The Supplementary Materials for Article "Corrections of Zipf's and Heaps' Laws Derived from Hapax Rate Models"

Łukasz Dębowski*

1 Introduction

This is the supplementary report for article "Corrections of Zipf's and Heaps' Laws Derived from Hapax Rate Models". The article introduces corrections to Zipf's and Heaps' laws based on systematic models of the hapax rate. The derivation rests on two assumptions: The first one is the standard urn model which predicts that marginal frequency distributions for shorter texts look as if word tokens were sampled blindly from a given longer text. The second assumption posits that the rate of hapaxes is a simple function of the text size. Four such functions are discussed: the constant model, the Davis model, the linear model, and the logistic model. This report contains all tables, all plots, and an instruction for rerunning the numerical experiment.

2 Running the experiment

We worked on Linux Ubuntu 20.04.4 LTS applying a mixture of Bash, Perl, and Gnuplot scripts. We processed 14 texts in English downloaded from Project Gutenberg [1] and listed in Table 1. These texts were projected to 26 letters and a space (27 distinct characters in total), compressed by `gzip`, and placed into directory `gutenberg/`. The scripts for running the experiment are located in directories `scripts/` and `TypeToken/`. To repeat the experiment, it suffices to run:

```
cd ./scripts/
./make.bash
```

*Ł. Dębowski is with the Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland (e-mail: ldebowsk@ipipan.waw.pl).

Table 1: The selection of texts from Project Gutenberg.

Title	Author	File
First Folio/35 Plays	W. Shakespeare	00ws110.txt
One of Ours	W. Cather	1ours10.txt
20,000 Leagues under the Sea	J. Verne	2000010.txt
Critical & Historical Essays	T. Macaulay	2cahe10.txt
Five Weeks in a Balloon	J. Verne	5wiab10.txt
Eight Hundred Leagues on the Amazon	J. Verne	800lg10.txt
The Complete Memoirs	J. Casanova	csnva10.txt
Memoirs	Comtesse du Barry	dbrry10.txt
The Descent of Man	C. Darwin	dscmn10.txt
Gulliver's Travels	J. Swift	gltrv10.txt
The Mysterious Island	J. Verne	milnd10.txt
Mark Twain, A Biography	A. Paine	mt7bg10.txt
The Journal to Stella	J. Swift	stlla10.txt
Life of William Carey	G. Smith	wmcry10.txt

Script `make.bash` calls other scripts in directories `scripts/` and `TypeToken/`, which apply Bash, Perl, and Gnuplot. Prior to running the experiment, make sure that you have installed these in your operating system. In particular, the final Latex arrays for Tables 2 and 3 are produced by scripts

`make_parameters_herdan_1.pl`,
`make_parameters_herdan_2.pl`.

The output files such as text tables and PDF figures are located in the respective subdirectories of directory `output/herdan/`. Each Project Gutenberg text has its own directory, named accordingly. Additionally, directory `output/` contains the PDF image for a plot of a U-shaped hapax rate function, Figure 1, which does not depend on empirical data.

3 Supplementary figures

For each of the 14 texts, we produced three plots depicting: the hapax rate function, the vocabulary size function (Herdan's law plot), and the rank function (Zipf's law plot), and three plots depicting the fitting residuals for each of the three forementioned functions. The respective PDF images are

Table 2: The parameters fitted by least squares to function $G(n)$.

File	Constant	Davis	Logistic			Linear		Length
	β	α	γ	β	α	γ	α	N
00ws110.txt	0.768	12.06	0.314	0.218	10.11	0.0509	2.14	835726
1ours10.txt	0.797	11.55	0.318	0.203	9.72	0.0507	1.7	128963
2000010.txt	0.801	11.48	0.323	0.008	10.62	0.0578	2.22	101247
2cahe10.txt	0.796	12.12	0.314	0	11.38	0.0576	2.79	298339
5wiab10.txt	0.808	11.64	0.315	0.001	10.86	0.0552	2.13	92558
800lg10.txt	0.799	11.43	0.327	0.162	9.77	0.0534	1.84	95493
csnva10.txt	0.732	11.39	0.308	0.157	9.94	0.0542	1.87	1268149
dbrry10.txt	0.787	11.39	0.325	0.065	10.31	0.0583	2.23	159710
dscmn10.txt	0.774	11.5	0.328	0	10.75	0.0629	2.71	312075
gltrv10.txt	0.796	11.4	0.322	0.001	10.62	0.0584	2.22	104909
milnd10.txt	0.773	11.14	0.347	0.127	9.63	0.0608	2.24	195064
mt7bg10.txt	0.775	11.91	0.296	0.001	11.45	0.0565	2.55	519886
stlla10.txt	0.757	10.91	0.333	0.231	8.87	0.0536	1.45	245882
wmcry10.txt	0.799	11.69	0.314	0	10.96	0.0567	2.34	145487
Mean	0.783	11.54	0.32	0.084	10.36	0.0562	2.17	321678

Table 3: The goodness of fit $\sqrt{\text{WSSR/ndf}}$ for function $G(n)$.

File	Constant	Davis	Logistic	Linear
00ws110.txt	1784.34	120.42	**11.82**	43.71
1ours10.txt	478.53	74.39	**7.02**	16.69
2000010.txt	439.57	117.31	**2.18**	24.14
2cahe10.txt	1118.75	255.29	**29.17**	86.71
5wiab10.txt	414	111.21	**4.15**	25.14
800lg10.txt	402.88	83.74	**3.12**	16.48
csnva10.txt	1721.89	107.98	**6.86**	34.72
dbrry10.txt	587.39	125.46	**5.65**	31.08
dscmn10.txt	982.09	215.02	**19.93**	63.73
gltrv10.txt	463.34	117.35	**6.86**	31.39
milnd10.txt	629.47	125.02	**1.88**	23.22
mt7bg10.txt	1433.58	194.1	**8.75**	73.41
stlla10.txt	603.45	45.14	9.67	**9.34**
wmcry10.txt	592.57	143.7	**5.25**	36.47
Mean	832.27	131.15	**8.74**	36.87

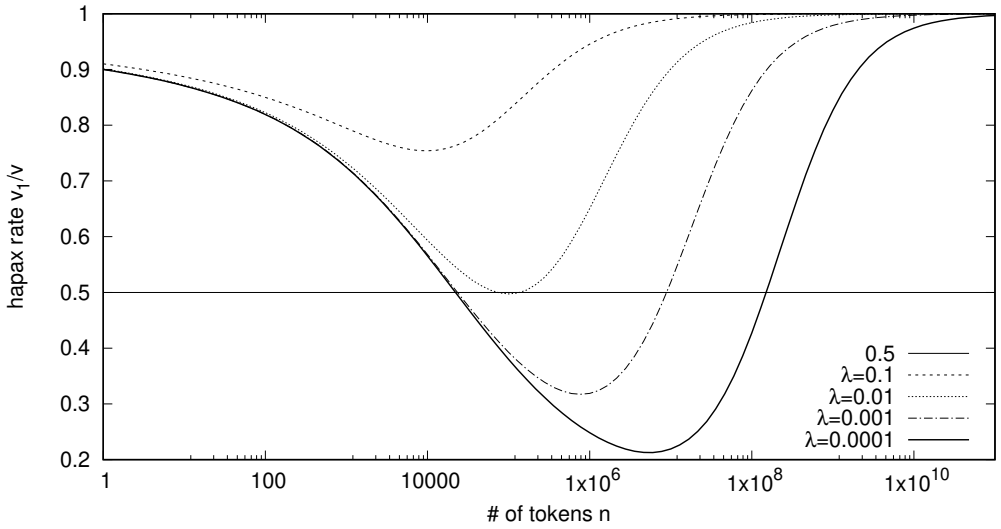


Figure 1: The U-shaped hapax rate function for a mixture of the Davis model with $\alpha = 10$ and the maximal model. The weight of the Davis model is $(1 - \lambda)$ and the weight of the maximal model is λ.

located in the proper subdirectories of directory `output/herdan/`. For convenience, we reproduce them in the present report as Figures 2–29. Eighteen of these plots have been produced in article "Corrections of Zipf's and Heaps' Laws Derived from Hapax Rate Models". The legends for all 84 plots in this supplementary report are analogous as in the main article.

References

[1] Project Gutenberg, (n.d.). Retrieved May 25, 2011, from `https://www.gutenberg.org/`.

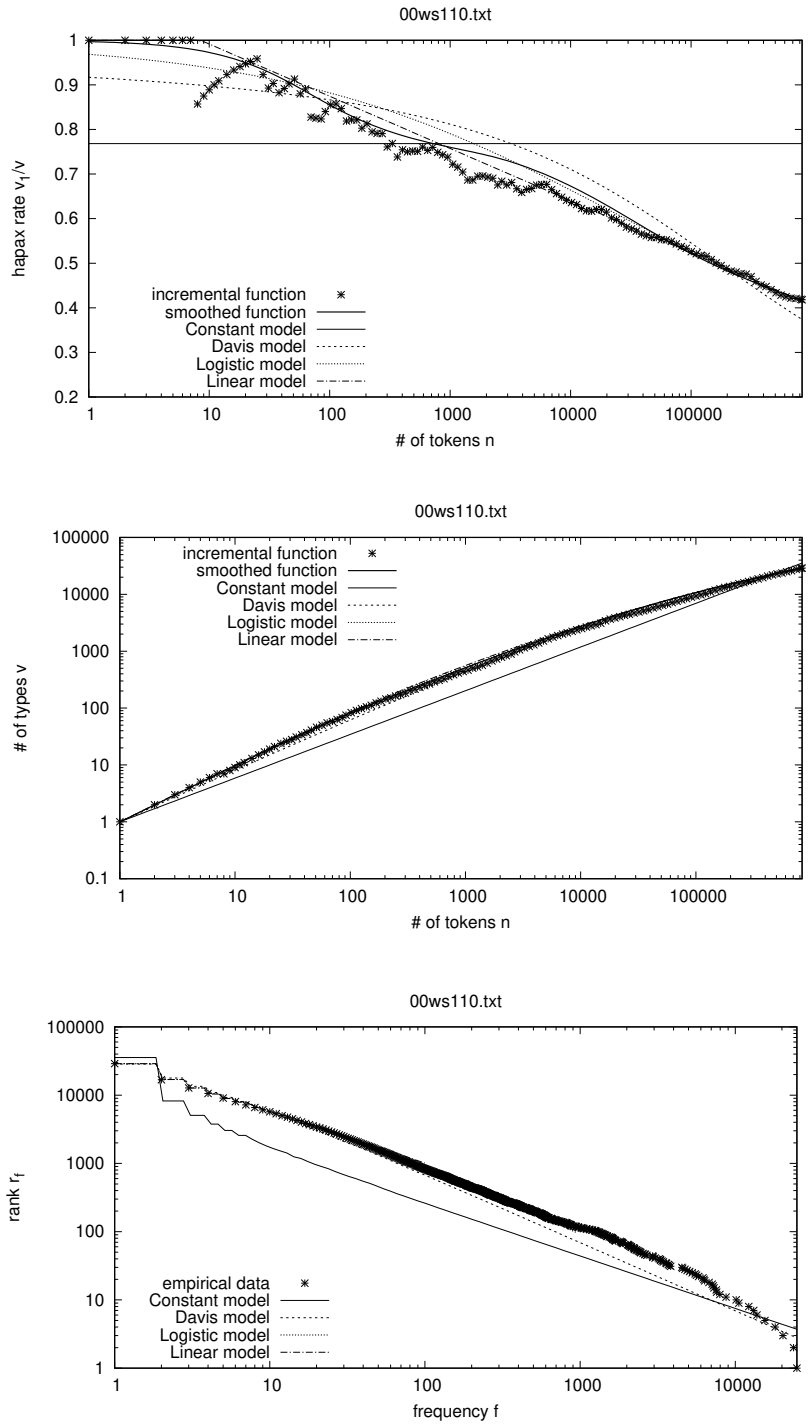
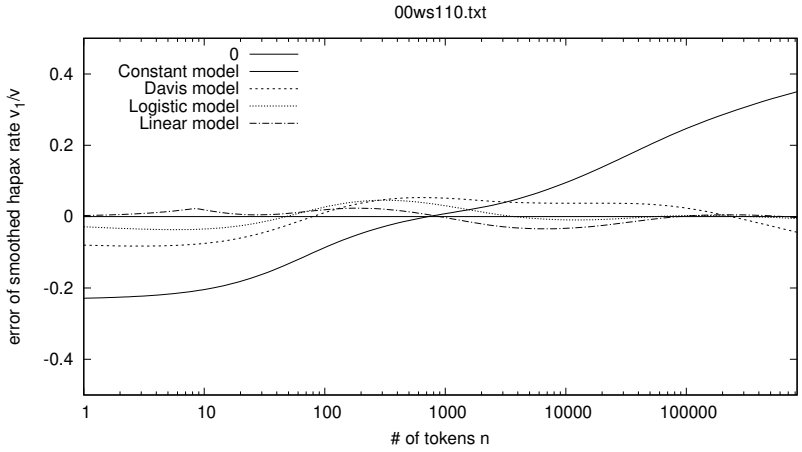
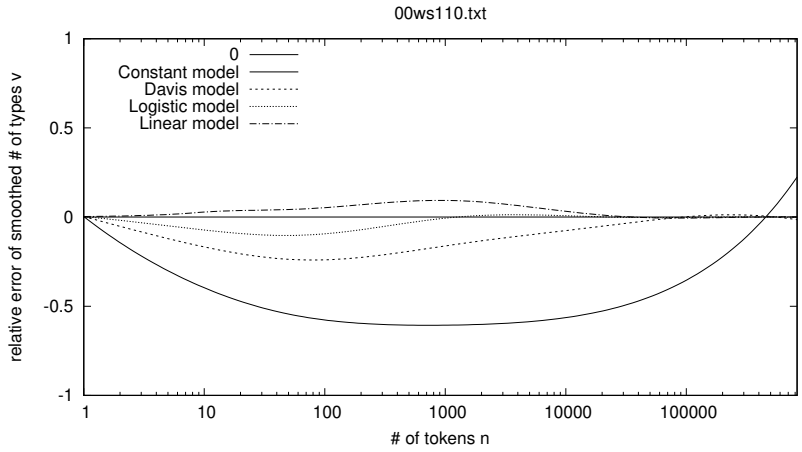


Figure 2: W. Shakespeare, *First Folio/35 Plays*.





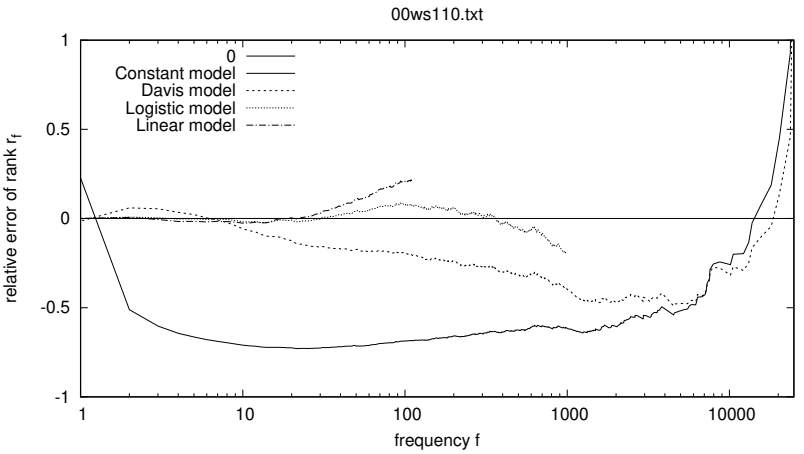
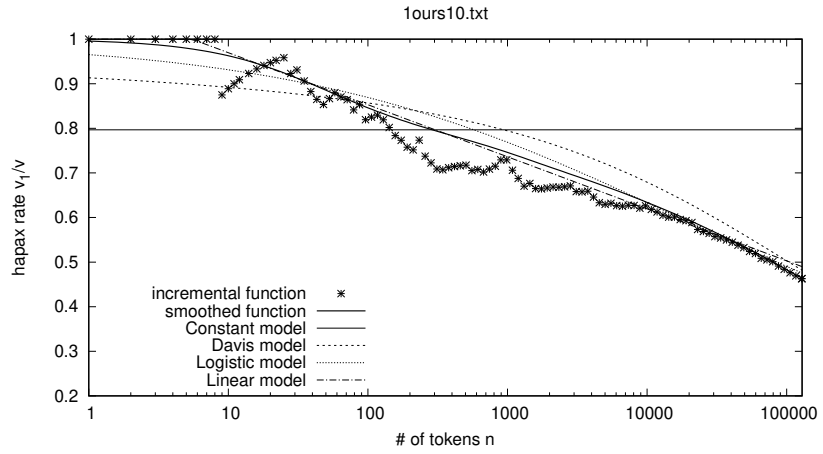
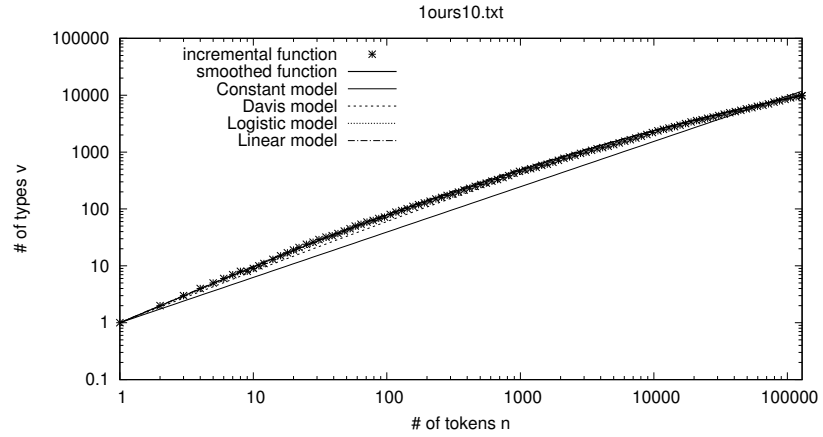


Figure 3: W. Shakespeare, *First Folio/35 Plays*.





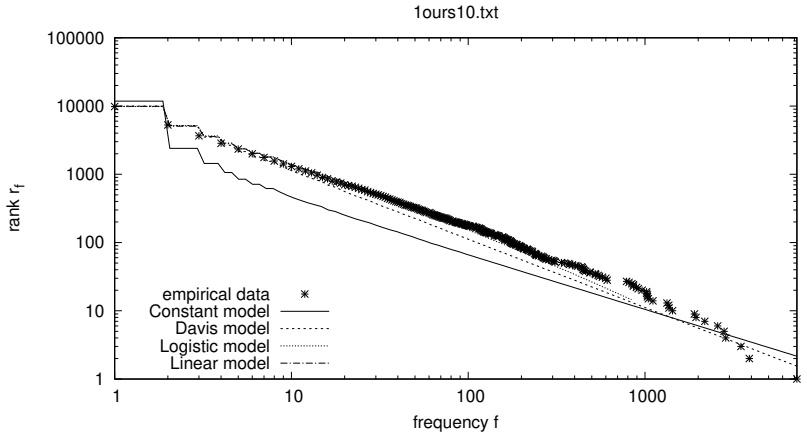


Figure 4: W. Cather, *One of Ours*.

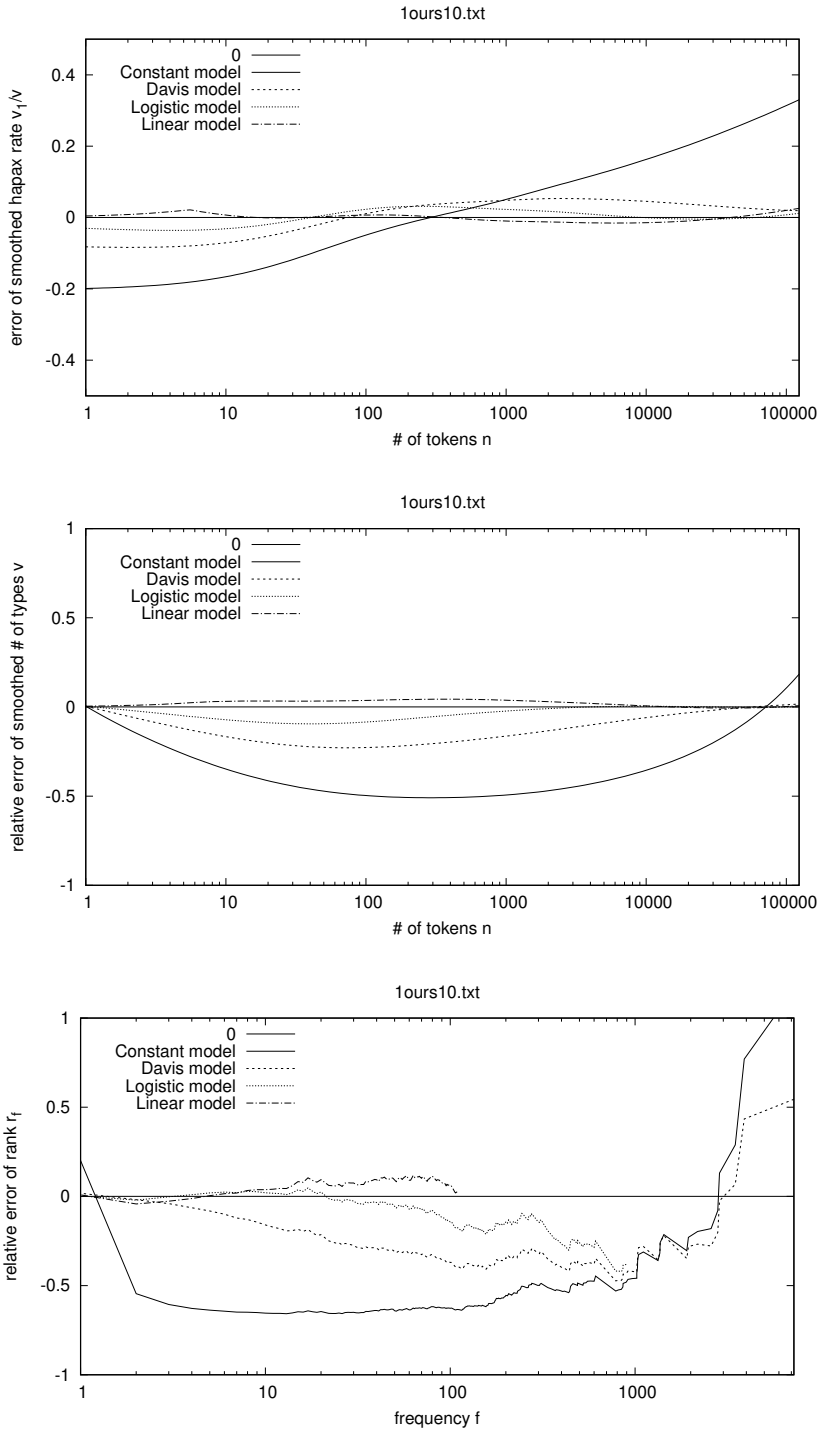


Figure 5: W. Cather, *One of Ours*.

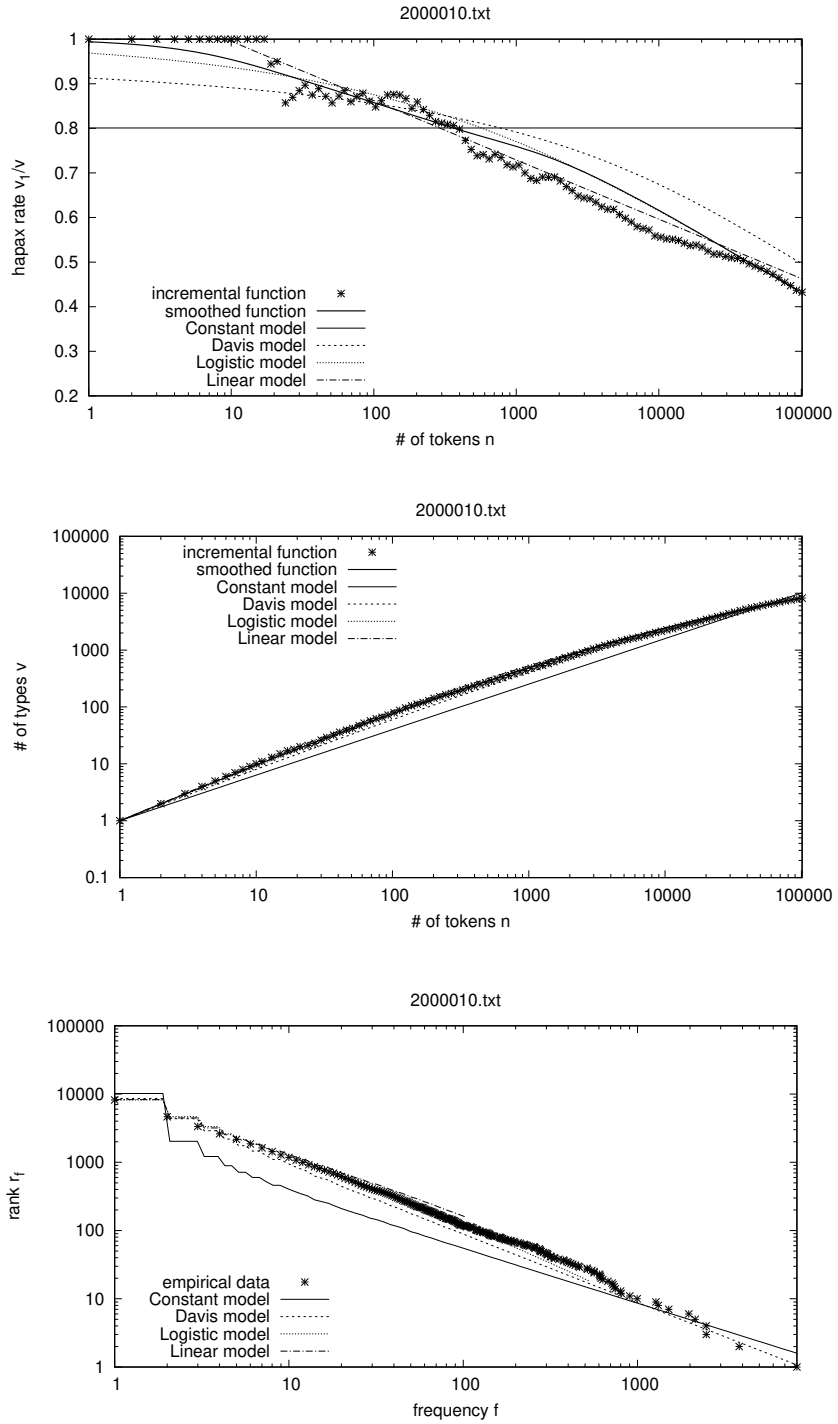


Figure 6: J. Verne, *20,000 Leagues under the Sea*.

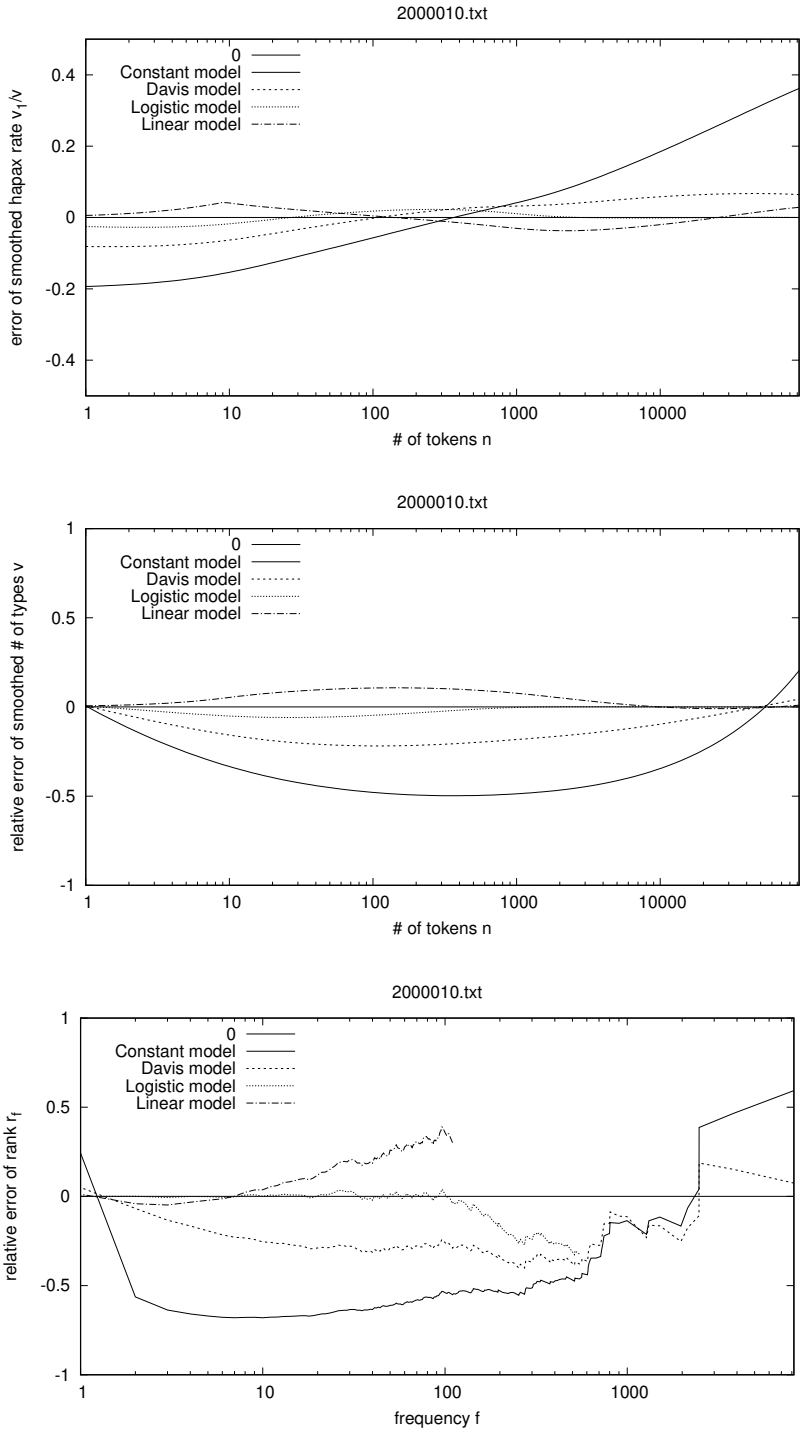


Figure 7: J. Verne, *20,000 Leagues under the Sea*.

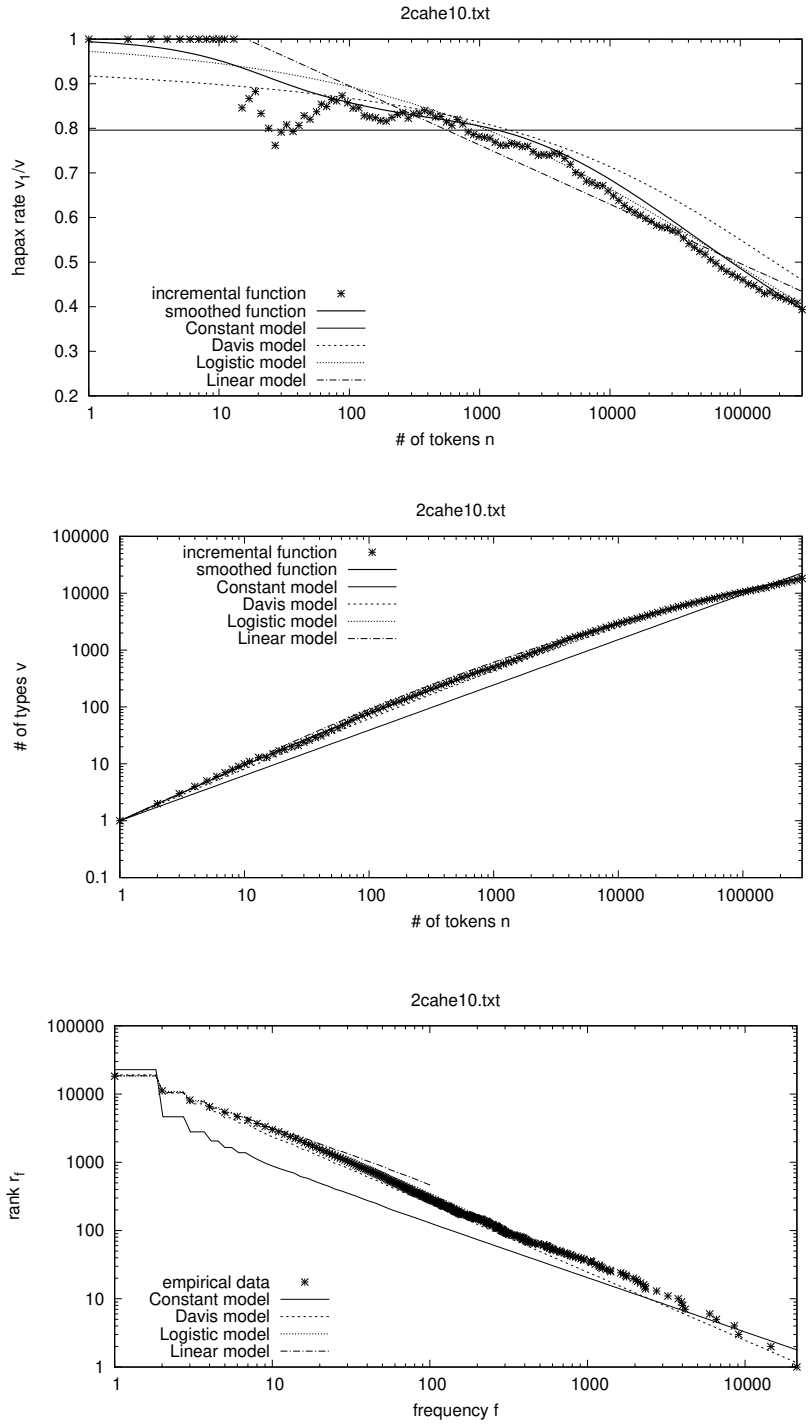
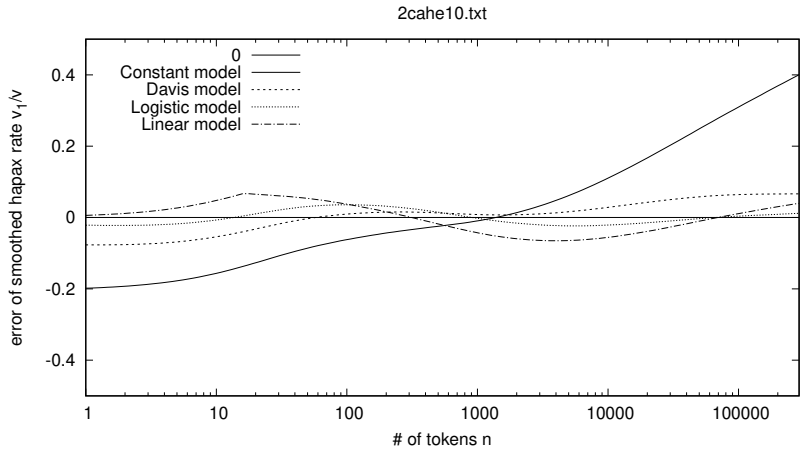
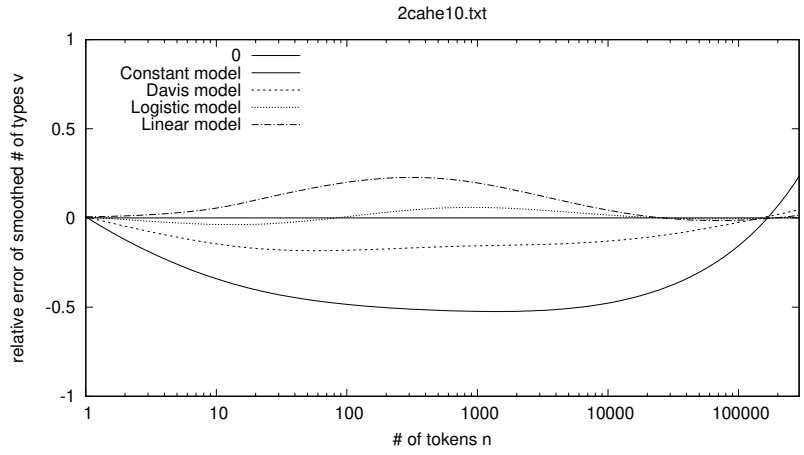


Figure 8: T. Macaulay, *Critical & Historical Essays*.





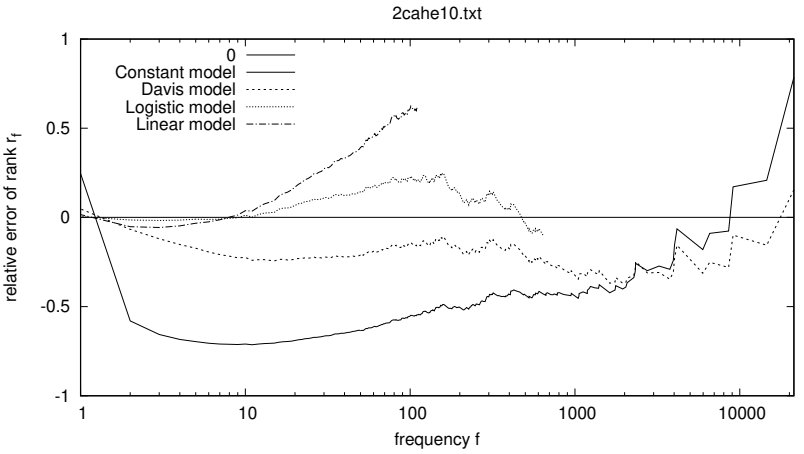
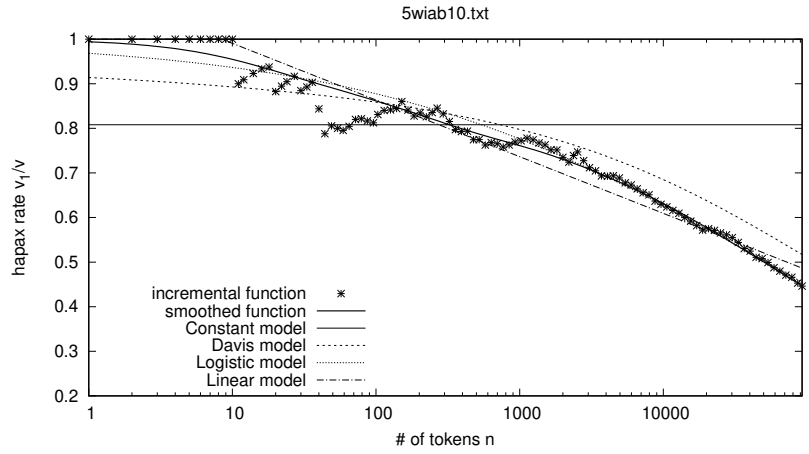
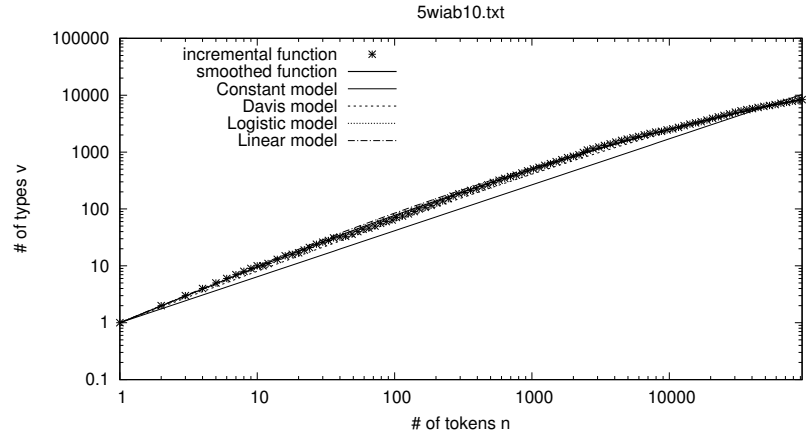


Figure 9: T. Macaulay, *Critical & Historical Essays*.





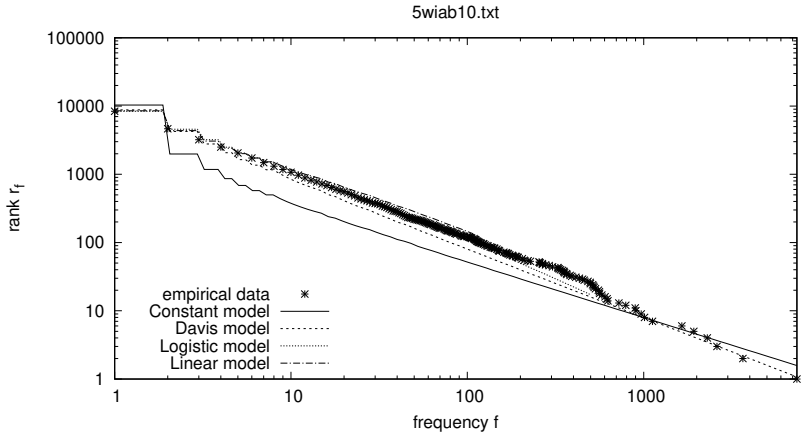


Figure 10: J. Verne, *Five Weeks in a Balloon*.

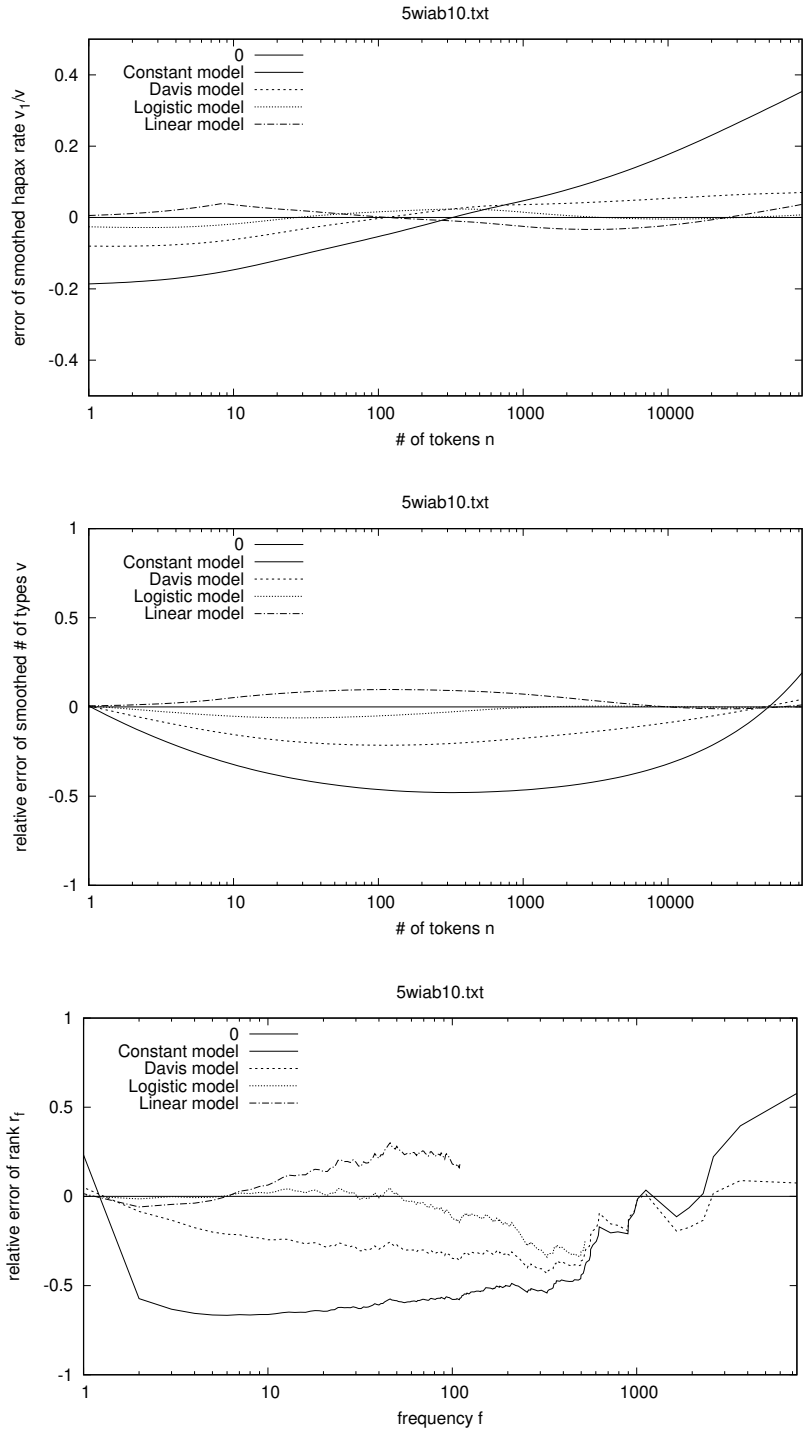
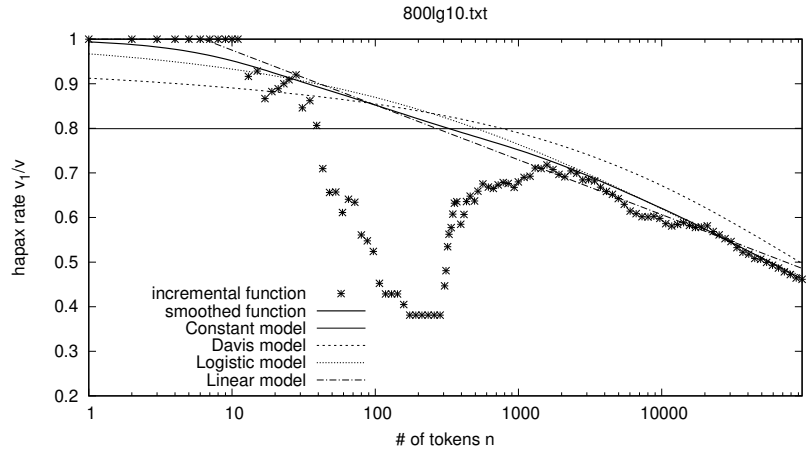
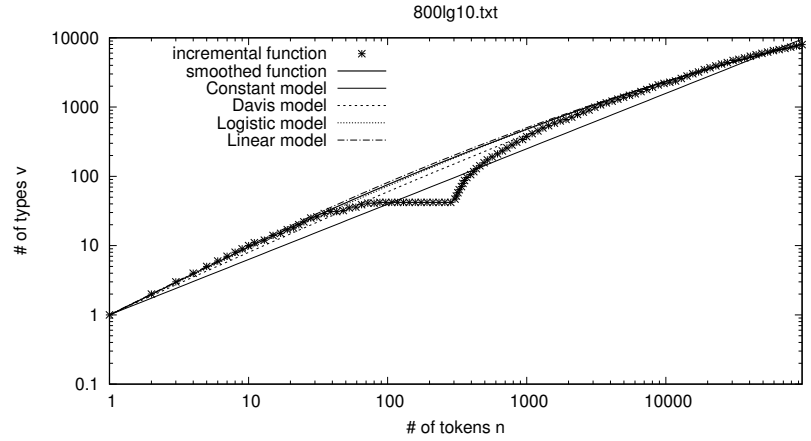


Figure 11: J. Verne, *Five Weeks in a Balloon*.





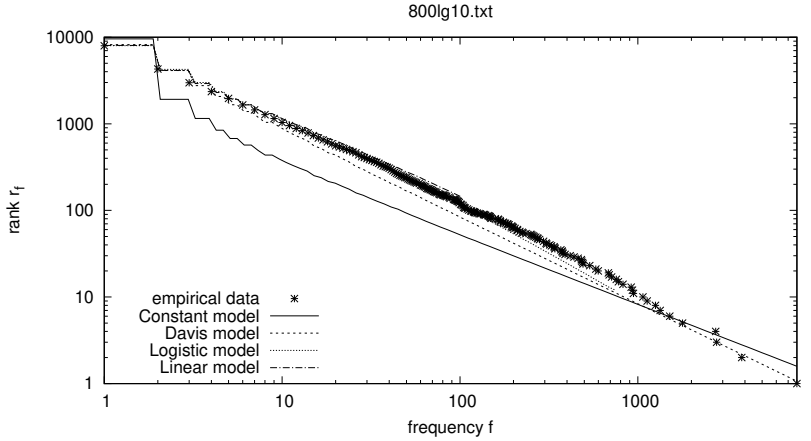


Figure 12: J. Verne, *Eight Hundred Leagues on the Amazon.*

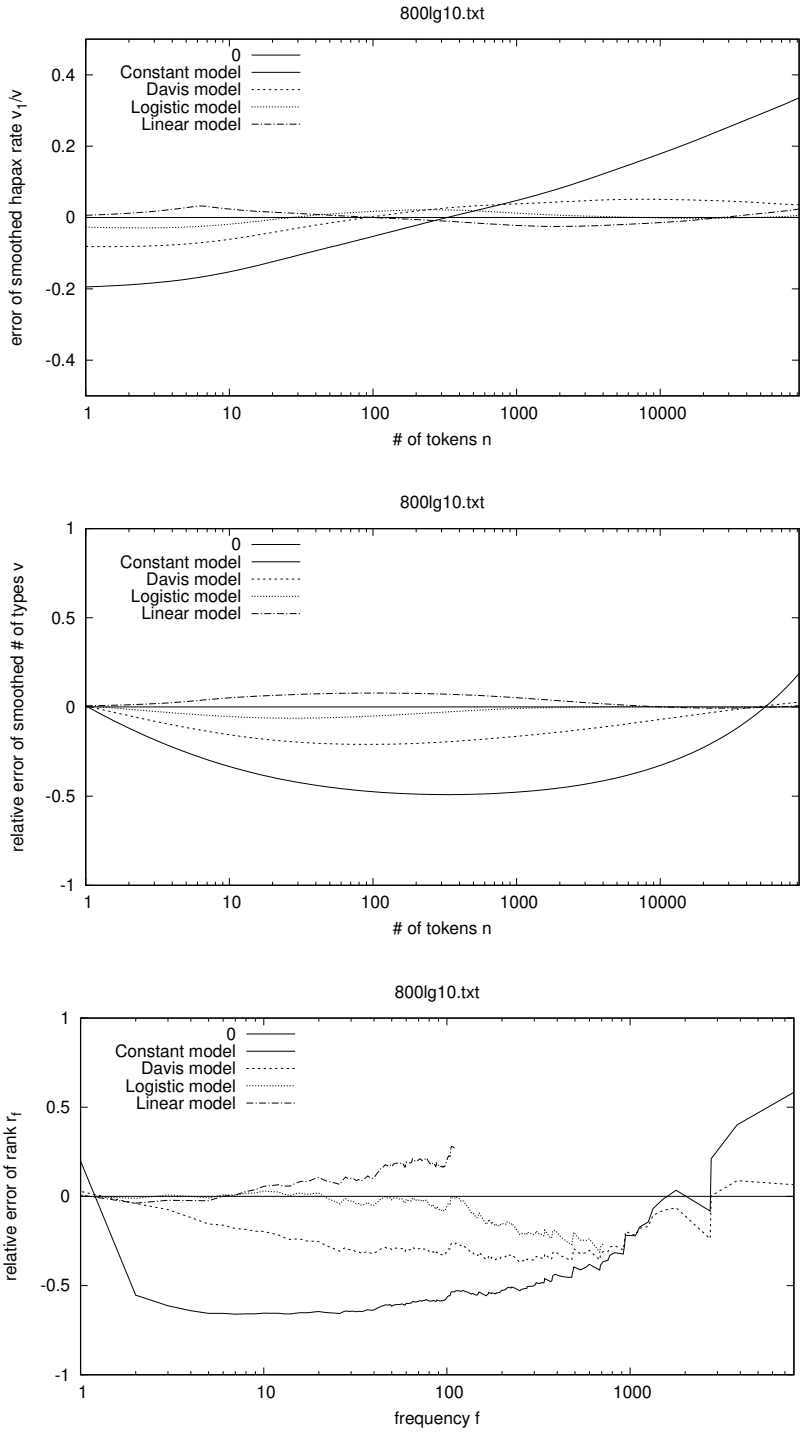


Figure 13: J. Verne, *Eight Hundred Leagues on the Amazon.*

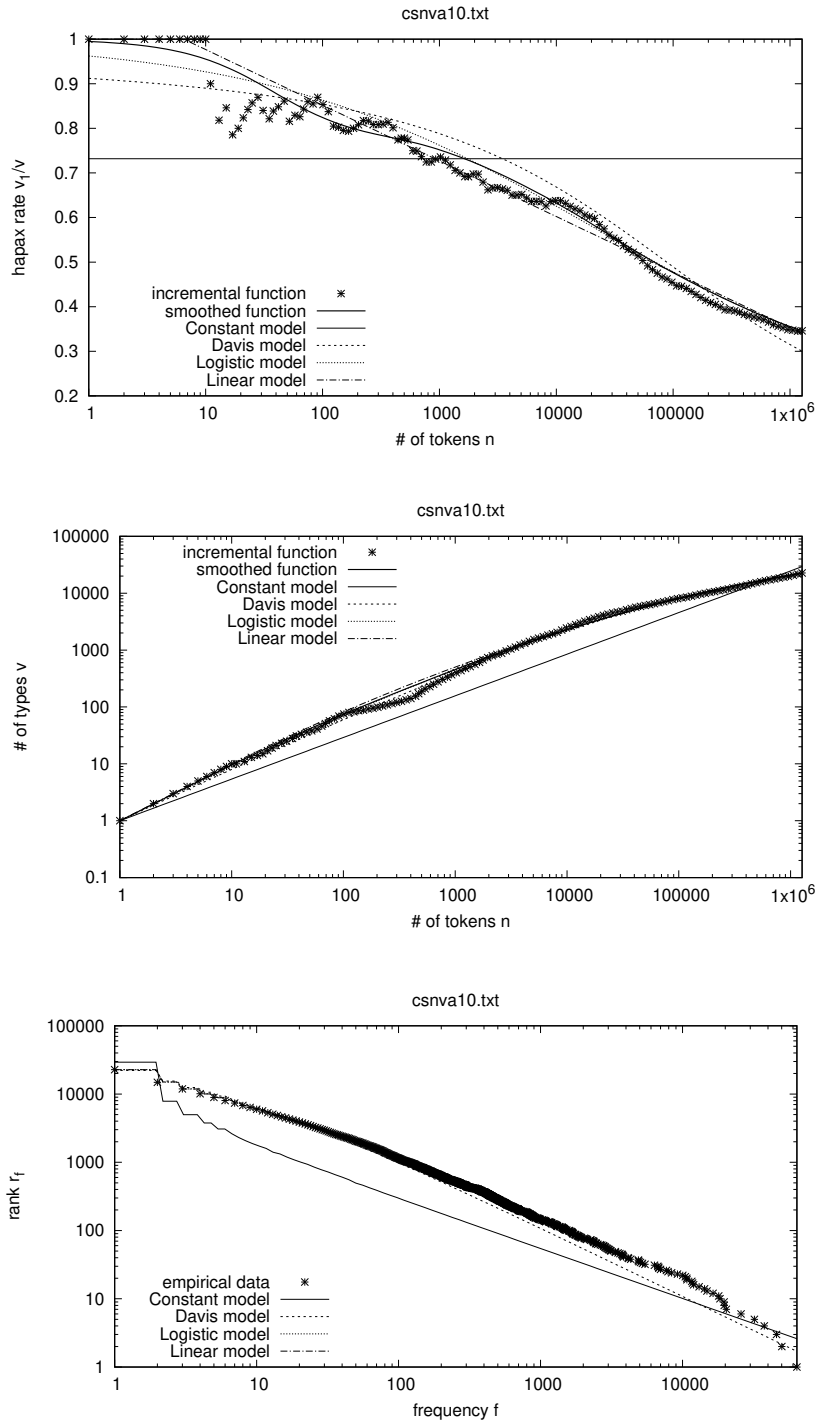


Figure 14: J. Casanova, *The Complete Memoirs*.

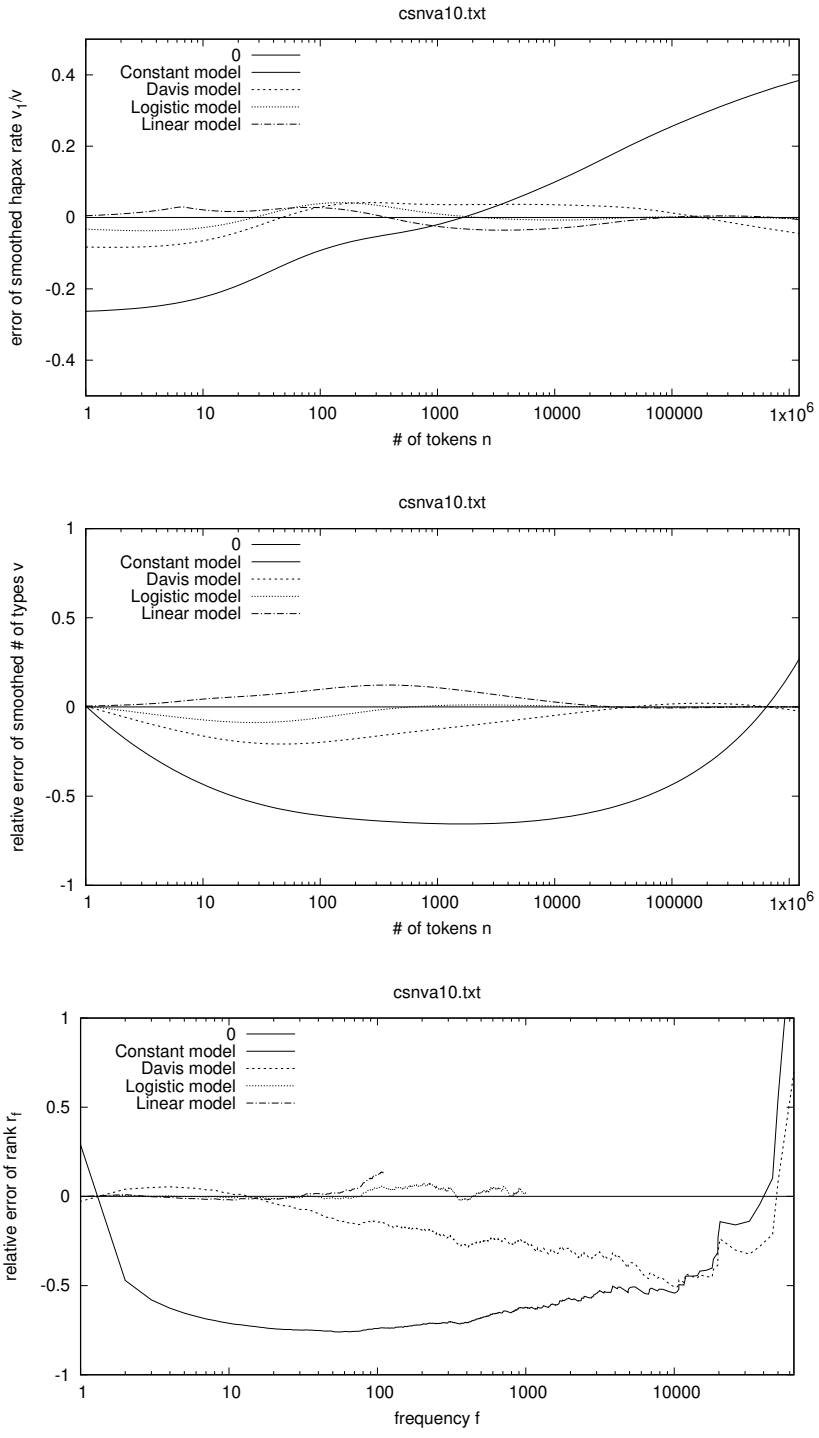


Figure 15: J. Casanova, *The Complete Memoirs*.

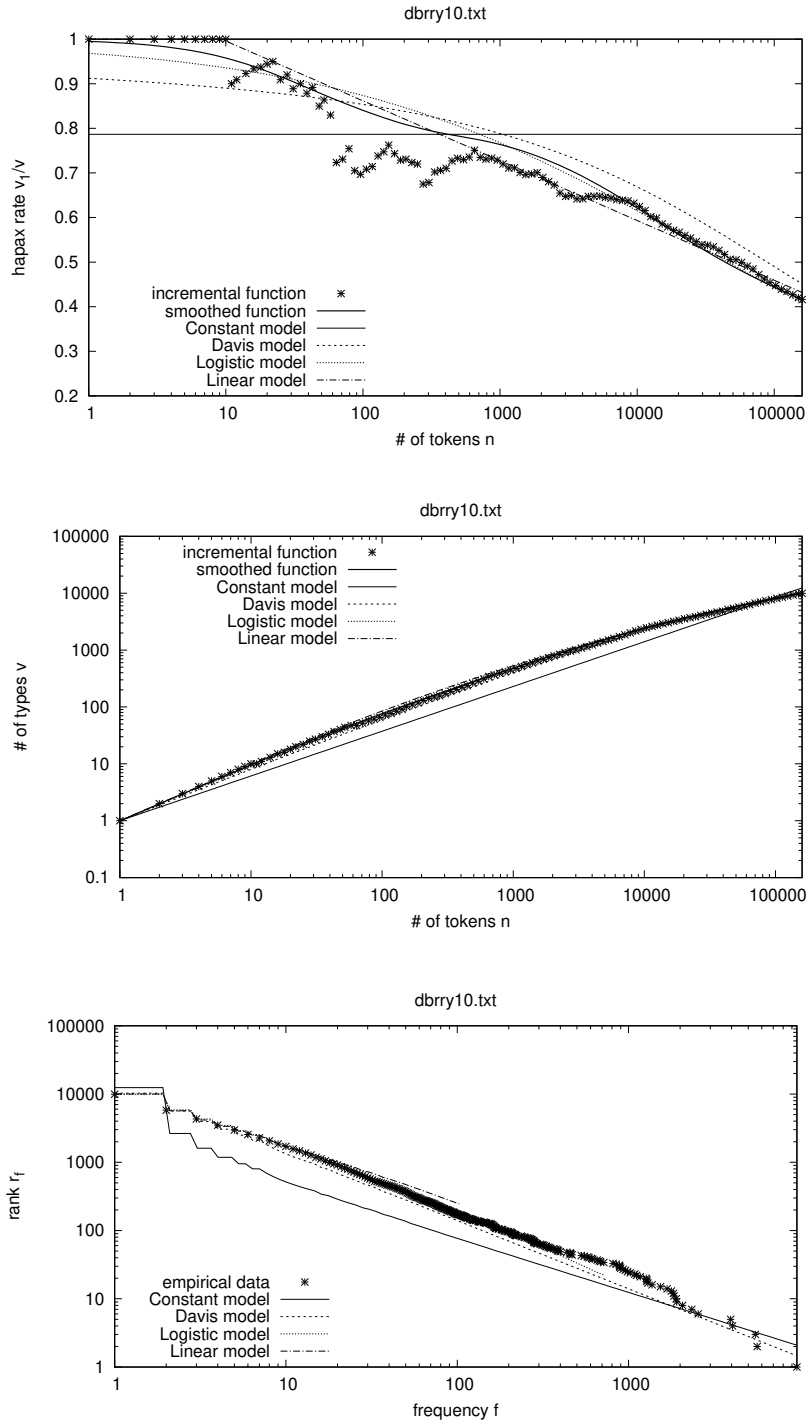


Figure 16: Comtesse du Barry, *Memoirs*.

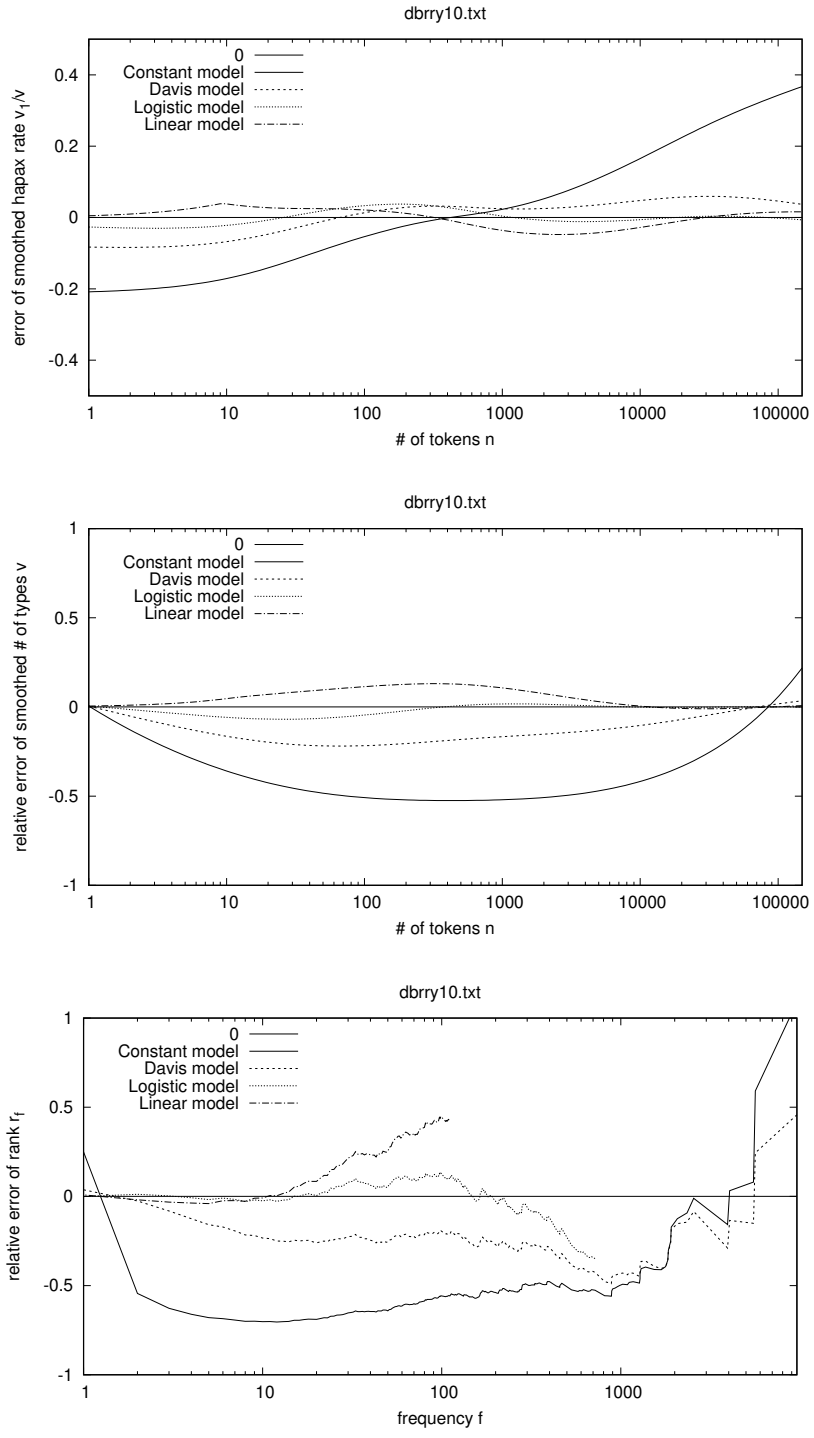


Figure 17: Comtesse du Barry, *Memoirs*.

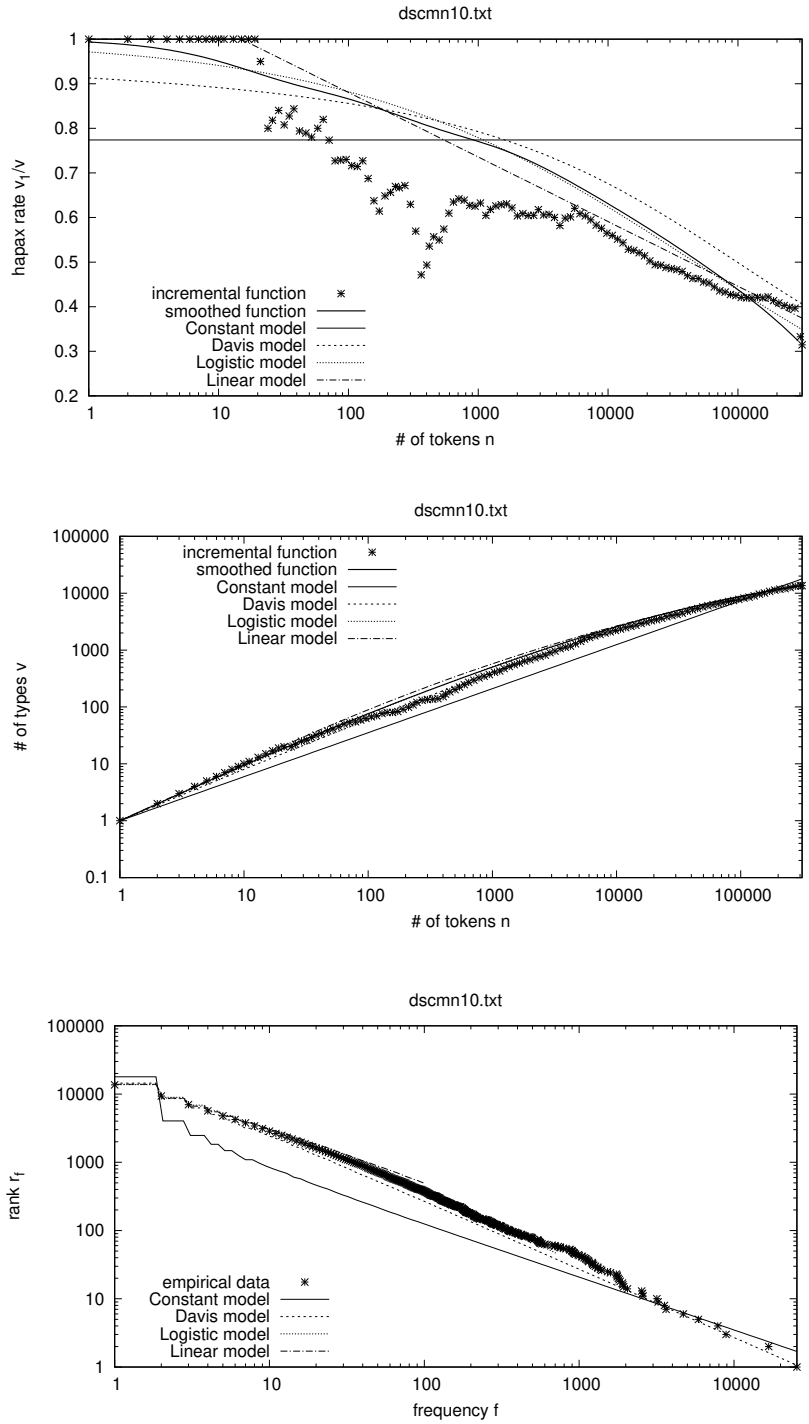


Figure 18: C. Darwin, *The Descent of Man*.

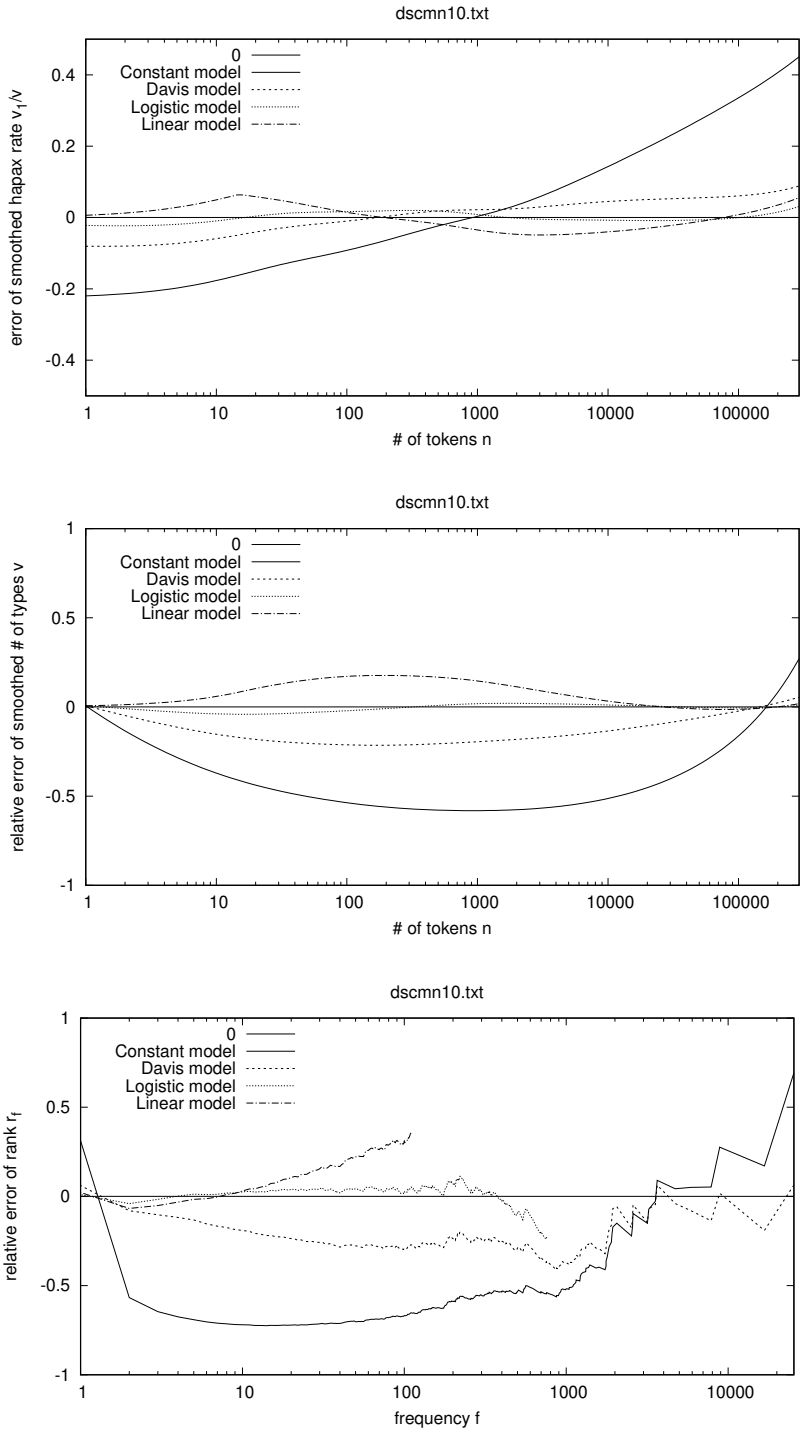


Figure 19: C. Darwin, *The Descent of Man*.

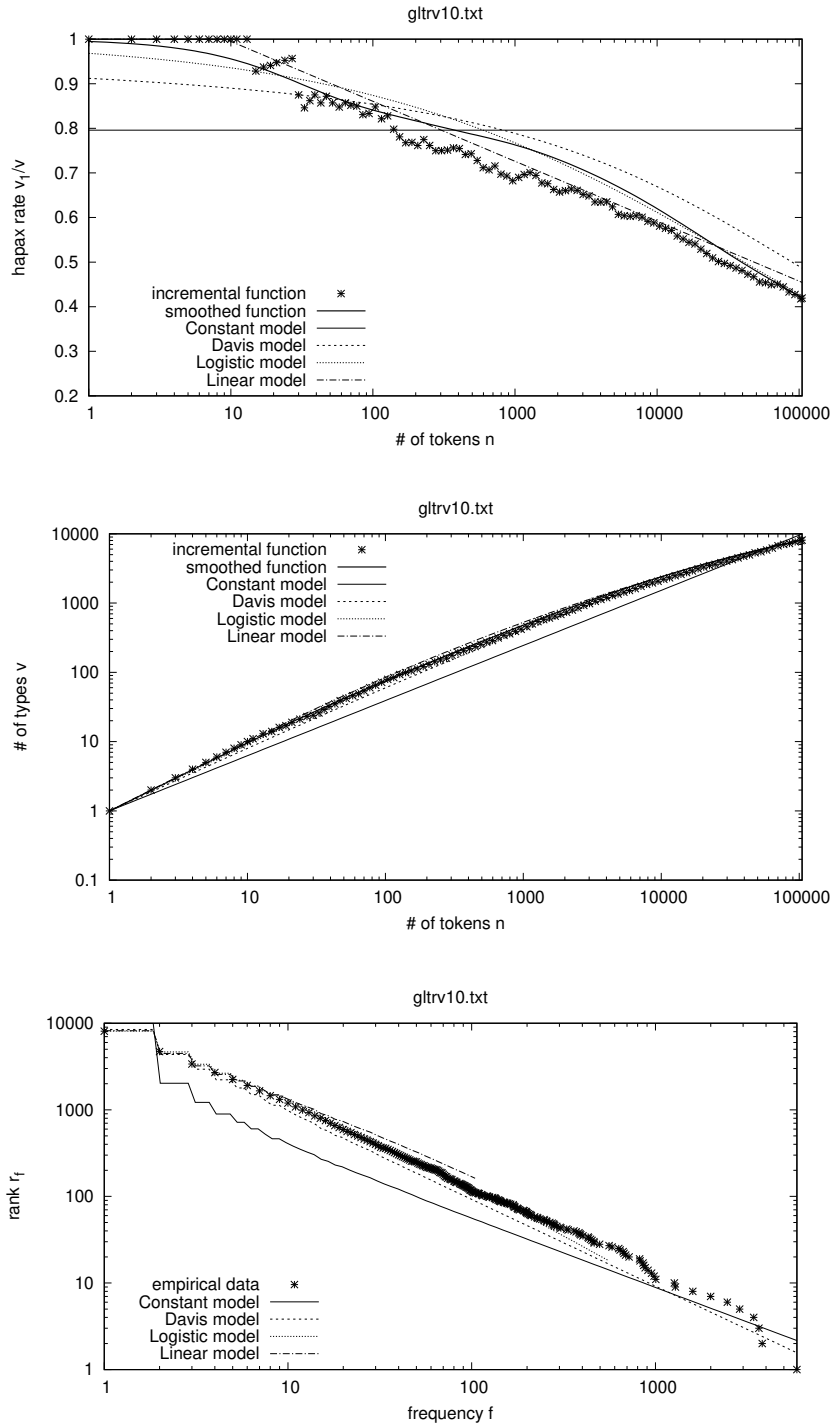
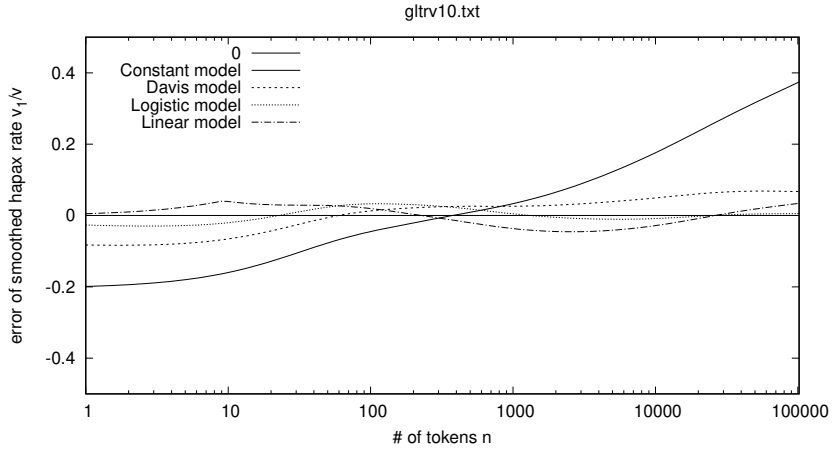
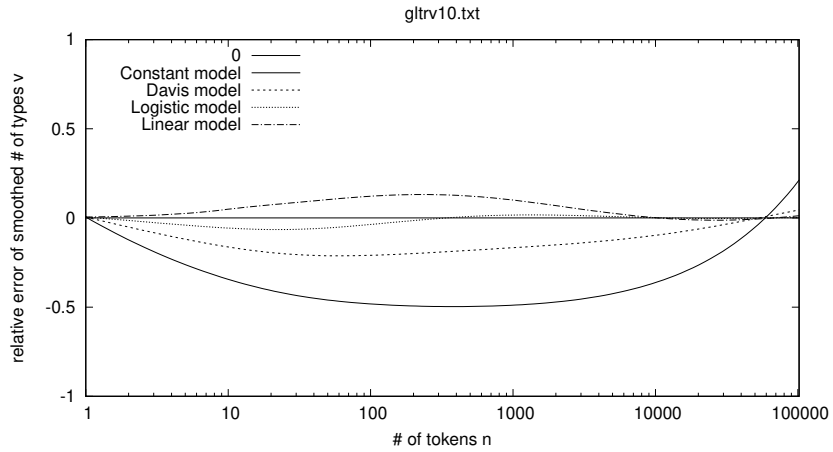


Figure 20: J. Swift, *Gulliver's Travels*.





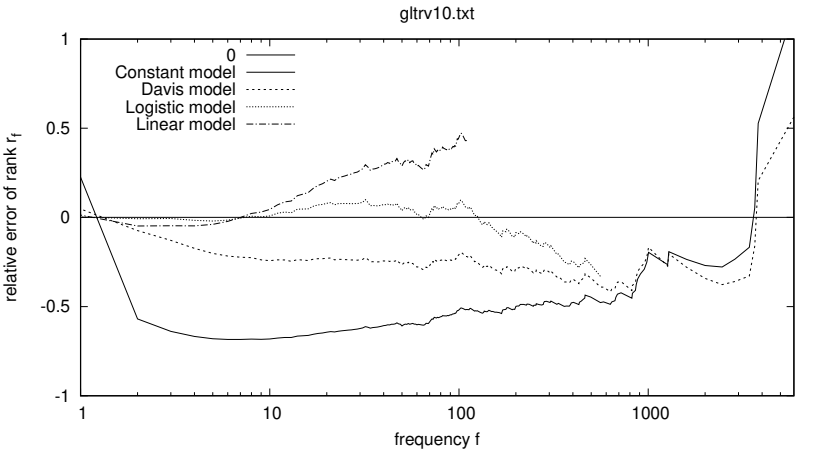


Figure 21: J. Swift, *Gulliver's Travels*.

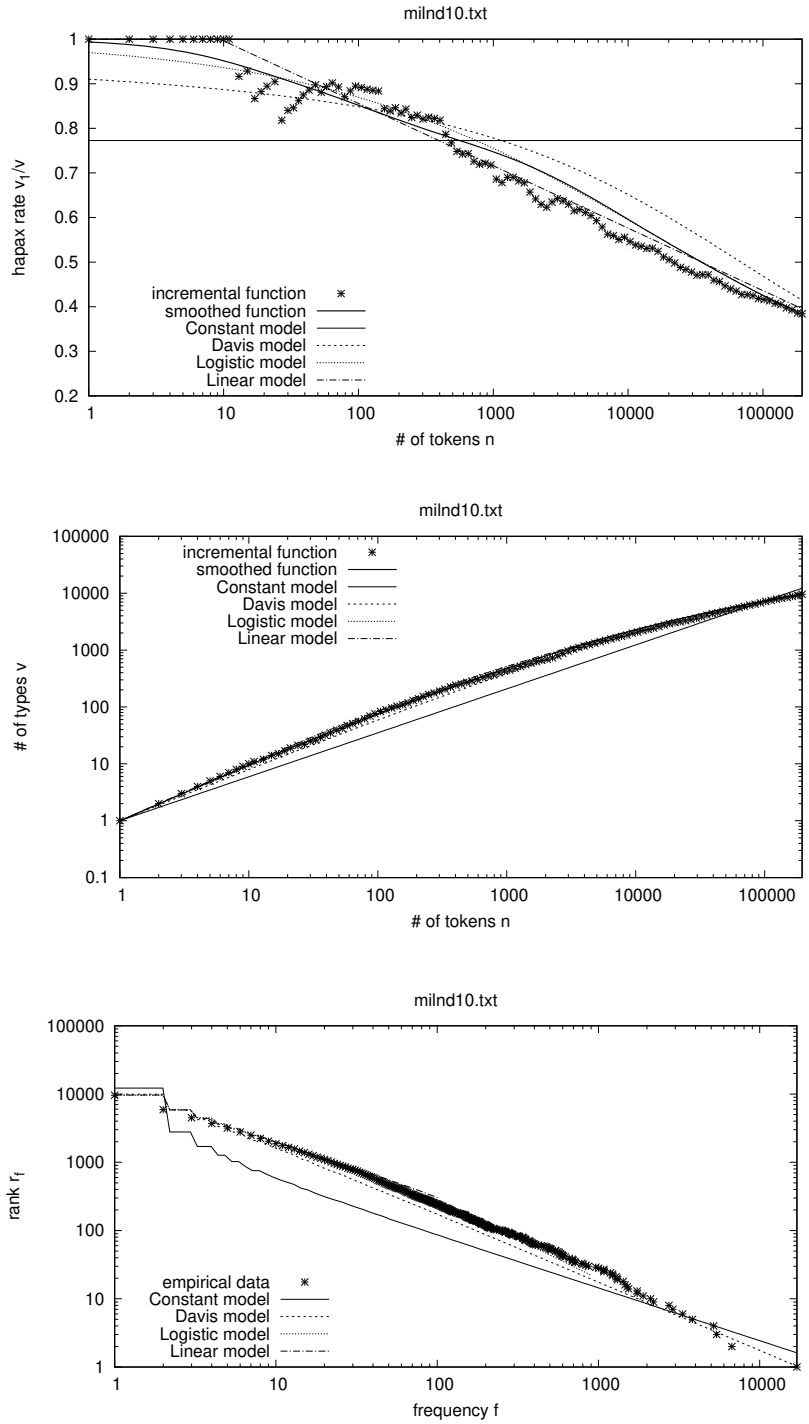
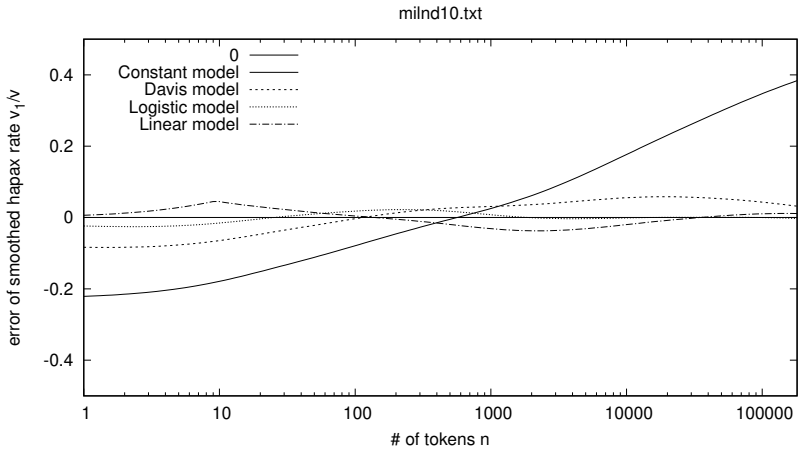
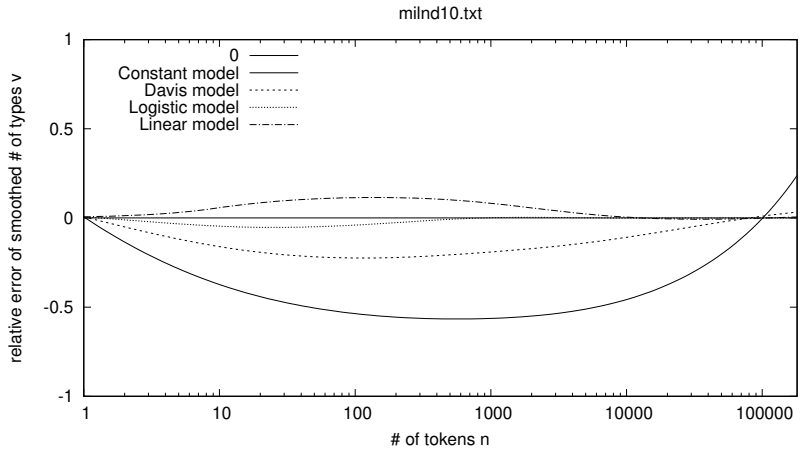


Figure 22: J. Verne, *The Mysterious Island*.





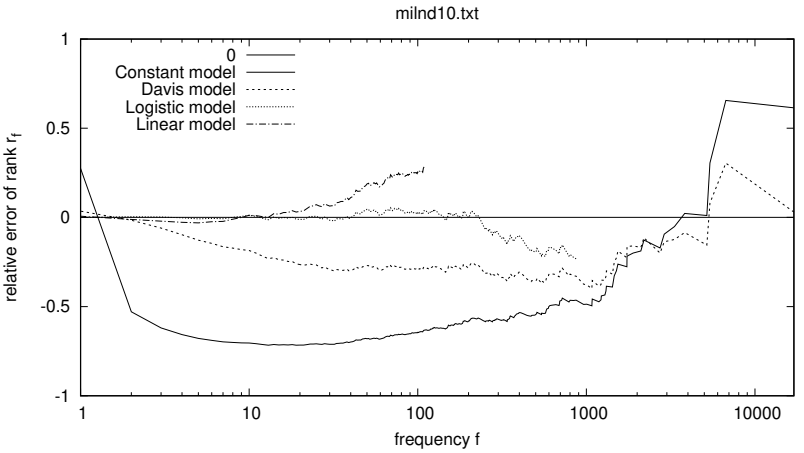
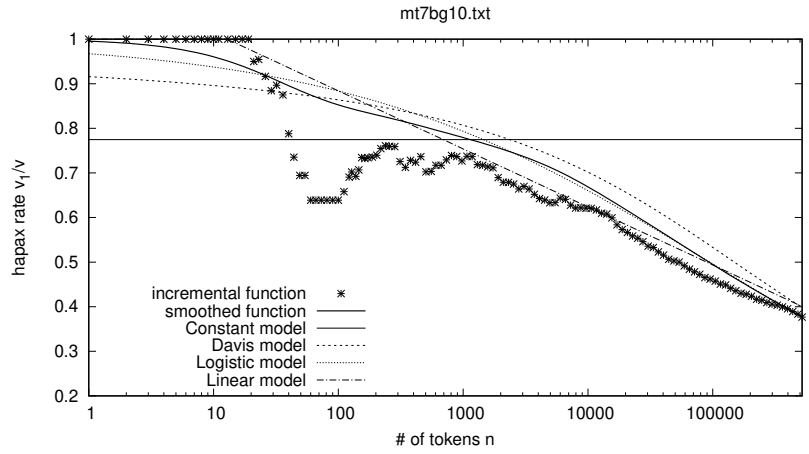
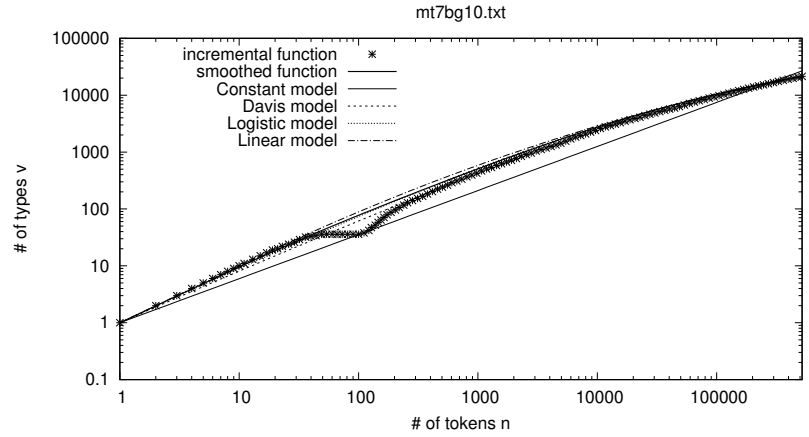


Figure 23: J. Verne, *The Mysterious Island*.





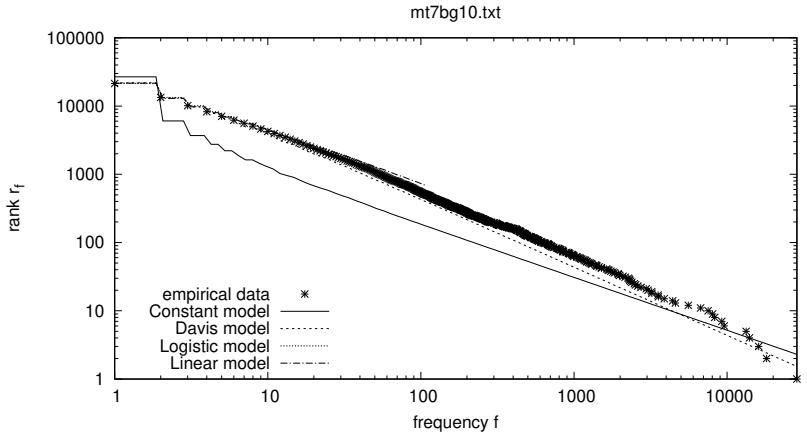


Figure 24: A. Paine, *Mark Twain, A Biography.*

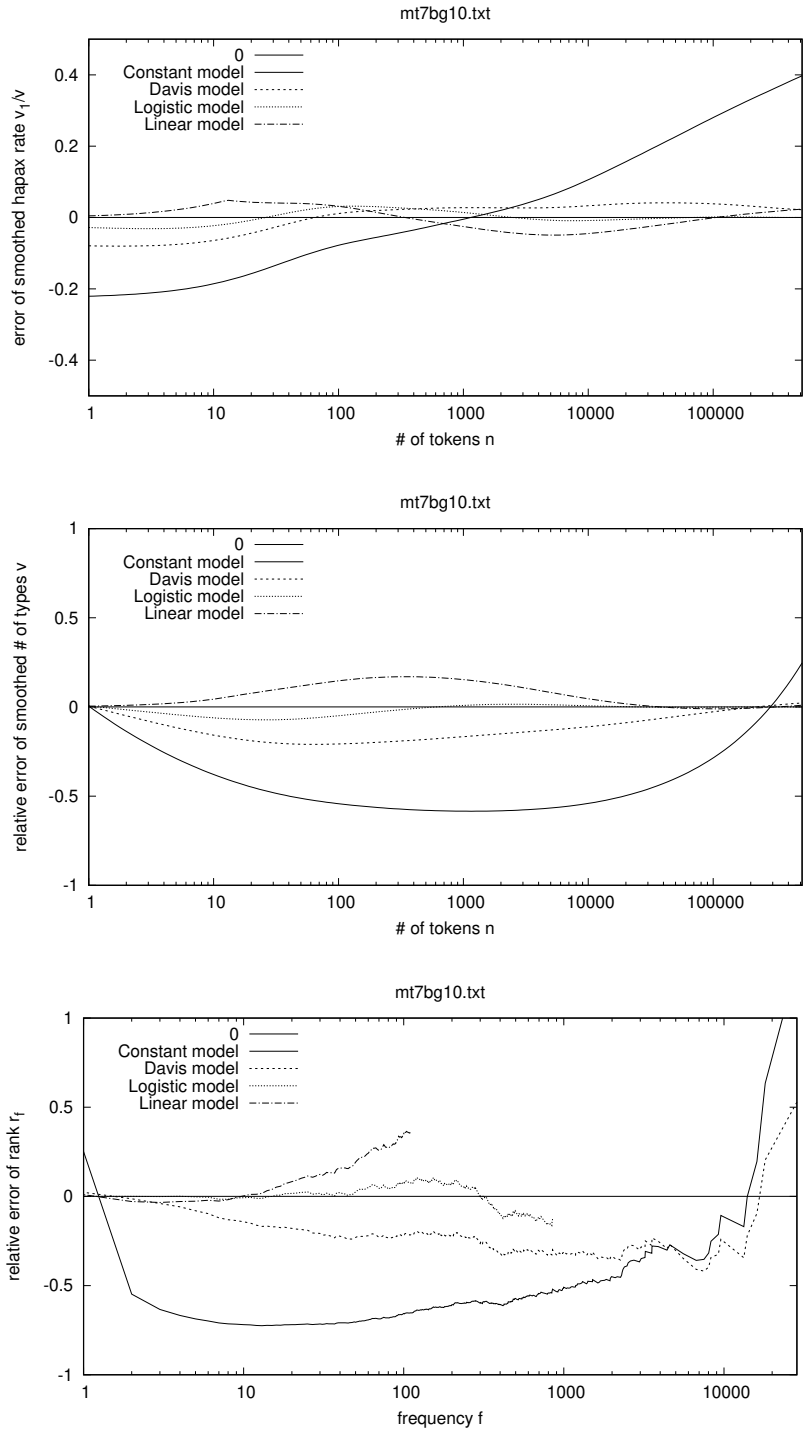


Figure 25: A. Paine, *Mark Twain, A Biography.*

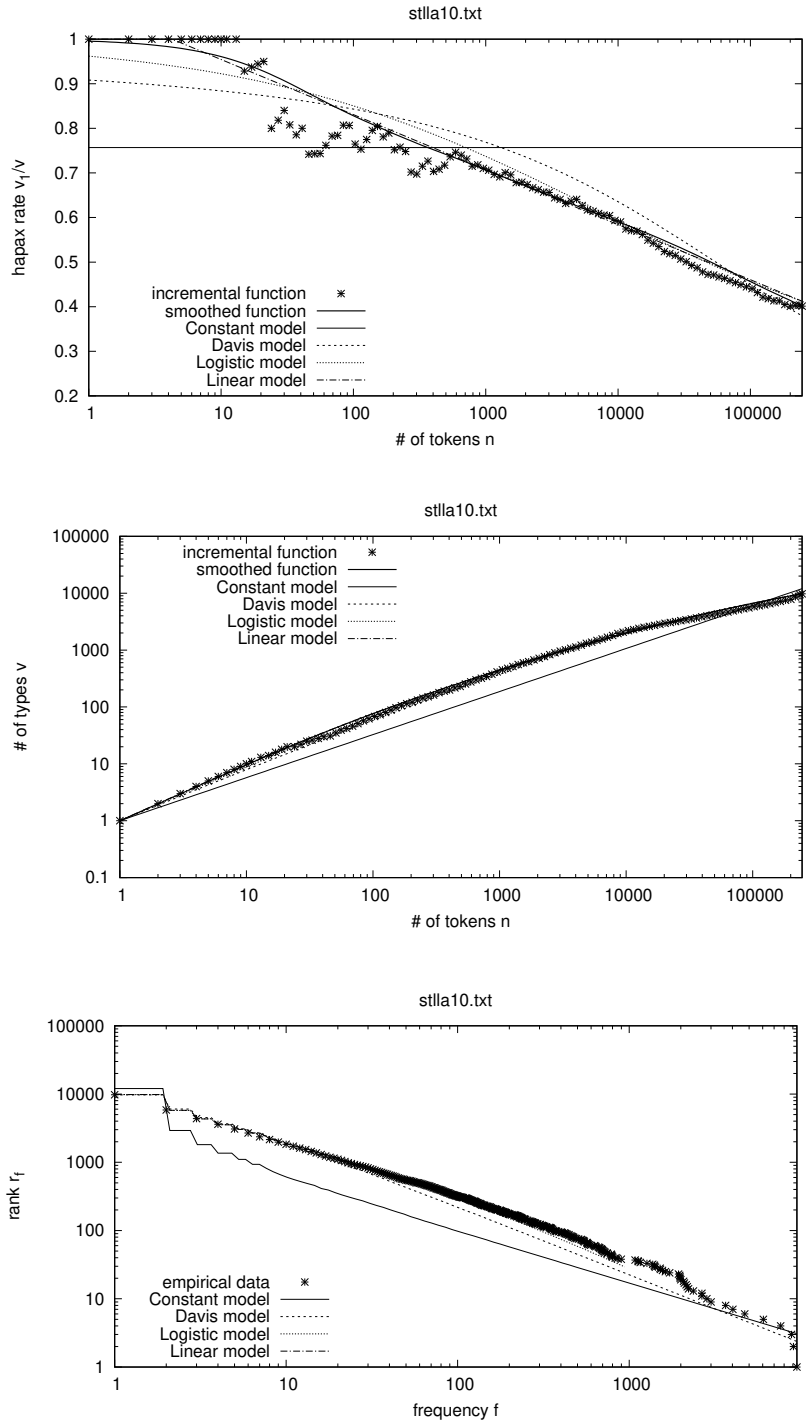


Figure 26: J. Swift, *The Journal to Stella*.

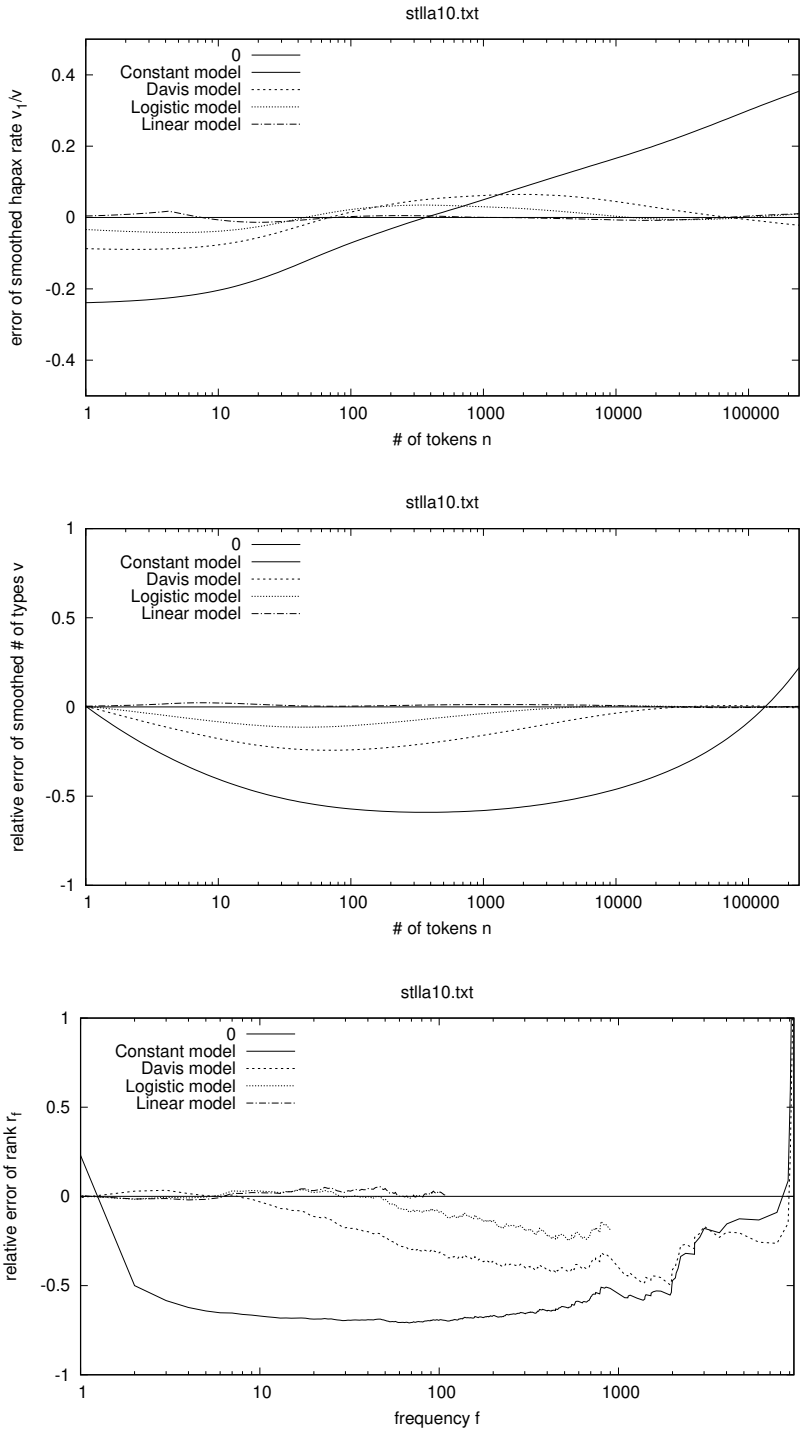


Figure 27: J. Swift, *The Journal to Stella*.

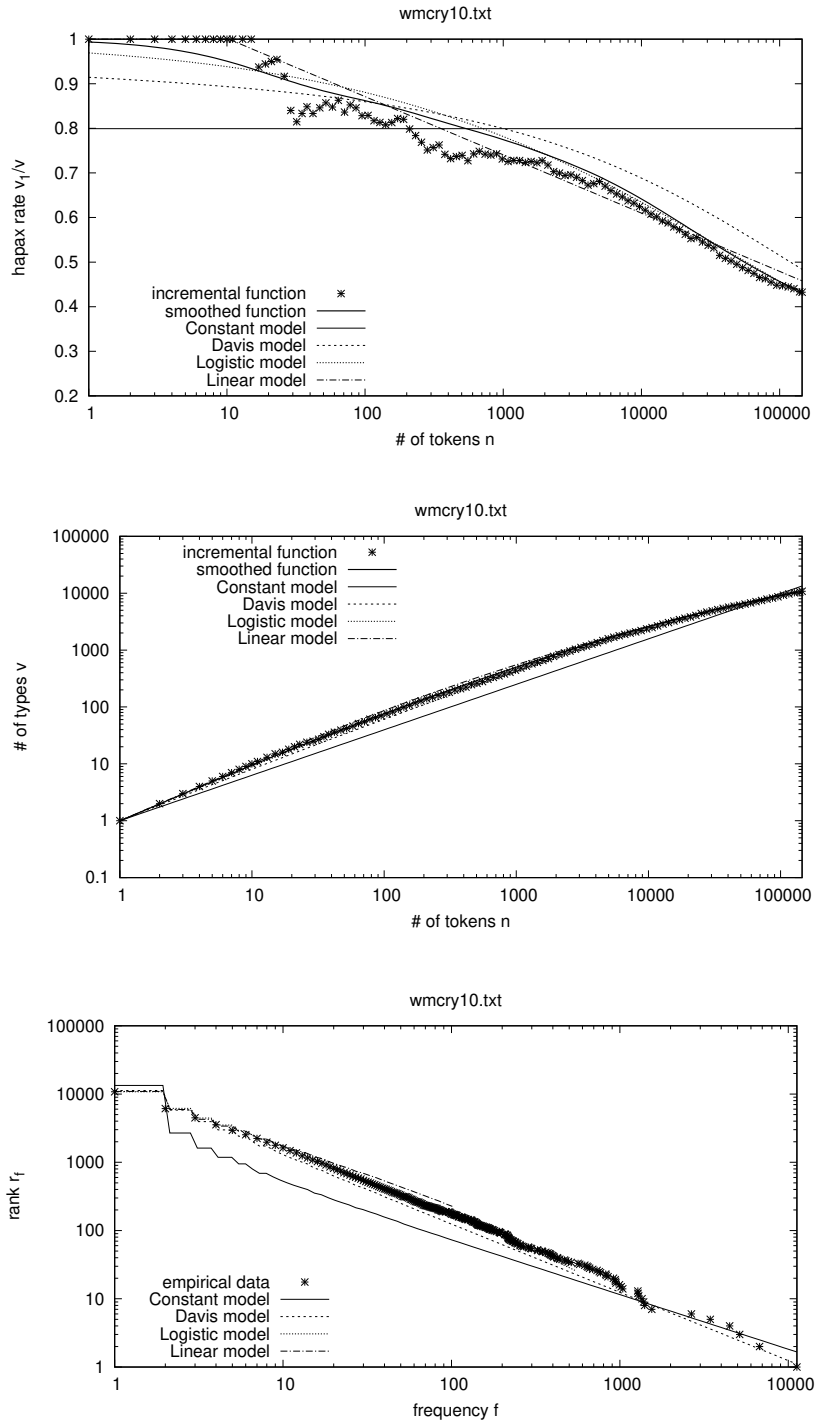


Figure 28: G. Smith, *Life of William Carey.*

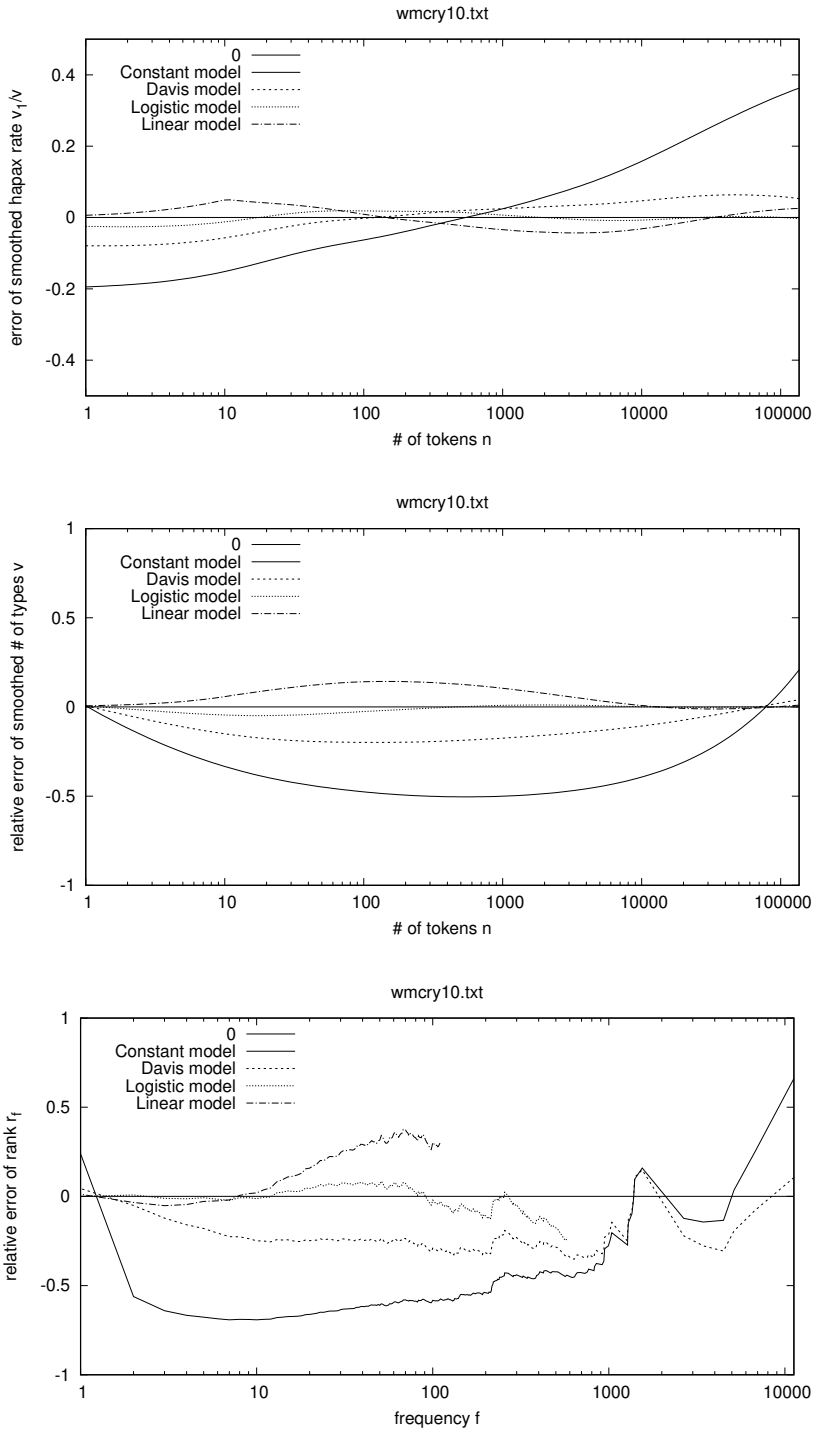


Figure 29: G. Smith, *Life of William Carey.*