A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text

Marco Baroni and Silvia Bernardini
SSLMIT, University of Bologna

Abstract

In this article we describe an approach to the identification of 'translationese' based on monolingual comparable corpora and machine learning techniques for text categorization. The article reports on experiments in which support vector machines (SVMs) are employed to recognize translated text in a corpus of Italian articles from the geopolitical domain. An ensemble of SVMs reaches 86.7% accuracy with 89.3% precision and 83.3% recall on this task. A preliminary analysis of the features used by the SVMs suggests that the distribution of function words and morphosyntactic categories in general, and personal pronouns and adverbs in particular, are among the cues used by the SVMs to perform the discrimination task. A follow-up experiment shows that the performance attained by SVMs is well above the average performance of ten human subjects, including five professional translators, on the same task. Our results offer solid evidence supporting the translationese hypothesis, and our method seems to have promising applications in translation studies and in quantitative style analysis in general. Implications for the machine learning/text categorization community are equally important, both because this is a novel application and especially because we provide explicit evidence that a relatively knowledge-poor machine learning algorithm can outperform human beings in a text classification task.

Correspondence:
Marco Baroni, SSLMIT,
University of Bologna,
Corso della Repubblica 136,
47100 Forlì (FC), Italy.
E-mail:
baroni@sslmit.unibo.it

1 Introduction

It is common, when reading translations, to feel that they are written in their own peculiar style. Translation scholars even speak of the language of translation as a separate 'dialect' within a language, which they call *third code* (Frawley, 1984) or *translationese* (Gellerstam, 1986).

Recently, attempts have been made to establish whether translationese really exists, i.e. whether translations do tend to share a fixed set of lexical, syntactic and/or textual features, and to identify such features (Puurtinen, 2003). This approach departs from a more traditional method of analysis in translation studies which consists in comparing a source text in language A with its translation in language B. Instead, it typically compares large bodies of translated text with large bodies of original text in the same language. The aim here is that of exploring how 'text produced in relative freedom

Literary and Linguistic Computing, Vol. 21, No. 3, 2006.
doi:10.1093/llc/fqi039 Advance Access Published on 5 August 2005

from an individual script in another language differs from text produced under the normal conditions which pertain to translation, where a fully developed and coherent text exists in language A and requires recoding in language B' (Baker, 1995). See Subsection 2.1 in this article and Puurtinen (2003) for a more extensive discussion of translationese.

In an unrelated line of research, various recent studies extend supervised machine learning techniques traditionally used for topic classification tasks to the categorization of texts by genre and style. See Subsection 2.2 in this article and Santini (2004) for a discussion of some of the relevant literature.

In this article, we show that text classification with support vector machines (Joachims, 1997) can be successfully applied to the task of telling high quality translated text from original (non-translated) text written in the same language (Italian), dealing with the same topics (geopolitics subdomains), and belonging to the same genre (journal articles) and publication (the *Limes journal*, http://www.limesonline.com). We also present the results of an experiment showing that the algorithm's performance is decidedly better than average when compared to that of human beings faced with the same task.

From the point of view of translation studies, our results are of interest because they bring clear evidence of the existence of translationese features even in high quality translations. In particular, they do so by showing that these features are robust enough to be successfully used for the automated detection of translated text. Moreover, the difference in performance obtained with specific combinations of features and models provides preliminary insights into the nature of these features. We believe that our study should be of interest from the methodological point of view as well, in that we use learnability by a machine as a new criterion to assess the significance of stylistic differences, a criterion that could possibly complement more typical statistical significance analyses.

From the point of view of automated text categorization, our results are interesting not only because of the novelty of the task and methodology but also because, as far as we know, this is the first study to provide experimental evidence that a relatively knowledge-poor machine learning algorithm can outperform human beings in a text classification task. This suggests that automated text categorization techniques are reaching a level of performance at which they can compete with humans not only in terms of cost-effectiveness and speed, but also in terms of quality of classification in difficult tasks.

From a practical point of view, an automated 'translationese spotter' could have some interesting applications. For example, it might become part of a (self-)assessment tool for translators and translation students on the one hand, and translation commissioners on the other. It could also help web-based parallel corpus extractors (Resnik and Smith, 2003) in the search and assessment of candidate parallel texts (one of their sides must contain translated text). Finally, one could envisage an application in multilingual plagiarism detection, if such an issue exists (for a general overview of plagiarism detection see e.g. Clough, 2000).

The remainder of this article is structured as follows: In Section 2, we briefly review previous work on the characterization of translationese and on automated genre/style categorization. We then describe our corpus (Section 3) and the machine learning algorithm we used (Section 4). In Section 5, we discuss the ways in which we represent documents in the automated categorization experiments, which are reported in Section 6. In Section 7, the results of these experiments are compared with the performance of humans on the same task. Section 8 concludes by summing up our main results and presenting suggestions for further work.

2 Related Work

2.1 Characterization of Translationese

Translationese has originally been described (Gellerstam, 1986) as the set of 'fingerprints' that one language leaves on another when a text is translated between the two. Thus, Gellerstam searches for fingerprints of English on Swedish texts, with the aim to describe 'the Swedish language variety used in translations from English' (Gellerstam, 1996). More recently, the hypothesis has also been put forward that any translated language variety, regardless of the source and target languages, might share characteristic features typical of translation 'as a mediated communicative event' (Baker, 1993).

The methodology adopted in studies of translationese is typically based on the construction of *monolingual comparable corpora*, which include original (non-translated) texts in a given language and translations into the same language. These corpora are then used to compute statistics about the distribution of manually selected features expected to be relevant to the translated/original distinction. A corpus of this kind has also been used to investigate human perception of translationese, i.e. whether subjects can tell originals and translations apart, and what cues they use for this purpose (Tirkkonen-Condit, 2002).

Preliminary hypotheses based on corpus evidence tentatively suggest that translated text might be more explicit, more conservative and less lexically dense than comparable original text; see Hansen (2003), Laviosa (1998), Olohan (2001) and the survey in Puurtinen (2003).

A number of more specific hypotheses have equally been put forward, e.g. that translations tend to underrepresent those linguistic features typical of the target language, which lack obvious equivalents in the source language (Mauranen, 2002; Tirkkonen-Condit, 2004).

Searching for typical collocations respectively in EU reports (in original and translated English) and geopolitics articles (in original and translated Italian), Baroni and Bernardini (2003) find that the bigrams most characteristic of translated text (as opposed to originals) are sequences of function words.

Borin and Prütz (2001) compare original news articles in British and American English with articles translated from Swedish into English. Focusing on syntactic aspects, they search for part-of-speech n-grams which appear to be under- or overrepresented in English translations vs. English originals. While some of the deviations they find can be explained by reference to cross-linguistic differences between English and Swedish, others are more intriguing, e.g. the overrepresentation of adverbs, infinitives, pronouns, sentence-initial verbs, and sentence-initial prepositions in translation. The authors suggest that some of these findings at least seem to be interpretable in terms of translationese effects.

The comparable corpus methodology provides interesting insights and has a number of practical advantages: monolingual corpora are easier to assemble than parallel ones, as text material is more abundant and copyright permission less burdensome to obtain; also, these corpora do not raise the cross-linguistic comparison issues or pose the alignment problems typical of parallel corpora. Yet, the methodology is not free from problems, which derive primarily from the difficulty of obtaining truly comparable corpora. While all the studies mentioned here highlighted differences between originals and translations which might be interpreted in terms of translationese, these effects are often weak, or, more worryingly, might also be due to confounding factors.

For example, Gellerstam (1996) finds differences in the use of reporting clauses in translated vs. original novels in Swedish. While these differences might hint at translationese, Gellerstam also mentions the possibility that they are due to a genre-based difference, i.e. to a higher incidence of detective stories in the translated corpus than in the original corpus (detective stories being often translated from English).

Borin and Prütz (2001) similarly hypothesize that the overrepresentation of verb-initial sentences in their translated newsarticle text corpus with respect to its original counterpart might be due to the presence of a more substantial number of 'letters to the editor' in the former than in the latter. The language here is likely to differ from that of the rest of the corpus, the authors suggest, because readers' letters are more likely to contain direct (yes/no) questions, and hence verb-initial sentences.

Among the limits of her translationese study, Puurtinen (2003) mentions the possibility that subgenres of children's literature (the genre under investigation) show different lexical and even syntactic features, and may be subject to different translation conventions.

One could go on to add, among the potential confounding variables, publishing and marketing policies, age and gender of the intended audience, and so forth. Ideally, the impact of these factors should be controlled during corpus construction, but this is not easy when working with authentic texts in general, and it is particularly difficult when dealing with translation (Bernardini and Zanettin, 2004).

We believe that the present study has two important contributions to make to research on translationese. First of all, the comparability of the corpus components is as good as it can be, and consequently risks related to potential confounding factors are far lower than in many translationese studies. This was achieved by constraining the typology of texts to be included in the corpus to articles published by a single journal and within the same time span, consistent in topic and subgenre. While this corpus has a number of drawbacks, first and foremost its limited 'representativeness', we believe it to be very well-suited to the search for translationese features. We shall come back to the (dis-)advantages of our corpus in Section 3.

Secondly, we introduce a new explicit criterion to prove the existence of translationese, namely learnability by a machine. The results of statistical significance tests applied to a few hand-picked features are often hard to interpret, as pointed out by Teich (2003) and Toury (2004). If differences between the two types of texts are robust enough to allow a machine, relying on very large sets of unfiltered features, to learn the distinction between an original and a translation, then this is, we suggest, a very strong argument in favour of the existence of translationese.

2.2 Automated Text Categorization by Genre and Style

In the last 15 years or so, substantial research has been conducted on text classification through supervised machine learning techniques (Sebastiani, 2002). The vast majority of studies in this field focus on classification by topic, where bag-of-content-word models turn out to be very effective. Recently, there has also been increasing interest in automated categorization by overall sentiment, degree of subjectivity, authorship and along other dimensions that can be grouped together under the loosely defined cover terms of 'genre' and 'style'. Among the studies representing this line of research, see, for example, Argamon *et al.* (1998), Stamatatos *et al.* (2000), Mayfield Tomokiyo and Jones (2001), Pang *et al.* (2002), Koppel *et al.* (2002), Finn and Kushmerick (2003), Kindermann *et al.* (2003). Some of the work on genre identification has been recently surveyed by Santini (2004).

Genre and style classification tasks cannot be tackled using only the simple lexical cues that have proved so effective in topic detection. An objective report and a subjective editorial about the Iraq war will probably share many of the same content words. Conversely, objective reports about Iraq and about football will share very few interesting content words. Thus, categorization by genre and style must rely on more abstract topic-independent features, such as sequences of morphosyntactic categories. At the same time, because of the typical empirical NLP constraints of rapid development, scalability and easy adaptation to new languages and domains, work in this area has concentrated on relatively shallow features that can be extracted from texts efficiently and with little resources.

The machine learning/text categorization approach to genre and style differs from more linguistically sophisticated methods, such as the one developed by Biber (1995) for register analysis. The latter relies on a detailed tagging/parsing system with human post-editing and disambiguation, on the selection of meaningful linguistic features and on the functional interpretation of the dimensions underlying the various factors, in order to develop a detailed multidimensional analysis based on a large set of co-occurring lexico-syntactic features. This means that, while possible, portability (e.g. to novel languages and very large-scale projects) is not straightforward.

In the tradition of genre and style classification by machine learning, on the other hand, popular choices of features (in part inspired by the classic literature on stylometry; e.g. Mosteller and Wallace, 1964) have been function words (which are usually discarded or down-weighted in topic-based categorization), textual statistics (e.g. average sentence length, lexical richness measures) and knowledge-poor surrogates of a full syntactic parse, such as n-gram and part-of-speech (pos) information.

Because of the different languages, experimental settings and performance measures generalizations are not straightforward. However, most genre/style categorization studies report accuracies around 80% or just above this threshold, which indicates that much more work is needed in this area to reach the performance level of topic-based categorization (where accuracy is often well above 90%).

We briefly discuss here four style/genre categorization studies that are relatively similar to ours in terms of method (Kindermann *et al.*, 2003), language and method (Baroni *et al.*, 2004), task definition (Mayfield Tomokiyo and Jones, 2001), and task difficulty (Koppel *et al.*, 2002).

Kindermann *et al.* (2003) successfully apply support vector machines, the same machine learning technique we used, to authorship attribution of German news articles. In their experiments, wordform unigrams achieve better recall, whereas models trained on word lengths and representations of uni- and bigrams similar to our 'mixed' features (see Section 5) achieve (slightly) better precision.

Baroni *et al.* (2004) report 90% accuracy (and similar precision and recall) for the categorization of Italian news articles into objective reports and subjective commentaries. They use support vector machines trained on a wordform unigram model. They hypothesize that the success of this simple model is due to the rich inflectional system of Italian, which carries explicit cues of different genres (e.g. conditional inflections cue subjective texts).

Mayfield Tomokiyo and Jones (2001) categorize transcriptions of speech produced by native and non-native English speakers in a variety of experimental settings. They obtain good results using the Naive Bayes classifier trained on word and pos n-grams, and relying on feature selection techniques favouring high frequency (function?) words. Our task (automated identification of translated text) is superficially similar to the nativeness detection task. However, written text produced by professional translators (who are normally native speakers of the target language) is likely to differ from original text in much more subtle ways than those in which speech elicited from non-native speakers with significant communication difficulties (Mayfield Tomokiyo and Jones' explicit targets) differs from native speech.

Koppel *et al.* (2002) train a variant of the Winnow algorithm to distinguish between English texts written by female and male authors. With a combination of function word and pos n-gram features, they achieve accuracy around 80%. While author gender identification has little in common with translated text detection, both are very difficult tasks, which would be challenging even for humans. Indeed, in both cases we train the machine to detect impalpable differences that we are not even sure exist.

To the extent that the latter two lines of research are successful, they represent a new, more ambitious application of text categorization, which is no longer used only as a cheaper and faster alternative to human labour, but also as a discovery procedure to find patterns that humans are not good at detecting.

3 Corpus Construction

The corpus we used for this project is a collection of articles from *Limes*, an Italian geopolitics journal. The complete text of the 1993–1999 collection was extracted from a CD-ROM (with the editors' permission). Tables, figures, and any 'suspicious' text exemplar were removed semi-automatically: for instance, we discarded all 'round-table' articles for the fear that they might be preponderantly original, and all interviews because they are likely to systematically feature original and translated language, i.e. when the interviewee speaks a language other than Italian. Translated articles (i.e. all the articles containing the pattern *translated by NAME*) were then automatically identified and manually checked; these formed the translation subcorpus. What was left became the *de facto* original subcorpus. Data about the *Limes* corpus are given in Table 1.

This corpus would seem to be very well-suited to the purpose of investigating translationese. It is very homogeneous in terms of genre and macro-topic (all articles deal with geopolitical issues), and it is well-balanced in terms of

Table 1 The *Limes* corpus

Parameter	Originals	Translations
n of articles	569	244
n of words	2 032 313	877 781
avg art length	3 572	3 597
n of authors	517	134
n of translators	NA	103
source languages	NA	Arabic, English, French, Russian, Spanish, . . .

micro-topics (each journal issue centers around one theme, and contains original and translated articles dealing with it). Lastly, all articles are likely to have gone through the same editorial process and the quality of translations is extremely high (as judged by the present authors, both native speakers of Italian).

A drawback of having a very uniform, very comparable corpus is that the results of our experiment may be true only for the specific genre and domain under analysis. However, the necessity to control external variables as closely as possible is arguably primary at these early stages of experimentation. We believe that the generality of results should be demonstrated through an accumulation of findings from several experiments, each based on a small homogeneous corpus, rather than through a single experiment with a large varied corpus, where confounding factors would be difficult to control. We intend to pursue this route in future studies.

Another advantage of our corpus is that translations are carried out from several source languages into Italian (the journal board tells us that they translate mainly from English, Arabic, French, Spanish, and Russian). This should be some guarantee that any effect we find is less likely to be due to the characteristics of a source language in particular than to more general translationese patterns. Unfortunately, there is no way of knowing exactly what the source language is for each article, or to evaluate the relative proportion of each source language out of the total articles.

The articles in the corpus were tagged with the combination of taggers described in Baroni *et al.* (2004) and lemmatized with the Italian TreeTagger (Schmid, 1994). To eliminate a potentially rich source of content-based information, all words tagged as proper nouns were replaced by a string of shape *NPRid*, where a unique, increasing id number is assigned to all distinct proper nouns of an article, in the order in which they appear (restarting from 1 at the beginning of each article). For example, if the first two proper nouns appearing in an article are *Italia* and *Roma*, the article is recoded by substituting all occurrences of *Italia* with NPR1 and all occurrences of *Roma* with NPR2. If, in another article, the first name is *Bush* and the second name is *Italia*, in that article all occurrences of *Bush* are replaced with NPR1 and all occurrences of *Italia* are replaced with NPR2. In other words, the same id will typically correspond to different names in different articles, and the same name will be mapped to different ids in different articles. In this way, we remove an important source of extra-linguistic information that could help to distinguish between translated and original text (for example, Italian writers could have a tendency to name Italian public personalities more often than foreign writers would). Of course, this is only an issue for classifiers based on content words.

4 Support Vector Machines

We use support vector machines as implemented in the SVMLight package (Joachims, 1999). Support vector machines (SVMs) are a classification technique that was first applied to text categorization by Joachims (1997). During training, this algorithm constructs a hyperplane that maximally separates the positive and negative instances in the training set (instances – in our case: documents – are represented as vectors of features). Classification of new instances is then performed by determining which side of the hyperplane they fall on. For an introduction to SVMs, see Cristianini and Shawe-Taylor (2000).

We chose SVMs because they provide state-of-the-art performance in text categorization, including promising results in style/genre-based classification (Kindermann *et al.*, 2003; Baroni *et al.*, 2004). Moreover, SVMs require neither preliminary feature selection (they are able to handle a very large number of features) nor heuristic parameter tuning (there is a theoretically motivated choice of parameter settings). Thus, we can concentrate on different featural representations of the documents without worrying about the potential combinatorial explosion of experiments to run that would be caused by the need to test different feature selection techniques and parameter values for each representation. At the same time, we reduce the risk of throwing away meaningful data (as we could end up doing if we focused on a single manual feature selection scheme).

5 Representation of Documents

We explore a number of different ways to represent a document (i.e. a journal article) as a feature vector, by varying both the size (unigrams, bigrams, and trigrams) and the type (wordform, lemma, pos tag, mixed) of units encoded as features, as shown in Table 2.

In the mixed representation, function words are left in their inflected wordform, whereas content words are replaced by the corresponding tags. Less frequent adverbs, with *frequency* < 30 (an arbitrary threshold), and adverbs of the open -*mente* class (corresponding to English -*ly*) are treated as content words, the other adverbs are treated as function words. To see why these classes should be treated differently, compare *stancamente* ('tiredly'), quite obviously a content-rich word, to *non* ('not'), obviously a functional element.

Unigram wordform and lemma representations mostly convey lexical cues, and they are akin to the representations typically used in topic-based categorization. All pos/mixed representations and, to some extent, wordform and lemma multiword representations convey grammatical information, and they have been used in other style-based categorization studies.

For each feature set, we build both unweighted and weighted frequency vectors representing the documents. Following standard practice, we use *tf*idf* weighting (term frequency times inverted document frequency). The value of a feature in a document is given by (a logarithmic transformation of) its frequency in the document divided by its overall document frequency (i.e., by the number of distinct documents in which the feature occurs). All features that occur in fewer than 3 documents are discarded, and all vectors are length-normalized. In the unweighted models, we discard features that occur in more than half the documents.

We also experiment with combinations of SVMs trained on different representations. We use two methods to combine the outputs of the single classifiers: *majority voting* (which labels an article as translated only if the majority of classifiers think it is translated, with ties broken randomly) and *recall maximization* (which labels an article as translated if at least one classifier thinks it is translated). We decided to try the recall maximizing method after observing in unrelated text categorization experiments (Baroni *et al.*, 2004) that, when the majority of training instances are negative (as in the current

Table 2 Units encoded as features

Unit size	Unit type	Example
unigram	wordform	Prendendo (*taking*)
unigram	lemma	PRENDERE (*TAKE*)
unigram	pos	V:geru
unigram	mixed	cont word: V:geru; func word: i (*the (pl.)*)
bigram	wordform	i fatti (*the (pl.) facts*)
bigram	lemma	IL FATTO (*THE FACT*)
bigram	pos	ART N
bigram	mixed	i N (*the (pl.) N*)
trigram	wordform	Prendendo i fatti (*taking the (pl.) facts*)
trigram	lemma	PRENDERE IL FATTO (*TAKE THE FACT*)
trigram	pos	V:geru ART N
trigram	mixed	V:geru i N (*V:geru the (pl.) N*)

case), SVMs behave conservatively on the test set, achieving high precision at the cost of very low recall. As far as we know, this is the first text categorization study that combines SVMs using the recall maximization method.

Since we have 24 distinct single classifiers (i.e. all those in Table 2, weighted and unweighted), it is not realistic to analyze all their possible combinations. Thus, we select a set of combinations that are plausible a *priori*, in two senses: first, they are composed only of sensible single classifiers; second, the classifiers in each combination are reasonably different from each other. As an example of the application of the first criterion, we only consider combinations with SVMs trained on trigram pos and mixed representations, since these are likely to be more informative than trigram wordform- and lemma-based features, which will suffer from data-sparseness problems. As an example of a choice based on the second criterion, we do not consider combinations of unigram wordform and lemma representations, since these are likely to be rather similar. We could have looked for the best combinations experimentally on a development test set. However, besides the fact that this would have complicated our cross-validation design, we feel that *a priori* plausible combinations are more likely to lead to easily interpretable results. The selected combinations are reported in Table 3.

Table 3 Combinations

Id	Unigrams	Bigrams	Trigrams
1	wform	wform + pos	pos
2	wform	wform + pos	mix
3	wform	lemma + mix	pos
4	wform	lemma + mix	mix
5	wform	lemma + pos	pos
6	wform	lemma + pos	mix
7	wform_tfidf	wform + pos	pos
8	wform_tfidf	wform + pos	mix
9	wform_tfidf	lemma + mix	pos
10	wform_tfidf	lemma + mix	mix
11	wform_tfidf	lemma + pos	pos
12	wform_tfidf	lemma + pos	mix
13	lemma_tfidf + mix	wform + pos	pos
14	lemma_tfidf + mix	wform + pos	mix
15	lemma_tfidf + mix	lemma + mix	pos
16	lemma_tfidf + mix	lemma + mix	mix
17	lemma_tfidf + mix	lemma + pos	pos
18	lemma_tfidf + mix	lemma + pos	mix
19	lemma_tfidf + mix_tfidf	wform + pos	pos
20	lemma_tfidf + mix_tfidf	wform + pos	mix
21	lemma_tfidf + mix_tfidf	Lemma + mix	pos
22	lemma_tfidf + mix_tfidf	lemma + mix	mix
23	lemma_tfidf + mix_tfidf	lemma + pos	pos
24	lemma_tfidf + mix_tfidf	lemma + pos	mix

We also trained single SVM classifiers using combined features from multiple representations (e.g. training a single model with unigram, bigram, and trigram features). However, the performance of these models was very low. We shall therefore not discuss them here.

6 Experiments with SVMs

We split the corpus into 16 sections, each made of 15 random original documents and 15 random translated documents. This left a remainder of 240 original texts and 4 translated texts. The 30-document sections were used in a series of 16-fold cross-validation experiments; the articles in the remainder were used as part of the training data in each fold, but never as test data. Thus, within each fold, the training set contained 229 translated texts and 465 original texts; the test sets contained 15 translations and 15 originals. All the results we report are averaged across the 16 folds and expressed in percentages. The usual performance measures were computed by treating translations as positives.

To put the results that are about to be reported into perspective, consider that, on the same data, a random classifier that assigns documents to the translated and original classes with equal probability would obtain 50% accuracy, 50% precision, 50% recall, and 50% F. Note that this random classifier, unlike our SVMs, knows the true proportion of translations and originals in the test sets (50–50). A trivial acceptor treating all documents as translated would have 50% accuracy, 50% precision, 100% recall, and 66.7% F.

6.1 Results with Single Classifiers

Table 4 reports the results obtained by the single (i.e. non-combined) models, ranked by F value. For each representation, we only report results

Table 4 Results of single classifiers

Unit size	Unit type	tfidf	Accuracy	Precision	Recall	F
unigram	wordform	no	77.1	94.5	57.5	71.5
bigram	mixed	no	77.1	94.5	57.5	71.5
unigram	mixed	yes	76.9	93.3	57.9	71.5
unigram	lemma	no	74.2	92.6	52.5	67.0
bigram	lemma	no	74.0	96.7	49.6	65.6
bigram	wordform	no	73.8	97.5	48.8	65.0
trigram	pos	no	71.5	93.3	46.2	61.8
trigram	mixed	no	70.4	97.1	42.1	58.7
trigram	lemma	no	65.4	98.7	31.2	47.5
bigram	pos	yes	63.1	92.0	28.8	43.8
trigram	wordform	no	62.5	98.4	25.4	40.4
unigram	pos	no	49.6	25.0	0.4	0.8

obtained either with unweighted vectors or with *tf*idf*-weighted vectors, depending on which scheme performed better.

These results are quite encouraging. Most models outperform the random baseline, and several models are also outperforming the trivial acceptor. As we expected, given that positive instances are a minority in the training sets, precision is consistently much higher than recall. This confirms that trying to maximize recall is a sensible strategy for classifier combination.

The three best performing models are the one based on unigram wordforms, the one based on the unigram mixed representation, and the one based on the bigram mixed representation. Although, as mentioned in Section 3, special care was taken to select a corpus where translations and originals pertain to similar topics, and all proper nouns were recoded as generic strings to minimize topical effects, the success of the first of these models could still be due to uninteresting content-based cues (e.g. perhaps adjectives referring to Italian-specific topics and locations are more frequent in originals than in translations). However, the success of the other two models (which do not rely on content words at all) shows that translations are recognizable on purely grammatical/syntactic grounds. The fact that, for the unigram wordform model, *tf*idf* scores performed worse than non-transformed frequencies could be another cue to the central role played by function words, since the weight of the latter is reduced in *tf*idf* vectors.

With reference to the trigram models, we observe that performance increases in function of the abstractness of the representation (best performance with pos sequences, worst performance with wordforms). This is almost certainly due to the data-sparseness problems that plague less abstract trigram representations (the vast majority of trigram wordforms and lemmas occur once or very few times). Overall, the trigram models do not perform particularly well, suggesting that the less abstract information that can be exploited effectively by unigrams and bigrams is important to the task.

Finally, we note that the strikingly low performance of the unigram pos model is not surprising, since this model is using the relative frequency of 50 pos tags as its only cue.

6.2 Results with Combinations

Combinations based on majority voting performed disappointingly. The best combination in this class (unigram wordform, bigram lemma, bigram mixed, trigram tagged) achieved 77.5% accuracy, 98.5% precision, 55.8% recall, and 71.3% F; i.e. it was more or less in the range of the best single classifiers (unsurprisingly, with higher precision but lower recall).

On the other hand, all the recall maximizing combinations outperformed the best single measures, showing that the gain in recall is well worth the cost incurred in terms of increase of false positives. The *worst* performing recall maximizing combination (unigram wordform with *tf*idf*

Table 5 Results of best recall maximizing combinations

Id	Accuracy	Precision	Recall	F
21	86.7	89.3	83.3	86.2
23	86.5	89.2	82.9	86.0
19	86.4	89.2	82.9	86.0
22	86.0	89.9	81.2	85.3
15	85.6	88.0	82.5	85.2
3	85.6	90.1	80.0	84.8
13	85.2	87.9	81.7	84.7
24	85.4	89.3	80.4	84.6
17	85.0	87.5	81.7	84.5
20	85.2	89.3	80.0	84.4

weighting, bigram lemma, bigram mixed, trigram mixed) attained 81.7% accuracy, 91.3% precision, 70% recall, and 79.2% F; i.e. it outperformed the best single measures in all respects, except in terms of precision (while still achieving high precision). Table 5 reports the results for the 10 recall maximizing combinations with the highest F values, ranked by F and keyed by their Table 3 ids.

The results in Table 5 show that high quality translations have enough features in common to be identifiable with precision close to 90% and recall above 80%. They also show that combining SVMs with a recall maximizing scheme really pays off, at least in this particular task.

In general terms, the best combinations are those involving both SVMs trained on unigram lemmas and SVMs trained on unigram mixed representations. This may be partly because these are the largest combinations in our set (being composed of 5 models). In general, adding more models would appear to improve, or at least not hurt performance much. Indeed, a recall maximizing combination of *all* 24 models leads to 85% accuracy, 80.9% precision, 91.7% recall and 85.9% F (the fourth highest F value). However, the success of the best models in Table 5 is probably also due to the goodness of the mixed representation models.

Taking a closer look at the models composing the best combination (unigram lemmas with *tf*idf* weighting, unigram mixed representation with *tf*idf* weighting, bigram lemmas, bigram mixed representation lemmas, and trigram pos), we can distinguish between the unigram and bigram models based on lexical information (unigram and bigram lemmas) and those based on grammatical/syntactic information (unigram and bigram mixed representations). Interestingly, if we remove the two lexical models from the combination, performance drops less dramatically than if we remove the two non-lexical models. Without the lexical models, we obtain the following results: 86% accuracy, 90.2% precision, 80.2% recall, and 85.2% F. This is still among the best combinations (the fifth best combination in terms of F value). On the other hand, without the unigram and bigram mixed representation models, we obtain 83.7% accuracy, 92.2% precision, 73.75% recall, and 81.9% F. In terms of F, this combination is in the lower half of the overall combined model ranking. Again, this provides evidence that, while lexical cues help, they are by no means necessary, and translated text can be identified purely on the basis of function word distributions and shallow syntactic patterns.

6.3 Preliminary Feature Analysis

6.3.1 *Background*

Our results suggest that linguistically shallow representations contain cues that suffice to tell translations apart from originals with accuracy well above chance level (and, consequently, that translations and originals are linguistically different). Moreover, as we observed in the previous subsections, the nature of the best performing models indicates that function words and pos n-grams are playing a role that is at least as important as that played by lexical cues. However, our current data do not tell us anything about the specific features/units that characterize translated vs. original text, e.g. about whether, say, the distribution of pronouns is playing a role in translation discrimination.

Unlike with rule-based techniques, there is no straightforward way to interpret the models built by the SVM algorithm directly, from a qualitative/linguistic point of view. However, we can assess the role played by a certain linguistically interesting class by training a model without all features belonging to the relevant class, and by comparing the impact that this has on performance to the impact of removing random features that have similar frequency to the features of the target class.

We chose to study four classes of features in this way: (1) non-clitic personal pronouns, (2) adverbs,

(3) punctuation marks, and (4) non-finite verbal forms (including non-finite auxiliary forms). The overall rationale for picking these particular classes is that they are easy to identify and to remove from our document representations, and that their distribution (or that of closely related categories/properties) has been proposed as a possible translationese cue.

By blanking out non-clitic (1), or ‘strong’ personal pronouns, we aimed to test the hypothesis that subject pronouns in particular tend to be over-represented in translation into Italian. Clitics cannot function as subject pronouns in Italian, for which reason they are irrelevant to our hypothesis. Besides, they cannot be confused with strong pronouns, thus making it straightforward to tease the two classes apart (this would not be the case with strong object pronouns, which are not distinguishable from strong subject pronouns without further syntactic parsing). It has been suggested that this is an area in which interference from non-pro-drop languages is felt in Italian, a pro-drop language (Cardinaletti, 2004), resulting in a higher frequency of optional subject pronouns in translated than in original texts. This insight is confirmed by human perception of the translated/original distinction (see Section 7).

Similarly, by obscuring adverbs (2) and non-finite verbs (4) we aimed to test whether adverbs and non-finite constructions are typically over-represented in translated texts, and consequently signal translationese. The first hypothesis is based on Borin and Prütz’s and Hansen’s findings regarding English (Borin and Prütz, 2001; Hansen, 2003), and the second on work by Puurtinen on Finnish (Puurtinen, 2003).

Lastly, by removing (3) punctuation marks we aimed to hide (indirect) information about sentence length, in order to check whether, as hypothesized by Laviosa (1998) with reference to English, translated text, being ‘simpler’, contains shorter sentences.

6.3.2 *Method*

As a baseline for the feature removal experiments, we chose a combination of the weighted unigram mixed model and the unweighted bigram and trigram mixed models. We chose a combination of mixed representations since it offered a good trade-off between performance (in the top half of the recall maximizing combinations) and ease of experiment construction (the pos representations do not distinguish between clitic and non-clitic pronouns; the low frequency items in the lexical representations would have caused problems in maintaining comparable word frequency distributions in the target and control models).

From the three models in the baseline combination, we remove in turn each of the classes of features under study. For each class, we build five controls. The controls are constructed by removing from the three baseline models random features whose combined token frequency matches that of the target features (e.g. in the pronoun experiments we remove from each of the control models – unigram, bigram, and trigram – features whose overall frequency is close to that of the corresponding pronoun features). In all experiments and for all models, the difference between the overall token frequency of the targets and the average overall token frequency of the controls is below 0.04% of the former. Type frequency varies more: the maximum difference between a target model and the average of the corresponding controls is 4.2% in the pronoun experiments, 62.8% in the adverb experiments, 20.0% in the punctuation experiments, and 8.3% in the non-finite verb form experiments. However, the results reported below suggest that there is no systematic correlation between differences in type frequency and differences in performance.

We ran experiments with each of the target combinations and the corresponding 5 controls using the recall maximization scheme and the same 16-fold cross-validation setting described in Section 6.

6.3.3 *Results*

The results of the experiments (averaged across the 16 folds) are presented in Table 6. The results are expressed in terms of percentage difference from the baseline combination. For each control set, we report median performance followed by the minimum and maximum values in parentheses.

Table 6 Results after removing certain feature classes

Models	Accuracy	Precision	Recall	F
Baseline	85.0	91.2	77.5	83.8
No pron	−1.9	−0.4	−3.8	−2.4
Control	+0.2 (−0.2, +0.4)	+0.5 (−1.2, +1.1)	+0.4 (−0.4, +1.3)	+0.2 (−0.3, +0.4)
No adv	−2.9	−3.5	−2.9	−3.2
Control	−1.3 (−1.5, −0.4)	−0.8 (−1.5, +0.3)	−1.3 (−2.5, −1.3)	−1.4 (−1.8, −0.6)
No punc	−1.5	−0.8	−2.5	−1.8
Control	−1.7 (−3.3, −1.1)	−2.3 (−3.6, −1.1)	−0.8 (−5.0, −0.4)	−1.7 (−3.7, −1.0)
No inf	+0.6	+0.5	+0.8	+0.7
Control	−1.1 (−1.7, −0.2)	−1.1 (−3.1, +0.7)	−1.3 (−1.7, +0.4)	−1.2 (−1.6, −0.4)

Before we discuss the results for the target features, it is interesting to observe that, in the pronoun experiments, removal of frequency-matched random controls is actually leading to a small improvement in performance. This suggests that some amount of feature filtering, even if random, can help the performance of the algorithm. Probably not coincidentally, pronouns, and thus their controls, are the class with the lowest overall frequency. It is reasonable to hypothesize that random feature filtering has a positive effect only below a certain frequency mass threshold, a conjecture that should be supported by further investigation.

Turning now to the effect of removing target features, the removal of pronouns and of adverbs has a negative impact on performance that is systematically stronger than that of removing random features with the same overall frequency.

Non-clitic pronouns have a higher relative frequency in the translated texts (0.49% of the overall number of tokens vs. 0.35% in the original texts). This difference provides some evidence of overrepresentation of pronouns in translation. A more fine-grained analysis is needed to evaluate this and other hypotheses. The differences we observe might be directly related to under- or overrepresentation of certain features, but also to the different ways in which these features are used, along the lines of Mauranen (2002). A greater effect would perhaps emerge from an analysis which singles out subject pronouns from other personal pronouns. A quick glance at the rank list for personal pronouns in translated and original articles would seem to confirm this hunch. The wordform *egli*, an unambiguous subject pronoun (*he*), is the 11th most frequent personal pronoun in the translation corpus, with a frequency of 228, corresponding to a relative frequency of ~0.26 per 1000 words. The same wordform is 18th in rank in the original corpus, with 222 occurrences, corresponding to a lower relative frequency of ~0.11 per 1000 words.

Adverbs have a higher relative frequency in the original texts (5.02% vs. 4.66% in translations), contrary to what has been observed by Borin and Prütz (2001) and Hansen (2003) for English. Clearly, this may be due to a number of reasons, including language- or genre-specific preferences. Furthermore, adverbs are a mixed bag, with some exemplars falling squarely at the functional end of the continuum, and others which are fully lexical, or even creative coinages. It is not unlikely that conflating these adverbial sub-classes might hide more local regularities. Yet adverbial use does seem to be a promising area of further study for the identification of translationese features, since the SVM performance decreases substantially when adverbs are blanked out.

Punctuation removal does not seem to have a stronger effect than the one observed in the matching controls. In the same direction, we observe that the difference in relative frequency of punctuation is small (13.00% in original texts vs. 13.06% in translations). These results are rather difficult to interpret. On the one hand, they might suggest that punctuation use (and indirectly sentence length) are less relevant than other textual aspects to the identification of translationese. On the other, they might be taken to suggest that punctuation removal is too rough a way of getting at sentence length, and that other artifices are needed. Or they might simply hint at the fact that the SVMs are not making use of these features in the first place, which does not imply that they are not relevant to the original/translated distinction. Again, further study is needed in this area.

Lastly, the results for non-finite verbal forms are surprising, in that removing them from the models actually produces an improvement in performance. This seems to suggest that these forms are in fact misleading for the SVMs, perhaps cueing a distinction along textual dimensions independent from the original/translated divide. Notice that, again, there is only a small difference in the relative frequency of the relevant elements in the two sets of texts (5.12% in the originals vs. 5.10% in translations).

To conclude, the experiments described in this subsection provide some evidence that non-clitic pronouns and adverbs are playing a role in the detection of translated texts by SVMs. Experiments with punctuation and non-finite verbs have returned inconclusive results, but they have provided some confirmation that the methodology used is sound (i.e. it is not the case that the effects we see with pronouns and adverbs are simply due to the fact that removing a coherent class consistently causes a decrease in performance greater than the one we get when we remove random features in the controls). However, we are still far from a complete qualitative analysis of the significant features. On the one hand, we need to carry out an exhaustive investigation of all possibly relevant feature classes in order to assess their relative significance. On the other, the analysis should be carried out at a more detailed level, studying the impact not only of broad categories but also of specific elements (e.g. a certain pronoun or a certain punctuation mark). Still, we believe that the methodology introduced in this subsection (removing target features and comparing the effect of their removal to that of the removal of frequency-matched random controls) can lead to interesting discoveries.

7 Comparison with Human Performance

Having seen that translated texts contain patterns robust enough to be picked up by SVMs, one wonders whether humans would be able to perform the task equally well or better.

Table 7 Human subject performance

subject Id	Translator	Accuracy	Precision	Recall	F
1	n	93.3	93.3	93.3	93.3
2	y	90.0	80.0	100	88.9
3	y	86.7	86.7	86.7	86.7
4	y	83.3	80.0	85.7	82.8
5	n	80.0	73.3	84.6	78.5
6	y	76.7	80.0	75.0	77.4
7	n	76.7	73.3	78.6	75.9
8	y	66.7	73.3	64.7	68.7
9	n	66.7	46.7	77.8	58.4
10	n	63.3	40.0	75.0	52.2
Average	NA	78.3	72.7	82.1	76.3
SVMs	NA	86.7	86.7	86.7	86.7

To investigate this, we asked 10 subjects to identify translations among the 30 texts (15 translated, 15 original) in one of the 16 sections used for the nfold experiments. Of the 10 subjects, 5 were specialists with higher education degrees in translation; the other 5 had different educational backgrounds.

As a test set, we chose the section in which our 'best' combined model (combination 21 in Table 3) featured the performance level closest to its 16-fold average (86.7% for all performance measures, with 2 false positives and 2 false negatives).

The original versions of the texts, with proper nouns preserved, were handed out in electronic format. The subjects received the texts in different random orders, and they were encouraged to add comments on the cues they used for guessing. No time limit was set.

All subjects completed the task, reporting varying completion times, from about three hours to a whole day. Table 7 reports the results for all subjects (ranked by F value), their averages, and the results of the best combined SVM model on the same data subset.

Inspection of the table may suggest that translators in general performed better than non-translators. However, the differences in accuracy, precision, and recall between subjects in the two groups were far from statistically significant, as attested by a series of *t*- and Wilcoxon-tests.

Also, agreement rates within and across groups were similar. In the analysis to follow, we therefore ignore the translator/non-translator distinction.

The human data confirm that identifying translations in our corpus is not a trivial task: even the subject with the highest performance (interestingly, not a translator) made two mistakes, and most subjects performed a lot worse than that. At the same time, all subjects can identify translated text above chance level (although not always by much).

Comparing the humans to the best combined SVM model, we see that the SVMs are performing decidedly better than average. Only one subject performed better than the SVMs with respect to all measures, while another surpassed them in terms of accuracy, recall, and F. A third subject reached exactly the same performance level as the SVMs. The remaining 7 subjects performed worse than the SVMs with respect to all measures.

The average pairwise agreement rate among subjects is at 70.6%. The average pairwise agreement rate between each subject and the SVMs is at 74.4%. Thus, we have no reason to think that the SVMs behave in a radically different way from humans. In fact, on closer inspection, they seem to behave in rather similar ways. The composite human success rate on 3 of the 4 texts misclassified by the SVMs is below average, and the corresponding reports suggest that the decision was hard to make even for subjects who gave the right answer. In the remaining case of SVM misclassification, 9 out of 10 subjects correctly identified the text as a translation. Interestingly, this text explicitly refers to the author's nationality (Russian) from the very first line. However, 4 out of 5 translation experts pointed out that, were it not for this cue, the text would be difficult to categorize: linguistically, it is impeccable, fluent and idiomatic, 'a very good literary translation'. Which might explain why the SVMs, having no access to extra-textual evidence, were misled into thinking this text was an original.

We also have some preliminary evidence that humans and SVMs are sharing at least one translationese cue, i.e. the distribution of pronouns. We saw in Subsection 6.3 that (non-clitic) personal pronouns are among the categories whose removal has a significant impact on the performance of SVMs. At least one subject remarked that s/he believes repetition of optional subject pronouns to be a cue of translated text.

The analysis and comparison of cues used by SVMs and perceived by humans is an area in which more investigation is needed to arrive at any meaningful generalization.

8 Conclusions and Further Work

This study has introduced a new approach to the study of translationese, the 'dialect' of a language unconsciously adopted by translators. We have shown that the difference between high quality translations into Italian and comparable Italian originals can be 'learned' by a computer using support vector machines in a relatively knowledge-poor supervised learning setting.

The results of our experiments show that, while lexical cues are also playing a role, the SVMs perform the task by relying heavily on the distribution of n-grams of function words and morpho-syntactic categories. This confirms the hypothesis put forth by translation scholars that translations have their own peculiar lexicogrammatical/syntactic characteristics (Borin and Prütz, 2001; Hansen, 2003; Teich, 2003). A preliminary investigation of the relevant features at a finer-grained level suggests that non-clitic personal pronouns and adverbs are among the categories whose distribution is used by SVMs to detect translated/original Italian texts. These are hypothesized, therefore, to be among the distinguishing features of translationese. Further research is clearly necessary to find out whether this finding holds for other genres and translation settings, as well as for other languages.

We have also shown that the performance of SVMs is well above that of humans (including professional translators) performing the same task. Going beyond the domain of translation studies, this suggests that machine learning is reaching a stage in which it is no longer to be considered simply as a cheaper, faster alternative to human labour, but also as a heuristic tool that can help to discover patterns that may not be captured by humans alone.

In future work, we would like first of all to extend our experiments to other languages. In order to pursue this line of research, we will have to construct monolingual comparable corpora such as the one we built for Italian. This will not be a trivial task since, as discussed, it is extremely hard to find original and translated texts that do not also differ along other stylistic/textual/topical dimensions.

Another important line of research concerns the analysis of the features that are used by SVMs to detect translationese. The results we presented here in this respect are of a very preliminary nature.

We should then determine to what extent the features picked up by SVMs are the same that humans respond to when they perceive a text as a translation. This is an important question from a theoretical point of view, but also for any application of our method to translation quality assessment.

We would also like to experiment with different featural representations based on cues that have been suggested in previous research (in translation studies, stylometry and style/genre-based categorization), as well as on those reported by our subjects. These cues include word and sentence length, lexical richness measures, and collocational/colligational patterns.

Of course, other learning algorithms, as well as different ensemble methods could be tried out on the task. In particular, rule-based algorithms, while perhaps less effective than SVMs, may lead to results that are easier to interpret in qualitative terms. Conversely, the recall maximizing combination method could be applied to other tasks to test its general validity.

In the meantime, we believe that these initial results should be of interest to translation scholars, to the machine learning community and more, in general, to all those interested in computational/quantitative approaches to the study of genre and style.

Acknowledgements

We would like to thank our subjects for their patience and insights, two anonymous reviewers for helpful comments, Thorsten Joachims for the SVMLight package, and Adam Kilgarriff, who gave us the original idea for this study.

References

Argamon, S., Koppel, M., and Avneri, G. (1998). *Routing Documents According to Style*. Proceedings of IIIS-98, Pisa, Italy, June.

Baker, M. (1993). Corpus linguistics and translation studies: implications and applications. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds), Text and Technology. Amsterdam: Benjamins, pp. 223–50.

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7(2): 223–43.

Baroni, M. and Bernardini, S. (2003). *A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora*. Proceedings of Corpus Linguistics 2003, Lancaster, UK, March.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., and Aston, G. (2004). *Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-compliant Corpus of Newsarticle Italian*. Proceedings of LREC 2004, Lisbon, Portugal, May.

Bernardini, S. and Zanettin, F. (2004). When is a Universal not a Universal? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they exist?* Amsterdam: Benjamins, pp. 51–62.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Borin, L. and Prütz, K. (2001). Through a glass darkly: part of speech distribution in original and translated text. In Daelemans, W., Sima'an, K., Veenstra, J. and Zavrel, J. (eds), *Computational Linguistics in the Netherlands 2000*. Amsterdam: Rodopi, pp. 30–44.

Cardinaletti, A. (2004). *Language Contact in Translations: Attrition and Language Change*. Article presented at GLOW 2004, Thessaloniki, April 18–21, 2004.

Clough, P. (2000). *Plagiarism in Natural and Programming Languages: An Overview of Current Tools and Technologies*. Research Memoranda: CS-00-05. Department of Computer Science, University of Sheffield, UK. Online at: http://www.dcs.shef.ac.uk/∼cloughie/plagiarism/HTML_Version (accessed 16 June 2005).

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.

Finn, A. and Kushmerick, N. (2003). *Learning to Classify Documents According to Genre*. IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, August.

Frawley, W. (1984). Prolegomenon to a theory of translation. In Frawley, W. (ed.), *Translation: Literary, Linguistic and Philosophical Perspectives*. Newark: University of Delaware Press, pp. 159–75.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In Wollin, L. and Lindquist, H. (eds), *Translation Studies in Scandinavia*. Lund: CWK Gleerup, pp. 88–95.

Gellerstam, M. (1996). Translations as a source for cross-linguistic studies. In Aijmer, K., Altenberg, B. and Johansson, S. (eds), *Languages in Contrast*. Lund: Lund University Press, pp. 53–62.

Hansen, S. (2003). *The Nature of Translated Text*. Saarbrücken: Saarland University.

Joachims, T. (1997). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Technical report, Department of Computer Science, University of Dortmund.

Joachims, T. (1999). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C. and Smola, A. (eds), *Advances in Kernel Methods – Support Vector Learning*. Cambridge (MA): MIT Press.

Kindermann, J., Diederich, J., Leopold, E., and Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*, **19**(1–2): 109–23.

Koppel, M., Argamon, S., and Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, **17**(4): 401–12.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, **43**(4): 557–70.

Mauranen, A. (2002). Where is cultural adaptation? *Intralinea*. Online at: http://www.intralinea.it/specials/eng_open1.php?id=C0_46_32 (accessed 16 June 2005).

Mayfield Tomokiyo, L. and Jones, R. (2001). *You're not from round here, are you? Naive Bayes Detection of Non-native Utterance Text*. Proceedings of NAACL 2001, Pittsburgh, PA, June.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Reading (MA): Addison-Wesley.

Olohan, M. (2001). *Spelling out the Optionals in Translation: A Corpus Study*. Proceedings of Corpus Linguistics 2001, Lancaster, UK, March.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification Using Machine Learning Techniques*. Proceedings of EMNLP 2002, Philadelphia, PA, August.

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, **18**(4): 389–406.

Resnik, P. and Smith, N. (2003). The Web as a parallel corpus. *Computational Linguistics*, **29**(3): 349–80.

Santini, M. (2004). *State-of-the-art on Automatic Genre Identification*. Technical report, ITRI, University of Brighton.

Schmid, H. (1994). *Probabilistic Part-of-speech Tagging Using Decision Trees*. Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, July.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1): 1–47.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, **26**(4): 471–95.

Teich, E. (2003). *Cross-linguistic Variation in System and Text*. Berlin: Mouton de Gruyter.

Tirkkonen-Condit, S. (2002). Translationese – a myth or an empirical fact? *Target*, **14**(2): 207–20.

Tirkkonen-Condit, S. (2004). Unique items – over- or under-represented in translated language? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they Exist*? Amsterdam: Benjamins, pp. 177–84.

Toury, G. (2004). Probabilistic explanations in translation studies: welcome as they are, would they qualify as universals? In Mauranen, A. and Kujamäki, P. (eds), *Translation Universals. Do they Exist*? Amsterdam: Benjamins, pp. 15–32.