Squibs

Stable Classification of Text Genres

Philipp Petrenz*
University of Edinburgh

Bonnie Webber**
University of Edinburgh

Every text has at least one topic and at least one genre. Evidence for a text's topic and genre comes, in part, from its lexical and syntactic features—features used in both Automatic Topic Classification and Automatic Genre Classification (AGC). Because an ideal AGC system should be stable in the face of changes in topic distribution, we assess five previously published AGC methods with respect to both performance on the same topic–genre distribution on which they were trained and stability of that performance across changes in topic–genre distribution. Our experiments lead us to conclude that (1) stability in the face of changing topical distributions should be added to the evaluation critera for new approaches to AGC, and (2) part-of-speech features should be considered individually when developing a high-performing, stable AGC system for a particular, possibly changing corpus.

1. Introduction

This article concerns Automated Genre Classification (AGC). *Genre* has a range of definitions, but for Language Technology, a good one is a class of documents that share a communicative purpose (e.g., Kessler, Nunberg, and Schütze 1997). Although communicative purpose may be difficult to recognize without document understanding, researchers have found low-level features of texts to correlate with genre, making it a useful proxy.

AGC can directly benefit Information Retrieval (Freund, Clarke, and Toms 2006), where users may want documents that serve a particular communicative purpose (instructions, reviews, user guides, etc.). AGC can also benefit Language Technology indirectly, where differences in the low-level properties that correlate with genre may impact system performance. For example, if a part–of–speech (PoS) tagger or Statistical Machine Translation system trained on a corpus of *editorials* was then used for PoS tagging or translating a corpus of *letters to the editor*, it would benefit from the knowledge that inter alia the likelihood of the word "states" being a verb is considerably higher in *letters* (∼20%) than in *editorials* (∼2%).[1]

* University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK.
E-mail: p.petrenz@sms.ed.ac.uk.
** University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK.
E-mail: bonnie.webber@ed.ac.uk.

[1] This holds in the NYT Annotated Corpus, whether the topics are as different as *Health* and *Defense*.

Submission received: 23 July 2010; revised submission received: 15 October 2010; accepted for publication: 8 December 2010.

Genre differs from *topic*, which is what a text is about. Theoretically, a text from any given genre can be about any given topic (Finn and Kushmerick 2006), yet it is clear that co-variances exist between *genre* and *topic*, with some genre–topic combinations more likely than others (cf. fiction vs. news reports about dragons).

Because both genre classification and topic classification exploit low-level features of text as a basis for their predictions, a feature indicative of topic might benefit a genre classifier through correlations in the training corpus. However, if the topics addressed in different genres can change unpredictably over time, such correlated features can then harm performance. Although domain adaptation techniques might remedy this, they typically require extensive data in the target domain, and the remedy may fail as soon as the distribution changes again.

To date, the correlation between topic and genre has not been quantified, nor has the extent to which it may change in an actively growing corpus. In order to motivate research on *stability* in AGC,[2] we analyzed a large, publicly available newspaper corpus and found that (1) genres and topics do correlate substantially and (2) these correlations vary substantially over time.[3]

This squib makes two points: (1) Low-level features that correlate with topic can degrade the performance of AGC systems and are best removed unless the genre–topic distribution is guaranteed to be fixed, and (2) PoS features should not be lumped together in AGC because they have different correlations with genre and topic. Although the experiments used to make these points reflect an extreme situation—a complete change in genre–topic distribution—they do allow us to make these points convincingly.

2. Method

The data for our experiments come from the *New York Times Annotated Corpus* (*NYTAC*) (Sandhaus 2008), covering 21 years of publication (1987–2007) and more than 1.8 million articles.[4] Articles are richly annotated with meta-data, including fields whose values can be used to infer their genre and topic.

Genre: Two meta-data fields are related to the notion of genre as communicative purpose: *Types of Material* and *Taxonomic Classifier*. The former, appearing with 41.5% of articles, specifies the editorial category of an article. Usually the field has a single value, sometimes more than one. Although these values are not drawn from a fixed set, they can be used to infer genre after spelling errors are corrected (e.g., from *Reivew* to *Review*) and similar values are merged (e.g., *Editorial*, *editorial*, *Op-Ed*, and *Editors' Note*).

The values in the second genre-related field (*Taxonomic Classifier*) are drawn from a hierarchy, some of whose divisions indicate the section of the newspaper in which a document appears. In total, 99.5% of documents in the corpus contain a *Taxonomic Classifier* field, with an average of 4.5 values per article. Although the hierarchy varies in depth, its second level comprises a set of four fairly high level genres—that is, `Top/Classifieds`, `Top/Features`, `Top/News`, and `Top/Opinion`.

Because the *Types of Material* field did not include *news reports*, we used the *Taxonomic Classifier* field to recognize documents from this genre. Specifically, we considered

[2] The term **stability** is used in machine learning to describe the repeatability of a learner's results (cf. Turney 1995). Here, we use it to describe the robustness of a method to (topical) domain changes or changes in the topic–genre distribution.

[3] The results of our analysis can be found at `http://homepages.inf.ed.ac.uk/s0895822/SCTG/`.

[4] `www.ldc.upenn.edu, Catalog Id=LDC2008T19.`

Table 1
Genre (columns) and topic (rows) distribution in the data sets used in the experiments.

	Training set (3 × 4,309 = 12,927 texts)				Test set 1 (3 × 2,155 = 6,465 texts)				Test set 2 (6 × 2,285 = 13,710 texts)		
	News	Edit.	LttE		News	Edit.	LttE		News	Edit.	LttE
Edu	x			Edu	x			Edu		x	x
Def		x		Def		x		Def	x		x
Med			x	Med			x	Med	x	x	

any document with no *Types of Material* tag as a *news report*, if at least one of its *Taxonomic Classifier* values started with `Top/News`.

Topic: Topic descriptors were drawn from the *General Online Descriptors* meta-data field. The field appears with 79.7% of documents, with 3.3 descriptors per document on average.[5] Whereas *General Online Descriptors* are structured in a hierarchy, a document tagged with the more specific *United States Politics and Government* will also typically be tagged with the less-specific (i.e., closer to the root) value *Politics and Government*, but not vice versa.

Framework: Our experiments use *news reports*, *editorials*, and *letters* as target variables because similar classes have been used elsewhere in AGC research (e.g., Finn and Kushmerick 2006; Karlgren and Cutting 1994). For topics, we chose three that occur frequently and that were distinct from each other, in order to maximize differences in the formal cues used by classifiers. We based distinctiveness on the percentage overlap of topic tags in the corpus: Topics were taken to be distinct if (for our three chosen genres) fewer than 5% of texts that had one of the tags had another of them. The degree of overlap in the three topic tags we chose, ⟨Education and Schools⟩, ⟨Armament, Defense and Military Forces⟩, and ⟨Medicine and Health⟩, ranges from 0.5% to 2.9%. For comparison, the degree of overlap of the pair ⟨Politics and Government⟩ and ⟨International Relations⟩ ranges from 28.4% to 38.1% for our three genres. For all experiments, we only used texts which were unambiguously about Education, Medicine, or Defense. The small proportion of documents with more than one of these tags was ignored.

In order to examine the impact on system performance of a complete shift in topics, we varied the topical distribution of the test sets with respect to the training set. The training set consisted of 12,927 texts: *News reports* (News) about *Education* (Edu), *Editorials* (Edit.) about *Defense* (Def), and *Letters to the Editor* (LttE) about *Medicine* (Med). (This pairing yields more articles in the corpus than any of the five other possible combinations.) Table 1 shows the genre–topic distribution of both the training set and the two test sets used in these experiments. The first test set (6,465 articles) had the same distribution as the training set. In the second test set (13,710 articles), genre–topic pairings were inverted (see Table 1). All sets were balanced with an equal number of texts for each genre–topic combination (where not zero). The difference in the size of the test sets reflects the number of articles available with the desired topic–genre pairing. The training set and test set 1 were created by a random 2:1 split within each genre class. Because test set 2 comes from a different distribution than the training set, we report

[5] Genre-related differences in the number of *General Online Descriptors* are described further at `http://homepages.inf.ed.ac.uk/s0895822/SCTG/`.

results on these large holdout sets rather than performing cross-validation. We inferred confidence intervals by assuming that the number of misclassifications is approximately normally distributed with mean $\mu = e \times n$ and standard deviation $\sigma = \sqrt{\mu \times (1 - e)}$, where e is the percentage of misclassified instances and n is the size of the test set. We took two classification results to differ significantly only if their 95% confidence intervals (i.e., $\mu \pm 1.96 \times \sigma$) did not overlap.

3. Assessing Performance on Static and Altered Genre–Topic Distributions

To make our first point—that low-level features that correlate with topic can degrade the performance of AGC systems and are best removed unless the genre–topic distribution is fixed—we show how five published approaches to AGC perform in our experimental framework (Section 3). The choice of methods was partly motivated by the study by Finn and Kushmerick (2006), which compares bag-of-words, PoS frequencies, and text statistics as document representations in genre classification tasks across topical domains. The methods we assessed were chosen so that all these features were represented to different degrees. All were implemented on the same platform (Petrenz 2009).

KC: Karlgren and Cutting (1994) use a small set of textual features and discriminant analysis to predict genres. Most of these features involve either PoS frequencies or text statistics. Counts based on the fixed length of texts used in their experiments were adjusted to represent frequencies rather than absolute counts.

KNS/KNSPOS: Kessler, Nunberg, and Schütze (1997) predict genre based on surface cues. Because the paper gives few details about the specific features they use, we communicated with the authors directly. The list they gave us included features that require PoS tagging. As their published experiments do not make use of such features, we included two versions of their method, one version with PoS-based features and one without. [6]

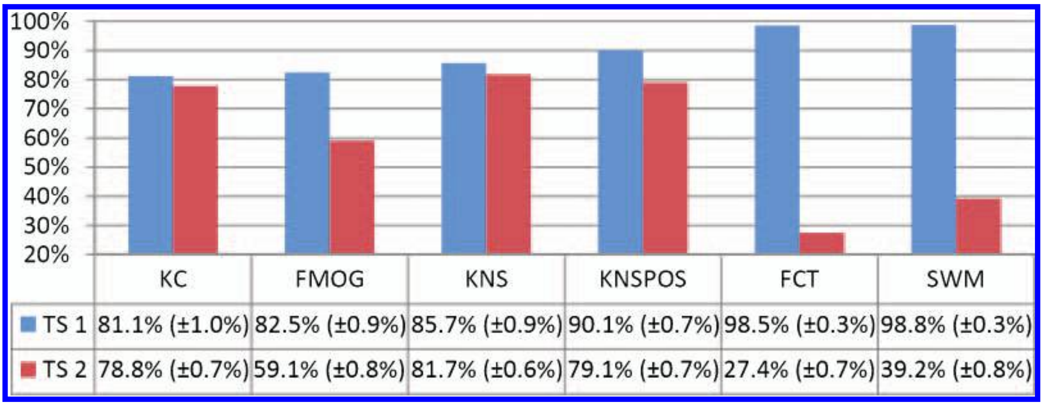
FCT: Freund, Clarke, and Toms (2006) predict genre using a support vector machine on a simple bag-of-words representation of a text. This feature set is not filtered using stop words or other techniques.

FMOG: Feldman et al. (2009) use part-of-speech histograms and principal component analysis to construct features. The authors classify genres using the QDA and Naive Bayes algorithms. We followed their decision to compute histograms on a sliding window of five PoS tags.

SWM: Sharoff, Wu, and Markert (2010) found a character *n*-gram based feature set to perform better than PoS *n*-grams and word *n*-grams in extensive AGC experiments using nine data collections and different choices of *n*. Although variable length *n*-grams as features for genre classification had previously been proposed by Kanaris and Stamatatos (2007), we followed Sharoff, Wu, and Markert in using fixed length 4-grams, which they found to yield higher accuracies.

Although a variety of Machine Learning (ML) methods were used in these approaches, here we just used the SVM implementation by Joachims (1999) because other ML methods produced similar, albeit poorer, results (Petrenz 2009). For PoS tagging, we used the Stanford maximum entropy tagger described in Toutanova et al. (2003).

[6] The features we used are listed at `http://homepages.inf.ed.ac.uk/s0895822/SCTG/`.



	KC	FMOG	KNS	KNSPOS	FCT	SWM
🟦 TS 1	81.1% (±1.0%)	82.5% (±0.9%)	85.7% (±0.9%)	90.1% (±0.7%)	98.5% (±0.3%)	98.8% (±0.3%)
🟥 TS 2	78.8% (±0.7%)	59.1% (±0.8%)	81.7% (±0.6%)	79.1% (±0.7%)	27.4% (±0.7%)	39.2% (±0.8%)

Figure 1
Classification accuracy of six different genre classification methods (cf. Section 3). Chart shows the percentage of correctly classified instances in both test sets (TS 1/2, cf. Table 1), with the confidence interval boundaries given in parentheses.

Figure 1 shows the results of training and testing on the same genre–topic distribution (test set 1). These results confirm the findings of SWM that binary character *n*-grams are good features in AGC. Both their approach and the bag-of-words approach used by FCT significantly outperform all other methods when the genre–topic distribution is the same for training and testing.

A different picture emerges from the second test set, however, whose genre–topic distributions differ from the training set. It shows that some feature sets owe their good results on test set 1 to the strong correlation between topics and genres. The performance of both SWM and FCT is significantly worse, with the latter even worse than the 33.3% that a random guess classifier would achieve in this balanced 3-class classification task. The performance drop for both KC and KNS is slight but still significant. Whereas KNSPOS had significantly outperformed KNS on test set 1, its performance is significantly worse than that of KNS on test set 2. (More on this shortly.)

Some of these results are not surprising: As bag-of-words and character *n*-grams reflect lexical differences of texts, systems that rely on them (SWM and FCT) will be misled by a major change in genre–topic distribution. Similar findings were reported in Finn and Kushmerick (2006) for bag-of-words and in (SWM) for character *n*-grams, although no explicit tests with topical distributions were carried out in the latter. More surprising are the results involving PoS tags: Two of the methods that used PoS tags—FMOG and KNSPOS—suffered when the genre–topic distribution changed, even though PoS frequencies had previously been reported to perform well as a feature set when used in new topical domains (Finn and Kushmerick 2006). However, the PoS frequencies in the KC feature set did not seem to harm stability much. This brings us to the second point of this squib.

4. Impact of PoS Features on Performance and Stability

We justify our second point—that PoS features should not be lumped together in AGC because they have different correlations with genre and topic—through a set of experiments that assess the effect of adding PoS features to a set of basic non-PoS

Table 2
Prediction accuracies for the baseline feature set and the same set with all 36 PoS tags added.

	Test set 1	Test set 2
13 surface features (Baseline)	72.9% (± 1.1%)	70.8% (± 0.8%)
13 surface + 36 PoS features	87.1% (± 0.8%)	72.5% (± 0.7%)

features similar to those used earlier by Karlgren and Cutting (1994), here normalized by document length (in words). These basic non-PoS features include character count per document, sentence count per document, average character count per sentence, average word count per sentence, average character count per word, type/token ratio, frequency of long words (ones with more than 6 letters), and the frequencies of the words *therefore*, *I*, *me*, *it*, *that*, and *which*.

The texts were PoS-tagged (Toutanova et al. 2003), using the same tag set as in the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993). All 36 non-punctuation tags were used, and counts of PoS-tags were normalized by document length.

Each experiment involved the 13 non-PoS features and a single PoS frequency feature (36 sets). Comparing performance with that on the non-PoS features alone (as a baseline) demonstrated the effect of adding each PoS frequency feature on classifier accuracy and stability.[7] All 36 feature sets as well as the baseline set were trained on the same training set. They were then tested on both test sets described in Section 2. As before, we use the same Support Vector Machines (SVM) classifier in all experiments, with the set of features as the experimental variable.

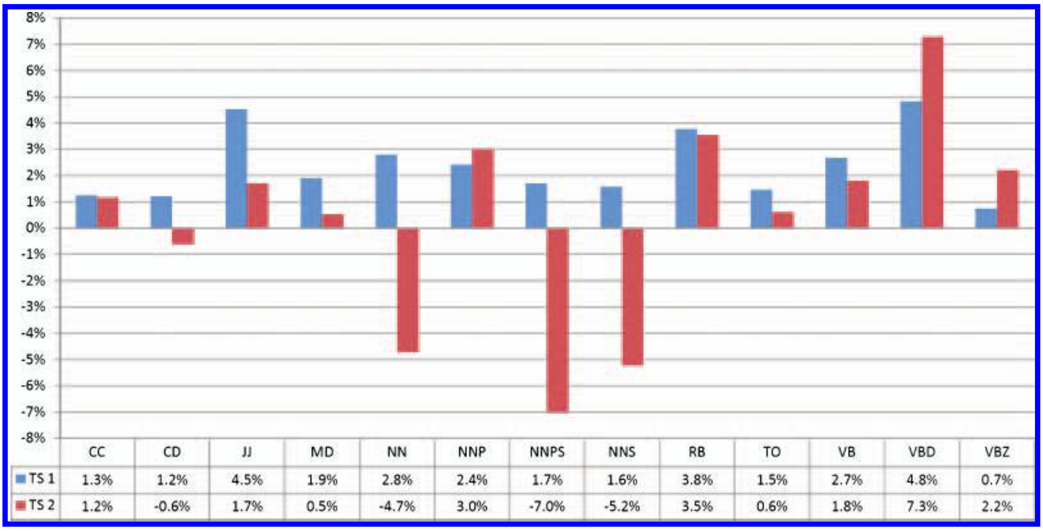
Table 2 shows both the accuracy of the baseline system on the two test sets as well as the accuracy of a system with the baseline features plus all 36 PoS features. Recall from Table 1 that the genre–topic distribution for test set 1 is the same as in the training set, whereas in test set 2 it is different. The first thing to note in Table 2 is that the classifier performs significantly better on the larger set of 49 features than on the smaller set of 13 basic features when the genre–topic distribution is not altered (column 2). When it is altered (column 3), the losses are less severe on our baseline than on the set that includes PoS features. This can be explained by looking at the contributions of each feature.

Figure 2 shows how accuracy changes when each PoS feature is added individually to the basic set of non-PoS features. To highlight the most interesting results, we only show features which cause a deviation of more than 1% from the baseline for at least one of the two test sets.[8] When the topic–genre distribution remains the same (i.e., test set 1), PoS frequencies appear to have a positive impact on prediction accuracy: When added to the basic feature set, accuracy increases. This is especially true for the tags VBD (past tense verb), JJ (adjective), RB (adverb), NN (singular noun), VB (base form verb), and NNP (plural proper noun).

The same is not true when the topic–genre distribution is changed (test set 2). Figure 2 shows that the PoS tags NN (singular noun), NNS (plural noun), and NNPS (plural proper noun) all have a toxic effect when added to the baseline set. This means the NN, NNS, and NNPS frequency features improve accuracy in stable conditions, whereas they severely harm it as topics change. This is not good for *stability*. A similar,

[7] The goal was not to select the best subset of PoS features—for that one would use a different method—but rather to show precisely how PoS features differ from each other with respect to AGC stability.

[8] The full results can be found at `http://homepages.inf.ed.ac.uk/s0895822/SCTG/`.



	CC	CD	JJ	MD	NN	NNP	NNPS	NNS	RB	TO	VB	VBD	VBZ
🟦 TS 1	1.3%	1.2%	4.5%	1.9%	2.8%	2.4%	1.7%	1.6%	3.8%	1.5%	2.7%	4.8%	0.7%
🟥 TS 2	1.2%	-0.6%	1.7%	0.5%	-4.7%	3.0%	-7.0%	-5.2%	3.5%	0.6%	1.8%	7.3%	2.2%

Figure 2
Deviation in percentage of correctly classified instances from the baseline feature set (cf. Table 2) for each added PoS tag frequency and test set (TS 1/2, cf. Table 1). CC = coordinating conjunction; CD = cardinal number; JJ = adjective; MD = modal; NN = singular noun; NNP = proper noun; NNPS = plural proper noun; NNS = plural noun; RB = adverb; TO = to; VB = base form verb; VBD = past tense verb; VBZ = third-person singular present verb.

if somewhat weaker effect, can be observed for the CD (cardinal number) frequency. Other PoS tags like CC (coordinating conjunction), RB (adverb), and VB (base form verb) increase predictive power while not impairing stability. Even more interesting are the results for the tags VBD (past tense verb) and VBZ (third-person singular present verb). Adding these features eliminates the significant difference between accuracies on test set 1 and 2, which we observed on the baseline feature set.

This makes sense if we consider how different genres and topics vary in their use of different parts of speech. In our data sets, for example, the frequency of plural noun (NNS) varies more by topic (on average, 8.0% of words are NNS in texts on *Education and Schools*, 7.6% in texts on *Medicine and Health*, and 6.3% in texts on *Defense and Military Forces*) than it does by genre (on average, 7.3% of words are NNS in news reports, 7.1% in letters, and 7.5% in editorials). The opposite holds for past tense verbs (VBD): Their average frequency varies less by topic (2.9%, 2.9%, and 3.3%) than by genre (4.8%, 1.8%, and 2.5%).

The odd result here is that singular proper noun (NNP) frequencies do not impair classifier stability: Unlike NN, NNS, and NNPS frequencies, they are topic-independent. This is because the most commonly tagged singular proper nouns in *news reports* are titles such as “Mr.”, “Ms.”, “Dr.”, etc., regardless of the topic. The frequency of titles among proper nouns is much lower in *editorials* and even lower in *letters*, across all three topical domains. NNP frequency in *news reports* is increased by the fact that titles are usually followed by one or more names, which are also singular proper nouns. We assume that this is the reason that the fluctuation between NNP frequencies is relatively low across topics (for the three topics and genres we examined) and hence a stable contributor to genre prediction.

Note that we are not making any claim about whether these specific results (e.g., that NN frequencies are bad for stability) hold for settings with different genres and topics. Rather, our point is that PoS tags should not be included or excluded wholesale for AGC. If one is going to the expense of PoS-tagging texts, only a subset of PoS tags should be used as features in AGC in order to maintain performance across changes in the topical distribution.

5. Conclusion

Our results suggest that prediction accuracy on a static topic distribution should not be the sole basis for assessing the quality of Automatic Genre Classification systems: In particular, approaches that perform well on a static topic distribution can be severely impacted by changes in topical distributions. The notion of topic independence for features has rarely been explored in the literature on genre classification. Nevertheless, this is an important issue, especially in dynamic environments like the World Wide Web, where new topics emerge rapidly and unpredictably. In topical domains with little or no labeled data to train on, instability can impede any useful application of classifiers. Because of this, we believe that *stability* should join *accuracy* as a criterion for assessing any new developments in genre classification. To this end, we introduced a cross-product methodology in Section 2 as a way of assessing stability.

Our results also suggest that, where the cost of PoS-tagging is acceptable, selective use of PoS-based features can yield high performance that is stable even when topical distribution differs from training to test sets. Although we have not identified a set of PoS-based features that supports classification among an arbitrary set of genres, we can say that it is crucial to evaluate and select PoS-based features carefully in order to achieve genre classification which is as topic-independent as possible.

Acknowledgments

We would like to thank Katja Markert, Mark Steedman, Maria Wolters, and four anonymous reviewers, each of whom gave us many valuable comments on earlier drafts of this article.

References

Feldman, S., M. A. Marin, M. Ostendorf, and M. R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784, Washington, DC.

Finn, Aidan and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

Freund, Luanne, Charles L. A. Clarke, and Elaine G. Toms. 2006. Towards genre classification for IR in the workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, pages 30–36, New York, NY.

Joachims, Thorsten. 1999. Making large-scale support vector machine learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, pages 169–184.

Kanaris, Ioannis and Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE International Conference on Tools with AI*, pages 3–10, Washington, DC.

Karlgren, Jussi and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Morristown, NJ.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, Morristown, NJ.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of

English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Petrenz, Philipp. 2009. Assessing approaches to genre classification. M.Sc. thesis, School of Informatics, University of Edinburgh.

Sandhaus, Evan. 2008. New York Times corpus: Corpus overview. LDC catalogue entry LDC2008T19.

Sharoff, Serge, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: Evaluating genre collections. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 3063–3070, Valletta.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL and Human Language Technology*, pages 173–180, Morristown, NJ.

Turney, Peter. 1995. Technical note: Bias and the quantification of stability. *Machine Learning*, 20(1–2):23–33.