Generating and Rendering Readability Scores for Project Gutenberg Texts

Ronald P. Reck
RRecktek LLC.
rreck@rrecktek.com
and
Ruth A. Reck
University of California
Davis, CA 95616 U.S.A.
rareck@ucdavis.edu

Abstract

Here the frequency distribution functions have been calculated for seven different types of readability measurements for over fourteen thousand texts from Project Gutenberg[1] (PG). Other supporting measurements were also obtained: the average characters per word, the words per sentence, and the syllables per word.

Three types of distributions have been demonstrated from the analysis of the metadata. While there are similarities among some of the scores, there is considerable interpretation yet to be made. The most complex and unique distribution function is found for the Flesch Reading Ease scores. Because of the computing intensity necessary to obtain these distributions it is only in the present age of information science that such a broad brush of characterization of a billion word data source can be made. It is essential that these be sorted by language to better interpret the meaning of the distributions.

1. Introduction

Various readability measurements can serve as indicators to quantify the relative accessibility of written information. However, domain specific attributes such as complex terminology or language can direct readability scores towards higher values than the actual complexity of the text warrants. For instance, scientific writing is likely to contain long words that may not significantly increase the complexity of the writing to those familiar with the terms but make the readability value appear greater. Despite this and other limitations, readability measurements remain useful attributes for describing text, especially when the values are regarded as relative measurements from within a specific type of writing or language. This paper reports the distribution of seven different readability measurements for over fourteen thousand texts from Project Gutenberg, a collection of free electronic books.

This effort creates the following types of readability measurements: (1) the Automated Readability Index; (2)Coleman-Liau formula; (3) Flesch Reading Ease Score; (4) the Gunning Fog Index; (5) the Flesch-Kincaid Score; (6) the Laesbarhedsindex (Lix) score; and (7) the SMOG score.

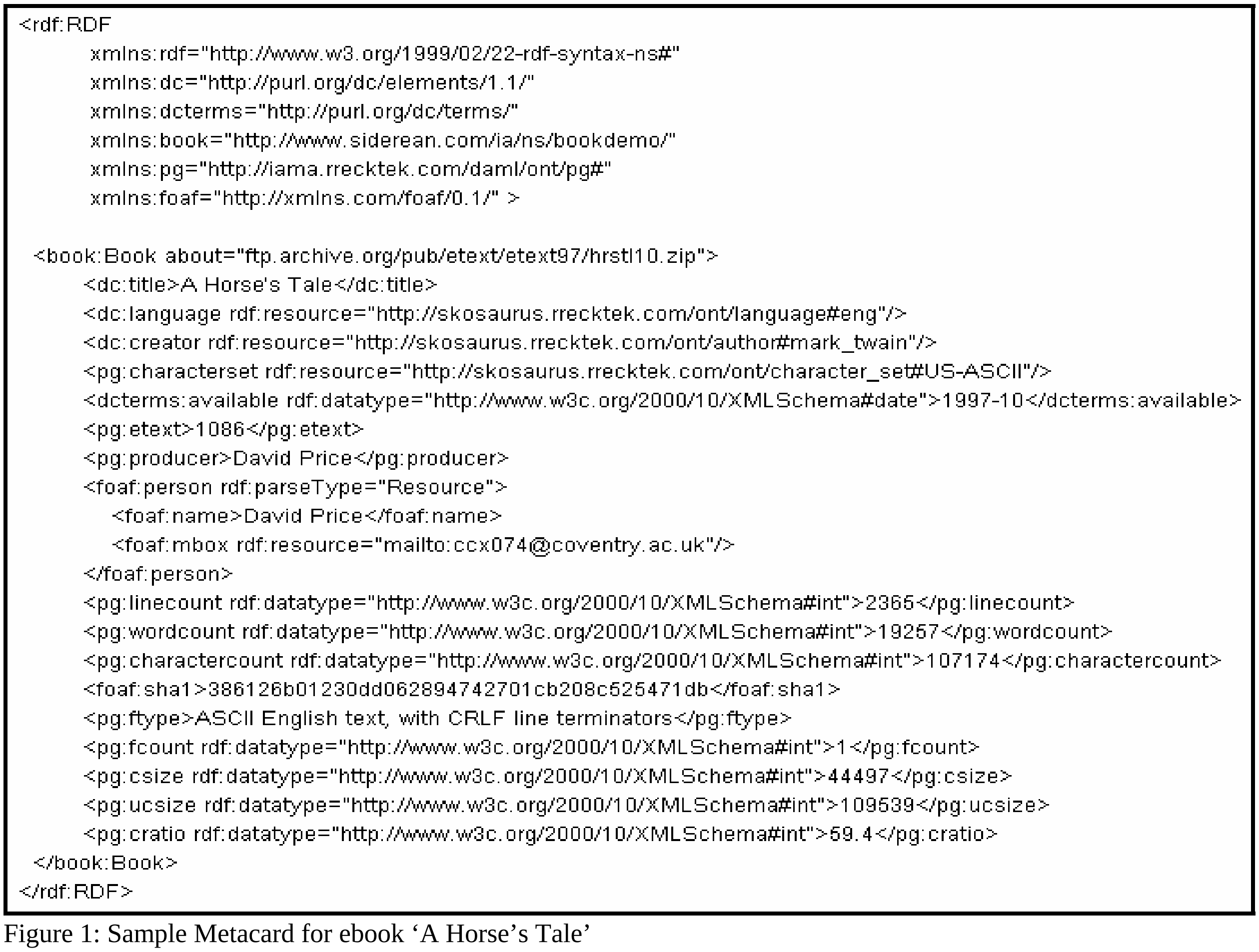
[1] Project Gutenberg is a library of thousands of free ebooks whose copyright has expired in the USA. It can be found at http://www.gutenberg.org

Work to identify readability began at least as far back as 1921 in The Teacher’s Word Book by Thorndike (Thorndike, 1921). Mathematical equations and word frequency were used to identify a measurement for book difficulty. This process was largely in response to teachers’ requests for science books that taught facts without being encumbered by vocabulary. As part of the ‘plain language movement’, it supported the idea that clear, unpretentious language can increase understanding.

In more recent times, efforts for readability have been used by the United States Navy (Kincaid et al, 1975) Enlisted personnel in training schools were tested to determine their comprehension level and then training manuals were designed to be within their comprehension levels.

2. Brief history of rendering Project Gutenberg metadata

Reck’s initial efforts for creating and articulating metadata that describes the Project Gutenberg repository were first documented in *Metadata Cards for Describing Project Gutenberg Texts* (Reck, 2006). That effort involved a process for creating as many as eighteen attributes for each of 15,511 PG texts thereby producing 912 thousand assertions. Substantial energy went to accommodating the wide range of variability and poor consistency of PG formats. A sample metacard, from that effort, which describes the attributes of “A Horse’s Tale” is shown in Figure 1.



```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:book="http://www.siderean.com/ia/ns/bookdemo/"
        xmlns:pg="http://iama.rrecktek.com/daml/ont/pg#"
        xmlns:foaf="http://xmlns.com/foaf/0.1/" >

  <book:Book about="ftp.archive.org/pub/etext/etext97/hrstl10.zip">
       <dc:title>A Horse's Tale</dc:title>
       <dc:language rdf:resource="http://skosaurus.rrecktek.com/ont/language#eng"/>
       <dc:creator rdf:resource="http://skosaurus.rrecktek.com/ont/author#mark_twain"/>
       <pg:characterset rdf:resource="http://skosaurus.rrecktek.com/ont/character_set#US-ASCII"/>
       <dcterms:available rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#date">1997-10</dcterms:available>
       <pg:etext>1086</pg:etext>
       <pg:producer>David Price</pg:producer>
       <foaf:person rdf:parseType="Resource">
          <foaf:name>David Price</foaf:name>
          <foaf:mbox rdf:resource="mailto:ccx074@coventry.ac.uk"/>
       </foaf:person>
       <pg:linecount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">2365</pg:linecount>
       <pg:wordcount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">19257</pg:wordcount>
       <pg:charactercount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">107174</pg:charactercount>
       <foaf:sha1>386126b01230dd062894742701cb208c525471db</foaf:sha1>
       <pg:ftype>ASCII English text, with CRLF line terminators</pg:ftype>
       <pg:fcount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">1</pg:fcount>
       <pg:csize rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">44497</pg:csize>
       <pg:ucsize rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">109539</pg:ucsize>
       <pg:cratio rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">59.4</pg:cratio>
  </book:Book>
</rdf:RDF>
```

Figure 1: Sample Metacard for ebook ‘A Horse’s Tale’

Soon after the initial metacards were complete, five thousand assertions describing the PG authors were leveraged in the XML2006 paper *Applying XQuery and OWL to The World Factbook, Wikipedia and Project Gutenberg* (Sall and Reck, 2006). That paper described a standards-based approach for answering the query “Who are the male European PG authors from the 19th century?” (Sall and Reck, 2006) The intersections between the three unrelated data sources can be seen in Figure 2. Figure 2 captures the intersection between the different sources of information and shows how a single search query can drive information interoperability. As a broad range of information is integrated, this becomes of greater value.

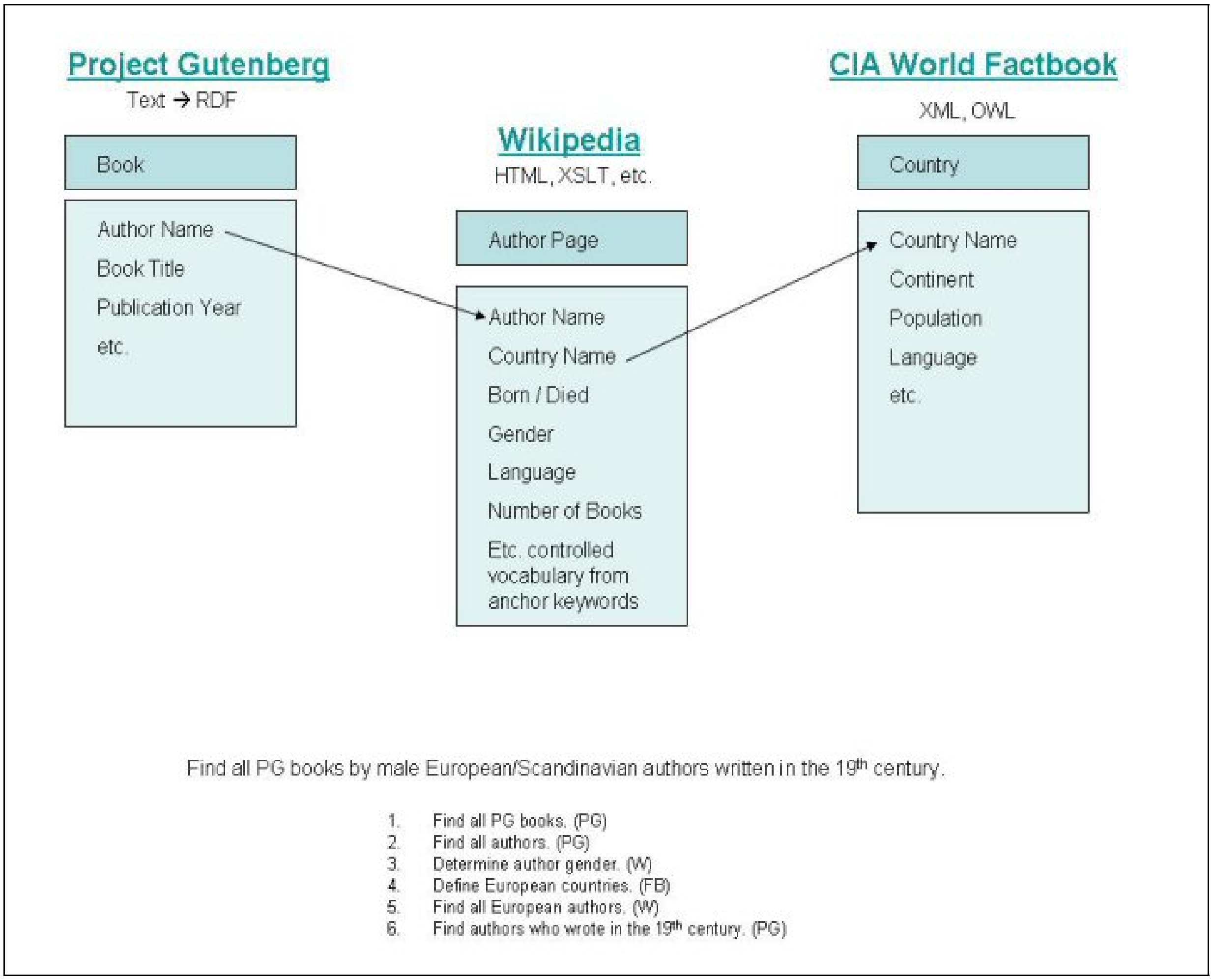
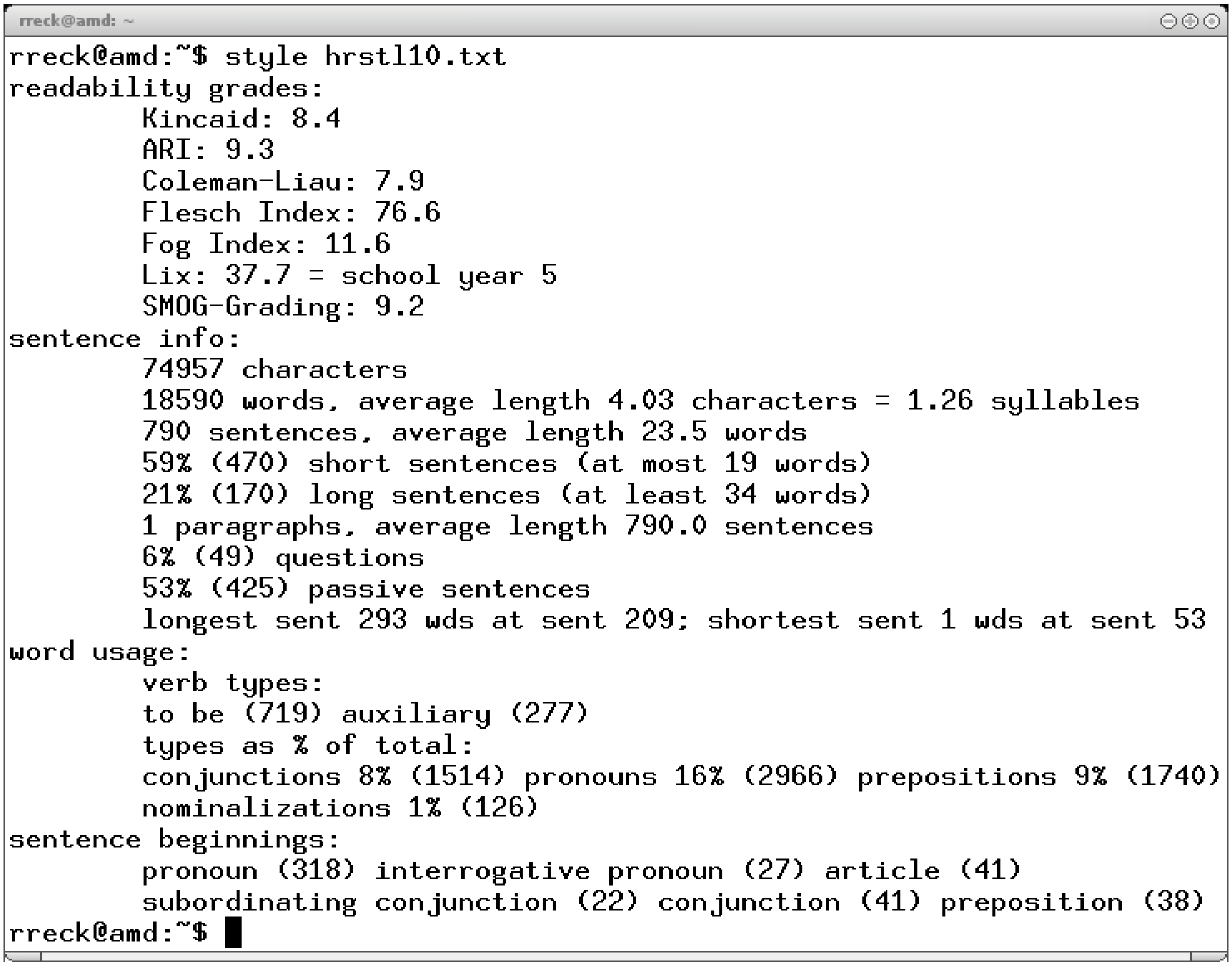

Project Gutenberg
Text → RDF
Book
Author Name
Book Title
Publication Year
etc.
Wikipedia
HTML, XSLT, etc.
Author Page
Author Name
Country Name
Born / Died
Gender
Language
Number of Books
Etc. controlled vocabulary from anchor keywords
CIA World Factbook
XML, OWL
Country
Country Name
Continent
Population
Language
etc.
Find all PG books by male European/Scandinavian authors written in the 19th century.
1. Find all PG books. (PG)
2. Find all authors. (PG)
3. Determine author gender. (W)
4. Define European countries. (FB)
5. Find all European authors. (W)
6. Find authors who wrote in the 19th century. (PG)

Figure 2: Associations among Three Unrelated Data Sources

After isolating the sixty-four European 19th century PG authors that answered the query, the metadata creation efforts were rekindled in the M.A. thesis, *Generating and Rendering String Frequency Measurements of Project Gutenberg Texts* (Reck, 2007). That effort yielded approximately 480 million machine readable assertions depicting the string frequencies for a billion string data set. The current effort built again on the metadata creation approach in a manner similar to the previous efforts.

3. Process and results

Readability scores for each of the data set were determined using the output of the UNIX ‘style’ command (version 1.02). This command is available under the GNU software license and the copyright is held by Michael Haardt Michael@moria.de. Currently the software’s homepage is http://www.gnu.org/software/diction/diction.html. The software was compiled and run under Ubuntu Linux. (Ubuntu is a derivative of Debian Linux). Sample output from the ‘style’ command being run on ‘A Horse’s Tale’ can be can be seen in Figure 3.

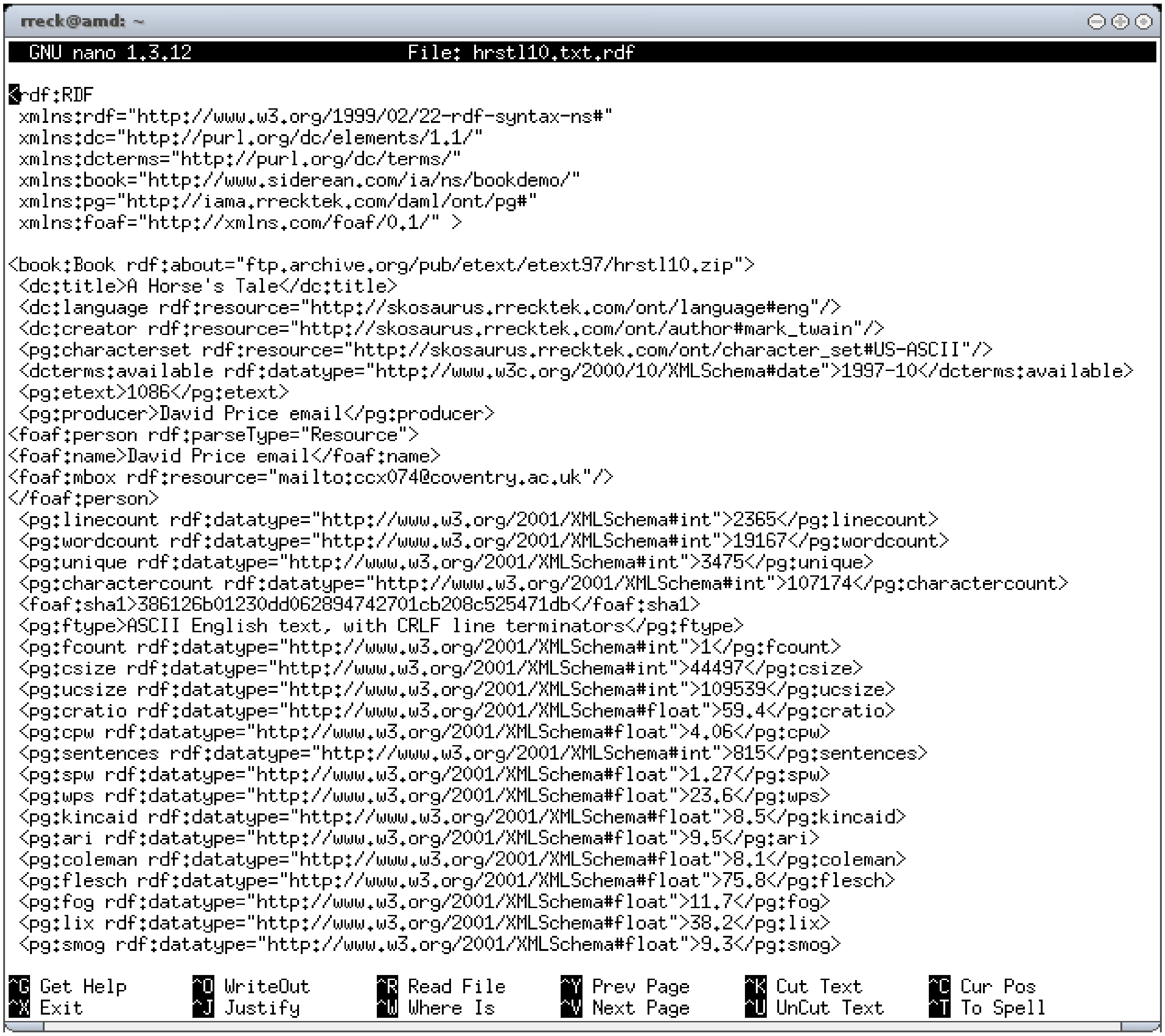


```
rreck@amd:~$ style hrstl10.txt
readability grades:
        Kincaid: 8.4
        ARI: 9.3
        Coleman-Liau: 7.9
        Flesch Index: 76.6
        Fog Index: 11.6
        Lix: 37.7 = school year 5
        SMOG-Grading: 9.2
sentence info:
        74957 characters
        18590 words, average length 4.03 characters = 1.26 syllables
        790 sentences, average length 23.5 words
        59% (470) short sentences (at most 19 words)
        21% (170) long sentences (at least 34 words)
        1 paragraphs, average length 790.0 sentences
        6% (49) questions
        53% (425) passive sentences
        longest sent 293 wds at sent 209; shortest sent 1 wds at sent 53
word usage:
        verb types:
        to be (719) auxiliary (277)
        types as % of total:
        conjunctions 8% (1514) pronouns 16% (2966) prepositions 9% (1740)
        nominalizations 1% (126)
sentence beginnings:
        pronoun (318) interrogative pronoun (27) article (41)
        subordinating conjunction (22) conjunction (41) preposition (38)
rreck@amd:~$ █
```

Figure 3: Output from the ‘style’ command’s analysis of ‘A Horse’s Tale’.

Here we see the information generated in an easy to read format beginning with measures of the readability of the text and going on to cover a number of other physical aspects describing the text. The results from the ‘style’ command were next parsed using Perl and then formatted in a version of XML designed for metadata presentation called Resource Descriptive Framework[2] (RDF). RDF is a data model design specifically for the articulation of metadata for machine comprehension. One format for presenting RDF is in XML (also called RDFXML). An example of this presentation is shown in Figure 4.

[2] Resource Descriptive Framework - The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. http://www.w3.org/RDF/



```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:book="http://www.siderean.com/ia/ns/bookdemo/"
 xmlns:pg="http://iama.rrecktek.com/daml/ont/pg#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" >

<book:Book rdf:about="ftp.archive.org/pub/etext/etext97/hrstl10.zip">
 <dc:title>A Horse's Tale</dc:title>
 <dc:language rdf:resource="http://skosaurus.rrecktek.com/ont/language#eng"/>
 <dc:creator rdf:resource="http://skosaurus.rrecktek.com/ont/author#mark_twain"/>
 <pg:characterset rdf:resource="http://skosaurus.rrecktek.com/ont/character_set#US-ASCII"/>
 <dcterms:available rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#date">1997-10</dcterms:available>
 <pg:etext>1086</pg:etext>
 <pg:producer>David Price email</pg:producer>
<foaf:person rdf:parseType="Resource">
<foaf:name>David Price email</foaf:name>
<foaf:mbox rdf:resource="mailto:ccx074@coventry.ac.uk"/>
</foaf:person>
 <pg:linecount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2365</pg:linecount>
 <pg:wordcount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">19167</pg:wordcount>
 <pg:unique rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3475</pg:unique>
 <pg:charactercount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">107174</pg:charactercount>
 <foaf:sha1>386126b01230dd062894742701cb208c525471db</foaf:sha1>
 <pg:ftype>ASCII English text, with CRLF line terminators</pg:ftype>
 <pg:fcount rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</pg:fcount>
 <pg:csize rdf:datatype="http://www.w3.org/2001/XMLSchema#int">44497</pg:csize>
 <pg:ucsize rdf:datatype="http://www.w3.org/2001/XMLSchema#int">109539</pg:ucsize>
 <pg:cratio rdf:datatype="http://www.w3.org/2001/XMLSchema#float">59.4</pg:cratio>
 <pg:cpw rdf:datatype="http://www.w3.org/2001/XMLSchema#float">4.06</pg:cpw>
 <pg:sentences rdf:datatype="http://www.w3.org/2001/XMLSchema#int">815</pg:sentences>
 <pg:spw rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.27</pg:spw>
 <pg:wps rdf:datatype="http://www.w3.org/2001/XMLSchema#float">23.6</pg:wps>
 <pg:kincaid rdf:datatype="http://www.w3.org/2001/XMLSchema#float">8.5</pg:kincaid>
 <pg:ari rdf:datatype="http://www.w3.org/2001/XMLSchema#float">9.5</pg:ari>
 <pg:coleman rdf:datatype="http://www.w3.org/2001/XMLSchema#float">8.1</pg:coleman>
 <pg:flesch rdf:datatype="http://www.w3.org/2001/XMLSchema#float">75.8</pg:flesch>
 <pg:fog rdf:datatype="http://www.w3.org/2001/XMLSchema#float">11.7</pg:fog>
 <pg:lix rdf:datatype="http://www.w3.org/2001/XMLSchema#float">38.2</pg:lix>
 <pg:smog rdf:datatype="http://www.w3.org/2001/XMLSchema#float">9.3</pg:smog>
```

Figure 4: Sample Metacard for ‘A Horse’s Tale including readability measurements.

Seven readability scores can be seen at the bottom of Figure 4. The measurements that were captured and preserved are briefly explained here:

3.1 The Automated Readability Index.

The Automated Readability Index (ARI) is an approximate representation of the U.S. grade level needed to comprehend the text. It relies on characters per word instead of syllables per word which distinguishes this measurement from other types of readability measurements. It is easier to calculate accurately since determining the number of characters is easier than determining syllables. It is typically higher in value than the Kincaid and Coleman-Liau measures, but lower than the Flesch. The formula used for determining the ARI is:

$$ARI=4.71*characters/words+0.5*words/sentences-21.43$$

The XML tag <pg:ari> was used to denote the ARI value as a floating point value in the metacard. Approximately 1500 texts had ARI scores of nine. A plot of the distribution of ARI scores in the Project Gutenberg data set is shown in Figure 5.

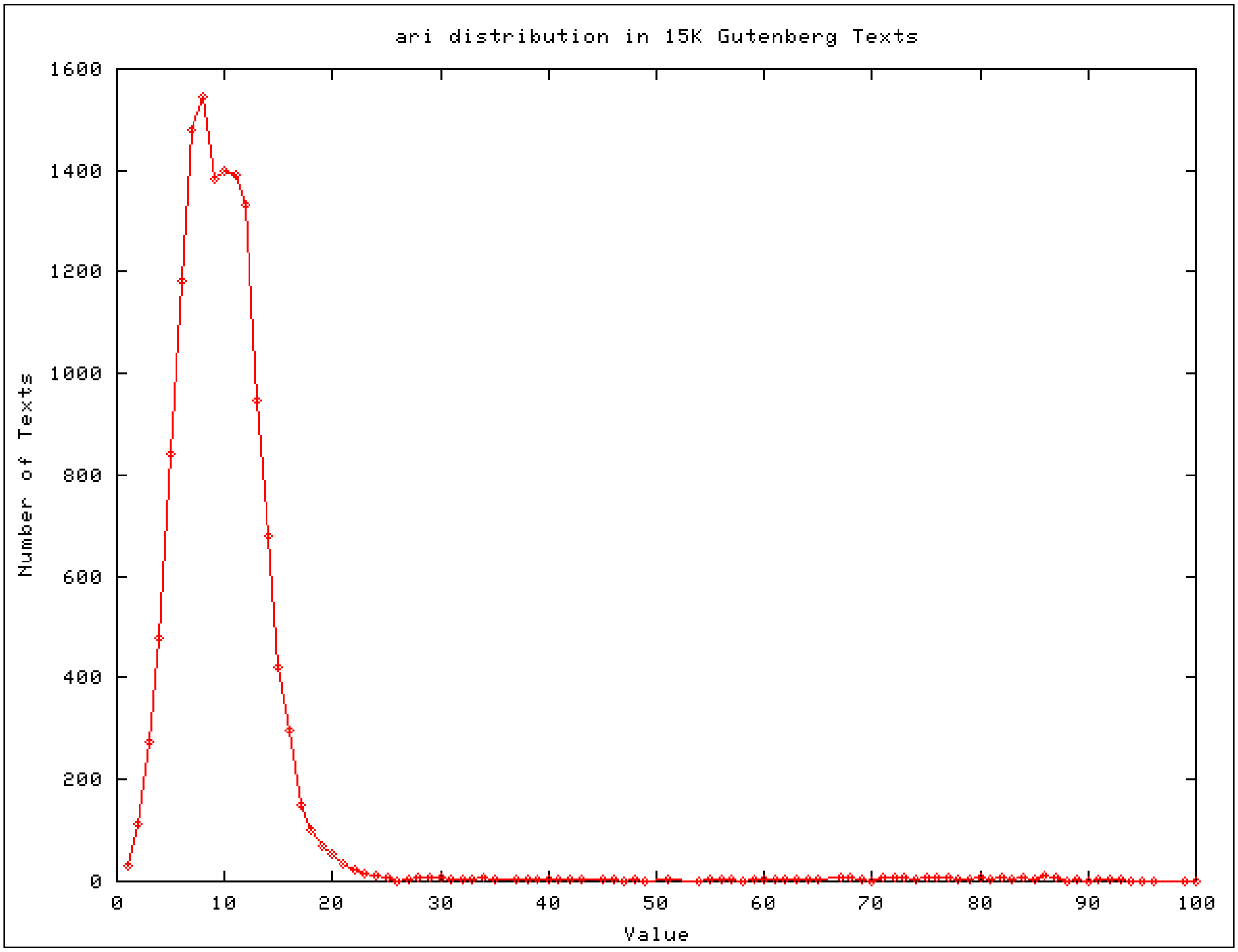


Figure 5: Distribution of ARI scores across the Project Gutenberg data set.

Here we see a narrow range of ARI scores from one to approximately twenty-five for the 15K Gutenberg texts. The distribution of ARI values appears to have a maximum at ARI around eight, followed by a slight secondary extremum at eleven with a sharply falling tail for larger values of ARI, appearing to reach an asymptotic limit of zero at about twenty-five. The significance of the details of this distribution has not been identified yet.

3.2 The Coleman-Liau Formula

This readability test was designed by Meri Coleman and T. L. Liau. The measurement is an approximate U. S. grade level needed to comprehend the text. It relies on the number of characters in the words instead of syllables, exactly like the Automated Readability Index. The Coleman-Liau Formula usually gives a lower grade value than any of the Kincaid, ARI and Flesch values when applied to technical documents. The formula used for calculating the Coleman-Liau Index is:

Coleman-Liau=5.89*characters/words-0.3*sentences/(100*words)-15.8

The XML tag <pg:coleman> was used to denote the Coleman-Liau value as a floating point value in the metacard. The majority of Project Gutenberg texts had a Coleman-Liau score of ten. A plot of the distribution of Coleman-Liau scores in the Project Gutenberg data set is shown in Figure 6.

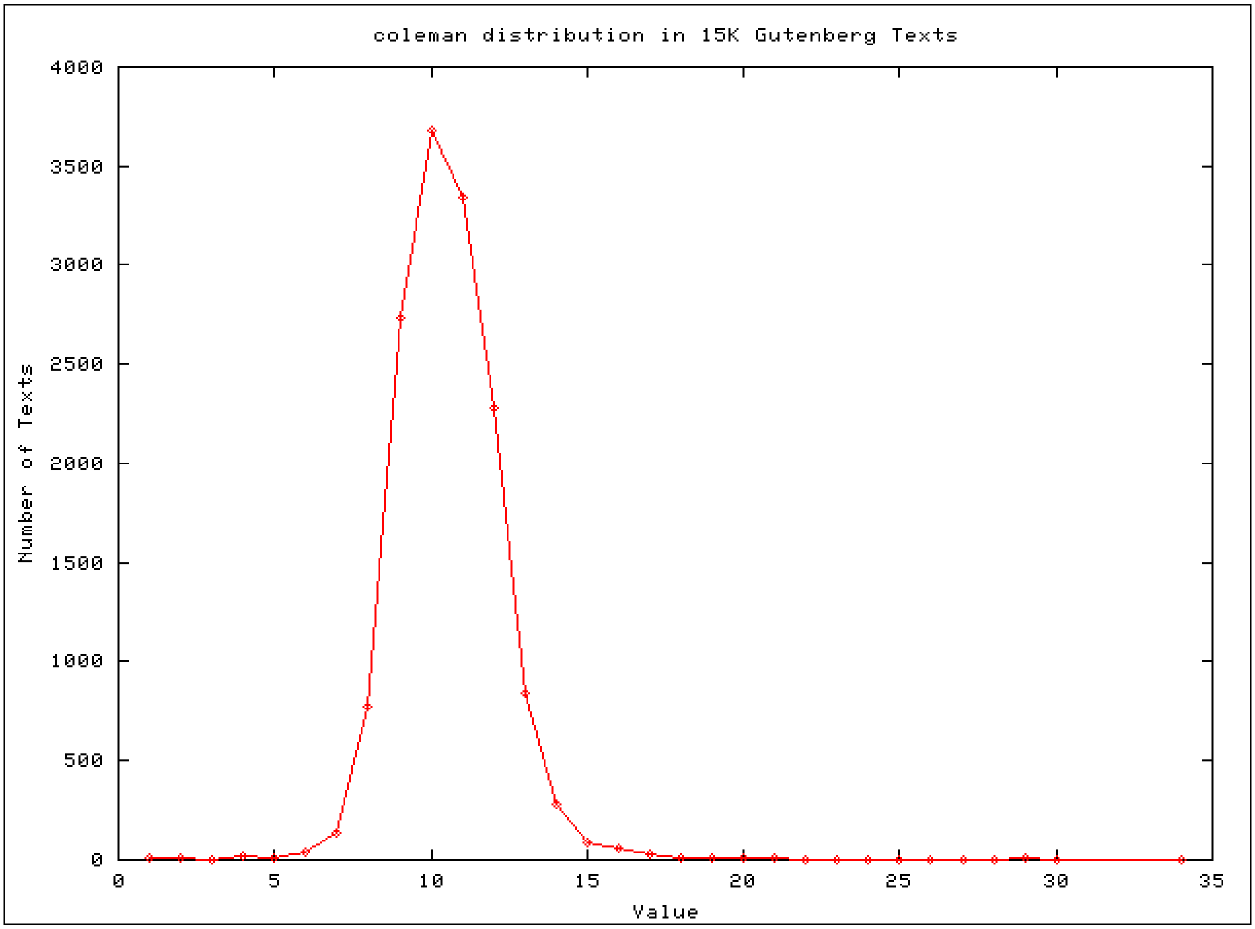


Figure 6: Distribution of Coleman-Liau scores across the Project Gutenberg data set.

Here we see a rather smoothly distributed set of Coleman-Liau scores for all Gutenberg texts, with a range of values from five to eighteen. This distribution has but one extremum. The equation used to define this score contains two parameters having values 5.89 and 15.8 and it is these values which restrict the range of values within the distribution.

3.3 The Flesh Reading Ease Formula

The Flesh Reading Ease Formula has been developed by Flesh in 1948 and it is based on school texts covering grades three to twelve. The index is usually between 0 (hard) and 100 (easy). This orientation contrasts with some of the other readability measurements since higher scores mean easier reading. This test is often used to assess adult reading materials; in fact it is used by some United States government agencies and the United States Department of Defence as an indicator of readability. The formula used for the Flesch Reading Ease is:

Flesch Index=206.835-84.6*syllables/words-1.015*words/sentences

The XML tag <pg:flesch> was used to denote the Flesch Reading Ease value as a floating point value in the metacard. The majority of Project Gutenberg texts had a Flesch Reading Ease scores near eighty-one. A plot of the distribution of Flesch Reading Ease scores in the Project Gutenberg data set is shown in Figure 7.

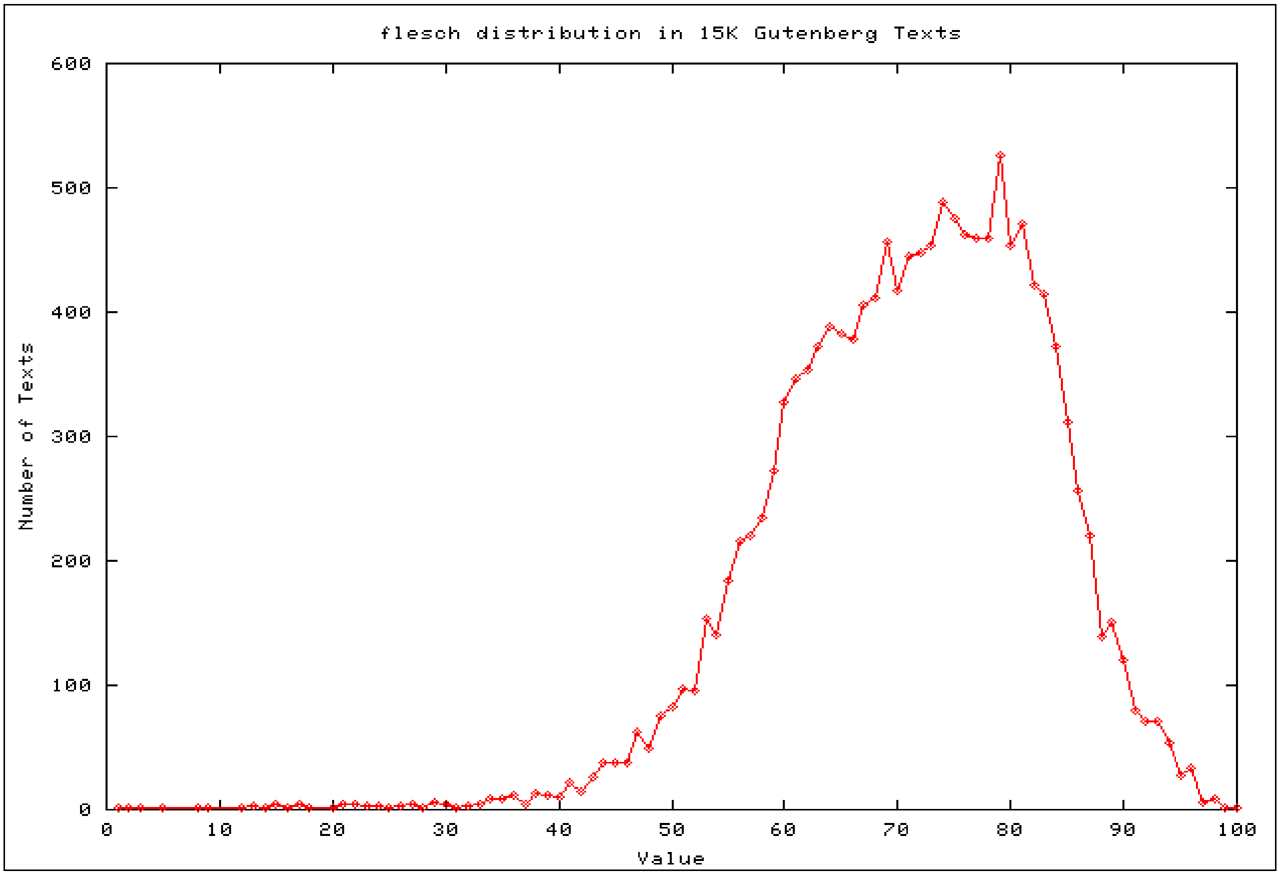


Figure 7: Distribution of Flesch Reading Ease scores across the Project Gutenberg data set.

This distribution is considerably broader than the others that were obtained and is probably a better measure for some particular uses. Note the steepness of the right side of the distribution and the not so steep left side of the distribution. Very few manuscripts had a value below fifty.

3.4 Gunning Fog Index

The Gunning Fog Index has been developed by Robert Gunning in 1952. The Gunning Fox number is the number of years of formal education that a person requires to easily comprehend the text on an initial reading. The “ideal” Fog Index level is seven or eight, and a score of twelve indicates the writing is too hard for most people to read. The Fog value is often used in the health care and insurance industries for analyzing business publications. The formula used to calculate the Gunning fox index is:

Fog Index = 0.4*(words/sentences+100*((words >= 3 syllables)/words))

The XML tag <pg:fog> was used to denote the Gunning Fog Index value as a floating point value in the metacard. The majority of Project Gutenberg texts had a Gunning Fog Index score between three and twenty-three. A plot of the distribution of Gunning Fog Index scores in the Project Gutenberg data set is shown in Figure 8. Again we see a doubly peaked distribution with the higher peak to the lower Fog Index value and secondary extremum three units over towards higher values. In this sense, the Gunning Fog Index distribution is very similar to the ARI index seen previously, both distributions having very sharply defined left and right tails.

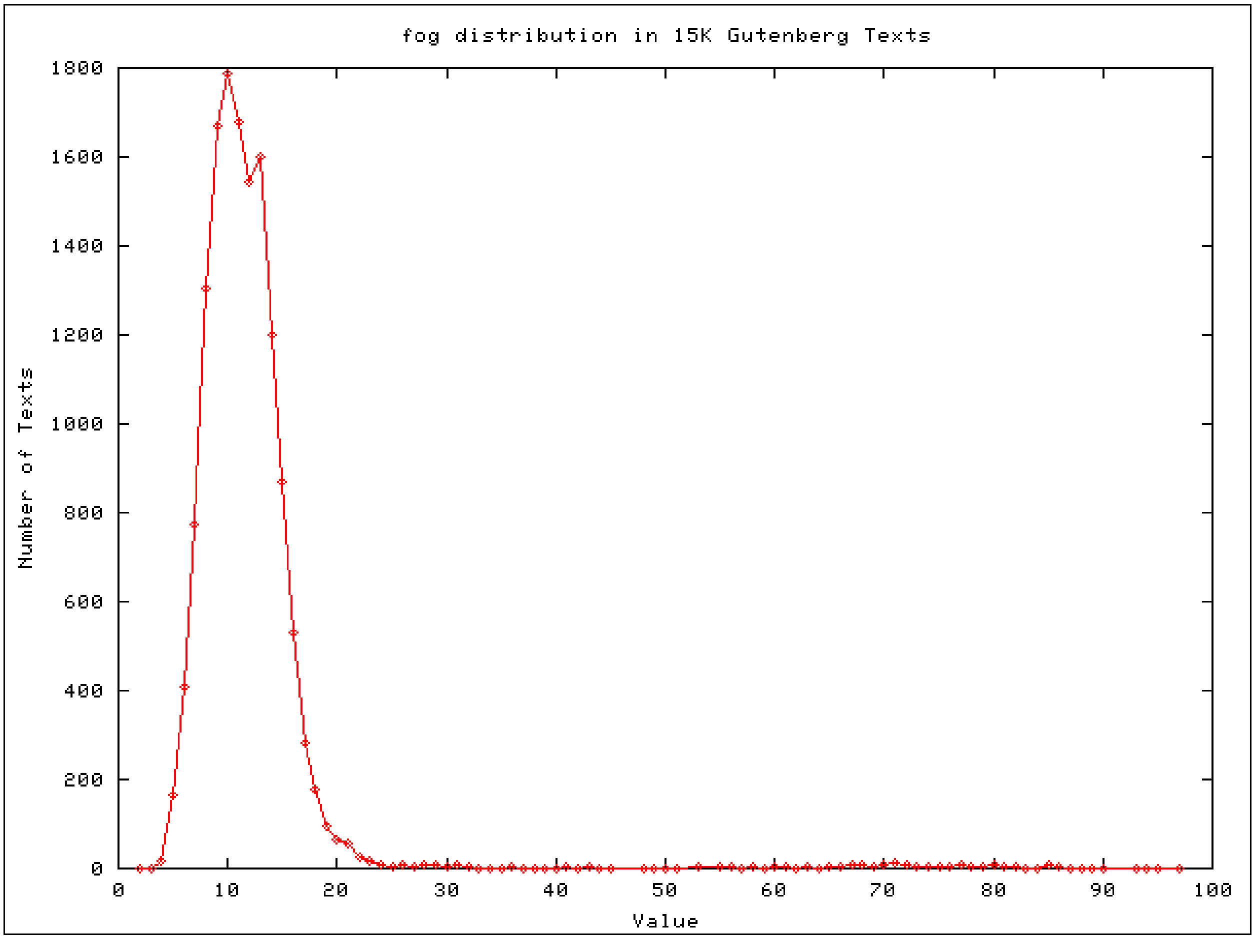


Figure 8: Distribution of Gunning Fog Index scores across the Project Gutenberg data set.

3.5 Flesch-Kincaid Grade Level Formula

The Flesch-Kincaid Formula results in a score between one and one hundred. It relates to a U.S. grade level, or the number of years of education required for understanding the text. It is considered most reliable when used on materials that are for upper elementary school and secondary level education. It has been used for Navy training manuals which were found to range in difficulty from 5.5 to 16.3. The formula used to calculate the Flesch-Kincaid score is:

Kincaid=11.8*syllables/words+0.39*words/sentences-15.59

The XML tag <pg:kincaid> was used to denote the Flesch-Kincaid Grade Level score as a floating point value in the metacard. The majority of texts in the Project Gutenberg data set had a Flesch-Kincaid Grade Level score of nine. A plot of the distribution of Flesch-Kincaid scores in the Project Gutenberg data set is shown in Figure 9. Yet again, we see a steep double peaking of the distribution with the lesser peak on the side of the larger values of the Flesch-Kincaid Grade Level score.

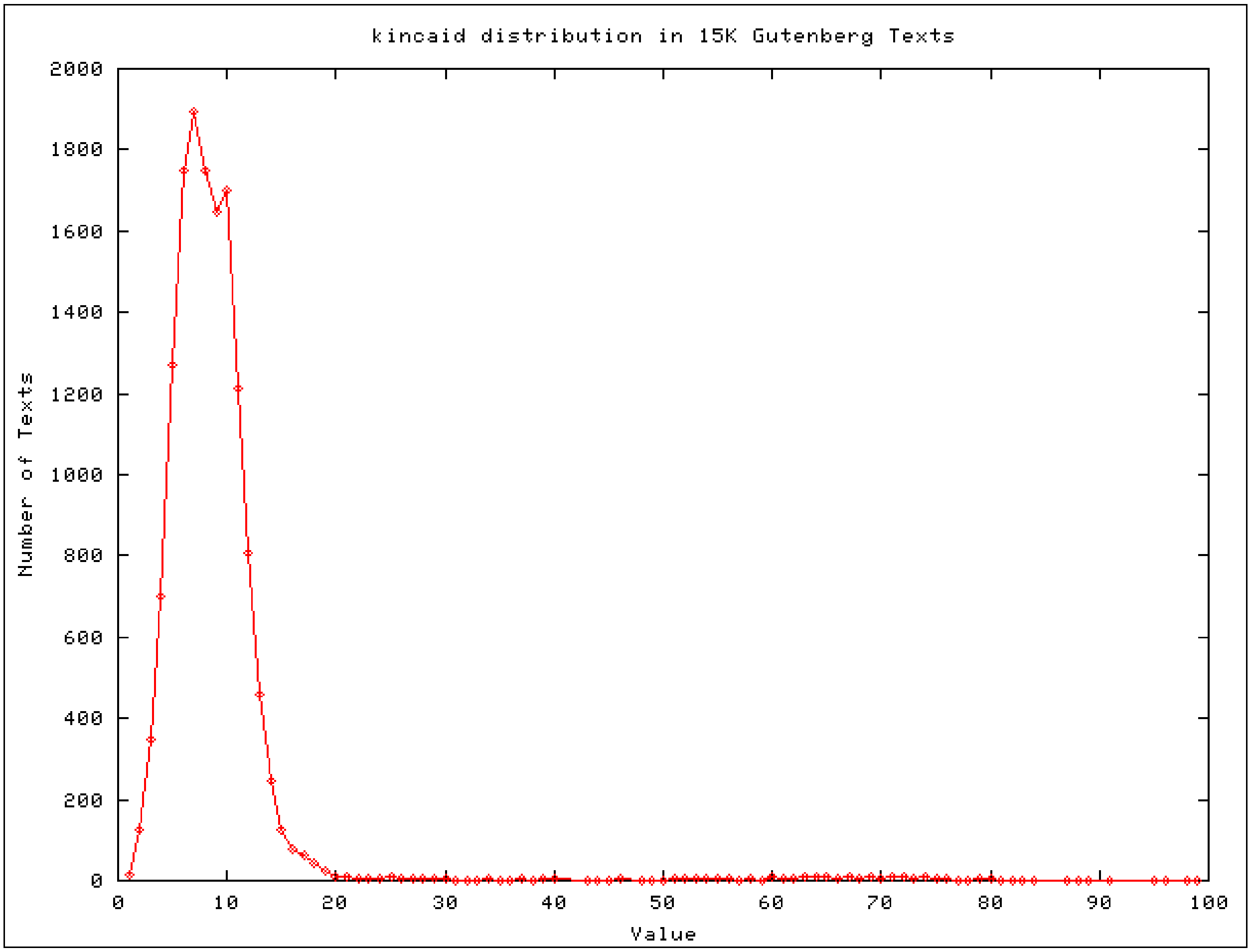


Figure 9: Distribution of Flesch-Kincaid scores across the Project Gutenberg data set.

3.6 The Laesbarhedsindex Formula (Lix)

The Lix formula, developed by Bjson from Sweden, is very simple and uses a mapping table for its scores. It is useful for documents of Western European languages. The score is based on sentence length and the number of long words (long words are words over six characters). The formula used to calculate the Lix index is:

Lix = words/sentences+100*(words >= 6 characters)/words

The XML tag <pg:lix> was used to denote the Lix score as a floating point value in the metacard. A plot of the distribution of Lix in the Project Gutenberg data set is shown in Figure 10. In this figure we see a much less smooth distribution with again one apparent maximum with several lesser peaks. It is difficult to identify how meaningful the shape of this distribution actually is.

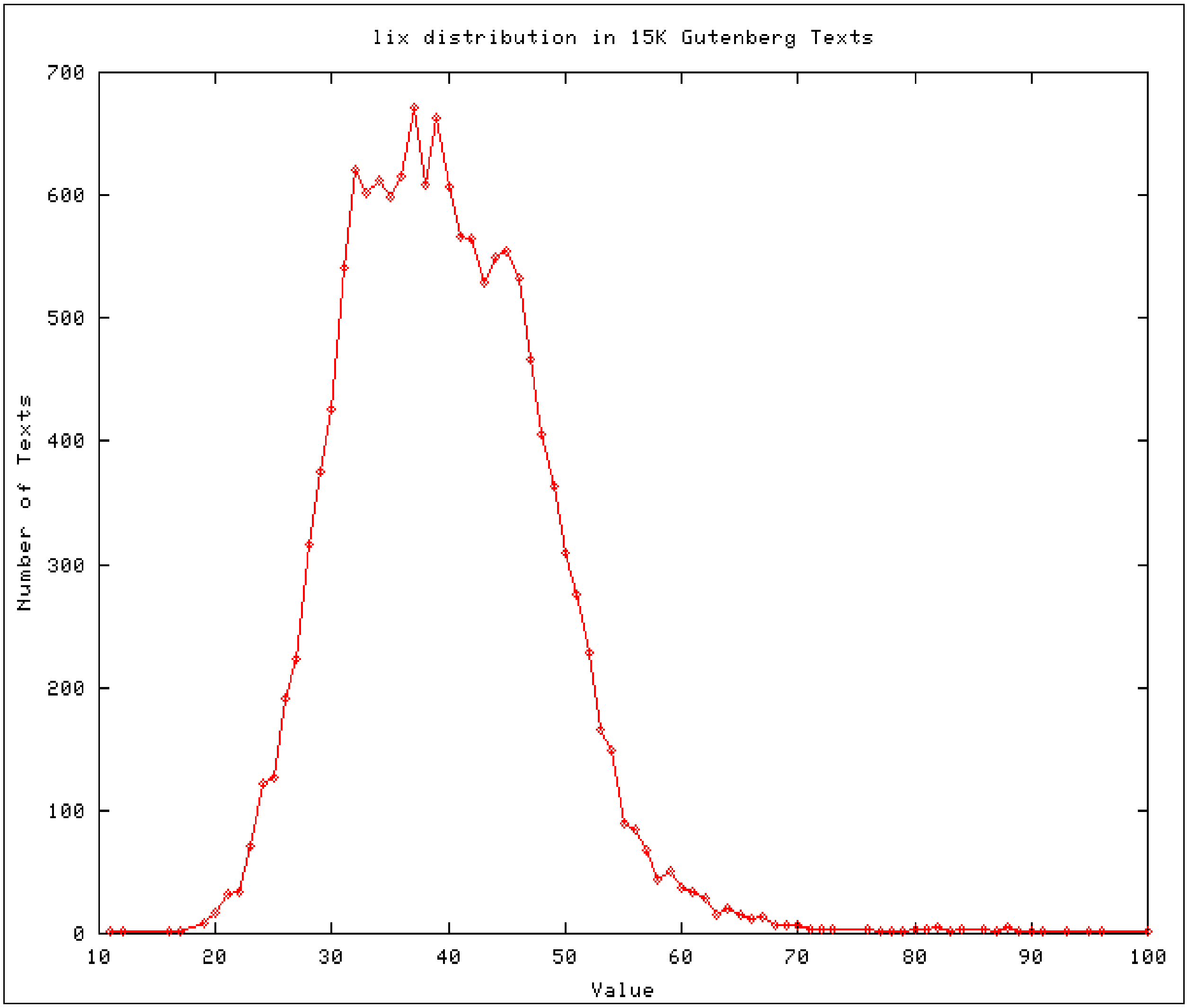


Figure 10: Distribution of Lix scores across the Project Gutenberg data set.

3.7 Simple Measure of Goggledygook (SMOG)

The SMOG index for English texts was developed by McLaughlin in 1969. The index is an estimate of the number of years of United States education needed to fully comprehend the text. This emphasis on full comprehension distinguishes this measurement from other readability scores. The formula used to calculate the SMOG index is:

SMOG-Grading = square root of (((words >= 3 syllables)/sentences)*30) + 3

The XML tag <pg:smog> was used to denote the SMOG score as a floating point value in the metacard. The majority of Project Gutenberg texts had a SMOG index score in the range of ten. A plot of the distribution of SMOG scores in the Project Gutenberg data set is shown in Figure 11.

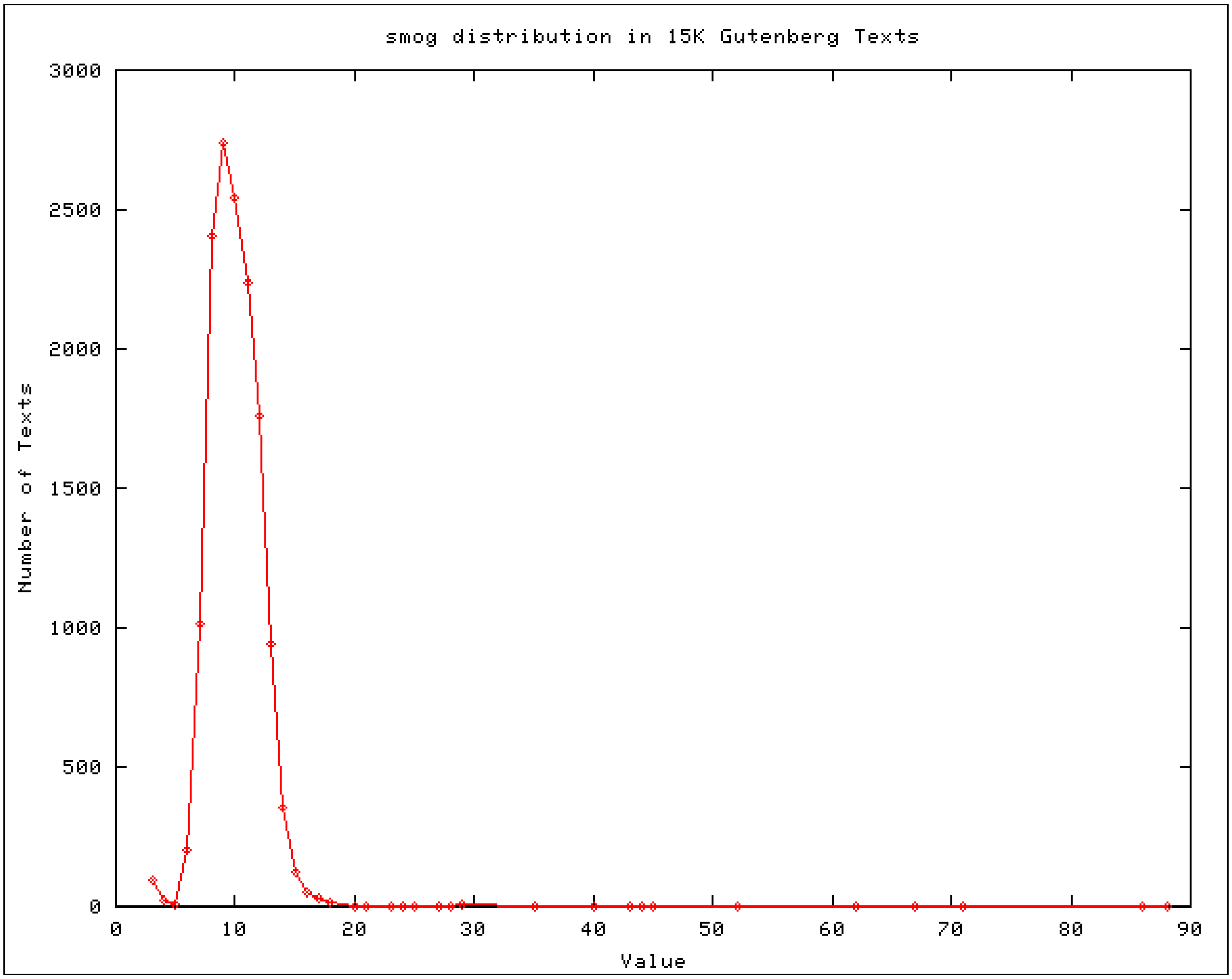


Figure 11: Distribution of SMOG scores across the Project Gutenberg data set.

Again we find a smooth and steep distribution with but one maximum, similar to the Coleman-Liau distribution curve.

3.8 Characters per word

The ‘characters per word’ is an average score for the words in a text. The results come from the style command’s output line stating ‘XX words, average length’. The XML tag <pg:cpw> was used to denote the characters per word as a floating point value in the metacard. The majority of Project Gutenberg texts had a length of four characters. A plot of the distribution of ‘characters per word’ values in the Project Gutenberg data set is shown in Figure 12.

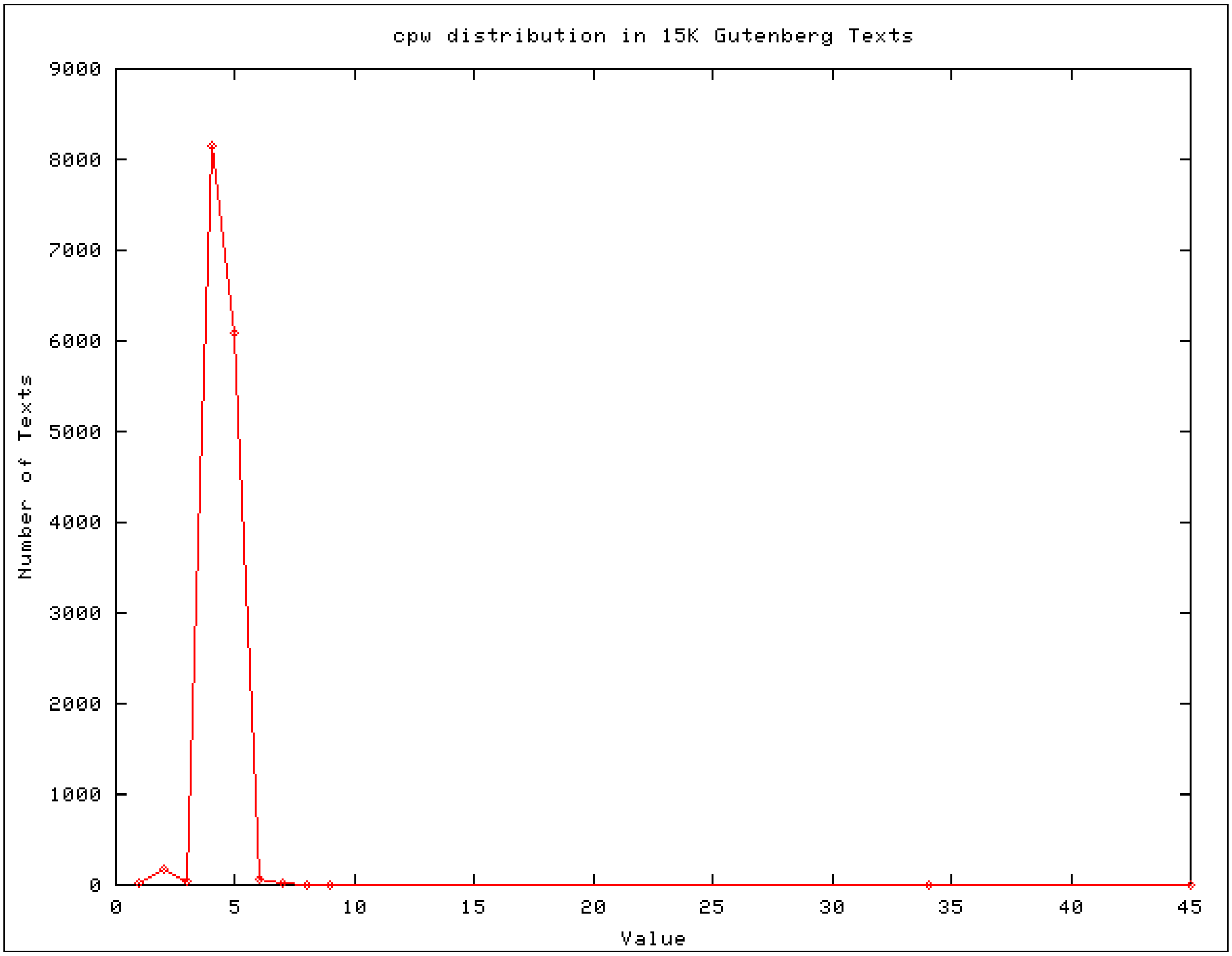


Figure 12: Distribution of 'characters per word' values in the Project Gutenberg data set.

Again we fine mostly a smooth and steeply defined distribution. It is interesting to note here one very small peak in the left tail of the distribution. At this point there is no significance given to this peak except to find it a peculiarity of some texts.

3.9 Words per sentence

The 'words per sentence' score is an average for the number of words in a sentence. The results come from the style command's output line stating 'XX sentences, average length'. The XML tag <pg:wps> was used to denote the words per sentence as a floating point value in the metacard. The majority of Project Gutenberg texts had a mean sentence length around nineteen words. A plot of the distribution of 'words per sentence' values in the Project Gutenberg data set is shown in Figure 13.

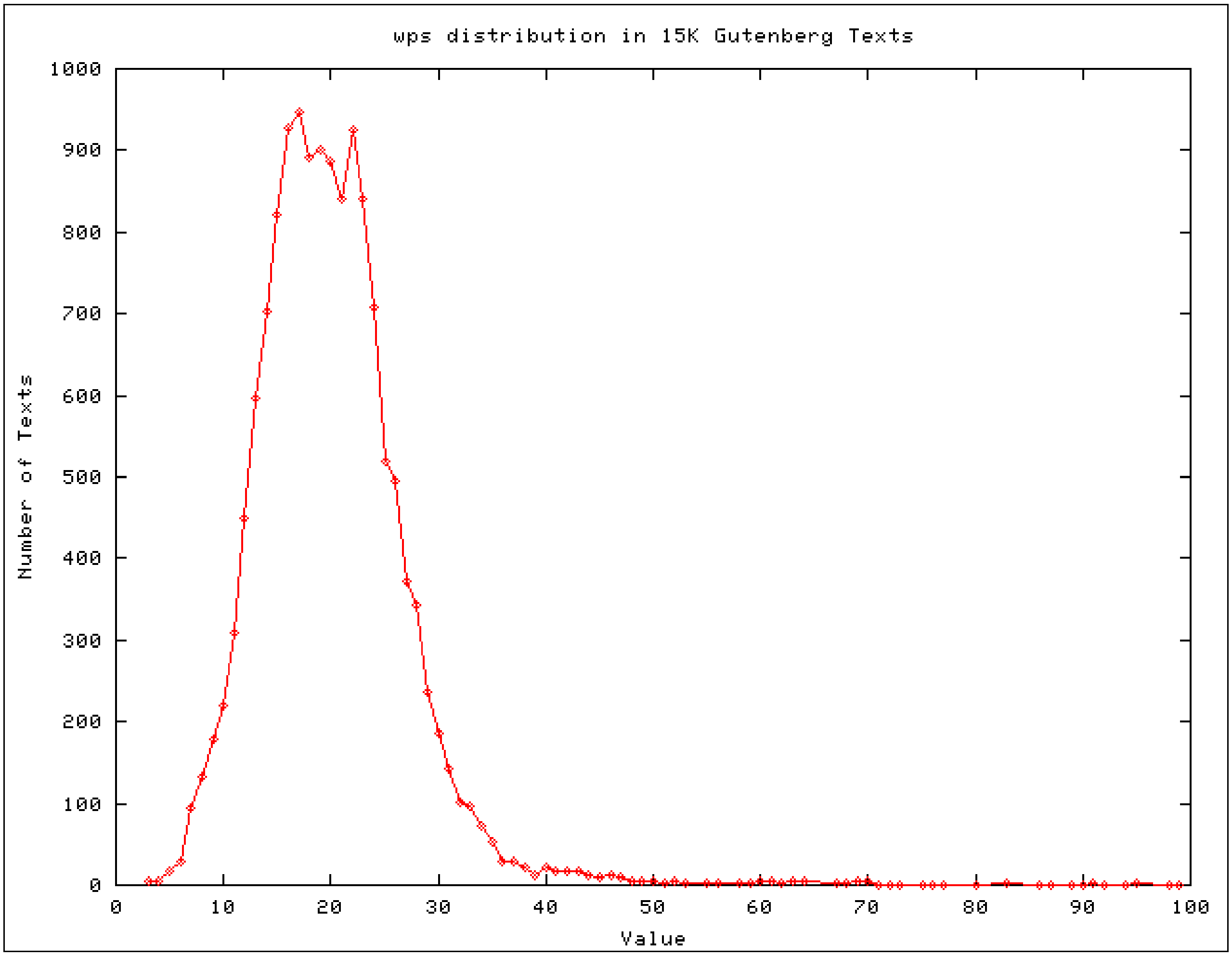


Figure 13: Distribution of ‘words per sentence’ values in the Project Gutenberg data set.

In the ‘words per sentence’ distribution we now find three extremums with the highest one appearing first, which has been the case for all distributions obtained in this study. This maximum is followed by two secondary peaks, the lesser of the two extremes being flanked by the two other extremes.

3.10 Syllables per word

The ‘syllables per word’ score is an average for the number of syllables per word inside a specific text the results come from the style command’s output line stating ‘XX sentences, average length’. The XML tag <pg:spw> was used to denote the words per sentence as a floating point value in the metacard. The majority of Project Gutenberg texts had on average 1.3 syllables per word. A plot of the distribution of ‘syllables per word’ values in the Project Gutenberg data set is shown in Figure 14.

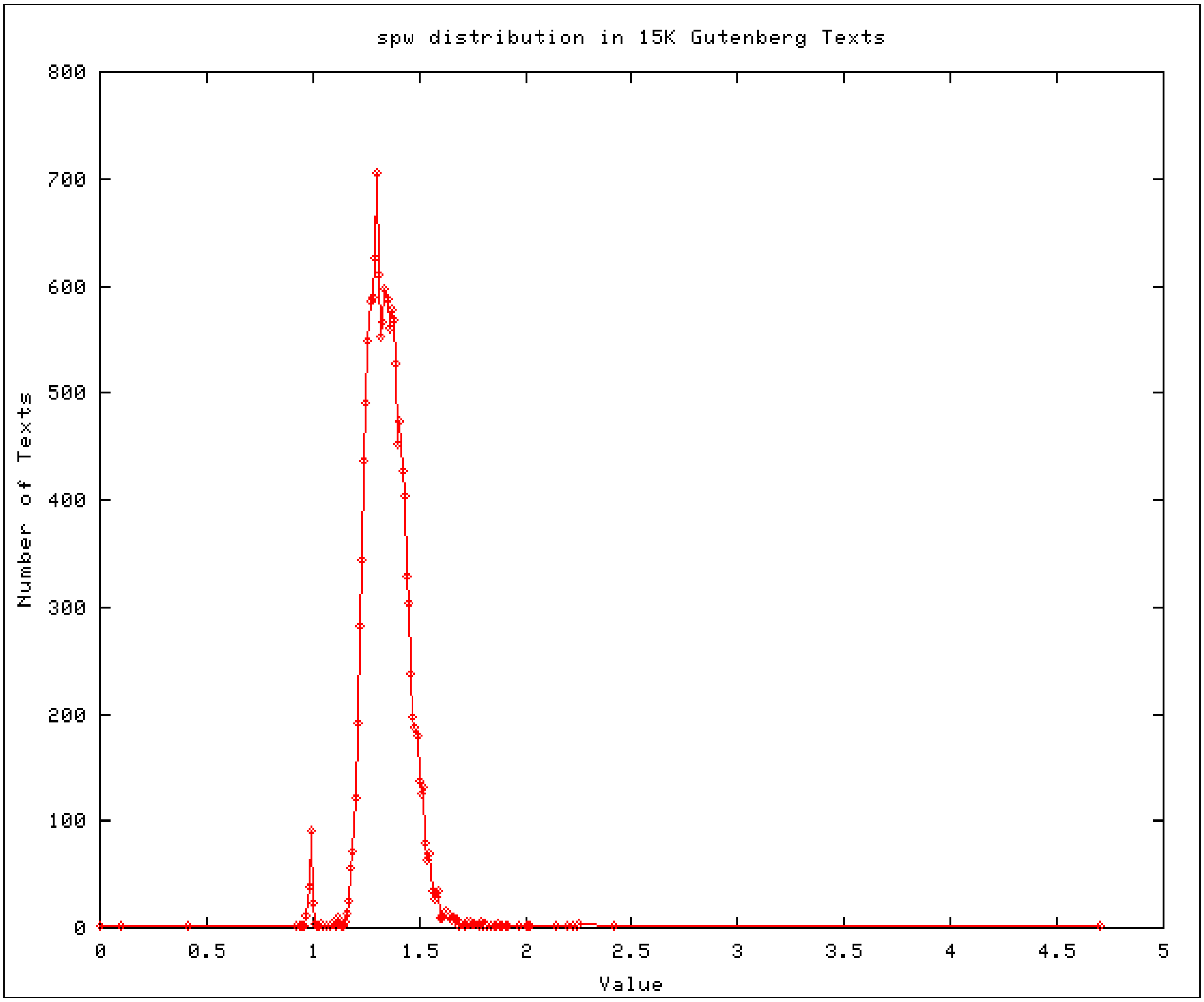


Figure 14: Distribution of ‘syllables per word’ in the Project Gutenberg data set.

Here we see the distribution function having much more structure than has been seen in the distributions defined previously. There is a small peak around the value of one and two additional peaks between 1 and 1.5 with the larger peak to the left of the lesser peak. It is interesting that although many multi-syllables words appear in the texts they do not play any significant role in the distribution of the entire text.

Results in RDF

The readability scores from the analysis were rendered in an XML format intended for machine consumption called Resource Descriptive Framework. Reck (2007) describes this format and its merits at length, hence that will not be reviewed again here. Suffice to say, results in RDF provide a means for scores to be interpreted and leveraged without further description behind the XML format. The RDF format captures the semantics behind the readability values and their relationship to the text they represent. RDF assertions convey the same information no matter what context they occur in. This encourages information interoperability for any clients or users that are RDF aware.

4. Challenges relative to the results

There are several significant problems with the current analysis. There is a major problem in that the Project Gutenberg data set includes texts from nineteen different languages. The majority of the readability measurements used here, are intended exclusively for the English language, therefore adding scores for non-English texts confounds the results. Clearly, some logic should be added to the analysis so that when a language other than English is encountered, the readability analysis is not performed. Alternatively, given that the values are already in the metacards, logic could be used to ignore those values in non-English texts.

A second problem with the analysis is that the definition for the unit "word" for the style command, and the definition for the unit "word" for the analysis differs. In Figure 4 the element expressing the number of words pg:wordcount indicates that there are 19,167 words in 'A Horse's Tale'. This sharply contrasts the output from the "style" command in Figure 3 which shows that 'A Horse's Tale' has 18,590 words. The exact reason for this discrepancy is not known at this time. Interestingly, as yet a third measurement of word count by the UNIX "wc" command indicates that there are 19,175 words in "A Horse's Tale". Clearly, a more thorough analysis would rectify these discrepancies.

Another concern involves the formulas used to render the readability measurements themselves. Fortunately, the documentation presented in the manual page for the "style" command clearly states the precise formula to be used to render each of the readability measurements. Upon closer inspection those formulas slightly differ from other published articulations of the formulas. Additional effort could be spent investigating and determining the formulas from an authoritative source and then implementing the authoritative formulas into the process.

Simple inspection of the plots for the distribution of readability scores reveals another glaring concern. Clearly, there are several situations where the score values fall outside the expected range of values. The exact nature of these situations is not understood at this time, but further work need to be done.

The files from Project Gutenberg span several types of discourse, for example, narratives, poetry, speeches, translated texts, just to name a few. A more precise analysis would attempt to direct the analysis toward a specific range of discourse.

5. Summary

In this work the distribution functions have been calculated for seven different types of readability measurements for approximately fifteen thousand texts from Project Gutenberg. Other supporting measurements were also included, the average characters per word, the words per sentence, and the syllables per word. Three types of distributions have been demonstrated. The simplest type is a single, highly peaked distribution which is shown for the Coleman-Liau scores, the SMOG scores as well as for the 'characters per word'. A slightly more complex distribution is with a double extremum for the ARI scores, the Gunning Fog Index, and the Flesch-Kincaid scores. In some sense both these first two types may be embodying the same type of information and could be in some cases thought to be redundant. The third type of distribution is more complex and unique in its characterization of the texts. For example the ‘words per sentence’ is a three-extremum distribution. The ‘syllables per word’ is doubly peaked but has a small peak as well on the left tail of the distribution around the value one. The Lix scores are much less peaked, that is, the distribution is broader and has considerable character to its distribution function. The most complex and unique distribution function is found for the Flesch Reading Ease scores.

6. Conclusions and next steps

Whether or not readability scores indicate the accessibility of the information they represent, they clearly define a quantitative measure of the texts.

There are several possible improvements mentioned in the results section above. Assuming the limitations discussed there were properly addressed, there are at least a few exciting next steps for synthesizing the results from this analysis. One direction for future work would correlate the authors and the readability scores. Readability measurements could be used to discriminate authors from each other and to rank the difficulty one would expect in reading the work from different authors. In general, is Mark Twain easier to read than Charles Dickens?

It would be interesting to determine if there were any trends in the readability scores over time. Are there any authors that wrote increasingly difficult books as they got older? Do all authors generally write books of increasing difficulty as time progresses? Are books from a certain time period such as the nineteenth century render scores showing they are more difficult to read? Determining trends in text complexity could yield insight into the writing process.

It would be interesting to contrast other attributes of authorship with readability. Are there generalities between gender and readability? One would expect nationality to play a role in influencing non-native author’s ability to write. Do Russian authors write more difficult books than those from Western European countries?

8. Bibliography

Anderson, J. (1983) LIX and RIX: variations on a little-known readability index, Journal of Reading, 26, no. 6, 490-96

Björnsson, H. and B. Hård at Segerstad (1979) Lix på frnska och tio andra språk. Stockholm: Pedagogiskt centrum, Stockholms skolförvaltning.

Daconta, M., L. Obrst and K. Smith (2003) The Semantic Web. Indianapolis, Indiana. Wiley Publishing, Incorporated.

Kincaid, J. P., R. P. Fishburne, R. L. Rogers and B. S. Chissom (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix+html&identifier=ADA006655, (accessed: 24 June 2007)

McEnery, T. and A. Wilson (2001) Corpus Linguistics. 2nd Ed. Edinburgh: Edinburgh University Press.

McLaughlin, G. H. (1969) SMOG Grading - A New Readability Formula. Journal of Reading, 12, 8, 639-46.

PG - Project Gutenberg (2006) Free eBooks. http://www.gutenberg.org/ (accessed: 24 June 2007)

Powers, S. (2003) Practical RDF. Sebastopol, CA O'Reilly & Associates Inc.

RDF (2004) Resource Description Framework. http://www.w3.org/RDF/ (accessed: 24 June 2007)

Reck, R. P. (2006) Metadata Cards for Describing Project Gutenberg Texts , OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies. Genoa, Italy.

Reck, R. P. (2007) Generating and Rendering String Frequency Measurements of Project Gutenberg Texts, Master's thesis, Eastern Michigan University, Ypsilanti, Michigan.

Redish, J. C. and J. Selzer (1985) The Place of Readability Formulas in Technical Communication, Technical Communication, Fourth Quarter, Arlington, VA

Sall, K. and R. P. Reck, (2006) Applying XQuery and OWL to The World Factbook, Wikipedia and Project Gutenberg, XML 2006. Boston, Massachusetts.

Schultheis, R. A. and R. Anderson (1982) The Effectiveness of the Smog Index in Determining the Reading Levels of Business and Distributive Education Texts, Delta Pi Epsilon Journal, 24 no.2, 53-59.

Thorndike, E. J. (1921) The Teacher's Word Book. New York: Teachers College Columbia University.

Wilson, W. M., L. H. Rosenberg and L. E. Hyatt, Conference Paper, Software Engineering, (1997) Proceedings of the 1997 (19th) International Conference on, ISBN; 0-89791-914-9, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=610237 (accessed: 24 June 2007)