Copyright:

Lesson2:
Modelling the Web with Simple Statistical Descriptive Text Models
Unit5:
Compare the sentence lengths and word lengths of Simple and English Wikipedia

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST

People and Knowledge Networks

Completing this unit you should

- Get a feeling for interdisciplinary research
- Know the Automated Readability Index
- Have a strong sense of support for our research hypothesis
- Be able to critically discuss the limits of our models

How would linguists tackle this problem?

- Flesch-Kincaid readability test

$$fkt = 206.835 - 1.015\left(\frac{\texttt{total words}}{\texttt{total sentences}}\right) - 84.6\left(\frac{\texttt{total syllables}}{\texttt{total words}}\right)$$

- Wherever the weights and coefficients drop from the idea is clear:
  - first term is low if sentences are shorter
  - second term is low if words have fewer syllables

- Knowing syllables is a non trivial problem for a computer

- Hard to automatically calculate

Interpreting the results of the FKRT

$$fkt = 206.835 - 1.015\left(\frac{\texttt{total words}}{\texttt{total sentences}}\right) - 84.6\left(\frac{\texttt{total syllables}}{\texttt{total words}}\right)$$

Score	School Level	Notes
90 - 100	5th grade	Very easy to read for average 11 year old
80-90	6th grade	Easy to read. Conversational English for consumers
70-80	7th grade	Fairly easy to read
60-70	8th & 9th grade	Plain English. Easily understood by 13 – 15 year old students
50-60	10th to 12th grade	Fairly difficult to read
30-50	college	Difficult to read
0-30	College graduate	Very difficult to read.

Automated Readability Index

$$ari = 4.71\left(\frac{\texttt{total characters}}{\texttt{total words}}\right) + 0.5\left(\frac{\texttt{total words}}{\texttt{total sentences}}\right) - 21.43$$

- Wherever the weights and coefficients drop from the idea is clear:
  - first term is low if words have fewer characters
  - second term is low if sentences are shorter

- Counting words, sentences and characters is easy for a computer

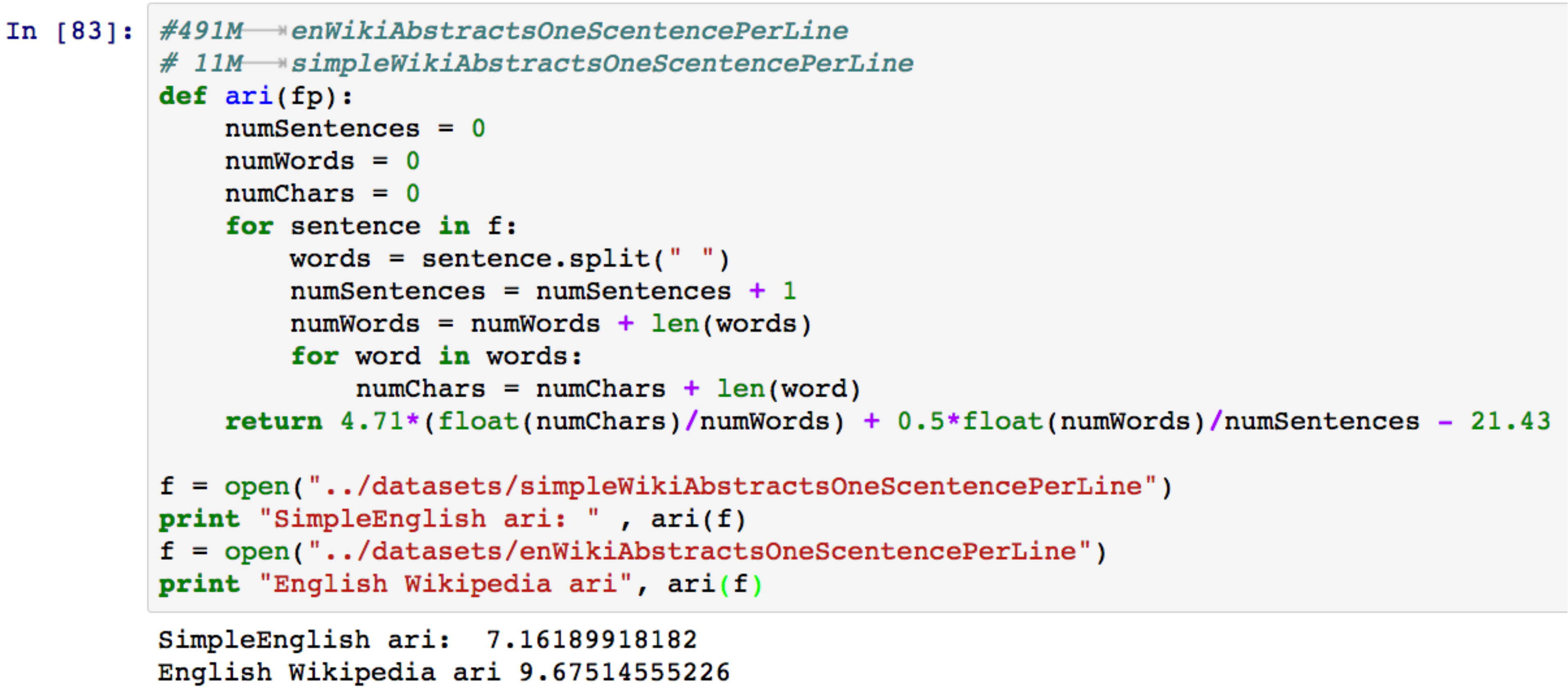
- Formula corresponds to our testable prediction

Interpreting the results of the ARI

$$ari = 4.71\left(\frac{\texttt{total characters}}{\texttt{total words}}\right) + 0.5\left(\frac{\texttt{total words}}{\texttt{total sentences}}\right) - 21.43$$

Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First grade
3	7-8	Second grade
4	8-9	Third grade
5	9-10	Fourth grade
6	10-11	Fifth grade
7	11-12	Sixth grade
8	12-13	Seventh grade
9	13-14	Eighth grade
10-13	15-18	High school
> 14	18-22	College

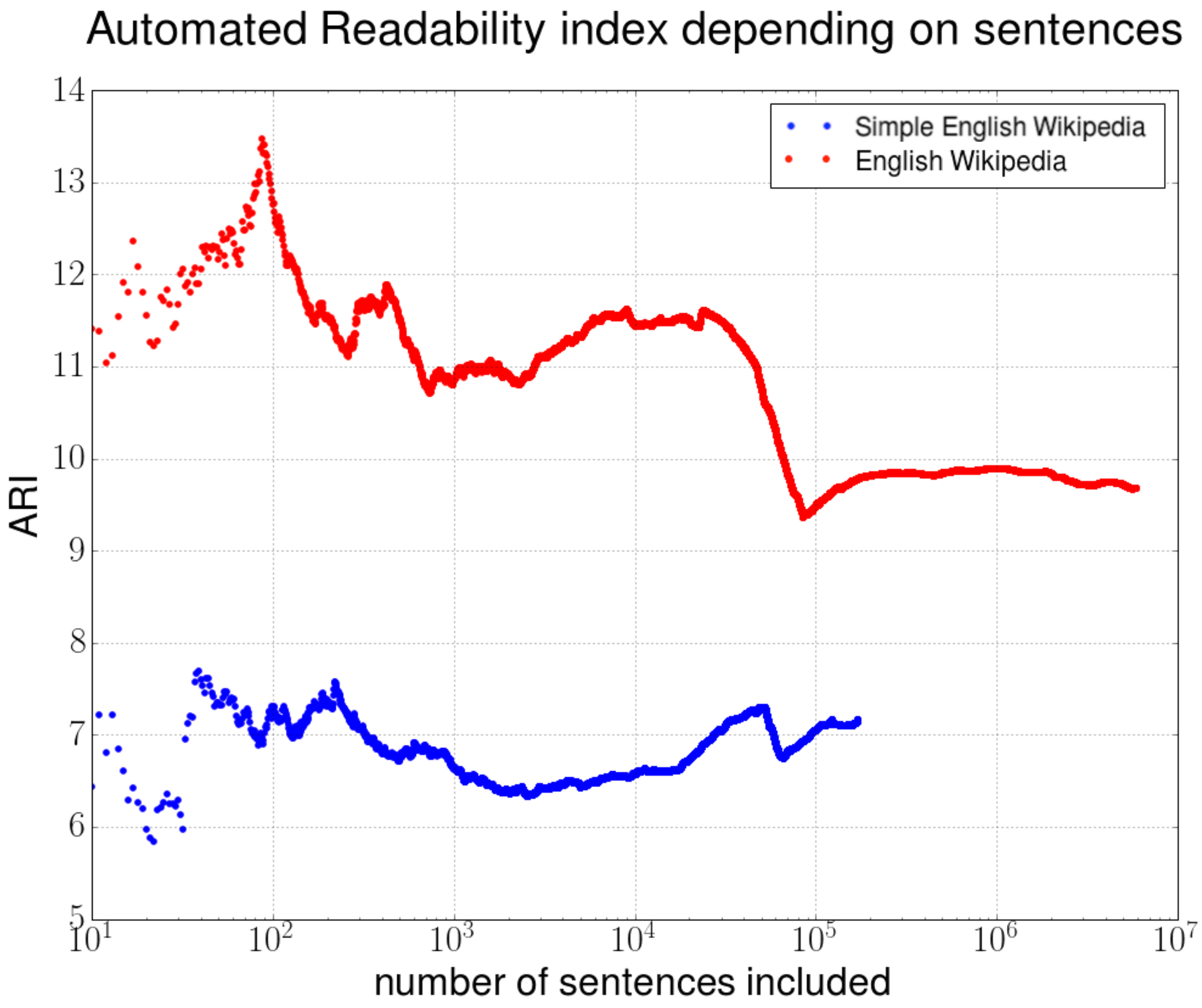
What does the ARI for Wikipedia look like?



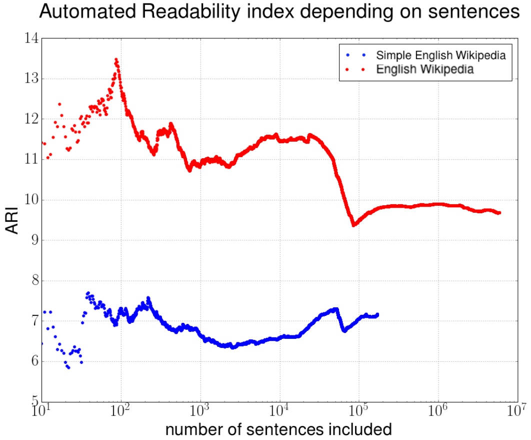
```
In [83]: #491M    enWikiAbstractsOneScentencePerLine
         # 11M    simpleWikiAbstractsOneScentencePerLine
         def ari(fp):
             numSentences = 0
             numWords = 0
             numChars = 0
             for sentence in f:
                 words = sentence.split(" ")
                 numSentences = numSentences + 1
                 numWords = numWords + len(words)
                 for word in words:
                     numChars = numChars + len(word)
             return 4.71*(float(numChars)/numWords) + 0.5*float(numWords)/numSentences - 21.43

         f = open("../datasets/simpleWikiAbstractsOneScentencePerLine")
         print "SimpleEnglish ari: " , ari(f)
         f = open("../datasets/enWikiAbstractsOneScentencePerLine")
         print "English Wikipedia ari", ari(f)

         SimpleEnglish ari:  7.16189918182
         English Wikipedia ari 9.67514555226
```

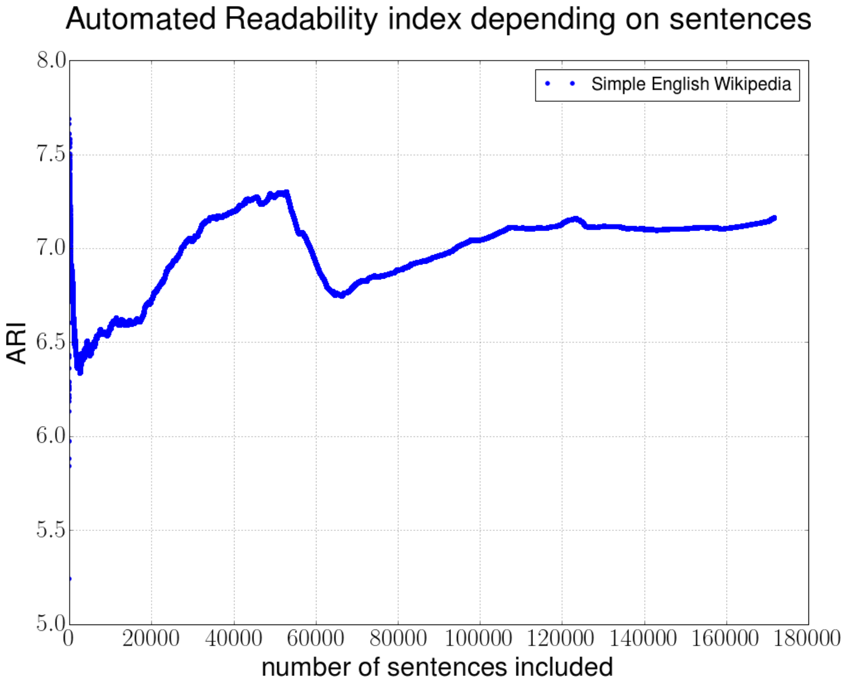
- Can we depend on the result?

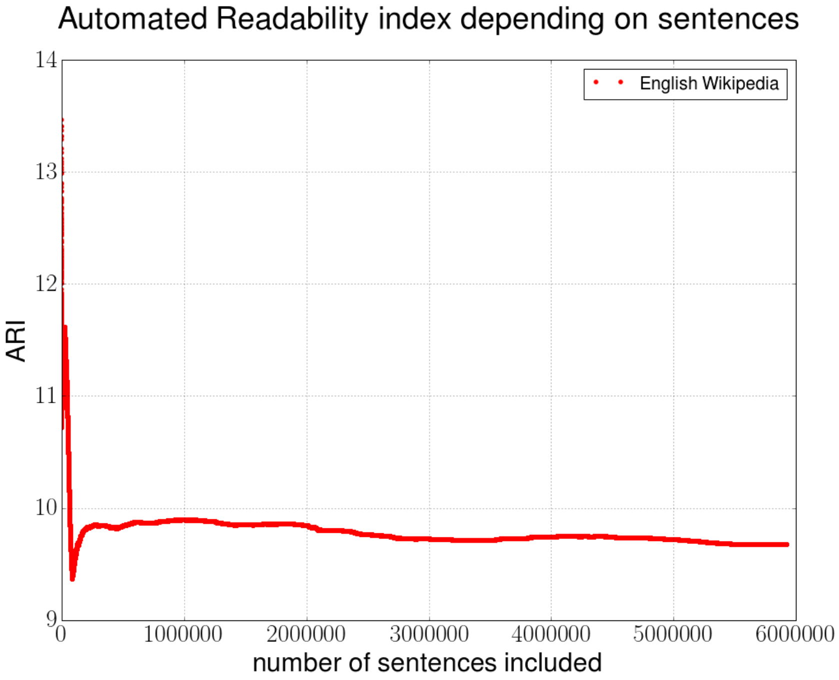
Lots of fluctuation for the readability index



Remember! Logarithmic scales are tricky


Automated Readability index depending on sentences


Automated Readability index depending on sentences


Automated Readability index depending on sentences

The full cycle of research… Making new observations asking questions

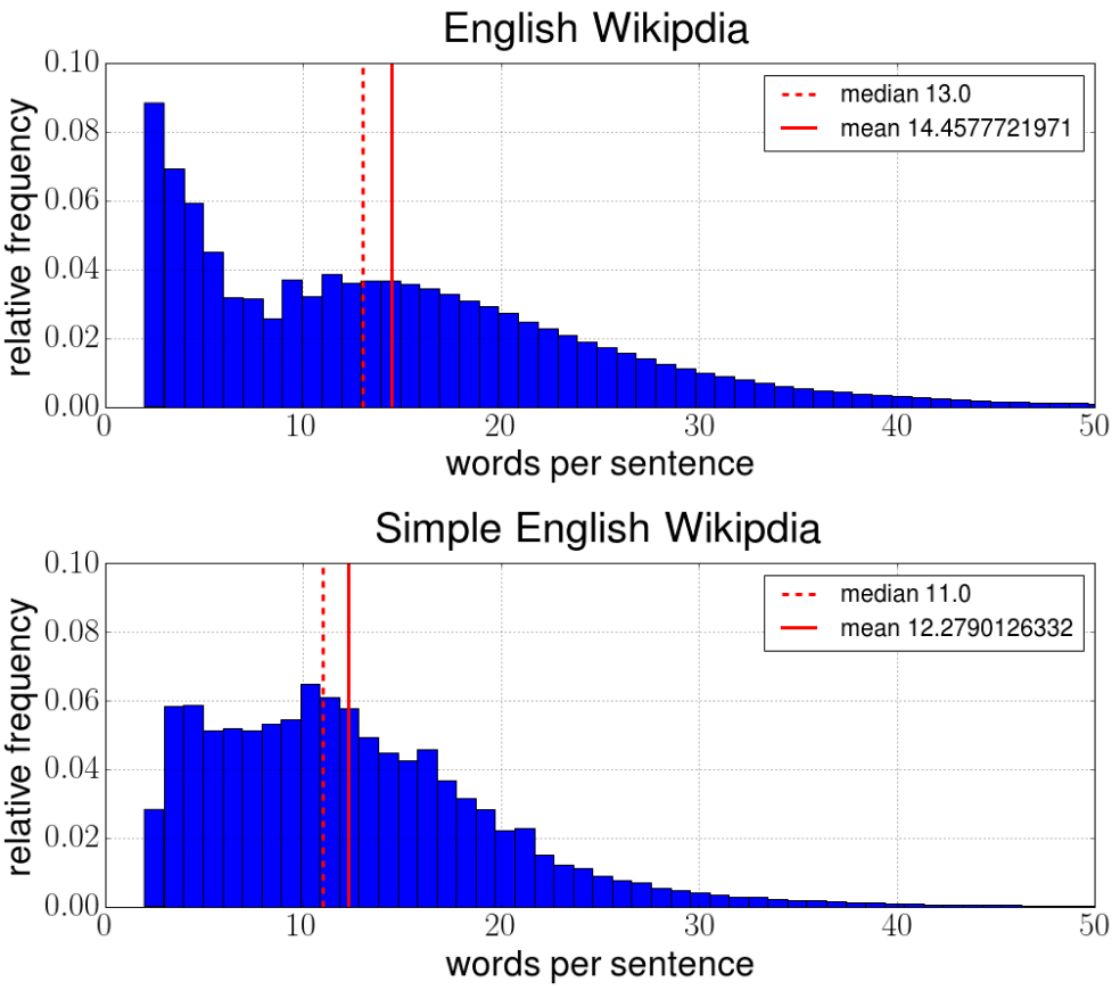
More words are needed to understand 50% of sentences in English Wikipedia than in Simple Wikipedia

The ARI in Simple English is lower than in English Wikipedia

Could the distribution of sentence lengths be the reason?

Research starts over again with new question

Histogram of Sentence lengths on abstracts of Wikiedia data sets



Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright:

The instantiated model reflects a particular situation in the world

- When we take a collection of web pages in order to build a text model

- Model characterizes how the world might work in general

- But the models we study only have a special snapshot of a special situation

- also das Modell charakterisiert wie ein Ausschnitt der Welt im Allgemeinen funktioniert und das spezielle Modell instantiiert eine spezielle Situation in der Welt