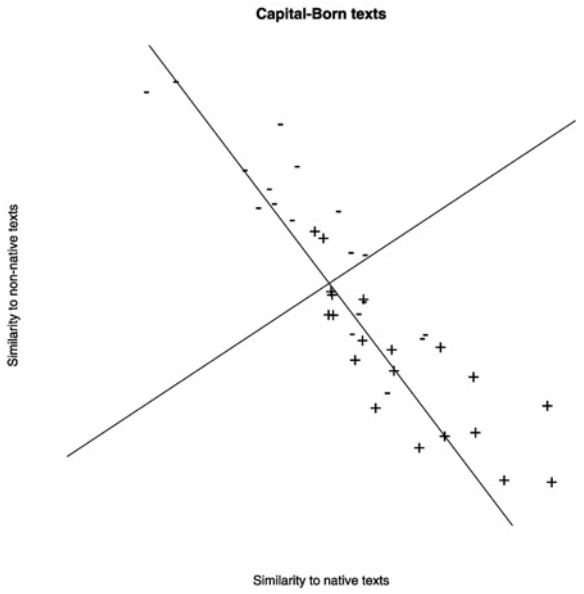

Figure 6. Text classification of internet texts (for a description see section 4.2)

We did train such a new classifier, using only the odd-numbered Capital-Born texts and classified the even-numbered ones, using the same parameters D and S as above. We repeated the process with the two sets switching roles. Figure 6 shows a superposition of the classifications in the two experiments. The native texts appear as plus signs (+), the non-native texts as minus signs (–). Note that we adjusted the separation and support lines in order to bring them in line with the data. Only a rough separation is visible, with 2 out of 20 native texts misclassified and 6 out of 18 non-native texts. Still, given the extremely small size of the training sets and the variety of non-native nationalities, these results are rather promising. It appears that even internet texts can be filtered for nativeness, as long as a restricted, and more sizeable, seed corpus can be constructed.

5 Conclusion

The results show that language verification is indeed possible, as long as we accept that near-

native texts produced by non-natives will not be filtered out.

Furthermore, whenever a verification filter is needed, it will be necessary to create a new filter, based on a seed corpus which contains both native and non-native texts as similar as possible in type to the texts which are to be filtered.

There are now two avenues open for future research. First of all, we would like to explore the classification procedure linguistically: a) examine the distinguishing features in more detail and compare our findings with those in the literature, and b) examine the correlation of the nativeness score of the various texts to extra-linguistic text variables such as mother tongue and learner level.

Secondly, once more insight is gained into the linguistic workings of the procedure, the classification process can be refined. At this point, we would also like to examine the effects of domain shift in more detail, and attempt to estimate a minimum size for seed corpora for use in filtering internet material.

6 Acknowledgements

Thanks are due to Sylviane Granger and Sylvie De Cock (Centre for English Corpus Linguistics, Université Catholique de Louvain, Belgium) for making the LOCNESS and ICLE data available to us.

References

Douglas Biber. 1995. *Dimensions of register variation. A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Douglas Biber 1998. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Douglas Biber, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

BNC. 2002. The BNC sampler. Web page: www.natcorp.ox.ac.uk/getting/sampler.html

Susan Conrad and Douglas Biber (eds.) 2001. *Variation in English: Multi-dimensional studies*. Harlow, England: Longman.

Alan Davies. 2003. *The Native Speaker: Myth and Reality*. Clevedon: Multimingual Matters Ltd.

Sylviane Granger (ed.) 1998. *Learner English on Computer*. London and New York: Longman.

Sylviane Granger. 1998. The computer learner corpus. In Sylviane Granger (ed.): 3-18.

Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson (eds.) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins.

Sylviane Granger, E. Dagneaux, and Fanny Meunier (eds.) 2002. *International Corpus of Learner English.* Louvain: UCL Presses Universitaires de Louvain.

Hans van Halteren. 2000. The detection of inconsistencies in manually tagged text. *Proc. Workshop on Linguistically Interpreted Corpora (LINC2000)*. 48-55.

Hans van Halteren. 2003. New feature sets for summarization by sentence extraction. *IEEE Intelligent Systems*, July/August 2003: 34-42.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. Proc. ACL 2004.

Marianne Hundt, Andrea Sandt and Rainer Siemund. 1998. *Flobman. Manual of Information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Freiburg: Englisches Seminar.

Stig Johansson with Geoffrey Leech and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Dept. of English, University of Oslo.

M. van der Laaken, R. Lankamp, and Michael Sharwood Smith. 1997. *Writing Better English*. Bussum: Coutinho.

LOCNESS. http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Icle/LOCNESS.htm