

A Shallow Approach To Syntactic Feature Extraction For Genre Classification

Marina Santini

ITRI

University of Brighton, UK

Marina.Santini@itri.brighton.ac.uk

Abstract

In this paper, the shallow and computationally inexpensive approach to syntax suggested by Argamon et al. (1998) is explored, enhanced and applied to ten different genres included in the BNC. Their approach to syntax uses POS trigrams. The rationale behind this choice is that trigrams are large enough to encode useful syntactic information, and small enough to be computationally manageable. The sets of experiments described in this paper show that features representing syntactic structure have strong discriminating power. Results are extremely encouraging and deserve further investigations.

1 Introduction

Automatic classification of documents is a challenging task and can be carried out according to different criteria: topic (text categorization, information filtering), author (authorship attribution), and genre (genre identification/detection/classification/categorization). Text categorization, authorship attribution and genre classification are similar in a way, because they all aim at determining a category, broadly speaking, for a document by using features. The difference is that features for text categorization are orthogonal to authorship attribution and genre identification features. Features for text categorization are the terms that occur in the document, therefore the classification is done according to topic (an exhaustive overview of several approaches and the specific problems connected to text categorization can be found in Sebastiani (2003)); features for authorship attribution and genre identification are topic independent. More specifically, authorship attribution seeks features that are invariant within the

documents of a given author but variant from author to author (McEnery and Oakes (2000), Baayen et al. (1996), Koppel et al. (2003a), Argamon et al. (2003b)); features for genre classification, instead, must capture similarities within texts of the same type, for example, many nominalizations and passives usually co-occur in scientific prose, while conversation abounds in 1st and 2nd person pronouns and contractions (Biber, 1988, 55). Interestingly, it has been noted that texts written by one author in different genres can differ more than texts written by different authors in the same genres (Baayen et al., 1996).

Another kind of document classification has emerged recently, namely author gender categorization, which uses mainly function words, n-grams of parts-of-speech and machine learning techniques (Argamon et al. (2003a), Koppel et al. (2003b)).

This paper focuses on genre classification and aims at determining the contribution of a particular class of features, namely those representing syntactic structure, for discriminating among different genres. As in any other classification processes, in genre classification an informed choice of features can make all the difference, and we claim that features representing syntactic structure are good discriminators.

1.1 Background

The idea that certain genres or writing styles favour certain syntactic constructions is not new (Biber (1988, 229-230), Baayen et al. (1996), Stamatatos et al. (2001b), etc.). However, even if syntax is acknowledged to have discriminating power, (though reluctantly sometimes (Aaronson, 1999)), it has often been neglected in genre categorization studies, because the extraction of syntactic features is considered to be

computationally expensive and time-consuming (Kessler et al., 1997). The 67 linguistic features selected by Biber more than 15 years ago (Biber, 1988, 73-75, 221-245) are based mainly on word identification, even when the features are really syntactic, because NLP tools were quite limited at that time. For example, the identification of adverbial clauses is based on the presence of specific subordinators, such as "although" and "though" for concessive clauses, and "because" for causative clauses. However, the lexically-based approach to syntax is quite limited, because subordinators can be ambiguous. To overcome the ambiguity issue, Biber used only unambiguous subordinators; for example "because" is the only causative subordinator included in his features, being the only one "to function unambiguously as a causative adverbial. Other forms, such as *as*, *for*, and *since*, can have a range of functions, including causative" (Biber, 1988, 236).

Nowadays, even though more sophisticated linguistic tools are available, syntactic information extraction is still troublesome. For example, parsers are becoming more reliable, but they often fail on long sentences and complex constructions, which are very common in certain genres, like academic prose or editorials. Owing to these limitations, we elude, for the time being, a deep syntactic analysis of texts and explore the adaptability and extensibility of the shallow approach to syntax suggested by Argamon et al. (1998).

1.2 Purpose and Rationale

In this paper, the shallow and computationally inexpensive approach to syntax suggested by Argamon et al. (1998) is explored, enhanced and applied to ten different genres included in the BNC. Their approach to syntax uses POS trigrams. The rationale behind this choice is that trigrams are large enough to encode useful syntactic information, and small enough to be computationally manageable. Grammatical n-grams (i.e. sequences of part-of-speech, or dependency, or functional tags) to measure syntactic differences among texts have already been used in stylometric studies, for example by Baayen et al. (1996), who use the term "pseudo-word sequence"; by Koppel et al. (2003b), who use the term "quasi-syntactic features"; by Argamon et al. (2003a), by Koppel and Schler

(2003), etc. Their results are promising. If a set of POS trigrams is particularly discriminating among a collection of genres, it could be a helpful addition to the features currently used in genre classification.

This paper is organized as follows: next section, describe the corpus and the features used in our sets of experiments; section 3 accounts for the methodology used in our approach; section 4 includes the setting up of the experiments and the classification results; in section 5 some evaluation measures are reported; section 6 contains a short discussion, the conclusions we have drawn and suggestions for future work.

2 Corpus and Features

In our experiment we used a subset of the BNC documents belonging to ten different genres. The genres selected¹ for our sets of experiments included four spoken genres (conversation, interview, public debate, planned speech) and six written genres (academic prose, advert, biography, instructional, popular lore, reportage)². Each genre collection contained 15 documents (see Biber (1993) for the number of documents needed to represent a stylistic category); each document was cut to 300 sentences (in order to have a manageable corpus), and the number of words per document ranged between 1500 and 7500. This variation in document length is a good assessment for the practical use of this classification model: if the model performs well with highly uneven document length, real world documents have a better chance of being classified correctly.

The BNC was tagged with the CLAWS5 tagset (Aston and Burnard, 1998, 230-234). Four different sets of features were tried. The first was a list of 835 POS trigrams, not including punctuation. The second was a list of 1033 POS trigrams, including punctuation. In both cases, we selected those POS trigrams with a frequency of occurrence between 30 and 100 in a

¹We used the *BNC Indexer* to select the genres involved in these experiments. The detailed description of this tool is in Lee (2003). The codes used by Lee for genre classification are included in the headers of the BNC documents under the label *classCode*.

²The official codes are: S_conv, S_interview, S_pub_debate, S_speech_scripted, W_ac_tech_engin, W_advert, W_biography, W_instructional, W_pop_lore, W_newsp_brdshst_nat_reportage.

single genre collection; after this first selection, we collected together all the POS trigrams coming from the ten genres and weeded out those trigrams with a frequency of occurrence greater than 3. The rationale behind this selection was to single out POS trigrams that were not too common and not too rare. The third set was a list of 65 features filtered out from the 835 POS trigrams without punctuation, and the fourth set was a list of 74 features filtered out from the 1033 POS trigrams with punctuation.

3 Methodology

3.1 Data Representation

In a typical supervised machine learning task, data is represented as a dataset of examples. Each example is described by a fixed number of measurements, and by a label that denotes the category, or class, that the example belongs to. This is a vectorial representation. The dimension of the vector depends on the number of measurements, or features, used to represent the documents. As our data representation is based on 835-, 1033-, 65- and 74 POS trigrams, we used respectively 835-, 1033-, 65- and 74-dimensional vectors.

3.2 Learning Method

The learning algorithm chosen for our experiments is a Naïve Bayes classifier. Preliminary investigations on a restricted subset of BNC genres showed that the Naïve Bayes classifier performed much better than the decision tree classifiers (Weka J48³ and See5⁴), and better than instance-based or memory-based classifiers (respectively Weka Ibk⁵ and TiMBL⁶).

Naïve Bayes models are based on two assumptions: attributes are equally important and statistically independent, given a class value. This method is "naïve" because it assumes independence, i.e. the value of an attribute gives no information about the value of another attribute. This means that evidence can be split into independent parts, i.e.:

$$Pr[H|E] = \frac{Pr[E_1|H][E_2|H] \dots [E_n|H]Pr[H]}{Pr[E]}$$

³Witten and Frank (1999, 269, passim).

⁴Demo version available at:

<http://www.rulequest.com/see5-info.html>.

⁵Witten and Frank (1999, 283, passim).

⁶Daelemans et al. (2002).

The assumption that attributes are independent, given the class, in real life is over-simple. In practice, the Naïve Bayes learners perform well in many document classification problems despite the incorrectness of this independence assumption (Domingos and Pazzani (1997) provide a detailed analysis of this phenomenon).

A common assumption, not intrinsic to the Naïve Bayesian approach but included in many actual implementations, is that the values of numeric attributes are normally distributed within a single class. When features are not normally distributed (linguistic data does not have a normal distribution), there are several remedies to overcome this problem. One solution is to discretize the data. Another solution is to use "kernel density estimation"⁷, which does not assume any particular distribution for attribute values. This is the solution we adopted in these sets of experiments.

Another drawback of the Naïve Bayes classifier is the so-called "zero-frequency problem". If an attribute value does not occur with every class value, its probability will be zero, and consequently the posteriori probability will also be zero. Probabilities that are zero override the other ones. The remedy is to add 1 to the count for every attribute⁸. This is a standard technique called the "Laplace estimator" (Witten and Frank, 1999, 85). With this addition, probabilities will never be zero and probability estimates are stabilized.

Despite the restrictions described above, but easily fixed with minor adjustments, Naïve Bayes works very well when tested on actual datasets, particularly when combined with attribute selection procedures. In fact, although this method deals well with random attributes, it has the potential (under certain conditions) to be disoriented when there are dependencies between attributes, and especially when redundant ones are added. This negative potential

⁷Kernel density estimation is a nonparametric technique for density estimation in which a known density function (the kernel) is averaged across the observed data points to create a smooth approximation. John and Langley (1995) showed that the Bayesian classifier's performance can be much improved if the traditional treatment of numeric attributes, which assumes normal (Gaussian) distributions, is replaced by kernel density estimation.

⁸In some cases adding a constant different from 1 might be more appropriate.

can be bypassed by selecting a subset of features. If the input features are selected prior to induction, then the feature selection algorithm is called "filter"; if the induction algorithm is bound to the process of searching, evaluating and selecting features, then it is called "wrapper" (Holmes and Nevill-Manning (1995), Hall and Smith (1997)). The feature selector used in our sets of experiments is a filter: in real datasets where there are a large number of features, filters seems the most appropriate choice (Hall and Smith, 1997)⁹.

A standard technique to ensure that results are representative and reliable is "cross-validation" (where a fixed number of folds can be decided) with "stratification" (random sampling of both training and test sets). In cross-validation, each fold, or partition of data, in turn is used for testing while the remainder is used for training. The standard evaluation technique in situations where only a small amount of data is used to build a classifier (as in our case) is stratified tenfold cross-validation. A single tenfold cross-validation might not be enough to get a reliable error estimate: it is a standard procedure to repeat the cross-validation process 10 times, and then average the results (Witten and Frank, 1999, 126-127). This is the solution we chose to get an accurate error estimates in our experiments.

4 Experiments

4.1 Setup

As mentioned above, the genres selected for the experiments from the BNC included four spoken genres and six written genres. Each dataset contained 150 records, where each record represented a document. A class (or genre) was assigned to each record, and each class accounted for 15 records. Four groupings of documents were tried: the ten genres altogether, spoken genres only, written genres only, spoken genres vs. written genres. For each grouping, four sets of features were tested: 835 features (full set of POS trigrams without punctuation); 1033 features (full set of POS trigrams with punctuation); 65 features (filtered

⁹One risk with feature selectors is that they can lead to over-fitting. Classifiers that overfit are very good at classifying data with which they have been trained, but they are bad with other data.

out from the 835 features); 74 features (filtered out from the 1033 features). The total number of datasets included in this first set of experiments was 16. The machine learning algorithm used was the Naïve Bayes classifier from the Weka package¹⁰ with the kernel density estimation option (-K). By default, this algorithm always adds 1 (Laplace estimator) to the number of different values for a particular attribute in order to bypass the zero-frequency problem. The filter used to create the selected features for the third and fourth dataset is included in the Weka package (the command *weka.filters.AttributeSelectionFilter* was launched with *BestFirst* as search technique, and *CfsSubsetEval* as evaluation class, which evaluates subsets of features by the correlation among them). To predict the accuracy of the learning technique on each single dataset, stratified tenfold cross-validation was used, repeated 10 times, with random seed values set from 1 to 10.

4.2 Classification results

In the tables below, column 1, "Genres", shows how genres have been grouped together (all genres, written genres only, etc.), column 2, "Features", tells us if the full set of features or selected (filtered) features have been used, column 3 and 4 refer to the presence or absence of punctuation in the sets of features.

Genres	Features	No Punct	Punct
All (10)	All	82.6%	81.1%
All (10)	Selected	87.0%	85.8%
Written (6)	All	78.9%	78.6%
Written (6)	Selected	88.9%	84.1%
Spoken (4)	All	94.6%	91.6%
Spoken (4)	Selected	87.8%	88.1%
Sp vs Wr	All	98.5%	99.3%
Sp vs Wr	Selected	98.4%	98.4%

Table 1: Average Accuracy for Trigrams

From Table 1, it turns out that the set of features without punctuation tends to perform better than the set of features with punctuation, with some exceptions. For Spoken → Selected, and Sp vs Wr → All, the average accuracy with punctuation is slightly higher (re-

¹⁰Weka is an open source machine learning software package, available at:

<http://www.cs.waikato.ac.nz/ml/weka/index.html>.

spectively 88.1% with punctuation, 87.8% without punctuation, and 99.3% with punctuation, 98.5% without punctuation). In one case, the average accuracy is the same with and without punctuation (Sp vs Wr \rightarrow Selected \rightarrow 98.4%).

The datasets with selected features tend to perform better than those with the full set of features, with the exception of Spoken genres (94.6% vs. 87.8% without punctuation, and 91.6% vs. 88.1% with punctuation) and for Spoken vs. Written genres (98.5% vs. 98.4% without punctuation, and 99.3% vs. 98.4% with punctuation).

The worst performance comes out from the full set of features for written genres, both with and without punctuation: when using the full set of features, three written genres (advert, popular lore and instructional) show higher error rates. Interestingly, when selected features are used, the classification results for these three genres are very good, and the overall accuracy reaches 88.9% without punctuation and 84.1 with punctuation.

As a whole, accuracy results of POS trigrams are extremely promising. But are POS trigrams the best combination to detect underlying syntactic structure? In order to have a deeper insight into this problem, we set up the same experiments using bigrams and unigrams.

For POS bigrams, we used the same set of BNC documents, and the same criteria to select the features (see the section Corpus and Features above). We ended up with 16 datasets, but their dimensionality was different from POS trigrams datasets. Data representation using POS bigrams was based on 451 bigrams without punctuation, 568 with punctuation, 36 selected features without punctuation, 41 selected features with punctuation. Classification results are shown in Table 2.

Genres	Features	No Punct	Punct
All (10)	All	77.6%	77.3%
All (10)	Selected	84.9%	85.4%
Written (6)	All	76.1%	70.5%
Written (6)	Selected	86.8%	86.8%
Spoken (4)	All	87.5%	89.1%
Spoken (4)	Selected	84.6%	84.6%
Sp vs Wr	All	98.0%	98.3%
Sp vs Wr	Selected	97.6%	99.0%

Table 2: Average Accuracy for Bigrams

From Table 2, we can see that the average accuracy results tend to be lower than the results achieved using POS trigrams, but not always. Notably, Sp vs Wr \rightarrow Selected \rightarrow Punct reaches 99% of average accuracy.

For unigrams, we applied the same criteria adopted for the other two sets of experiments, but we ended up with 8 datasets, because we used only the full set of features, being the number of these features quite low: 20 unigrams without punctuation, and 24 unigrams with punctuation. Average accuracy results are definitely lower than the ones achieved for POS trigrams.

Genres	Features	No Punct	Punct
All (10)	All	77.3%	78.5%
Written (6)	All	72.2%	74.4%
Spoken (4)	All	86.2%	85.8%
Sp vs Wr	All	88.3%	89.0%

Table 3: Average Accuracy for Unigrams

5 Evaluation measures

5.1 K statistic

The K statistic is usually used to measure within-group reliability (Carletta, 1996), but there are ways of applying it as an inter-group measure. We want to compare the best accuracy result achieved on POS trigrams by Argamon et al. (1998), i.e., 79.6% for pairwise distinguishability (Daily News vs. Newsweek), with our results. K is computed as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times that the model value is equal to the actual value and P(E) is the expected proportion by chance. When all the predictions are correct, K is 1; when the predictions are equal to which it would be expected by chance, K is 0. But it is possible to test whether or not K is significantly different from chance, therefore interpretation of the scale of accuracy is possible. A kappa statistic of 0.7 or higher is generally regarded as good statistic correlation, and the higher the value, the better the correlation¹¹.

¹¹Content analysis researchers generally think that $K \geq 0.8$ is a good reliability but $0.67 \leq K \leq 0.8$ allows tentative conclusions to be drawn (Carletta, 1996).

We computed K statistic for Argamon et al. (1998), and their K is around 0.6. The K statistic of our results is mostly around 0.8 and above for all the groupings of genres. Our K statistic values show high level of agreement, and consequently good reliability of the results achieved so far.

5.2 Percentage

Argamon et al. (1998)'s results are not directly comparable with our results, because the genres used in their experiment were different from ours. But it would have been interesting to see the performance they would have achieved on POS bigrams and POS unigrams on their collections. We can only say that their best result on POS trigrams is 79.6% for pairwise distinguishability (Daily News vs. Newsweek), that is 29.6% more than the random expectation of 50%. This is quite a good result, if we consider that their collections all belong to the journalistic style (could we dare to say "press genre"?) and the language/style variation is certainly less pronounced than, say, academic prose vs. conversation. With our wider range of differentiated genres, for the set of selected POS trigrams without punctuation, we get:

- For all genres, 77% more than the random expectation of 10% (average accuracy 87.0%, see Table 1 above).
- For pairwise distinguishability, 48% more the random expectation of 50% (average accuracy rate 98.4%, see Table 1 above).

Results are extremely encouraging and deserve further investigations.

5.3 Chi square

On the basis of percentages of a sample's behaviour, we can make claims about the sample itself, but we cannot generalize about the population from which we drew our sample, unless we submit our results to a test of statistical significance. To test the significance of our results, Chi-square has been computed using the worst classification results of the "All genres" grouping for: trigrams vs. bigrams, and bigrams vs. unigrams. Results suggest that there is a substantial advantage in using bigrams instead of unigrams, while there is only a slight benefit in using bigrams instead of trigrams.

6 Discussion, Conclusion and Future Work

As stated above, our experiment is based on that of Argamon et al. (1998) with some differences:

- Our experiment only focused on frequencies of POS n-grams in documents, whereas Argamon et al. (1998) used also a set of 500 function words.
- We used ten collections of different BNC genres: four spoken genres (interview, public debate, planned speech, conversation) and six written genres (advertisement, biography, popular magazines, reportage, instructional, academic prose). Each genre collection included 15 documents; each document was cut to 300 sentences, and the number of words per document ranged between 1500 and 7500. Instead, Argamon et al. (1998) used four text collections (NY Times news, NY Times editorial, NY Daily News, Newsweek). Each collection included 200 documents, and documents contained between 300 and 1300 words.
- In our experiment, four different sets of POS trigrams (CLAWS5 tagger) were tried. The first was a list of 835 POS trigrams, not including punctuation. The second was a list of 1033 POS trigrams, including punctuation. In both cases, we selected those POS trigrams with a frequency of occurrence between 30 and 100 in a single genre collection; after this first selection, we ruled out those trigrams with a frequency of occurrence greater than 3. The third set of features was a subset of 65 filtered out from the 835 POS trigrams, and the fourth set was a subset of 74 features filtered out from the 1033 POS trigrams with punctuation. Then we tried bigrams and unigrams. Argamon et al. (1998) merely used a list of 685 POS trigrams (Brill's tagger), including punctuation. They selected those trigrams that appear in between 25% and 75% of the documents in their corpus, and counted the frequencies of occurrence of each trigram.
- We used a Naïve Bayes classifier, together with "kernel density estimation", which

does not assume a normal distribution or any other particular distribution for the attribute values. Argamon et al. (1998) used a decision-tree learning algorithms (Ripper).

- In our experiment, the collections were tested in the following ways: ten genres altogether, spoken against written, only spoken, only written. In Argamon et al. (1998), the collections were tested for pairwise distinguishability, which means that two collections at time were compared.
- We used a 10-fold cross-validation repeated 10 times, in order to get the average accuracy. In Argamon et al. (1998), a single five-fold cross-validation was used.

The sets of experiments presented in this paper showed that features representing syntactic structure have strong discriminating power. A selected set of POS n-grams could easily (because it is computationally inexpensive) be included in the traditional bunch of features used for genre classification, for example, they could be added to the 55 "generic cues" suggested by Kessler et al. (1997), or to the classical 67 linguistic features suggested by Biber (1988, 73-75). It would be interesting to see if the accuracy of document classification would increase, or instead if some kind of noise would be created, thus confusing the statistical methodologies used for the classification task. In any case, the claim that structural features are computationally expensive and unstable does not hold any more. Taggers are quite accurate and parsers are becoming more reliable. The approach to text categorization in terms of stylistically homogeneous categories (text genres, authors, etc.) taking advantage of NLP tools (and therefore more sophisticated linguistic features, especially at syntax level) provides discriminating power without additional cost, and usually outperforms lexically based methods (Stamatatos et al. (2001a), Stamatatos et al. (2001b)).

It's hard to make comparisons with previous lexically based results because they all use different corpora and different classification methods. Future work is a busy agenda. The most urgent tasks include experiments using different sets of features on the same corpus (or corpora),

using the same statistical methodologies. The sets of features we have in mind are:

- Common Word Frequencies (cf. Stamatatos et al. (2000)).
- Word Class Frequencies (cf. Rayson et al. (2002)).
- Word n-grams. This is another important area that needs to be investigated comparatively with POS n-grams (some work has been done with character-level n-grams by Peng et al. (2003)).
- Further experiments to get deeper insight into the influence of punctuation in n-grams (for the time being we cannot see a clear-cut difference between results with or without punctuation).

If a coherent and consistent body of experiments could be set up, where all the items listed above could be measured on the same corpus (or corpora) and with the same classification methodologies, results could provide a benchmark for subsequent experiments on genre classification.

References

- S. Aaronson. 1999. *Stylometric Clustering. A comparative analysis of data-driven and syntactic features.* project report available at <http://www.cs.berkeley.edu/aaronson/sc/report.doc>.
- S. Argamon, M. Koppel, and G. Avneri. 1998. Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems (IIS-98)*.
- S. Argamon, M. Koppel, J. Fine, and A. Shimoni. 2003a. Gender, genre, and writing style in formal written texts. *Text*, 23(3).
- S. Argamon, M. Saric, and S. Stein. 2003b. Learning algorithms and features for multiple authorship discrimination. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.
- G. Aston and L. Burnard. 1998. *The BNC Handbook*. Edinburgh University Press, Edinburgh, UK.

- H. Baayen, H. Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11.
- D. Biber. 1988. *Variations across speech and writing*. Cambridge University Press, Cambridge, UK.
- D. Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8:1–15.
- J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.
- W. Daelemans, J. Zavrel, K. van der Sloot, and A. van der Bosh. 2002. *TiMBL: Tilburg Memory Based Learner, version 4.3. Reference Guide*. ILK Technical Report 02-10, available from <http://ilk.kub.nl/>.
- P. Domingos and M. Pazzani. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.
- M. Hall and L. Smith. 1997. Feature subset selection: a correlation based filter approach. In *Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858. Dunedin.
- G. Holmes and C. Nevill-Manning. 1995. Feature selection via the discovery of simple classification rules. In *Proceedings of the Symposium on Intelligent Data Analysis (IDA-95)*, pages 75–79, Baden-Baden, Germany.
- G. John and P. Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.
- B. Kessler, G. Numberg, and H. Shutze. 1997. Automatic detection of text genre. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.
- M. Koppel and J. Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.
- M. Koppel, A. Akiva, and I. Dagan. 2003a. A corpus-independent feature set for style based text categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.
- M. Koppel, S. Argamon, and A. Shimoni. 2003b. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4).
- D. Lee. 2003. Genres, registers, text types and styles: clarifying the concepts and navigating a path through the bnc jungle. *Language Learning and Technology*, 5(3).
- T. McEnery and M. Oakes, 2000. *Authorship Identification and Computational Stylometry*, pages 545–562. Marcel Dekker.
- F. Peng, D. Schuurmans, and S. Wang. 2003. Language and task independent text categorization with simple language models. In *Proceedings of HLT-NAACL*, pages 110–117, Berfield.
- P. Rayson, A. Wilson, and G. Leech. 2002. *Grammatical word class variation within the British National Corpus Sampler*. Rodopi, Amsterdam and New York.
- F. Sebastiani, 2003. *Text Categorization*, page Forthcoming. Available online at: <http://faure.iei.pi.cnr.it/fabrizio/Publications/Publications.html>. IOS Press.
- E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000. Text genre detection using common word frequency. In *Proceedings of the 18th Int. Conference Computational Linguistics (COLING2000)*, Saarbruecken, Germany.
- E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001a. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.
- E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001b. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214.
- I. Witten and E. Frank. 1999. *Data Mining Practical machine learning tools with Java implementations*. Morgan Kaufmann Publishers, San Francisco, California.