Vocabulary size and email authentication

Robert Warren
School of Computer Science, University of Waterloo
rhwarren@uwaterloo.ca

Technical Report CS-2005-17
May 12, 2005

Abstract

This paper explores the performance of the method proposed by Efron and Thisted to predict vocabulary sizes based on sampled text. The primary objective of this research is to determine whether this simple and quick test can be used as a coarse indicator of authorship. Three sets of emails, as well as other texts are analyzed in order to collect performance data. The conclusion is that the test is at best a lower bound indicator within the T$<$1.0 region and is not sufficient as an authentication method.

1 Introduction

Email tends to be an impersonal medium which is difficult to authenticate: even with proper transmission safe-guards it is simple for someone to mis-represent themselves. While most security measures concentrate on system-level security, such as passwords, another possibility is to attempt to authenticate the message based on its contents alone. For the purposes of this paper we will use the term authorship attribution and authentication interchangeably.

The work of Efron and Thisted [ET76] in 1976 was originally designed to extrapolate an author's full vocabulary count based on a sample of his writings. By comparing the actual size of the vocabulary within a new communication with the size predicted from previous ones, we could achieve a crude method for authorship verification. Such an approach was undertaken by [ET87] for verifying the authorship of an anonymous poem attribute to Shakespeare.

2 Previous work

Author authentication and text analysis is a mature field and a wider view of the possible research areas can be obtained in [Yul68] and [MW64]. The original Efron-Thisted paper [ET76] was written using Shakespeare text data, the model [ET87] providing three different tests: a prediction of the length of the text, of the number of unseen words and the pattern of rare words use within the new text. The method remained untested until 1987 when an attempt was made to authenticate an anonymous poem attributed to Shakespeare. While the predicted values matched those of the poem, the current consensus is that the poem, while written in the style of Shakespeare, is written by another person. The method was theoretically tested by [Val91] in 1991 using a randomized word generator to benchmark the methods according to their assumptions. Later, [EV96] also used the methods as part of massive effort to analyze patterns within the Shakespeare corpus. Generally, it is recognized that the method predicts the length of text poorly while the predictions on the use of rare words is the most reliable, with the predictor of new words having a medium performance.

3 Experiments

In our experiments we used the Efron-Thisted test for new words represented in Equation 1, where $\mathcal{N}_x$ represents the number of words which occurred x times within the training corpus and t is the number of words in the new text divided by the number of words within the corpus. The fundamental assumption of the test is that words occur in texts as Poisson processes. By using the observed counts and occurrences of words as rate estimators, it should be possible to estimate the number of new words to be observed in the future. For space and time considerations, we do not derive the formula which is reviewed in detail in [ET76].

$$\triangle(t) = \sum_{x=1}^{\infty} (-1)^{x+1} \mathcal{N}_x t^x \tag{1}$$

This particular test was chosen because of it's simplicity of implementation and the ease with which incremental data can be added to its model. A log of email written by three different authors was used as experimental data. In the case of the third set of email, a sizeable amount of additional

text written by the same author was used as a base set of words to make a separate prediction of $\triangle(t)$. In all cases, the emails are added to the training corpus as they are processed in order to generate a predictive model incrementally.

The parsing rules for the text are similar to those prescribed by [ET76] in their original paper (eg: any sequence of characters constitute a word). To cut down on the clutter of parsing a sizeable amount of raw text, any non alphanumeric character except for apostrophe (') and dash (-) are replaced with a blank space. Any sequence of characters longer than 1 character and shorter than 25 character is considered a word.

Table 1 tabulates some statistics about all three data sets. The specific questions that we wanted answered were: the precision that could be expected from the test, what was the effect of the size of the underlying sample and the effect of the size of the individual email processed.

Set	#	# of words		Avg. words per email	
Set		Total	Unique	Total	Unique
1	340	30,806	20,538	90	60
2	145	11,562	9,006	80	62
3	2,945	156,481	117,780	53	40

Table 1: Statistics on all email datasets. Note that Set 3 has a sizeable advantage with a pre-build corpus of 282,194 words.

Three Java applications were written to parse the emails and generate reports based on the datasets. These were custom written to support this research and a more complete analysis suite will be written later. A database management system was used to store and manipulate the different datasets, table descriptions and commented source code is joined to this document.

Each set of email is processed according to its sent date starting at the earliest one. This is done in an attempt to satisfy the requirements stated in [ET76] that the distribution of new words is binomial. For each email, the number of words and the number of words unseen so far are parsed into the database and the value of $\triangle(t)$ calculated based on the size of the email. The words in the email are then aggregated into the base set and the next email is computed. In the case of set 3, an additional value of $\triangle(t)$ is computed based only on the initial training set.

4 Results

The analysis was run for all three data sets and the results plotted for the number of new words versus the number of words parsed for both observed and predicted counts. Only the most interesting results are presented here, while the rest are presented in Appendix B. Note that whenever the predicted values are invalid, or out of a reasonable range, the values are not plotted.

Figure 1 is a plot of the predicted and actual word counts for Set 3. From the graph it is obvious that both predicted values of new words are severely underestimated. The prediction based on incremental data is especially low because the rise in the coefficient values is too shallow when compared with the continual drop in the value of T. The results of Figure 1 are consistent with the results for sets 1 and 2 (Figures 4 and 5).

Possibly we can explain these results with too small of an initial corpus to calculate $\triangle(t)$, too small a number of words per email or possibly because email is too terse of a communication to properly analyze with this test. As an experiment, the training corpus of set 3 was incremented in blocks of 50,000 words to verify the effects of a rising corpus size. The end result was an unwarranted marked decrease in the number of new words. The test was attempted with several other data sets in order to clarify these issues. Whenever possible, detailed bibliographical sources are provided to point to common text sources.

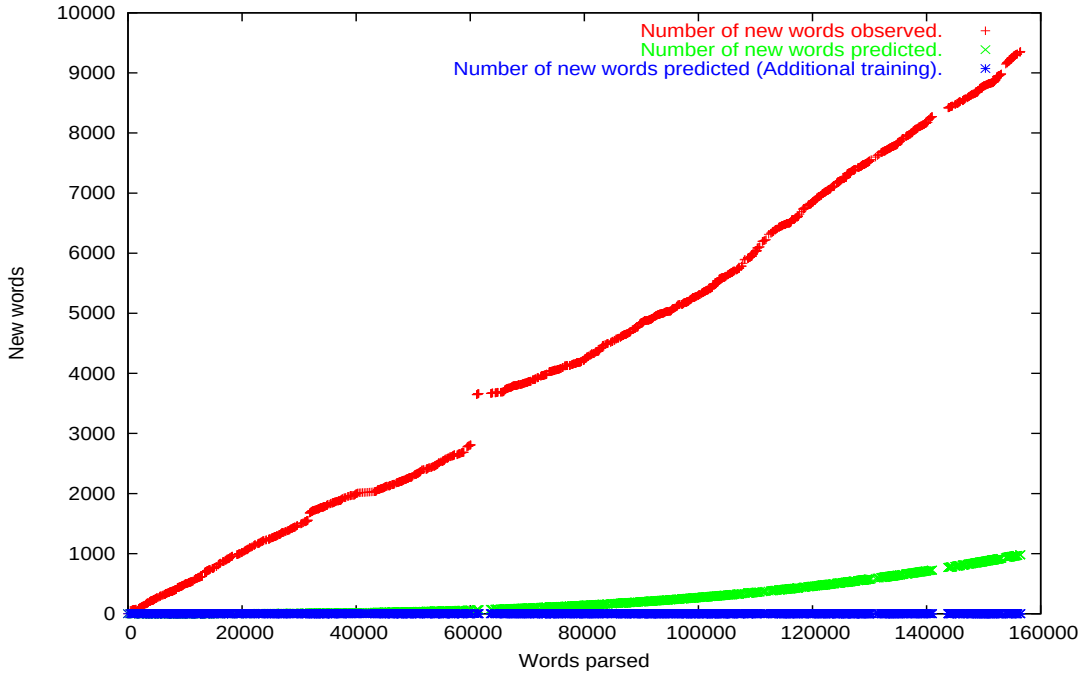

Figure 1: Plot of observed news words versus predicted for Set 3.

4.1 Sample sizes: Shakespeare revisited and hello Tolstoy

In the [ET87] paper, most of the Shakespeare works were used to generate the base corpus from which the $\triangle(t)$ were computed. In our case, we made use of the 1623 Folio [Sha00] as a base source of data with 781,600 words. The first 75% of the folio were used as a base set to predict the number of new words in the leftover 25% and the full folio used to predict the number of new words within "The Sonnets" [Sha97] and "The Phoenix and the Turtle" [Sha98]. It should be noted that our own email parser was used to prepare the texts for analysis. Also, the texts used here are provided in ASCII format by the Gutenberg project and reflect some typographical "weirdness" of that era (for example, v's would sometimes be substituted for u's in text because of the high cost of type). This resulted in a mismatch between our counts and those of previous efforts [ET87], making hard number comparisons of little use. This is much the same problem as [EV96] experienced.

Test	Size	# New words	
	of sample	$\triangle(t)$	Actual
Shakespeare Folio	184,784	377	4,508
The Sonnet	19,020	0.2	1,113
Phoenix	2,298	0	46

Table 2: Predictions on Shakespeare texts.

The results presented in Table 2 are again consistent with those found in the email data sets: the predicted values being severely understated compared with their actual values. As another check, some of the works of Tolstoy were processed in the same manner. While the texts are originally in Russian, we processed them while comparing only texts which were translated by the same person. A prediction of new words in the last 25% of "War and Peace" [Tol01c] computed at 236 words with the actual count being 2,546 new words. A similar test with "Anna Karenina" [Tol98], translated by Constance Garnett, predicted 178 new words when the actual was 1,998.

Text	Length	# New words	
		$\triangle(t)$	Actual
Hadji - 1	4,545	0	271
Hadji - 2	9,738	1	573
Hadji - 3	15,539	4	928
Hadji - 4	23,787	12	1,324
Hadji - 5	30,460	26	1,777
Hadji - 6	45,280	82	2,615
Resurrection	148,443	2,530	5,852

Table 3: Predictions on Tolstoy texts.

A corpus of some of Tolstoy's works, all of which were translated by Louise and Aylmer Maude[1], was used as a reference to predict the number of new words within "Resurrection" [Tol99] and the chapters of "HadJi Murad" [Tol00]. The results tabled in Table 3 are no more encouraging, expect for Resurrection which almost achieves the 50% mark.

Thus, we conclude from these tests that the poor performance of the test does not seem to be related to the type of communication or to the size of the prediction window. At best, the test seems to behave as a sort of lower bound for the number of new words expected. The lower bounding behavior is in line with the original [ET76] paper which used a modified model to calculate the lower bound on Shakespeare's total vocabulary.

4.2 Discussion: Author Authentication

If the evidence suggests that the test is indeed acting as a lower bound for new words, is this sufficient for use as a coarse authentication mechanism? Figure 2 represents all three sets along with the two working predictors for the overlapping ranges. Ideally, the predicted values of new words would match those observed within a certain error range. This rate estimation could have been used as a crude indicator.

In this case, both predictors are not only under-estimating the number of new words, but their values are too similar for us to make use of them for authentication purposes. An interesting item is the slopes of the observed number of new words for each data set. Set 3 is clearly differentiable from sets 1 and 2, which is expected due to it's initial training corpus. But even sets 1 and 2 have different shapes to their near-instantaneous rates of new word discoveries; perhaps this can used to our advantage in determining authorship.

5 A proposal: Time Series

Efron and Thisted had as an underlying assumption that the arrival rate of each individual word type was an independent Poisson process. We propose instead that the rate of new words is a process which can be determined by averaging the last n th rates. New words are likely to occur in three cases: a change in topic, the everyday learning of new words and through the random arrival of words within the flow of text. Change in topic was partially covered by [ET76] through their assumption of a Poisson process where the event must equally be probable throughout the period. A change in topic violates this assumption, but it should be possible to average out their occurrence thought some kind of seasonality calculation.

The authors learning process is an interesting idea which is difficult to model. We submit that it is a continuous process that is inherently predictable as it is a long term process: it it unlikely that the author will learn an entirely new vocabulary overnight and use it all in a single text without referencing

[1] "The Cossacks" [Tol02], "The death of Ivan Ilynch" [Tol01a], "The kreutzer sonata" [Tol01b] and "Master and Man" [Tol97]

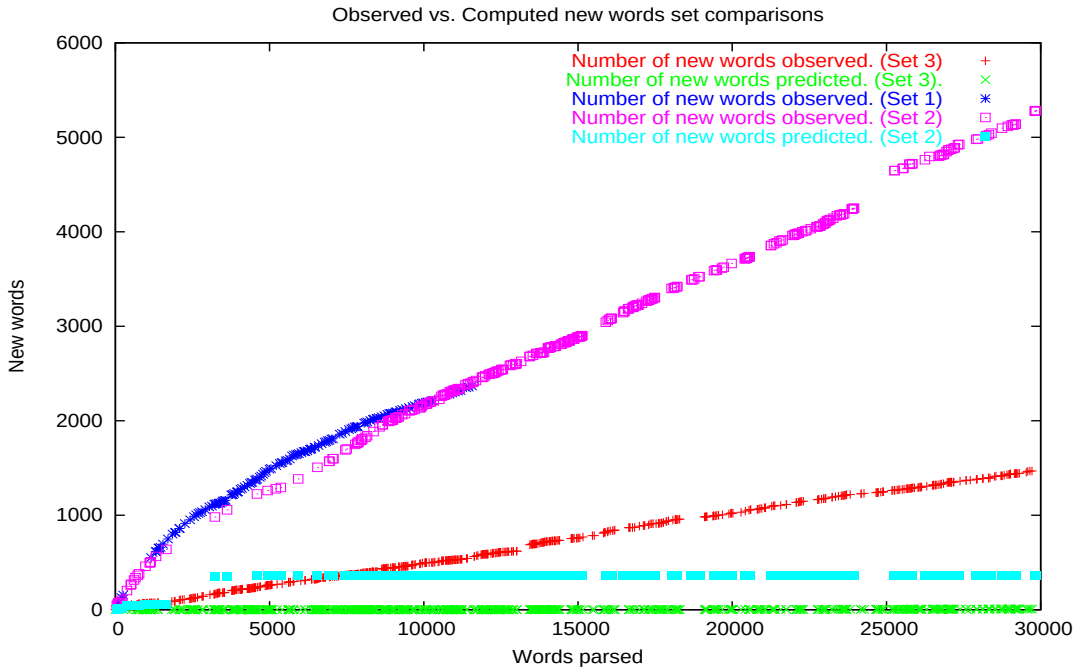

Figure 2: Comparing the behavior of all three sets.

some of the earlier words. Finally, the normal arrival of random words seems to be behave in a rather uniform rate, poorly modeled by a Poisson process.

Based on the above and the plots of Figure 2, we propose the use of a limited memory time-series as a predictor of new words for authentication purposes. The rate of new word observations does change in all three plots but in a fashion which can be predicted from previous values. We will harvest several targeted mailing lists over the coming months in the hopes of building a wider sample of author tagged texts to investigate this potential new method. The wider sample will allow us to average out statistical blips while ensuring a more robust challenge for authorship attribution.

Finally, the events described above are all sources of new words which should occur at some empirically defined frequencies. It would be interesting to investigate the use of Digital Signal Processing tools with new word observations being treated as a noisy signal.

6 Conclusion

In this work one of the Efron-Thisted tests was used with three sets of email data in order to evaluate the possibility of using it as a coarse authorship test. The data so far indicates that the test is too weak for this purpose, while it does seem to behave as a lower bound for the number of new words expected. It may still be possible to use the other Efron-Thisted tests for authorship authentication as they are reported to be stronger.

A Notes on series expansion of $\triangle(t)$

Since the goal of this paper was to review the possibility of using the Efron-Thisted test to authenticate email's, speed of processing was one of the considerations. As part of the analysis a few experiments were attempted to verify the tolerance of the $\triangle(t)$ function to the truncation of the series. The plot in Figure 3 is typical of the results, in that for values of T less or equal to one, the first few terms are the dominant one. For values larger than 1.0, the higher terms have an increased weight and truncating the

series has a direct effect on the $\triangle(t)$ function. This property can be used to significantly speedup the computation of the $\triangle(t)$ value in high throughput environments by computing the first 100th values or so.

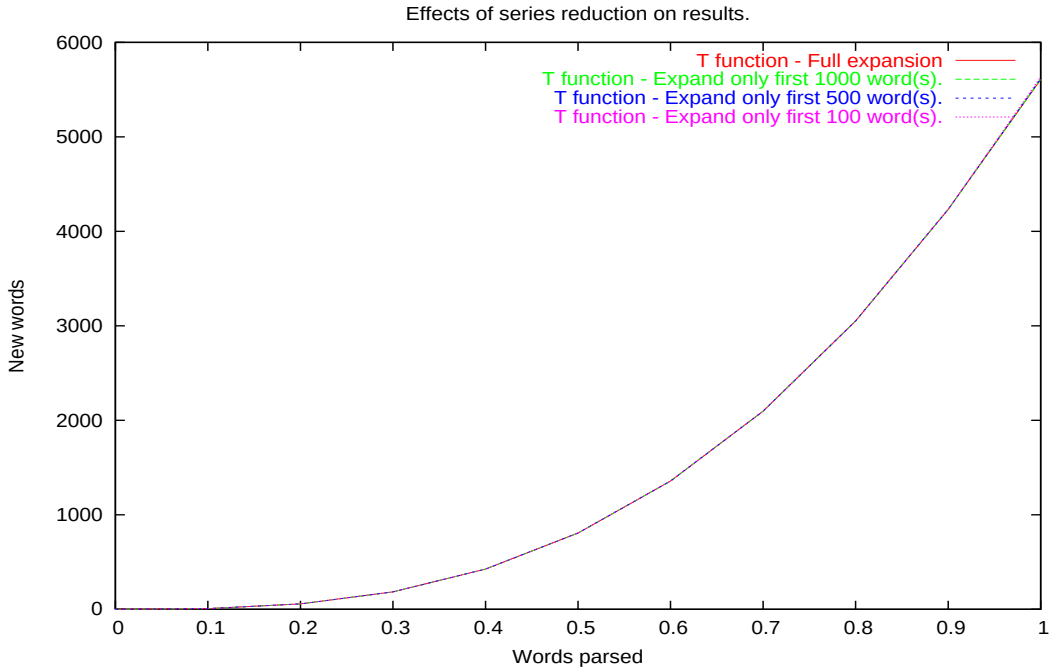

Figure 3: T function is tolerant of expansion cutoff for T<1.0, values above 1.0 vary wildly and are not plotted.

B Set 1 and 2 plots

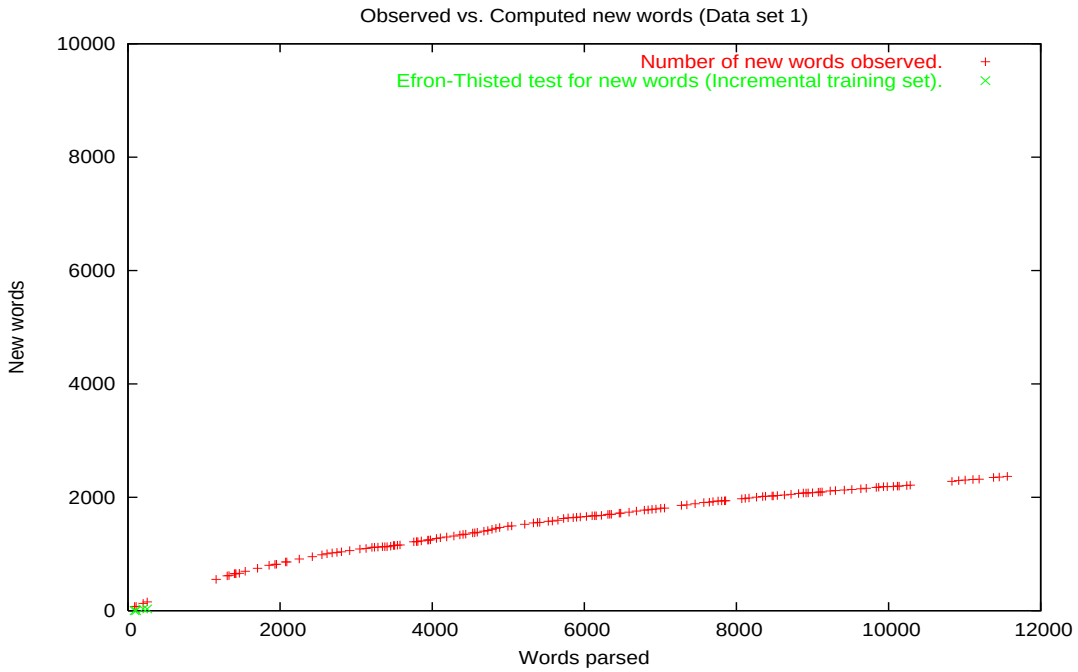

Figure 4: Plot of observed news words versus predicted for Set 1. The predicted value fails after about 300 words.

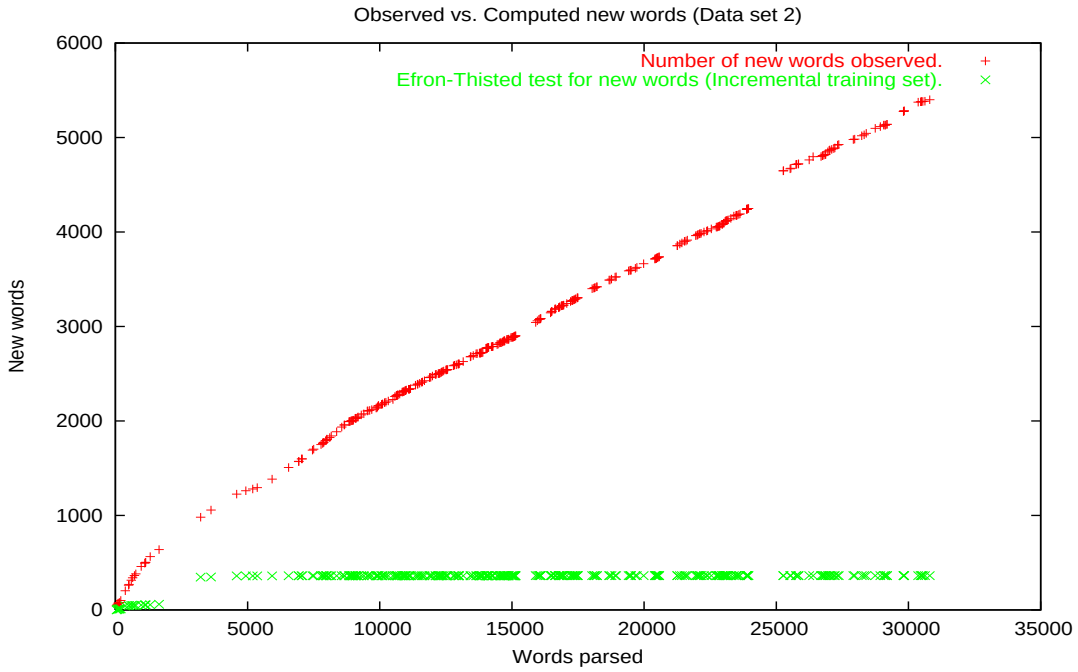

Figure 5: Plot of observed news words versus predicted for Set 2.

References

[ET76] Bradley Efron and Ronald Thisted, *Estimating the number of unseen species: How many words did shakespeare know?*, Biometrika **63** (1976), no. 3, 435–447.

[ET87] ———, *Did shakespeare write newly-discovered poem?*, Biometrika **74** (1987), no. 3, 445–455.

[EV96] Ward E. Y. Elliot and Robert J. Valenza, *And then there were none: Winnowing the shakespeare claimants*, Computers and the Humanities **30** (1996), 191–245.

[MW64] Frederick Mosteller and David L. Wallace, *Inference and disputed authorship: The federalist*, Addison-Wesley Publishing Company Inc., 1964.

[Sha97] William Shakespeare, *The sonnets*, Project Gutenberg, 1997, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext97/wssnt10.txt`.

[Sha98] ———, *The phoenix and the turtle*, Project Gutenberg, 1998, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext98/2ws2710.txt`.

[Sha00] ———, *Mr. william shakespeares comedies, histories, & tragedies (based on the first folio 1623)*, Project Gutenberg, 2000, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext00/00ws110.txt`.

[Tol97] Leo Tolstoy, *Master and man*, Project Gutenberg, 1997, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext97/mramn10.txt`.

[Tol98] ———, *Anna karenina*, Project Gutenberg, 1998, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext98/nkrnn11.txt`.

[Tol99] ———, *Resurrection*, Project Gutenberg, 1999, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext99/resur10.txt`.

[Tol00] ———, *Hadji murad*, Project Gutenberg, 2000, `http://www.ccel.org/tolstoy/hadij/`.

[Tol01a] ———, *The death of ivan ilych*, Project Gutenberg, 2001, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext01/wrnpc10.txt`.

[Tol01b] ———, *The kreutzer sonata*, Project Gutenberg, 2001, `http://www.ccel.org/tolstoy/kreutzer/kreutzer.txt`.

[Tol01c] ———, *War and peace*, Project Gutenberg, 2001, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext01/wrnpc10.txt`.

[Tol02] ———, *The cossacks*, Project Gutenberg, 2002, `ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext03/cossk10.txt`.

[Val91] Robert J. Valenza, *Are the thisted-efron authorship tests valid?*, Computers and the Humanities **25** (1991), 27–46.

[Yul68] G. Udny Yule, *The statistical study of literary vocabulary*, Archon Books, 1968.