Artigos Curtos

Legibilidade de Artigos Científicos: Análise de Dados da RCC

Luís Cavique
DCET, Univ. Aberta
lcavique@univ-ab.pt

Resumo

A legibilidade de uma revista de divulgação científica é um problema central das comissões editoriais, dos revisores e em especial dos leitores. O vocabulário de cada artigo relaciona-se em larga medida com as palavras-chave próprias de cada área científica, contudo a legibilidade também depende de outros factores. Neste artigo apresentam-se métricas de legibilidade que são função do comprimento das palavras e do comprimento das frases. Propõe-se que as métricas de legibilidade sejam balizadas por um limite superior e por um limite inferior. Finalmente, encontra-se uma relação entre as duas métricas, com base nos dados extraídos dos artigos publicados na RCC.

Palavras-chave: legibilidade, métricas em texto

Title: Readability of scientific papers : data analysis of RCC

Abstract

The readability of a scientific review is a very important issue for the editorial committees, for the reviewers and especially for the readers. The vocabulary of each article is largely related to the keywords of the specific scientific area; however, the readability depends also on other factors. In this article we present two readability metrics that are based on the length of the words and on the length of the sentences. We propose that the readability metrics must be bounded by a superior limit and an inferior limit. Finally, the correlation between the two metrics is presented, given the data extracted from the articles published in the RCC.

Keywords: readability, text metrics

1 Introdução

A legibilidade de um documento é usualmente associada ao tipo de vocabulário utilizado. São muitos os exemplos na literatura de textos de difícil leitura, causados pela utilização de vocabulário invulgar. Por exemplo, na obra de Aquilino Ribeiro, com uma forte influência popular, cheia de provincianismos e neologismos, um leitor comum só

terá possibilidade de desfrutar da sua excepcional riqueza com a ajuda não só de um dicionário mas também de um glossário adequado às obras do autor.

Este artigo centra-se na leitura de textos de divulgação científica deixando de lado a escrita literária. Os artigos científicos tratam assuntos de elevada complexidade, com um vocabulário próprio, onde a legibilidade é um assunto central. O vocabulário de cada artigo encontra-se restringido às palavras-chave e a termos próprios da área ou sub-área científica, contudo a legibilidade dependerá também de outros factores.

A par do vocabulário, a construção frásica tem uma importância determinante na legibilidade. Vejamos três tipos de frases categorizadas em "simples", "média" e "difícil".

- Uma frase "simples" com uma legibilidade alta: "Esta frase contém palavras vulgares e conceitos simples."

- Uma frase com legibilidade "média": "Apesar desta frase ser um pouco mais difícil que a anterior, e apesar da sua complexidade, o leitor não terá dificuldade em compreendê-la."

- E por fim, uma frase "difícil" ou de legibilidade baixa: "Nesta precisa frase, contida aqui, tem uma complexidade desigual, poderá tornar-se até, para o leitor, sem preparação prévia, plausivelmente difícil a decomposição frásica realizada ao nível do texto, embora não denote necessariamente um conteúdo intricado, o que permite gozar o estatuto de mais erudita, ou mais complexa, tendo sem dúvida uma baixa legibilidade."

Nas três frases anteriores, a legibilidade não depende do vocabulário utilizado. Os documentos escritos com o nível de legibilidade alta podem ser lidos por audiências com qualquer nível de educação. A maior parte dos artigos científicos devem cair na categoria da legibilidade média. A legibilidade baixa é causada pelas palavras e frases longas, sendo contudo possível alterar a sua desnecessária complexidade.

Para medir a legibilidade de um texto existe felizmente um conjunto de métricas, que permitem, de forma automática, avaliar o documento. No editor MS-Word em "Ortografia e Gramática" ou no Google-Docs em "Tools", "Count Words" é possível conhecer a média de palavras por frase, ou a média de caracteres por palavra e ainda ter acesso aos índices de Facilidade de Leitura Flesch (Flesch Reading Ease) e o índice de Legibilidade de Flesch-Kincaid em Anos de Escolaridade (Flesch-Kincaid Grade Level).

Na secção 2, descrevemos algumas das métricas mais conhecidas. Na secção 3, é discutida a necessidade de existirem dois limites para as métricas. Na secção 4, é apresentada uma relação entre as duas métricas que foram referidas, utilizando os dados extraídos dos artigos publicados na RCC. Finalmente, na secção 5, propomos algumas conclusões.

2 As métricas existentes

2.1 Facilidade de Leitura de Flesch (Flesch Reading Ease)

O resultado da fórmula de Facilidade de Leitura Flesch (FLF) [Flesch 1948] cai no intervalo de 0 a 100, o valor de 0 indica uma baixa legibilidade, enquanto que 100 indica que o texto tem uma alta legibilidade. Contudo, a métrica devolve o valor de 121 para uma frase com uma única palavra, o que não revela grande aplicabilidade. A expressão é construída à custa do comprimento médio da frase, CMF, e do número médio de sílabas por palavra, MSP.

$$FLF = 206{,}835 - (1{,}015 \times CMF) - (84{,}6 \times MSP)$$

onde:
CMF = comprimento médio da frase (número de palavras dividido pelo número de frases)
MSP = número médio de sílabas por palavra (número de sílabas dividido pelo número de palavras)

Como exemplo de uma revista com grande legibilidade, a "Reader's Digest" tem um FLF igual a 65 [Wikipedia]. Por outro lado a "Harvard Business Review" um FLF de 32, podendo ser encontradas revistas científicas com FLF iguais a 20 [Armstrong 1980].

2.2 Flesch-Kincaid Anos de Escolaridade (Flesch-Kincaid Grade Level)

A fórmula de Flesch-Kincaid Anos de Escolaridade (FK) [Kincaid e al. 1975] converte a legibilidade em anos de escolaridade dos EUA. Tal como em Portugal, o 4º ano de escolaridade corresponde a estudantes com 9 a 10 anos, o 9º ano de escolaridade a estudantes com 14 a 15 anos e o 12º ano de escolaridade a estudantes com 17 a 18 anos. Para estudantes no ensino superior correspondem anos de escolaridade superiores a 12.

O resultado da fórmula de Flesch-Kincaid Anos de Escolaridade tem como limite inferior o valor de 0 e como limite superior valores entre 30 e 35. O valor de 0 indica uma baixa escolaridade, enquanto que os valores entre 30 e 35 indicam uma muito alta escolaridade, que corresponde a uma baixa legibilidade.

A expressão seguinte é também construída à custa do comprimento médio da frase, CMF, e do número médio de sílabas por palavra, MSP.

$$FK = (0{,}39 \times ASL) + (11{,}8 \times ASW) - 15{,}59$$

onde igualmente:
CMF = comprimento médio da frase (número de palavras dividido pelo número de frases)
MSP = número médio de sílabas por palavra (número de sílabas dividido pelo número de palavras)

2.3 Outras Métricas para a Legibilidade

Muitas têm sido as abordagens subsequentes sobre métricas de legibilidade [Dubay 2004], [Dubay 2007]. A maior parte das métricas de legibilidade consideram a estrutura frásica (comprimentos das frases e palavras) e a percentagem de palavras menos vulgares do vocabulário, podendo enquadrar-se em modelos aditivos ou multiplicativos como apresentado nas seguintes expressões, onde os literais (a, b, c, d) são constantes:

Legibilidade = a + b (comp. frase) + c (comp. palavra) + d (palavras difíceis)

Legibilidade = a + b (comp. frase) × c (comp. palavra) + d (palavras difíceis)

Ao contrário das fórmulas FLF e FK que utilizam factores aditivos, a fórmula de SMOG [McLaughlin 1969] vai utilizar factores multiplicativos.

Dentro das métricas que têm em consideração as palavras difíceis dos vocabulários, podemos referir a Fórmula de Legibilidade de Dale-Chall [Zakaluk e Samuels 1988] e o Índice Fog [Gunning 1952]. As fórmulas FLF e FK foram estudadas para a língua inglesa, contudo como não estão dependentes de um dicionário, podem perfeitamente ser utilizadas em português.

2.4 Exemplo

Dando continuidade ao exemplo anterior e utilizando os índices de Facilidade de Leitura Flesch e o índice de Legibilidade de Flesch-Kincaid, obtemos os seguintes valores.

Tabela 1- Valores de FLF e FR para as frases ditas fácil, média e difícil

nível	*frase*	*(FLF) Facilidade Leitura Flesch (0..100)*	*(FK) Flesch-Kincaid Anos de Escolaridade*
fácil	Esta frase contém palavras vulgares e conceitos simples.	30	11
média	Apesar desta frase ser um pouco mais difícil que a anterior, e apesar da sua complexidade, o leitor não terá dificuldade em compreendê-la.	20	16
difícil	Nesta precisa frase, contida aqui, tem uma complexidade desigual, ……. , tendo sem dúvida uma baixa legibilidade.	0	31

A frase dita fácil, com FLF de 30 tem um FK de 11 anos de escolaridade. A frase média tem um FLF de 20 e um FK de 16. Finalmente, a frase difícil tem 0 de FLF e um FK de 31 anos de escolaridade. Na tabela vemos que FLF e FK são inversamente proporcionais, enquanto FLF decresce, o valor de FK é crescente.

3 Limites para as Métricas

Se pensarmos num qualquer adjectivo que denote mérito, por exemplo, uma “pessoa económica”, e o repensarmos por excesso ou por defeito, encontramos outros adjectivos sem qualquer mérito. Esquematicamente teremos: perdulário < económico < avarento.

Armstrong [Armstrong 1980], ao comparar 10 das mais prestigiadas revistas de gestão encontra uma correlação positiva de 70% entre o prestígio da revista e o índice de

dificuldade de leitura de Fog. Armstrong [1989] acrescenta ainda que o prestígio deve estar relacionado à complexidade até um certo nível, a partir do qual não existe ganho.

Neste artigo, propomos que as métricas de legibilidade sejam balizadas por um limite superior e por um limite inferior. Procuramos, assim, encontrar o meio-termo entre a escrita demasiado simples, ao nível do ensino básico, e a escrita demasiado elaborada, com uma complexidade desnecessária e de difícil leitura.

Para analisar a legibilidade na língua portuguesa, podemos utilizar as métricas complementares de FLF e FK para limites, já que estão disponíveis no MS-Word e no Google-Docs.

A métrica FLF devolve a facilidade de leitura, enquanto que a métrica FK está associada aos anos de escolaridade ou à dificuldade da leitura. Como limite da facilidade de leitura, dado que podemos encontrar revistas científicas com FLF ≥ 20 [Armstrong 1980], vamos adoptar esse valor. Como limite de dificuldade de leitura propomos o usualmente aceite FK ≥ 12.

4 Análise de Dados da RCC

Foram analisados os artigos da Revista de Ciências da Computação, RCC, disponível on-line em http://www.moodle.univ-ab.pt/moodle/course/view.php?id=31, dos números 1, 2 e 3 relativos aos anos de 2006, 2007 e 2008.

Para cada artigo foram determinados o valores de FLF e o FK, tendo sido encontrados valores médios de: média(FLF)= 29,9 e média(FK)=13,9. Note-se que ambos os valores médios são superiores aos limites propostos anteriormente. Os desvios padrão encontrados para as métricas foram de desvioP(FLF)=15,5 e desvioP(FK)=3,9.

Na figura 1, no gráfico de dispersão de dados Facilidade de Leitura de Flesch (FLF) versus Flesch-Kincaid Anos de Escolaridade (FK) podemos encontrar uma recta obtida através de uma regressão linear onde FLF=80,7 – 3,7 FK, com r^2=85%.

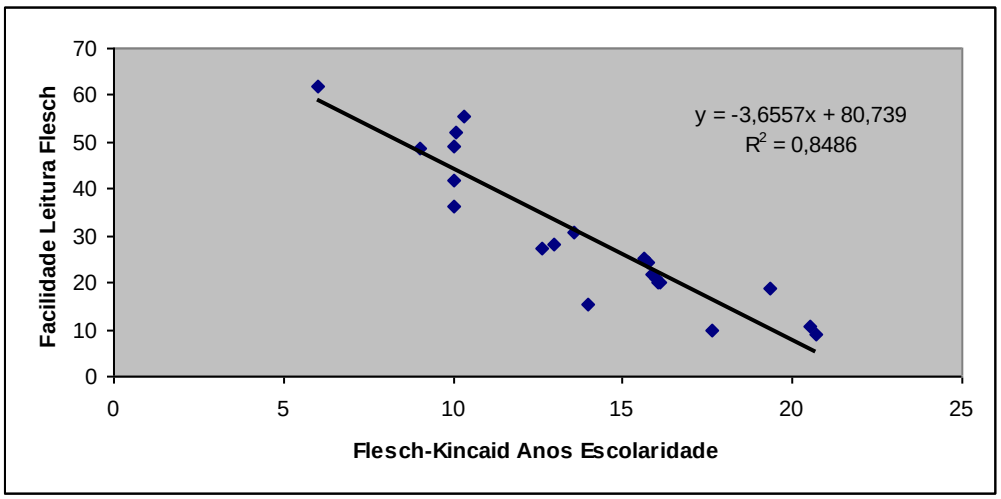

Figura 1 – Métrica FK versus FLF para os dados da RCC

Apesar dos valores médios serem superiores aos limites definidos (FLF ≥ 20 e FK ≥ 12), existem alguns artigos com legibilidade inferior a 20. Existem ainda, artigos em língua inglesa, escritos por portugueses, com índice de anos de escolaridade inferior a 12.

Dado que FLF varia entre 0 e 100 e FK entre 0 e cerca de 30, encontramos uma recta da regressão FLF=100 – 3,3 FK. Se associarmos os limites com as duas métricas (FLF ≥ 20 e FK ≥ 12), podemos encontrar limites superiores e inferiores para FLF e FK. Assim teremos como limites utilizando uma única métrica, em que: 20 <FLF<60 e 12<FK<24. É de notar que os limites com duas métricas são mais restritivos que os limites com uma única métrica. Utilizando os limites relaxados, com uma única métrica, todos os artigos da RCC estão dentro dos limites definidos.

5 Conclusões

O objectivo deste artigo consistiu em dar a conhecer métricas disponíveis no MS-Word e no Google-Docs de modo a incentivar o uso generalizado dos índices de legibilidade.

Julgamos que este tema tem sido pouco tratado em Portugal, ao contrário do que acontece com a língua inglesa, que promove campanhas contínuas de “Plain English” nos últimos 30 anos.

Embora os estudos tenham sido realizados para língua inglesa, as métricas Facilidade de Leitura de Flesch (FLF) e Flesch-Kincaid Anos de Escolaridade (FK) podem ser adaptadas à língua portuguesa.

Ao propor, neste artigo, que as métricas de legibilidade sejam balizadas por um limite superior e por um limite inferior, procurámos, assim, encontrar o meio-termo entre a escrita demasiado simples, ao nível do ensino básico, e a escrita demasiado elaborada, com uma complexidade desnecessária e de difícil leitura.

Se utilizar uma única métrica, o autor opta por uma abordagem mais relaxada, tendo como limites propostos os valores de (20 <FLF <60) ou de (12 <FK <24). Usando a via mais restritiva, ao utilizar as duas métricas, fica balizado por (FLF ≥ 20 e FK ≥ 12).

Para exemplificar o estudo, este artigo tem FK=11 e FLF= 34, caindo no intervalo dos limites com uma única métrica (20 <FLF <60).

Bibliografia

Armstrong, J.S., Readability and Prestige in Scientific Journals, Journal of Information Science, 15, 123-124, (1989).

Armstrong, J.S., Unintelligible Management Research and Academic Prestige, Interfaces, Vol. 10, No. 2, pp. 80-86, (1980).

DuBay, W., Smart Language: Readers, Readability, and the Grading of Text, Impact Information (2007).

DuBay, W., The Principles of Readability, Impact Information (2004).

Flesch, R., A new readability yardstick, Journal of Applied Psychology, Vol. 32, pp. 221-233 (1948).

Gunning, R., The Technique of Clear Writing, New York, McGraw-Hill (1952).

Kincaid, J.P., Fishburne, R.P.Jr., Rogers, R.L. e Chissom, B.S., Derivation of new readability formulas for Navy enlisted personnel, Research Branch Report 8-75, U. S. Naval Air Station, Memphis, TN (1975).

McLaughlin, G.H., SMOG Grading: a New Readability Formula, pp. 639-646, Journal of Reading (1969).

Zakaluk, B.L., Samuels, S. J., Eds, Readability: Its Past, Present, and Future, International Reading Association, Newark, Del. (1988).