Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: avaliação da Coleção Literatura para Todos

Text readability analysis with Natural Language Processing Tools: assessment of the "Literatura para Todos" Collection

Erica dos Santos Rodrigues
Cláudia Freitas
Violeta Quental
Pontifícia Universidade Católica do Rio de Janeiro – Rio de Janeiro – Rio de Janeiro – Brasil

Resumo: O presente trabalho apresenta resultados de pesquisa referente à inteligibilidade dos livros da *Coleção Literatura para Todos 1*, publicada pelo MEC/SECAD (2006) e distribuída para jovens e adultos recém-alfabetizados. A investigação da inteligibilidade dos textos buscou conjugar pressupostos da Psicolinguística e ferramentas de processamento automático da língua portuguesa. Utilizamos critérios de inteligibilidade referidos na literatura psicolinguística, e tentamos capturar de maneira objetiva o grau de complexidade linguística dos livros através de ferramentas computacionais: o analisador morfossintático PALAVRAS e o programa Coh-Metrix Port. Nossos resultados sugerem que os livros dessa coleção são complexos para o público pretendido. Assumindo que os neoleitores estão na etapa inicial do processo de alfabetização, é nítido que esses livros exigem um esforço de decodificação da escrita que está além de sua capacidade.

Palavras-chave: Legibilidade; Inteligibilidade; Coleção Literatura para Todos; Avaliação de inteligibilidade; Ferramentas para processamento computacional do Português

Abstract: This paper presents results of a research on readability of the *Literatura para Todos 1* Collection, published by MEC/SECAD (2006) and distributed to newly literate youth and adults. The investigation of text readability combined assumptions from Psycholinguistics and natural language processing tools. We used readability criteria referred in psycholinguistics literature and tried to evaluate objectively the degree of linguistic complexity of the books by means of computational tools: the morphosyntactic analyzer *PALAVRAS* and the program Coh-Metrix Port. Our results suggest that these books are complex for the intended audience. Assuming that new readers are in the initial stage of literacy process, it is clear that reading the books of the Collection takes an effort which is beyond their ability.

Keywords: Readability; Coleção Literatura para Todos; Readability assessment; Natural language processing tools

Introdução

O presente trabalho apresenta resultados de pesquisa referente à inteligibilidade dos livros que integram a *Coleção Literatura para Todos 1*, doravante CLPT, publicada pelo MEC/SECAD (2006) e distribuída para grupos de jovens e adultos recém-alfabetizados, chamados neoleitores. O objetivo da Coleção é democratizar o acesso à leitura, constituir um acervo bibliográfico literário específico para jovens, adultos e idosos recém-alfabetizados.[1] O estudo inseriu-se em um projeto mais amplo cujo objetivo era avaliar a recepção da referida coleção junto aos neoleitores e aos mediadores envolvidos no Programa.[2]

[1] http://portal.mec.gov.br/index.php?option=com_content&view=article&id=12313&Itemid=629

[2] Projeto Percursos da Leitura, coordenado pelo Professor Júlio Diniz (PUC-Rio).

O estudo foi conduzido a partir de uma perspectiva interdisciplinar, em que a investigação da inteligibilidade dos textos da Coleção buscou conjugar (i) pressupostos da Psicolinguística e (ii) ferramentas desenvolvidas para o processamento automático da língua portuguesa.

O desconhecimento da existência de material nos moldes da *Coleção Literatura para Todos* para a língua portuguesa se, por um lado, corrobora a necessidade deste tipo de iniciativa, por outro, impossibilita a comparação ou mesmo a utilização de critérios já estabelecidos para a aferição da inteligibilidade do material oferecido. Assim, o trabalho consistiu na seleção de critérios de inteligibilidade referidos na literatura psicolinguística que pudessem atuar como parâmetros na avaliação da adequação linguística dos livros ao público de neoleitores. Tentamos capturar, de maneira objetiva, o grau de complexidade linguística dos livros, com o auxílio de ferramentas da Linguística Computacional. O objetivo do presente trabalho, portanto, é triplo: (a) apresentar a análise e os resultados da avaliação; (b) apresentar as potencialidades da pesquisa interdisciplinar entre a Psicolinguística e a Linguística Computacional, direcionadas a uma aplicação educacional; (c) propor critérios objetivos para a avaliação quantitativa de textos cuja elaboração é, sem dúvida, desafiadora, uma vez que devem aliar a simplicidade gramatical e sintática à complexidade das experiências de vida dos neoleitores.

O artigo está organizado da seguinte maneira: na seção 2, discutimos o conceito de inteligibilidade, tendo em vista a justificativa dos parâmetros linguísticos escolhidos para a avaliação dos textos, bem como detalhamos os conceitos propriamente; na seção 3, apresentamos o perfil dos neoleitores e livros da Coleção; a seção 4 trata da análise dos textos; na seção 5, tecemos algumas considerações finais.

1 Conceituação de inteligibilidade

A leitura é uma atividade extremamente complexa que envolve um conjunto de subprocessos. Estes compreendem, entre outros, o reconhecimento de símbolos gráficos, a recuperação de palavras do léxico mental, o processamento sintático de enunciados, com mobilização de um mecanismo de *parsing*, e o processamento semântico local, do qual resultam proposições que são integradas em representações semânticas de partes maiores do texto (macroproposições) (COSCARELLI, 1996; KLEIMAN, 1993; KATO, 1999; PERFETTI, 1999; PERFETTI, LANDI e OAKHILL, 2005). Para a implementação desses processos, o leitor acessa várias bases de conhecimento (linguístico, enciclopédico, relativo a gêneros e tipos textuais...) e lança mão de inferências, predições, aplica estratégias metacognitivas, etc. (PEREIRA, 2002; BORBA, 2005)

Trabalhos que se voltam para a investigação de fatores que podem afetar a compreensão da leitura procuram distinguir três grupos de fatores – fatores associados ao texto, ao leitor e à intervenção leitora (LEFFA, 1996).

No que tange ao primeiro grupo de fatores – associados ao texto –, costuma-se estabelecer uma distinção entre legibilidade e inteligibilidade textual. O termo legibilidade tem sido utilizado, de modo geral, para caracterizar fatores tipográficos, tais como o tamanho de letras, tipo de fonte, diagramação do texto. Para fazer referência a estruturas linguísticas complexas e vocabulário pouco frequente como elementos que podem afetar o grau de compreensão de um texto, costuma-se adotar o termo inteligibilidade (LEFFA, 1996; SCARTON et al., 2010; SCARTON e ALUÍSIO, 2010). O emprego deste termo em referência apenas a fatores ligados ao texto não é, não obstante, consensual, e há autores que estendem o termo para indicar fatores associados ao leitor, como seu grau de motivação e interesse pelo assunto (BARBOZA e NUNES, 2007; RIBEIRO et al. 2011). Nessa acepção mais ampla, o termo inteligibilidade confunde-se com leiturabilidade (*readability*), entendido como "tudo o que torna um texto mais fácil de ler do que outros" por oposição ao termo legibilidade (*legibility*), empregado em referência a "características de tipografia e layout" (DUBAY, 2004; LIMA, 2007).[3]

Dada a dificuldade de uma convergência terminológica, a qual reflete, pelo menos em parte, mudanças na própria concepção de leitura, inicialmente centrada nos aspectos textuais e posteriormente incorporando o leitor e sua bagagem linguístico-cultural (KLEIMAN, 2004), optamos pelo emprego do termo inteligibilidade na nomeação dos critérios que elegemos para verificar a complexidade dos textos lidos. Essa opção foi motivada pelo fato de ser o termo mais comumente usado pelos pesquisadores que têm trabalhado na área de leitura e simplificação textual (SCARTON et al., 2010; SCARTON e ALUÍSIO, 2010).

1.1 *Parâmetros utilizados*

Para a caracterização da complexidade sintática dos textos da Coleção, foram considerados os seguintes parâmetros:

(i) total de verbos por período;
(ii) presença, no período, de elementos explicativos intercalados;
(iii) quantidade de vírgulas por período;

[3] Remetemos o leitor a Dubay (2004) para diferentes referências ao termo *readability*.

(iv) presença de orações reduzidas de gerúndio;
(v) quantidade de palavras antepostas ao verbo principal.

O item (i) foi tomado como indicativo da densidade do período e, por conseguinte, do grau de inteligibilidade textual, considerando-se restrições relativas à manutenção e integração de informações na memória de trabalho por unidade estrutural (período) no processamento sintático. Para a construção da coerência temática do texto, o leitor precisa relacionar o significado das sentenças, isto é, precisa integrar proposições (KINTSCH e VAN DIJK, 1978). Considerando-se que o verbo é, por excelência, do ponto de vista semântico, um elemento predicador que instancia proposições, pode-se afirmar que, quanto maior for o número de verbos em um período, maior será o número de estruturas sintáticas a ser analisado, maior o número de proposições a ser construído e mais complexa a integração dessas proposições em macroproposições. Assim, em (a), se analisarmos apenas as sentenças com predicados verbais, podemos considerar que haveria pelo menos seis proposições a serem integradas:

(a) "Porém, da última vez que visitei a casinha delas, enquanto preparava um café doce no fogão de lenha, elas riam e contavam que, quando moravam na roça, cozinhavam em latas, e que o fazendeiro era muito ruim." (CLPT, livro *Léo, o pardo*: 115)

O item (ii) é indicativo de quebras na ordem canônica da estrutura da sentença, e compreende estruturas apositivas, orações adjetivas explicativas, orações adverbiais deslocadas e sintagmas adverbiais deslocados. Em (b), os trechos sublinhados indicam estruturas potencialmente complicadas:

(b) "A minha Bia, a abelha, seria: 1) na verdade um coelho, que nasceu num corpo de abelha e, por ter natureza de roedor, rói o caule das plantas e é expulso da colmeia." (CLPT, livro *Cobras em Compota*: 63)

Todas essas estruturas têm em comum algum tipo de interrupção que cria demandas associadas à manutenção de informação pela memória de trabalho no processo de integração de informação sintática (COSCARELLI, 1996).

O item (iii) está relacionado à quantidade de informações por período e, é importante lembrar, uma leitura efetiva depende de alta proficiência na interpretação dos usos da vírgula, que pode indicar tanto a presença de uma enumeração quanto a intercalação de estruturas e o deslocamento de constituintes, como pode ser visto em (c):

(c) ali só mora índio, e se você bota reparo, tem gente com cara de índio que não sabe que é índio, mas é, e preto, e filho de branco com preto, preto com índio, e branco mais preto e índio. (CLPT, livro *Léo, o pardo*: 49)

O critério (iv) assume que as orações reduzidas de gerúndio podem ser, em princípio, complexas em termos de compreensão – uma vez que a relação semântica entre a oração principal e a reduzida de gerúndio não é explicitada pelo conectivo, ela precisa ser inferida. Os conectivos funcionam como instruções para o leitor estabelecer relações de coerência entre segmentos do texto e desempenham papel importante na construção do sentido global. Um texto sem a presença dessas marcas torna-se mais difícil para a leitura (MILLIS e JUST, 1994; SANDERS e NOORDMAN, 2000; BEN-ANATH, 2005; CAIN e NASH, 2011). No caso de orações reduzidas de gerúndio, como o gerúndio é uma forma nominal do verbo, também desaparecem informações relativas a tempo, modo, pessoa, o que pode dificultar o estabelecimento de relações temporais entre os eventos e também a identificação do sujeito do verbo, comprometendo, inclusive, o estabelecimento/manutenção da referência, como pode ser observado em (d).

(d) Falavam em nome da ordem, da justiça, do povo, e até de Deus, atendendo a denúncias anônimas dando conta de que Almerinda estava morta, vítima de maus tratos. (CLPT, livro *Cabelos Molhados*: 13)

Por fim, o critério (v) baseia-se na ideia de que um grande número de termos antes do verbo pode indicar a presença de sujeito complexo ou mesmo de outros termos argumentais a ele associados. O elemento inicial da sentença, para que possa ser corretamente analisado e interpretado, depende da informação codificada no verbo: é necessário manter esse elemento inicial na memória de trabalho até que o verbo seja identificado. Logo, em princípio, quanto maior a sequência de elementos antes de verbos principais, maior o custo em termos de processamento da leitura, como pode ser observado em (e), em que o verbo é antecedido pelo seu complemento e pelo sujeito:

(e) A quarta grande diferença entre a vida com a Dinda e a vida normal Cláudia descobriu na última noite: a avó falava com os bichos. (CLPT, livro *Madalena*: 1)

2 Caracterização dos neoleitores e da *Coleção Literatura para Todos* 1

Para avaliar a adequação dos livros ao neoleitores, é fundamental a caracterização desse grupo relativamente à capacidade de leitura. Segundo o documento "Perfil dos

neoleitores no Brasil", disponível na página do MEC,[4] o neoleitor é assim caracterizado:

> Os neoleitores possuem uma concepção de leitura associada à oralização do texto escrito. Fazem uma leitura lenta, entrecortada, com interrupções, cometem omissão de palavras, de trechos, trocam de palavras, fazem pseudoleitura (procuram adivinhar o que está escrito). Evocam conhecimentos prévios para preencher lacunas na leitura e, nesse processo, muitas vezes ouvem mais o que já sabem sobre o tema do que o que o texto diz. Não costumam reler, retomar o texto em busca de informações não retidas na memória. Apreendem o tema, mas têm dificuldade de reproduzi-lo oralmente, falando de experiências próprias relacionadas ao tema.

A CLPT é composta por 10 livros de gêneros diversos, premiados em um concurso literário anunciado em edital público. Os livros foram (ou deveriam ser) escritos especialmente para os neoleitores, em uma iniciativa de grande relevância, tendo em vista o perfil específico do grupo: habilidades de leitura muito iniciais associadas a uma vasta experiência de vida. Desse modo, os autores tinham o desafio de escrever de maneira simples, por um lado, mas com o cuidado de não infantilizar o leitor, por outro.

A CLPT abrange sete gêneros literários – teatro, novela, conto, crônica, biografia, tradição oral e poesia – distribuídos da seguinte maneira (Tabela 1). Notamos que os livros também estão disponíveis no portal Domínio Público, do MEC, para *download*.

Tabela1 – Distribuição dos livros da CLPT por gênero

Gênero	Título do livro
Biografia	– *Léo, o pardo*
Contos	– *Cabelos molhados* – *Cobras em compota*
Crônicas	– *Tubarão com a faca nas costas*
Novela (romance)	– *Madalena*
Poesia	– *Abraão e as frutas;* – *Caravela* [redescobrimentos]; – *Entre as junturas dos ossos*
Teatro	– *Família composta*
Tradição oral	– *Batata cozida, mingau de cará*

Além de um glossário, todos os livros são ilustrados e contêm um prefácio que, por sua vez, é em geral pouco esclarecedor tendo em vista o público pretendido, com o uso frequente de termos difíceis: Um exemplo: *A crônica, com seu característico de mensagem pessoal, humaniza o veículo.*

[4] <http://portal.mec.gov.br/index.php?option=com_content&view=article&id=12313&Itemid=629>.

Devido a limitações das ferramentas utilizadas, não consideramos os livros de poesia e teatro, como será detalhado próxima seção.

3 Análise da Coleção

Para a avaliação, utilizamos ferramentas computacionais capazes de processar textos automaticamente com o objetivo de mensurar os graus de dificuldade dos textos. Tais ferramentas, ao permitirem o processamento automático da informação linguística, possibilitam observação e análise de dados de uma perspectiva quantitativa, o que dificilmente seria conseguido se dependêssemos de um processamento manual.

Como já mencionado, não foram submetidos à análise os livros de poesia, visto que a avaliação da sua complexidade não poderia ser feita com base nos mesmos parâmetros usados para a análise dos textos em prosa. A ausência de pontuação, principalmente, é um problema para o analisador sintático, que considera a mudança de linha como marca de sentença – como seria o caso, por exemplo, de títulos em jornais e artigos científicos. Assim, na poesia, o procedimento de análise automática resultaria em erro.

Da mesma forma, o texto do gênero teatro (*Família Composta*) contém especificidades de sua organização textual-discursiva que dificultam o processamento automático: as convenções do texto teatral (indicação de personagem e rubricas de cenário ou de informação para os atores), expressas em geral na forma de frases nominais, quebram a expectativa do analisador automático, e podem resultar em erros de análise. Optamos, mesmo assim, por apresentar os resultados da análise desse livro, conscientes de que sua interpretação deve ser vista com muitas ressalvas.

3.1 *Ferramentas*

Para a investigação dos parâmetros mencionados na seção 2.1, utilizamos o analisador morfossintático PALAVRAS (BICK, 2000) e o programa Coh-Metrix Port (ALMEIDA e ALUÍSIO, 2009).

O analisador PALAVRAS foi fundamental para a análise detalhada das estruturas sintáticas dos livros da CLPT. Baseado no modelo de Gramática Constritiva (KARLSSON, 1990; KARLSSON et al., 1995), o programa é capaz de realizar uma análise gramatical e sintática de textos da língua portuguesa com um alto grau de precisão – 99% em termos de morfossintaxe (classe de palavras e flexão) e 97-98% em termos de sintaxe (BICK, 2005).

No Quadro 1 apresentamos, a título de ilustração, a saída do PALAVRAS, no formato de "árvores deitadas".

Para cada frase, o programa disponibiliza a informação linguística em pares do tipo Função & Forma, separados por dois pontos (F:f).[5]

Quadro 1 – Exemplo de uma frase analisada pelo programa PALAVRAS

Não devia existir colesterol naquela época, e é aí que começou o problema.

```
STA: par
CJT: fcl
=ADVL: adv (“não” <left>) não
=P: vp
==VAUX: v-fin (“dever” <fmc> IMPF 3S IND VFIN) devia
==MV: v-inf («existir» <mv>) existir
=SUBJ: n(«colesterol» M S) colesterol
=ADVL: pp
==H: prp («em» <sam-> <right>) em
==P<:np
===>N: pron-det («aquele» <dem> <-sam> DET F S) aquela
===H: n(«época» F S) época
,
CO: conj-c(“e” <co-fin> <co-fmc>) e
=CJT: x
FOC: adv (“ser” <foc>)          é
=ADVL: adv (“aí” <kc> <left>) aí
=FOC: adv (“que” <foc>) que
=P: v-fin (“começar” <fmc> PS 3S IND VFIN) começou
=SUBJ: np
==>N: pron-det (“o” <artd> DET M S) o
==H:n(“problema”M S) problema
```

Além do PALAVRAS, os textos da CLPT foram analisados pelo programa Coh-Metrix Port (ALMEIDA e ALUÍSIO, 2009), desenvolvido a partir das métricas da ferramenta Coh-Metrix, criada na Universidade de Memphis. A versão 1.0 do Coh-Metrix Port utiliza 34 das 60 métricas disponíveis na versão livre Coh-Metrix. Essas métricas levam em consideração vários níveis de análise linguística: léxico, sintático e discursivo. A ferramenta disponibiliza também o índice Flesch, medida estatística considerada padrão quanto ao grau de inteligibilidade.[6] Embora considerado um índice superficial, pois leva em conta apenas características como o número de palavras em sentenças e o número de letras ou sílabas por palavra, o índice é utilizado por ser a única métrica de inteligibilidade já adaptada para a língua portuguesa (MARTINS et al., 1996) e por incorporar o conceito de séries escolares. A aplicação da fórmula Flesh permite categorizar os textos em

- textos muito fáceis (índice entre 75-100), adequados para a escolaridade até a 4ª série do ensino fundamental;
- textos fáceis (índice entre 50-75), adequados para a escolaridade até a 8ª série do ensino fundamental;
- textos difíceis (índice entre 25-50), adequados ao ensino médio ou universitário e;
- textos muitos difíceis (índice entre 0-25), adequados apenas para áreas acadêmicas específicas.

3.2 *Análise e resultados*

A Tabela 2 apresenta, por livro, os resultados obtidos para cada parâmetro. Os quatro primeiros parâmetros foram obtidos por meio da análise do PALAVRAS, e o último foi obtido pelo Coh-Metrix Port.[7] Destacamos em negrito os resultados mais significativos.

Para facilitar a visualização dos resultados, apresentamos também as Figuras 1 e 2, que representam, respectivamente, os índices de inteligibilidade relativos a cada livro e a comparação entre livros em relação a cada índice.

Tabela 2 – Índices de inteligibilidade para os textos em prosa, por livro.

	Léo, o pardo	*Cabelos molhados*	*Cobras em compota*	*Madalena*	*Tubarão com a faca nas costas*	*Família composta*
Verbos por período	**2,61**	2,22	1,86	1,71	2,24	2,09
Elementos explicativos (intercalados) p/ or.	**0,25**	0,12	0,04	0,09	0,11	0,06
Vírgulas por período	**1,92**	1,02	0,68	0,72	1,18	1,02
Or. reduzidas de gerúndio	**0,16**	0,12	0,05	0,06	0,08	0,08
Palavras antes de verbos principais	3,07	2,18	2,06	2,2	2,53	**4,07**

[5] Especificamente, a codificação da frase exemplo informa que a frase é declarativa (STA – “statement”) e Coordenada (“par”). Na segunda linha está a informação dos elementos coordenados: CJT (“elemento conjunto”): fcl (oração finita). Ou seja, estamos diante de uma frase declarativa, que por sua vez é um período composto por uma coordenação de orações finitas. Para cada palavra, além das informações de forma e função, o programa indica, entre parênteses, o lema e informações morfossintáticas, como número, gênero, tempo, modo e pessoa verbal. Para uma explicação da análise realizada pelo PALAVRAS, remetemos o leitor a Bick (2000), bem como à página do projeto VISL) e, para o leitor interessado no formato árvores deitadas, sugerimos a leitura de Freitas e Afonso (2008).

[6] A fórmula Flesh é ILF = 164.835 – [1.015 × (Nº palavras/sentença)] – [84.6 × (Nº sílabas/texto/Nº palavras/texto)]

[7] Lembramos que, diferentemente do Coh-Metrix Port, o PALAVRAS não oferece diretamente os valores para os parâmetros – estes são obtidos por meio de uma busca semiautomática nos resultados da análise automática.

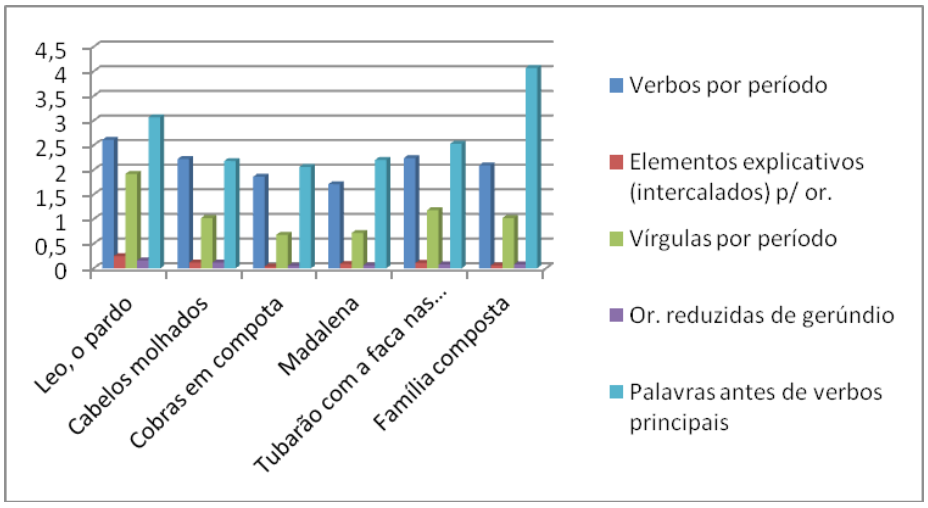

Figura 1 – Índices de inteligibilidade por livro analisado

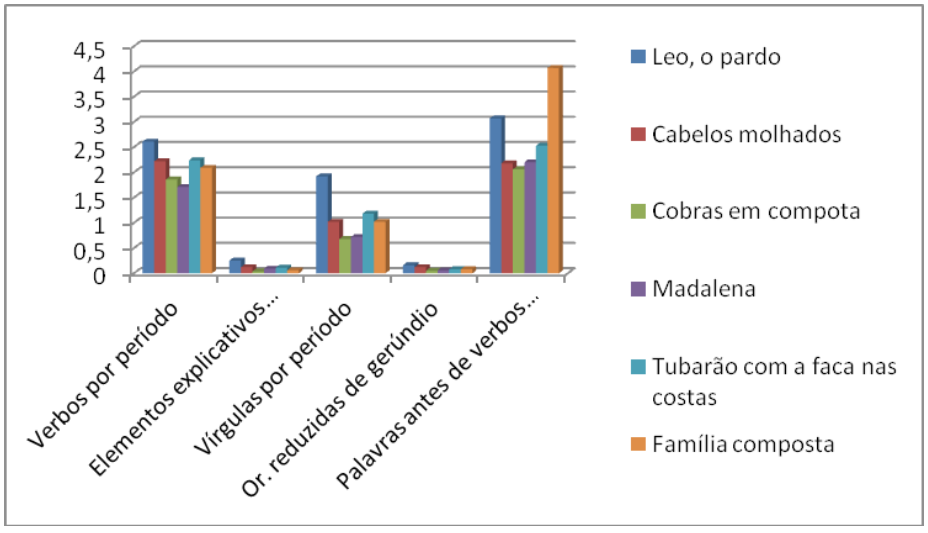

Figura 2 – Comparação dos livros em relação aos índices de inteligibilidade

Com base no parâmetro **total de verbos por período**, que permite verificar o número de períodos simples/compostos, os livros estruturalmente mais complexos seriam *Léo, o pardo* (2,61), *Tubarão com a Faca nas Costas* (2,24) e *Cabelos Molhados* (2,3). *Cobras em Compota* e *Madalena* apresentam níveis próximos, seriam os mais fáceis – em torno de 1,7 orações por período.

Tomando-se o parâmetro **elementos explicativos intercalados**, *Léo, o pardo* apresenta o maior valor – mais que o dobro de *Cabelos Molhados* e *Tubarão com a Faca nas Costas*, que aparecem em segunda posição. Nos outros livros, o total de intercalações tende a 0, indicando menor complexidade sintática.

Em relação ao parâmetro **número de vírgulas por período**, o livro *Léo, o pardo* novamente se destaca, com o dobro de ocorrências por período em comparação ao segundo colocado – *Cabelos Molhados*. Os demais livros não apresentam diferenças significativas entre si.

No parâmetro **orações reduzidas de gerúndio,** ainda que seu percentual seja bem pequeno em todos os livros, novamente *Léo, o pardo* e *Cabelos Molhados* aparecem como os mais complexos.

Relativamente ao parâmetro **total de palavras antes de verbos**, o livro *Léo, o pardo* aparece mais uma vez como o mais complexo.

A Tabela 3 apresenta os resultados do Índice Flesh, que permitem complementar os dados das análises anteriores com informação relativa à adequação dos livros às séries escolares (quanto maior o índice, mais fácil o livro). Excetuando-se o livro *Família Composta*, do gênero teatro, cuja análise deve ser vista com cautela devido à estruturação do texto (cf. seção 4), todos os livros são indicados para 5ª-8ª série, e, portanto, indicados para leitores com alguma proficiência, o que não é o caso do público alvo da CLPT.

Tabela 3 – Índice Flesh dos livros da CLTP.

Título do livro	Gênero	Índice Flesh	
Madalena	novela	63	(5ª - 8ª)
Cabelos molhados	conto	65	(5ª - 8ª)
Cobras em compota	conto	64	(5ª - 8ª)
Tubarão com a faca nas costas	crônica	63	(5ª - 8ª)
Família composta	teatro	75	(1ª - 4ª)
Léo, o pardo	biografia	63	(5ª - 8ª)

3.3 *Contextualização dos parâmetros*

Como mencionado na introdução deste artigo, não temos conhecimento de outros trabalhos que tenham realizado uma análise parecida com a que apresentamos. Logo, com base nos critérios estabelecidos, é possível uma comparação da inteligibilidade entre os livros da Coleção tomados isoladamente, mas, com exceção do índice Flesch, não temos como verificar a inteligibilidade da Coleção relativamente a livros indicados para recém-alfabetizados.

Assim, em uma tentativa de contextualizar a análise quantitativa, aplicamos os mesmos parâmetros a outro livro, que não foi escrito especificamente para neoleitores: *O Jardim do Diabo*, romance de Luis Fernando Veríssimo. A escolha da obra foi motivada pela necessidade de calibrar as medidas obtidas, já que L.F. Veríssimo é um escritor popular, de escrita acessível, que agrada a diferentes idades e classes, e cujo conteúdo é de reconhecida qualidade - portanto seria uma boa maneira de comparar o quão acessíveis seriam os livros da CLPT.

Subjacente à comparação, tínhamos a hipótese de que, idealmente, os livros da CLPT deveriam ser mais simples que *O Jardim do Diabo*, pois foram elaborados especialmente para serem lidos por recém-alfabetizados. Os resultados dessa segunda análise estão na Tabela 4, com os resultados mais significativos – isto é, aqueles que indicam mais facilidade de leitura – em negrito.

Como é possível observar, em todos os parâmetros investigados o livro *O Jardim do Diabo* está entre aqueles com o menor grau de complexidade, e, especificamente, nos parâmetros "verbos por período" e "vírgulas por período", chega a obter os índices mais baixos. Os resultados dessa comparação, somados aos dados do índice Flesh (Tabela 3) e à nossa impressão após a leitura completa e minuciosa da Coleção, reforçam a argumentação relativa à inadequação dos livros – em termos de inteligibilidade – ao público pretendido. A Figura 3 apresenta outra forma de visualização da comparação entre os livros da CLPT e o livro *O Jardim do Diabo*.

Quanto ao índice Flesh, *O Jardim do Diabo* obteve 73.67 (adequado para 1ª à 4ª série), pontuação que nos permite classificá-lo como mais fácil do que todos os da CLPT.

Tabela 4 – Comparação entre os livros da CLPT e o livro *O Jardim do Diabo.*

	Léo, o pardo	*Cabelos molhados*	*Cobras em compota*	*Madalena*	*Tubarão com a faca nas costas*	*Família composta*	*O Jardim do Diabo*
Verbos por período	2,61	2,22	1,86	1,71	2,24	2,09	**1,54**
Elementos explicativos p/ or.	0,25	0,12	**0,04**	0,09	0,11	**0,06**	**0,08**
Vírgulas por período	1,92	1,02	0,68	0,72	1,18	1,02	**0,66**
Or. reduzidas de gerúndio	0,16	0,12	**0,05**	**0,06**	0,08	0,08	**0,06**
Palavras antes de verbos principais	3,07	**2,18**	**2,06**	**2,2**	**2,53**	4,07	**2,93**

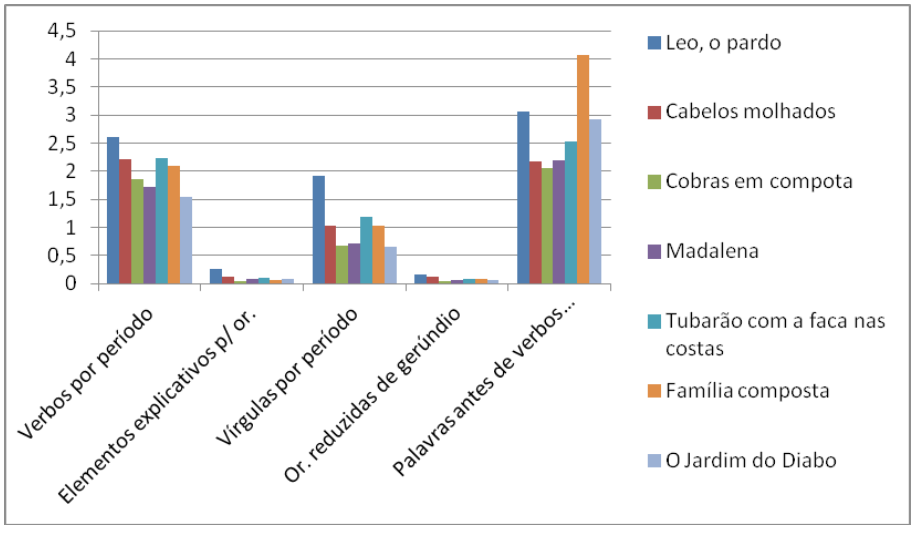

Figura 3 – Comparação entre os livros da CLPT e o livro *O Jardim do Diabo*

4 Considerações finais

Apresentamos aqui a análise de fatores de inteligibilidade dos livros da CLPT.

Ainda que os resultados obtidos apontem claramente para a inadequação dos livros de um ponto de vista linguístico, desconhecemos dados relativos à forma com que os livros foram/ são trabalhados – se lidos em voz alta pelo professor/mediador, por exemplo, o que pode diminuir o impacto de nossas observações.

De maneira geral, os resultados de nossa investigação sugerem que os livros da CLPT são, em sua maioria, complexos para o público pretendido, diferentemente do apresentado em MACIEL (2007). Assumindo que os neoleitores estão na etapa inicial do processo de alfabetização, é nítido que os livros da CLPT exigem um esforço de decodificação da escrita que está além de sua capacidade.[8]

A partir dos resultados, o livro *Léo, o pardo* aparece como mais difícil dentre os livros da CLPT, o que está em consonância com a análise baseada em nossa leitura. Trata-se de uma biografia cuja linguagem é altamente oralizada, o que exige a habilidade de decifrar frases muito longas, com inserções de discurso relatado (direto, indireto, indireto livre) sem a codificação tradicional. Da oralização decorre também a construção de sentenças com muitos elementos intercalados (apostos, adjuntos), coordenações longas ou ordem sintática canônica invertida. Por fim, a comparação entre os livros da CLPT e o livro *O Jardim do Diabo*, de L. F. Veríssimo, reforça a possibilidade de dissociação entre complexidade sintática, estrutural e complexidade quanto ao conteúdo, aspecto fundamental quando consideramos as especificidades de uma proposta como a CLPT: uma literatura gramaticalmente simples, mas que não infantiliza os leitores.

Novamente, consideramos louvável o reconhecimento de que os neoleitores são um público especial de leitores e que, portanto, merecem atenção especial no que se refere à constituição de um acervo bibliográfico. Mas acreditamos ser fundamental o reconhecimento – da parte de quem escreve e da parte de quem avalia a adequação dos livros ao público – da possibilidade de dissociação entre aspectos gramaticais e aspectos referentes ao conteúdo dos livros. No edital do concurso de 2010[9,10], não há qualquer menção a aspectos da estrutura linguística a serem considerados pelos autores; trata-se apenas do aspecto narrativo, como "narrativa literária atraente, destinada à captura do neoleitor, não se confundindo com objetivos escolares de ensino da língua e da gramática (...)". Relativamente à construção dos textos, o edital recomenda "na construção dos textos, em todos os gêneros, a leveza e a invenção poética, e assim aglutinar forças para o enfrentamento dos problemas e limites da realidade". A nosso ver, tais recomendações em nada consideram as habilidades de leitura dos neoleitores retratadas pelo próprio MEC, que apresentamos na seção 3.

Por fim, temos consciência de que nosso trabalho de análise é bastante inicial, tanto de um ponto de vista metodológico quanto de interpretação dos resultados obtidos. A exploração dos mesmos parâmetros em livros infantis, recomendados para crianças recém-alfabetizadas, pode nos dar pistas em termos da adequação linguístico-textual dos livros. Por outro lado, a investigação de outros aspectos mencionados na literatura psicolinguística como "dificultadores" da leitura pode sugerir novos pontos a serem considerados.

[8] Segundo o documento Brasil alfabetizado: marco referencial para avaliação cognitiva (BATISTA et al., 2006), os neoleitores são capazes de (a) identificar letras do alfabeto; (b) conhecer as direções da escrita; (c) diferenciar letras de outros sinais gráficos; (d) identificar, ao ouvir uma palavra, o número de sílabas; (e) identificar sons, sílabas e outras unidades sonoras; (f) distinguir, como leitor, diferentes tipos de letra; (g) demonstrar conhecimentos sobre a escrita do próprio nome; (h) escrever palavras ditadas, demonstrando conhecer o princípio alfabético.

[9] <http://portal.mec.gov.br/index.php?option=com_docman&task=doc_download&gid=6587&Itemid=>.

[10] Lembramos que nosso trabalho consistiu na análise da CLPT1.

Referências

ALMEIDA, Daniel Machado de; ALUÍSIO, Sandra Maria. *Manual de uso do Coh-Metrix Port 1.0.*Technical Report NILC-TR-09-05, 13 p. Agosto 2009, São Carlos-SP.

BARBOZA, Elza M.; NUNES, Eny M. de A. A inteligibilidade dos websites governamentais brasileiros e o acesso para usuários com baixo nível de escolaridade. *Inclusão Social*, Brasília, v. 2, n. 2, p. 19-33, abr./set. 2007.

BATISTA, Antônio Augusto Gomes; SILVA, Ceris Ribas da; CASTANHEIRA, Maria Lucia; ROCHA, Gladys; CAFIERO, Delaine. Matriz de Referência: avaliação de competências – Leitura e escrita. In: HENRIQUES, Ricardo; BARROS, Ricardo Paes; AZEVEDO, João Pedro. (Orgs.). *Brasil Alfabetizado*: marco referencial para a avaliação cognitiva. Brasília: Secretaria de Educação Continuada, Alfabetização e Diversidade, 2006. p. 12-27.

BEN-ANATH, Dafna. The Role of Connectives in Text Comprehension. Teachers College, *Columbia University Working Papers in TESOL & Applied Linguistics*, v. 5, n. 2, p. 1-27, 2005.

BICK, Eckhard. Gramática Constritiva na Análise Automática de Sintaxe Portuguesa. In: BERBER SARDINHA, Tony (Ed.). *A Língua Portuguesa no Computador*. Campinas: Mercado de Letras; São Paulo: FAPESP, 2005.

BICK, Eckhard. *The Parsing System "Palavras"*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. Dr.phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press, 2000.

BORBA, Valquíria Claudete M. *Preditibilidade de conjunções e compreensão leitora:* um estudo com crianças de 4ª série do Ensino Fundamental. Dissertação (Mestrado) – PUCRS, 2005.

BRASIL, MEC/SECAD. *Coleção Literatura para Todos 1*, 2006.

CAIN, Kate; NASH, Hannah M. The Influence of Connectives on Young Readers' Processing and Comprehension of Text. *Journal of Educational Psychology*, v. 103, n. 2, p. 429-441, 2011.

COSCARELLI, Carla. V. O ensino da leitura: uma perspectiva psicolingüística. *Boletim da Associação Brasileira de Linguística*, Imprensa Universitária, Maceió, p. 163-174, dez. 1996.

DuBAY, William H. *The principles of readability*. Califórnia: Impact Information, 2004. Disponível em: <http://www.impactinformation. com>. Acesso em: ago. 2012.

FREITAS, Cláudia; AFONSO, Susana. *Bíblia Florestal:* Um manual lingüístico da Floresta Sintá(c)tica. 2008. (Desenvolvimento de material didático ou instrucional – Manual/ Documentação).

KARLSSON, Fred. Constraint Grammar as a Framework for Parsing Unrestricted Text. *Proceedings of the 13th International Conference of Computational Linguistics*, v. 3, Helsinki, 1990. p. 168-173.

KARLSSON, Fred; VOUTILAINEN, Atro; HEIKKILÄ, Juha; ANTTILA, Arto. (Eds.) *Constraint Grammar:* A Language-Independent System for Parsing Running Text. Berlin; New York: Mouton de Gruyter, 1995. (Natural Language Processing, 4).

KATO, Mary A. *O aprendizado da leitura*. 5. ed. São Paulo: Martins Fontes, 1999.

KINTSCH, Walter; van DIJK, Teun A. Toward a model of text comprehension and production. *Psychological Review*, v. 85, n. 5, p. 363-394, 1978.

KLEIMAN, Ângela B. *Oficina de leitura, teoria e prática*. São Paulo: Pontes/Editora da Universidade Estadual de Campinas, 1993.

KLEIMAN, Ângela B. Abordagens da leitura. *SCRIPTA*, Belo Horizonte, v. 7, n. 14, p. 13-22, 1º sem. 2004.

LEFFA, Vison J. *Aspectos da leitura:* uma perspectiva psicolingüística. Porto Alegre: Sagra-Luzzatto, 1996.

LIMA, Vera L. de A. Legibilidade e leiturabilidade das bulas de medicamentos presentes no tratamento de pacientes cardíacos/ Vera Lopes de Abreu Lima; orientador: Anamaria de Moraes. 169 f. Dissertação (Mestrado em Artes e Design) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

MACIEL, Ira Maria. Coleção literatura para todos. *Rev. Bras. Educ.* [online], v. 12, n. 36, p. 537-540, 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-24782007000300014&lng=en&nrm=iso. Acesso em: 13 ago. 2012.

MARTINS, Teresa B. F.; GHIRALDELO, Claudete M.; NUNES, Maria das Graças V.; OLIVEIRA Jr., Osvaldo N. Readability Formulas Applied to Textbooks in Brazilian Portuguese. *Notas do ICMSC*, n. 28, 1996.

MILLIS, Keith K.; JUST, Marcel A. The Influence of Connectives on Sentence Comprehension. *Journal of Memory and Language*, v. 33, p. 128-147, 1994.

PEREIRA, Vera W. Arrisque-se... faça o seu jogo. *Letras de Hoje*, Porto Alegre, v. 37, n. 128, p. 47-63, jun. 2002.

PERFETTI, Charles A. Comprehending written language: A blueprint of the reader. In: BROWN, Colin M.; HAGOORT, Peter. (Eds.). The neurocognition of language. Oxford: Oxford University Press, 1999. p. 167-208.

PERFETTI, Charles A.; LANDI, Nicole; OAKHILL, Jane. The acquisition of reading comprehension skill. SNOWLING, Margaret J.; HULME, Charles. (Eds.). *The Science of Reading: A Handbook*. Oxford: Blackwell, 2005. p. 227-247.

RIBEIRO, Bruno; MODESTO, Débora; CAPRA, Eliane; FERREIRA, Simone B. L. Referencial Teórico sobre Analfabetismo Funcional. Relatórios Técnicos do Departamento de Informática Aplicada da UNIRIO n° 0008/2011. *Relatórios Técnicos de 2011*, v. 5, n. 1. Disponível em: http://www.seer.unirio.br/index.php/monografiasppgi/article/view/1498/1379. Acesso em: 13 ago. 2012.

SANDERS, Ted J. M.; NOORDMAN, Leo G. M. The Role of Coherence Relations and Their Linguistic Markers in Text Processing. *Discourse Processes*, v. 29, n. 1, p.37-60, 2000.

SCARTON, Carolina E.; ALUÍSIO, Sandra Maria. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, v. 2, n. 1, p. 45-62, 2010.

SCARTON, Carolina E.; OLIVEIRA, Matheus de; CANDIDO Jr., Arnaldo; GASPERIN, Caroline; ALUÍSIO, Sandra Maria. SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. *Proceedings of the NAACL HLT 2010*: Demonstration Session (p. 41-44). Los Angeles, Califórnia, junho 2010. Morristown, NJ, USA: Association for Computational Linguistics, 2010.

Recebido: 30 de agosto de 2012
Aprovado: 24 de agosto de 2012
Contato: ericasr@puc-rio.br; maclaudia.freitas@gmail.com;
violetaquental@gmail.com