Textual Characteristics of Different-sized Corpora

Robert Remus[†], **Mathias Bank**[‡]

†Natural Language Processing Group, University of Leipzig, Germany
‡Pattern Science AG, 63579 Freigericht, Germany
rremus@informatik.uni-leipzig.de, m.bank@cid.biz

Abstract

Recently, textual characteristics, i.e. certain language statistics, have been proposed to compare corpora originating from different genres and domains, to give guidance in language engineering processes and to estimate the transferability of natural language processing algorithms from one corpus to another. However, until now it is unclear how these textual characteristics behave for different-sized corpora. We monitor the behavior of 7 textual characteristics across 4 genres – news articles, Wikipedia articles, general web text and fora posts – and 10 corpus sizes, ranging from 100 to 3,000,000 sentences. Thereby we show, certain textual characteristics are almost constant across corpus sizes and thus might be used to reliably compare different-sized corpora, while others are highly corpus size-dependent and thus may only be used to compare similar- or same-sized corpora. Moreover we find, although textual characteristics vary from genre to genre, their behavior for increasing corpus size is quite similar.

Keywords: Textual Characteristics, Language Statistics, Corpus Comparison

1. Introduction

With the continuous development of natural language processing methods and machine learning algorithms, more and more approaches become available to assess various aspects of natural language text. Among these methods, many are highly text type-dependent and hence not easily transferable from one genre or domain to another, e.g. parsing (Sekine, 1997), word sense disambiguation (Escudero et al., 2000) and sentiment analysis (Aue and Gamon, 2005; Blitzer et al., 2007; Wang and Liu, 2011). Therefore, Bank et al. (2012) recently proposed to estimate the transferability of natural language processing methods from one genre or domain to another via the *textual characteristics* of the respective corpora. They found textual characteristics to vary greatly for different genres and pose the hypothesis, if textual characteristics of one corpus are similar to those of another, it is likely that algorithms working well on the former corpus also work well on the latter.

However, Bank et al. (2012) do not study the behavior of textual characteristics of *different-sized corpora*. Their analysis requires corpora of the same size in order to provide reliable results. As this requirement might not be applicable to real world scenarios, where one wants to compare different-sized corpora, we will address the following questions: Do textual characteristics vary not only across genres, but also across corpus sizes? If so, which textual characteristics are corpus size-dependent and which are not? Put differently, which textual characteristics may be used to compare both different-sized and same-sized corpora and which might only be used to compare similar- or same-sized corpora?

1.1. Related Work

To our knowledge, there has been only very little general work on comparing corpora based on their textual characteristics, and almost no work regarding potential corpus size-dependences of textual characteristics. Kilgarriff (2001) surveys several language statistics to measure corpus similarity and corpus homogeneity based on words and their distributions. Rayson and Garside (2000) propose to compare corpora using "frequency profiles" of words as well as syntactic and semantic tags.

With a specific goal in mind, several studies on textual characteristics have been carried out: Suzuki and Kageura (2007) explore Japanese prime ministers' Diet addresses, by focusing on the "quantity and diversity of nouns", to develop an understanding of changes in political content and the differences in 2 types of Diet addresses. Verspoor et al. (2009) investigate surface linguistic structures, sentence length distributions and term probability distributions in traditional and Open Access scientific journals to proof their similarity in order to ultimately be able to re-use previously proposed natural language processing algorithms. Na et al. (2010) analyze movie reviews from 4 online genres: critic reviews, user reviews, posts to discussion boards and blog posts. They analyze their vocabulary, average number of words, sentences and paragraphs, part of speech distributions, various movie aspects, as well as opinions expressed in the texts, partly automatically, partly manually. Goeuriot et al. (2011) analyze textual characteristics, e.g. posting lengths and part of speech distributions, of posts to 3 different drug review fora. Ghose and Ipeirotis (2011) measure amongst others readability and spelling accuracy of reviews to assess their helpfulness to other users and the reviews' economic impact. All studies mentioned above compare different-sized corpora, however without implicitly or explicitly addressing the potential difficulties these comparisons pose.

1.2. Outline

This paper is structured as follows: In the next Section, we describe the textual characteristics introduced in Bank et al. (2012). In Section 3. we apply them to corpora from different genres and monitor their behavior for different corpus sizes. Finally, we draw conclusions and point out possible directions for future work in Section 4.

2. Textual Characteristics

Bank et al. (2012) use only textual characteristics, i.e. language statistics, that can be easily and quickly calculated, without the need for advanced language processing modules, e.g. part of speech taggers or syntax parsers. This enables them to directly apply all measures to any corpus and ensures comparable results among them, without having to adapt those text type-dependent modules to previously unknown language properties. These textual characteristics are:

1. Shannon's *entropy* H measures the average amount of information in an underlying data structure. Applied in the field of language engineering, the mean amount of information of a token t_i can be calculated by approximating its probability $p(t_i)$ via its frequency in a given corpus. The entropy as given in Formula 1 is normalized to the vocabulary size $|V|$, i.e. the number of types in the corpus:

$$H = -\sum_{t_i \in V} p(t_i) \log_{|V|} p(t_i) \quad (1)$$

2. The *relative vocabulary size* R_{Voc} (Těšitelová, 1992, chapter 1.2.3.3) is given by the ratio of the vocabulary size $|V|$ and the total number of tokens N_m with respect to "meaningful" words. These are defined as words, that are not function words[1] ($N_m = \{t \mid t \notin N_f\}$), e.g. nouns, adjectives and verbs:

$$R_{\text{Voc}} = \frac{|V|}{N_m} \quad (2)$$

3. The *vocabulary concentration* C_{Voc}(Těšitelová, 1992, chapter 1.2.3.3) is defined by the ratio of the total number of tokens N_{top} with respect to the most frequent terms in the vocabulary V ($V_{top} = \{t \mid t \in V \wedge r(t) \leq 10\}$) and the total number of tokens N in a corpus

$$C_{\text{Voc}} = \frac{N_{top}}{N} \quad (3)$$

where rank $r(t)$ is defined as the position of a token t in a frequency-ordered list.

4. The *vocabulary dispersion* D_{Voc} expresses the relative amount of low frequency tokens ($V_{\text{low}} = \{t \mid t \in V \wedge f(t) \leq 10\}$) in the vocabulary V:

$$D_{\text{Voc}} = \frac{|V_{\text{low}}|}{|V|} \quad (4)$$

where frequency $f(t)$ is defined as the number of occurrences of the token t in a corpus.

5. The *corpus predictability* CP expresses the transition probabilities between tokens. For this, we need to calculate the entropy of a first-order Markov source $\mathcal{S}$ of two tokens t_i, t_j as given in Formula 5

$$H(\mathcal{S}) = -\sum_{t_i} p(t_i) \sum_{t_j} p_{t_i}(t_j) \log p_{t_i}(t_j) \quad (5)$$

where $p_{t_i}(t_j)$ denotes the probability of t_j given that it is preceded by t_i. CP is then calculated by normalizing the entropy of a first-order Markov source by its maximum possible entropy and subtracting it from 1:

$$CP = 1 - \frac{H(\mathcal{S})}{H_{max}(\mathcal{S})} \quad (6)$$

6. A rudimentary *grammatical complexity* GC can be calculated by the ratio of the number of function words N_f to the number of meaningful words N_m:

$$GC = \frac{N_f}{N_m} \quad (7)$$

Although this rather basic approach cannot state a real level of grammatical structure of a corpus, it still provides evidence for the amount of effort put into expressing syntax.

7. The *average sentence length* L_S influences parsing, relation extraction etc. The length $|s|$ of a sentence s is defined by the amount of tokens it contains, and the average sentence length of all sentences S is defined as in Formula 8:

$$L_S = \frac{1}{|S|} \sum_{s \in S} |s| \quad (8)$$

Additionally, Bank et al. (2012) measure *spelling accuracy* and *information density*. As both textual characteristics require manual intervention, we only compute the 7 measures described above.

3. Experiments

We now construct different-sized corpora and apply the textual characteristics described in Section 2. to them.

3.1. Constructing Different-sized Corpora

For our experiments we use 3 large English-language corpora provided by the *Wortschatz* project[2] (Quasthoff et al., 2006), each originating from a different genre: news articles, Wikipedia articles and general web text. To ensure comparability, all Wortschatz corpora are built in a standardized fashion (Quasthoff and Eckart, 2009). Their intended use is statistical corpus and language comparison (Eckart and Quasthoff, 2010). As an additional genre, we use a corpus of posts to the automotive web forum `benzworld.org`. Due to copyright reasons, this corpus is not publicly available.

To study the behavior of textual characteristics for different corpus sizes we construct sub-corpora C_k^g containing $k \in \{100, 300, 1000, 3000, \ldots, 3000000\}$ sentences for each genre $g \in \{\text{news}, \text{wikipedia}, \text{web}, \text{fora\_posts}\}$ so that

$$\forall l < m : C_l^g \subset C_m^g$$

i.e. any smaller corpus is always a real subset of any larger corpus. Table 1 provides an overview of the resulting news article, Wikipedia article, web text and fora post corpora.

[1] As function words N_f Bank et al. (2012) defined: *the, a, an, he, him, she, her, they, us, we, them, it, his, to, on, above, below, before, from, in, for, after, of, with, at, and, or, but, nor, yet, so either, neither, both, whether*

[2] `http://wortschatz.uni-leipzig.de/`

3.2. Results

Applying the textual characteristics described in Section 2. to these corpora leads to the results presented Figure 1(a), 1(b), 1(c) and 1(d). Interestingly, although the actual textual characteristics vary from genre to genre as expected (cf. also Table 1) and as it has been shown before (Bank et al., 2012), their behavior for different-sized corpora is very similar across all 4 genres: Not surprisingly, vocabulary concentration C_{Voc}, grammatical complexity GC and average sentence length L_S are almost "constant" for sufficiently large corpora, i.e. $k > 1000$. We note however, the larger the corpus, i.e. the larger k,

1. the lower its entropy H,
2. the lower its relative vocabulary size R_{Voc},
3. the lower its vocabulary dispersion D_{Voc} and
4. the higher its corpus predictability CP.

Across the 4 genres all pairwise *correlations* of H, R_{Voc}, D_{Voc} and CP are greater than 0.99 (significant at level $\alpha = 0.001$). Although this behavior may need further clarification in more experiments, it seems to signify invariant language properties, irrespective of the considered genres.

3.3. Discussion

The intuition behind the observed behavior of entropy, relative vocabulary size, vocabulary dispersion and corpus predictability is as follows: The entropy H is known to be dependent on the "message" length N (Manning and Schütze, 1999). The longer the message, i.e. the larger the corpus, the more redundant information it contains and hence the entropy decreases. The relative vocabulary size R_{Voc} decreases with increasing corpus size as the growth rate of "meaningful" tokens N_m is linear to the corpus size, whereas the growth rate of vocabulary size $|V|$ drops off for larger and larger corpora. As a result, the relative vocabulary size is almost zero for very large corpora. The vocabulary dispersion D_{Voc} also decreases with increasing corpus size, but with a lower rate than R_{Voc}. Its functional form is almost "s-shaped". This may be because the growth rate of low frequency terms $|V_{\text{low}}|$, e.g. spelling errors, is typically smaller then the growth of vocabulary size $|V|$. However, $|V|$'s growth rate drops off for larger and larger corpora and thus vocabulary dispersion decreases non-linearly. As text in a corpus typically follows language-internal rules, e.g. a grammar, and the vocabulary size $|V|$'s growth rate is smaller than the number of tokens N's growth rate, the number of possible term combinations is limited. Consequently, corpus predictability CP increases with increasing corpus size.

Coming back to our initial questions, we conclude: Vocabulary concentration, grammatical complexity and average sentence length are not corpus size-dependent given a sufficiently large corpus, i.e. more than 1000 sentences in our case. They may reliably be used to compare both same- and different-sized corpora. In contrast, entropy, relative vocabulary size, vocabulary dispersion and corpus predictability are corpus size-dependent and thus may *not* be reliably used to compare different-sized corpora.

To still compare different-sized corpora based on entropy, relative vocabulary size, vocabulary dispersion and corpus predictability, we suggest to *(under)sample* corpora to a common size and then apply the aforementioned textual characteristics. Alternatively, vocabulary dispersion may be used cautiously for corpora of a similar size and instead of entropy H we might calculate the *entropy rate* H_{rate} as shown in Formula 9:

$$H_{\text{rate}} = -\frac{1}{N} \sum_{t_i \in V} p(t_i) \log_{|V|} p(t_i) \tag{9}$$

Additionally to Formula 1's normalization to $|V|$, Formula 9 is also normalized to the number of tokens N and converges for $N \to \infty$ (Manning and Schütze, 1999).

4. Conclusions & Future Work

We studied the behavior of 7 textual characteristics for different-sized corpora in 4 genres. Although the actual textual characteristics vary from genre to genre as expected, we have shown their behavior for different-sized corpora is very similar across all 4 genres. We observed vocabulary concentration, grammatical complexity and average sentence length are not corpus size-dependent, whereas entropy, relative vocabulary size, vocabulary dispersion and corpus predictability are. Therefore, we suggest the former may reliably used to compare both same- and different-sized corpora and the latter may only be used to compare same- or similar-sized corpora.

Future research avenues include exploring the possibilities of fitting appropriate functions to the textual characteristics curves in order to interpolate between different-sized corpora and thereby avoid sampling. Additionally, we like to extend our study to more genres, e.g. novels, scientific essays, tweets and blog posts.

5. References

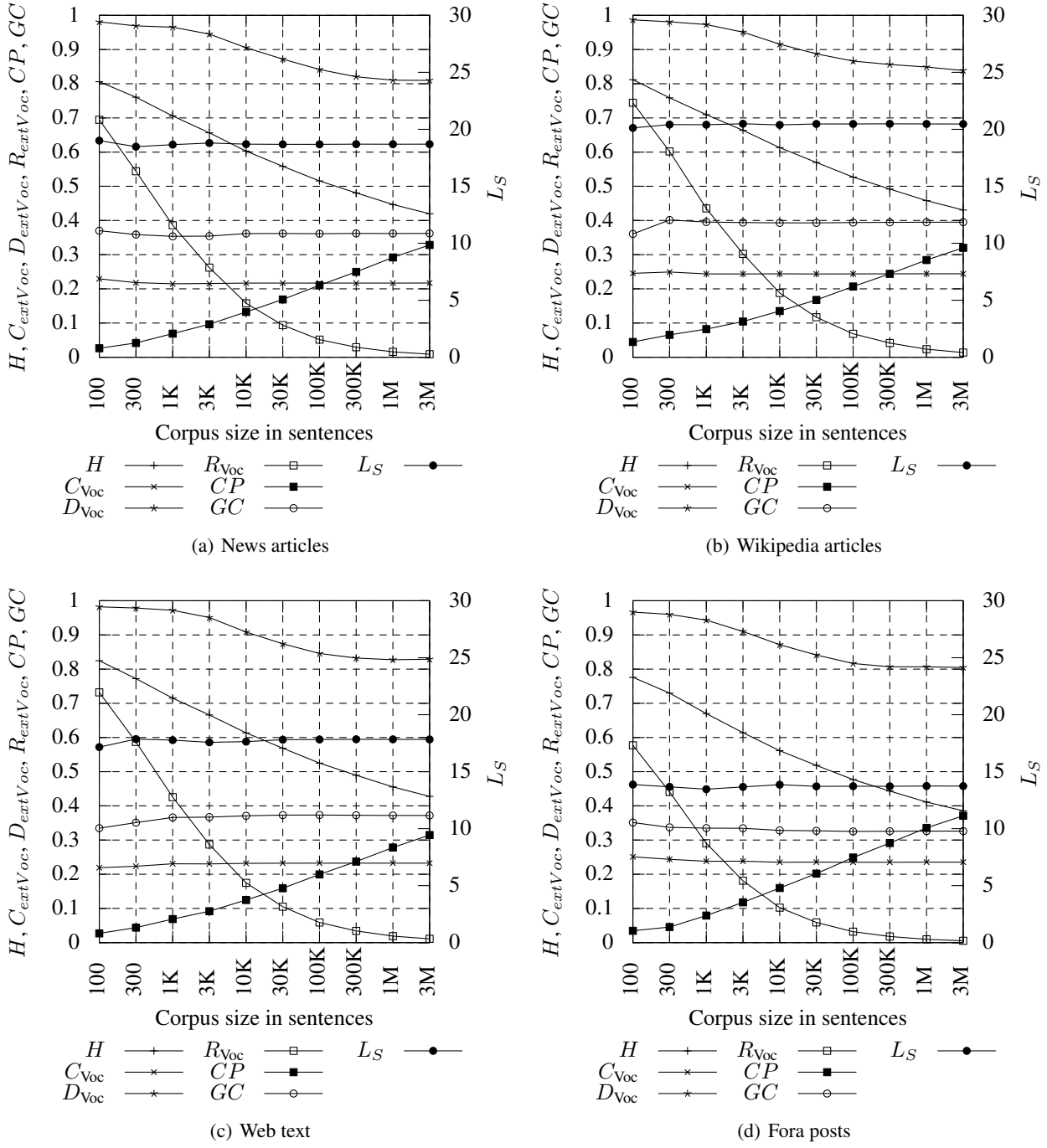
A. Aue and M. Gamon. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

M. Bank, R. Remus, and M. Schierle. 2012. Textual Characteristics for Language Engineering. In *8th International Conference on Language Resources and Evaluation (LREC'12), to appear*.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447.

T. Eckart and U. Quasthoff. 2010. Statistical Corpus and Language Comparison using Comparable Corpora. In *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 15–20.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 172–180.



(a) News articles

(b) Wikipedia articles

(c) Web text

(d) Fora posts

Figure 1: Behavior of textual characteristics of English-language corpora of increasing size.

A. Ghose and P. Ipeirotis. 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23:1498–1512.

L. Goeuriot, J.C. Na, W. Y. Min Kyaing, S. Foo, C. Khoo, Y.-L. Theng, and Y.K. Chang. 2011. Textual and Informational Characteristics of Health-related Social Media Content: A Study of Drug Review Forums. In *Proceedings of Asia-Pacific Conference on Library & Information Education & Practice (A-LIEP)*.

A. Kilgarriff. 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

C.D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

J.C. Na, T.T. Thet, and C. Khoo. 2010. Comparing Sentiment Expression in Movie Reviews from four online Genres. *Online Information Review*, 34(2):317–338.

Uwe Quasthoff and Thomas Eckart. 2009. Corpus Building Process of the Project "Deutscher Wortschatz". In *Proeedings of the Workshop on Lingustic Processing Pipelines*.

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceed-*

Corpus	Tokens	Types	H	R_{Voc}	C_{Voc}	D_{Voc}	CP	GC	L_S
news_100	1,866	946	0.8060	0.6946	0.2294	0.9810	0.0264	0.3700	19.01
news_300	5,428	2,173	0.7604	0.5439	0.2181	0.9692	0.0417	0.3587	18.48
news_1K	18,271	5,212	0.7058	0.3861	0.2149	0.9653	0.0691	0.3535	18.65
news_3K	55,178	10,676	0.6564	0.2620	0.2156	0.9452	0.0965	0.3543	18.80
news_10K	182,943	21,196	0.6028	0.1578	0.2170	0.9054	0.1324	0.3617	18.70
news_30K	548,395	37,806	0.5584	0.0939	0.2169	0.8720	0.1689	0.3619	18.68
news_100K	1,828,304	69,172	0.5157	0.0515	0.2168	0.8417	0.2106	0.3613	18.68
news_300K	5,491,076	117,964	0.4806	0.0293	0.2169	0.8203	0.2497	0.3622	18.70
news_1M	18,296,680	211,254	0.4468	0.0157	0.2170	0.8103	0.2919	0.3619	18.69
news_3M	54,903,309	357,955	0.4196	0.0089	0.2170	0.8098	0.3283	0.3620	18.70
wikipedia_100	1,947	1,065	0.8117	0.7442	0.2455	0.9869	0.0446	0.3606	20.12
wikipedia_300	5,943	2,553	0.7588	0.6020	0.2494	0.9812	0.0653	0.4013	20.42
wikipedia_1K	19,814	6,182	0.7096	0.4354	0.2438	0.9731	0.0823	0.3954	20.40
wikipedia_3K	59,699	12,944	0.6636	0.3022	0.2435	0.9509	0.1044	0.3940	20.49
wikipedia_10K	198,069	26,724	0.6130	0.1880	0.2440	0.9157	0.1353	0.3931	20.39
wikipedia_30K	597,049	50,208	0.5697	0.1171	0.2435	0.8873	0.1677	0.3930	20.40
wikipedia_100K	1,990,411	97,721	0.5269	0.0684	0.2437	0.8663	0.2068	0.3940	20.47
wikipedia_300K	5,975,787	177,640	0.4919	0.0415	0.2439	0.8563	0.2439	0.3945	20.48
wikipedia_1M	19,910,567	335,409	0.4580	0.0235	0.2441	0.8488	0.2844	0.3949	20.47
wikipedia_3M	59,719,241	588,673	0.4306	0.0138	0.2441	0.8386	0.3200	0.3950	20.47
web_100	1,643	901	0.8246	0.7319	0.2191	0.9822	0.0270	0.3347	17.16
web_300	5,192	2,256	0.7722	0.5870	0.2234	0.9787	0.0438	0.3510	17.86
web_1K	17,286	5,386	0.7152	0.4257	0.2305	0.9720	0.0689	0.3663	17.77
web_3K	51,253	10,754	0.6658	0.2868	0.2302	0.9509	0.0918	0.3671	17.57
web_10K	171,432	21,867	0.6136	0.1749	0.2320	0.9084	0.1247	0.3710	17.64
web_30K	519,253	39,720	0.5686	0.1050	0.2323	0.8741	0.1591	0.3730	17.81
web_100K	1,732,458	74,203	0.5251	0.0588	0.2325	0.8457	0.1994	0.3731	17.83
web_300K	5,204,182	129,709	0.4896	0.0342	0.2325	0.8324	0.2370	0.3726	17.84
web_1M	17,343,098	240,133	0.4554	0.0190	0.2324	0.8274	0.2783	0.3719	17.84
web_3M	52,014,020	421,318	0.4277	0.0111	0.2324	0.8288	0.3147	0.3720	17.83
fora_posts_100	1,339	572	0.7761	0.5772	0.2509	0.9668	0.0346	0.3512	13.87
fora_posts_300	3,958	1,305	0.7299	0.4409	0.2438	0.9602	0.0457	0.3372	13.67
fora_posts_1K	13,035	2,838	0.6699	0.2906	0.2380	0.9433	0.0791	0.3349	13.46
fora_posts_3K	39,714	5,383	0.6140	0.1809	0.2382	0.9099	0.1179	0.3346	13.66
fora_posts_10K	134,078	10,314	0.5614	0.1022	0.2351	0.8714	0.1596	0.3282	13.86
fora_posts_30K	397,647	17,729	0.5185	0.0592	0.2356	0.8412	0.2019	0.3271	13.71
fora_posts_100K	1,327,165	32,014	0.4769	0.0320	0.2350	0.8165	0.2489	0.3252	13.73
fora_posts_300K	3,979,848	53,994	0.4433	0.0180	0.2351	0.8068	0.2914	0.3258	13.72
fora_posts_1M	13,290,428	96,495	0.4112	0.0096	0.2354	0.8065	0.3355	0.3261	13.74
fora_posts_3M	39,851,933	160,608	0.3856	0.0053	0.2354	0.8050	0.3710	0.3262	13.74

Table 1: Textual characteristics of English-language corpora of increasing size.

ings of the 5th International Conference on Language Resources and Evaluation (LREC'06), pages 1799–1802.

P. Rayson and R. Garside. 2000. Comparing Corpora using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6.

S. Sekine. 1997. The Domain Dependence of Parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, pages 96–102.

T. Suzuki and K. Kageura. 2007. Exploring the Microscopic Textual Characteristics of Japanese Prime Ministers' Diet Addresses by Measuring the Quantity and Diversity of Nouns. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 459–470.

Marie Těšitelová. 1992. *Quantitive Linguistics*. John Benjamins Publishing Company.

K. Verspoor, K.B. Cohen, and L. Hunter. 2009. The Textual Characteristics of Traditional and Open Access Scientific Journals are Similar. *BMC bioinformatics*, 10(1):183.

D. Wang and Y. Liu. 2011. A Cross-corpus Study of Unsupervised Subjectivity Identification based on Calibrated EM. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 161–167.