Text Readability

Readability: Definition

Readability is a measure of how easy a text is to read, based on complexity, familiarity, legibility, typography. Readability formulas usually consider factors such as sentence length, syllable density, word familiarity.

textstat Python library

The readability scores of the NLP Suite are based on the textstat Python library (https://pypi.org/project/textstat/#files).

The NLP Suite implementation of textstat provides separate readability scores for an entire document and for each sentence in the document.

Readability formulas

textstat provides 8 different measures of readability, based on different formula: Flesch Reading Ease formula, Flesch-Kincaid Grade Level, Fog Scale (Gunning FOG Formula), SMOG (Simple Measure of Gobbledygook), Automated Readability Index, Coleman-Liau Index, Linsear Write Formula, Dale-Chall Readability Score. In addition, the algorithm provides an overall readability consensus score based on all these formulas.

These different measures of readability map on the U.S grade level (1 through 12 – 12 being the last year of high school); they refer to the level of education, in number of years, needed to comprehend a text. Thus,

a 12 readability score requires HIGHSCHOOL education;
a 16 readability score requires COLLEGE education;
a 18 readability score requires MASTER education;
a 24 readability score requires DOCTORAL education;
>24 readability score requires POSTDOC education.

Coleman-Liau Index, **Linsear Write** Formula, **Flesch-Kincaid** Grade, **SMOG** index, **Fog Scale** (**Gunning FOG** Formula) are grade formula in that a score of 9.3 means that a ninth grader would be able to read the document. **Automated Readability Index** (ARI) outputs a number that approximates the grade level needed to comprehend the text; for example, if the ARI is 6.5, then the grade level to comprehend the text is 6th to 7th grade.

The **Flesch Reading Ease** formula has the following range of values (the maximum score is 121.22; there is no limit on how low the score can be, with a negative score being valid):
0-30 College
50-60 High School
90-100 Fourth Grade

The **Dale-Chall** index has the following range of values (different from other tests, since it uses a lookup table of the

most commonly used 3000 English words and returns the grade level using the New Dale-Chall Formula):
4.9 or lower easily understood by an average 4th-grade student or lower
5.0–5.9 easily understood by an average 5th or 6th-grade student
6.0–6.9 easily understood by an average 7th or 8th-grade student
7.0–7.9 easily understood by an average 9th or 10th-grade student
8.0–8.9 easily understood by an average 11th or 12th-grade student
9.0–9.9 easily understood by an average 13th to 15th-grade (college) student

Input

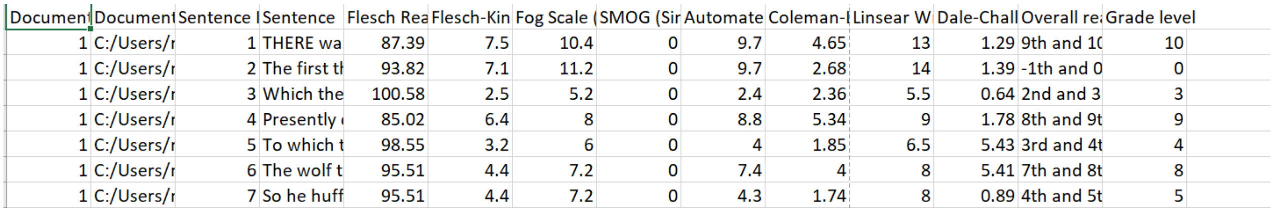
The NLP Suite text readability algorithm can process a single txt file or all the txt files in a directory.

Output

The script produces two types of output:

1. A txt file with the results for an entire document
2. A csv file with sentence by sentence results
3. A set of Excel charts

Example of csv file results



Document	Document	Sentence I	Sentence	Flesch Rea	Flesch-Kin	Fog Scale (	SMOG (Sir	Automate	Coleman-I	Linsear W	Dale-Chall	Overall re	Grade level
1	C:/Users/r	1	THERE wa	87.39	7.5	10.4	0	9.7	4.65	13	1.29	9th and 1(	10
1	C:/Users/r	2	The first tl	93.82	7.1	11.2	0	9.7	2.68	14	1.39	-1th and 0	0
1	C:/Users/r	3	Which the	100.58	2.5	5.2	0	2.4	2.36	5.5	0.64	2nd and 3	3
1	C:/Users/r	4	Presently (	85.02	6.4	8	0	8.8	5.34	9	1.78	8th and 9t	9
1	C:/Users/r	5	To which t	98.55	3.2	6	0	4	1.85	6.5	5.43	3rd and 4t	4
1	C:/Users/r	6	The wolf t	95.51	4.4	7.2	0	7.4	4	8	5.41	7th and 8t	8
1	C:/Users/r	7	So he huff	95.51	4.4	7.2	0	4.3	1.74	8	0.89	4th and 5t	5

Example of txt file results

RESULTS --

Syllable count 1029
Lexicon count 936
Sentence count 34

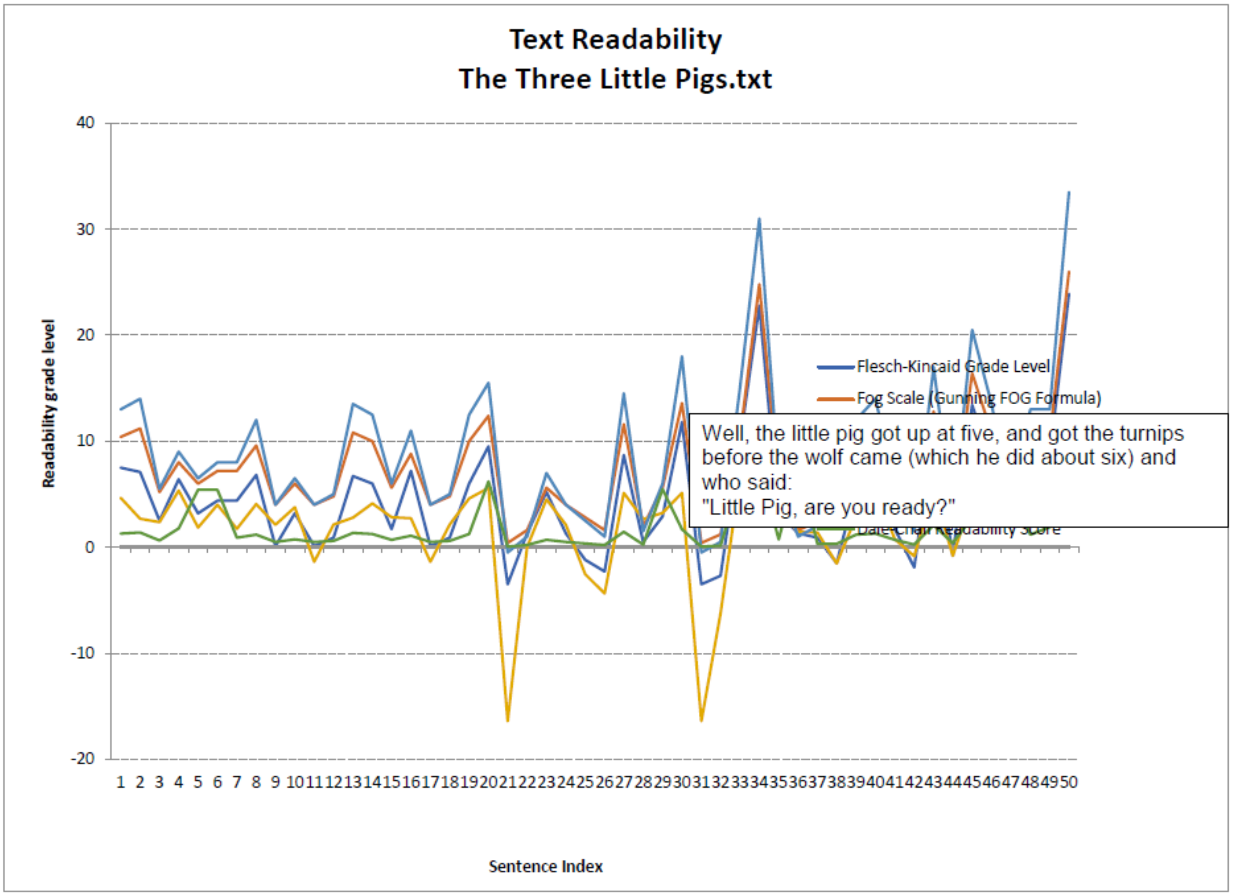
Flesch Reading Ease formula 85.86
Flesch-Kincaid Grade Level 8.1
Fog Scale (Gunning FOG Formula) 11.0
SMOG (Simple Measure of Gobbledygook) Index 6.1
Automated Readability Index 11.1
Coleman-Liau Index 4.48
Linsear Write Formula 9.0
Dale-Chall Readability Score 1.6

Readability Consensus Level based upon all the above tests: 8th and 9th grade.

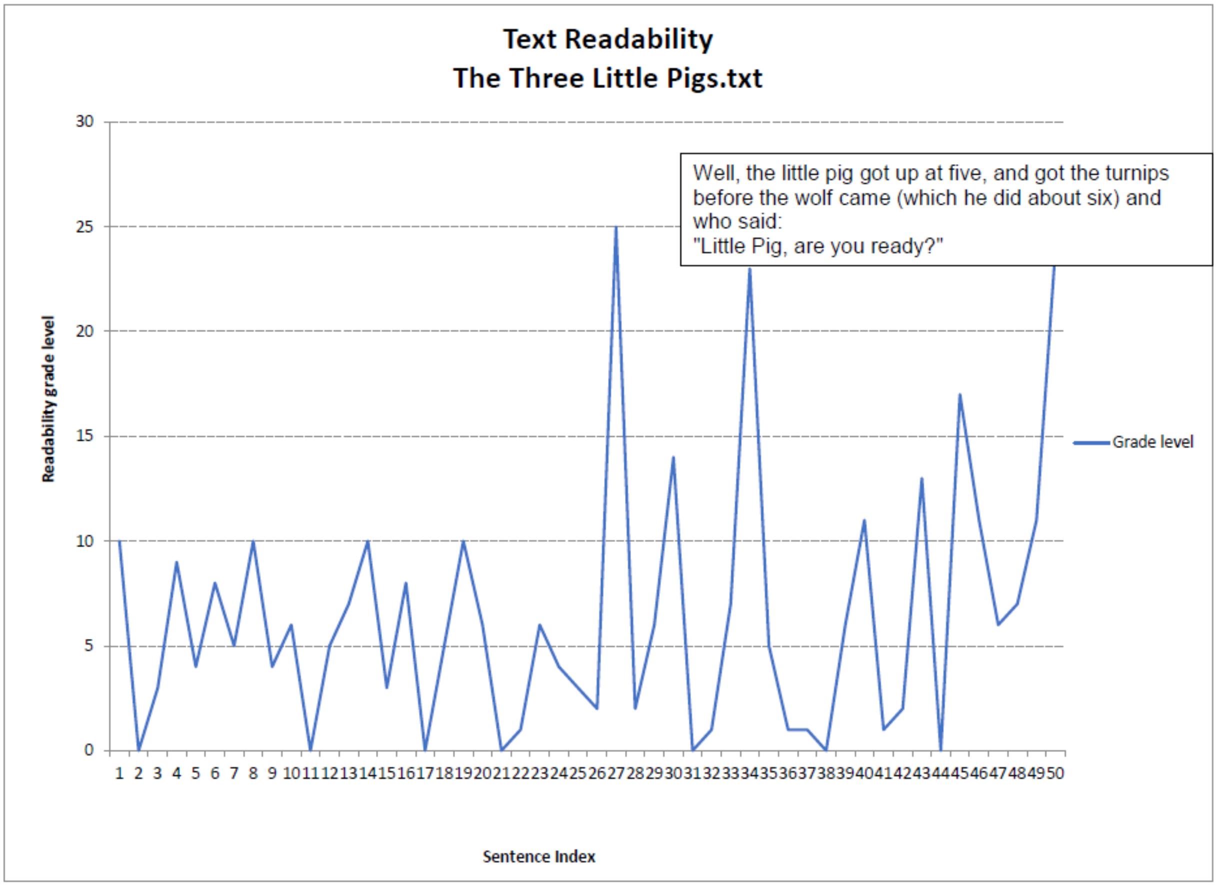
Example of Excel charts

The script produces three types of Excel charts with hover-over capability:

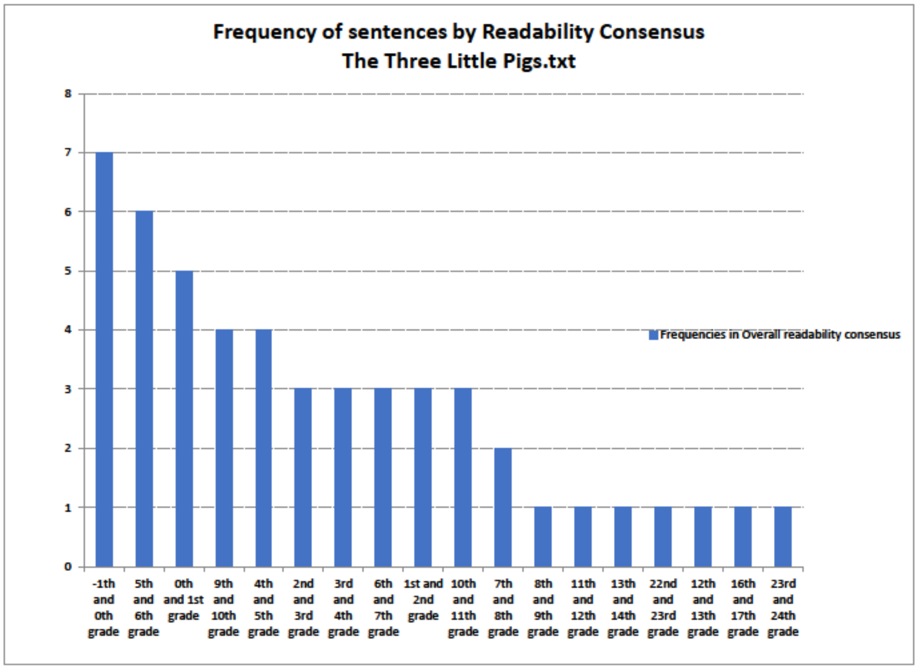
1. An Excel line chart by sentence index for the main readability formulas;
2. An Excel line chart by sentence index for the readability consensus score;
3. A bar chart with the frequency distribution of sentences by their readability consensus grade.



Well, the little pig got up at five, and got the turnips before the wolf came (which he did about six) and who said:
"Little Pig, are you ready?"



Well, the little pig got up at five, and got the turnips before the wolf came (which he did about six) and who said:
"Little Pig, are you ready?"


Frequency of sentences by Readability Consensus
The Three Little Pigs.txt
8
7
6
5
4
3
2
1
0
-1th and 0th grade
5th and 6th grade
0th and 1st grade
9th and 10th grade
4th and 5th grade
2nd and 3rd grade
3rd and 4th grade
6th and 7th grade
1st and 2nd grade
10th and 11th grade
7th and 8th grade
8th and 9th grade
11th and 12th grade
13th and 14th grade
22nd and 23rd grade
12th and 13th grade
16th and 17th grade
23rd and 24th grade
Frequencies in Overall readability consensus