COMPLEXIDADE TEXTUAL EM ARTIGOS CIENTÍFICOS: CONTRIBUIÇÕES PARA O ESTUDO DO TEXTO CIENTÍFICO EM PORTUGUÊS[1].

Maria José Bocorny Finatto[*]

Resumo: *This text presents the basis of a research project that deals with the issue of textual complexity (TC), examining aspects of Pediatrics papers against newspaper articles. In the literature review, it is mobilized works related to the subject of TC in Applied Linguistics, Terminology studies that follow a textual point of view, Computational Linguistics and Corpus Linguistics. Some highlighted elements in the contrast between the examination of texts and the literature reviewed are the measures to TC ratio and degrees of text specialization, which would preview differences between specialized language and everyday language. The paper concludes with the presentation of prospects for the treatment of the issue of TC among the studies on scientific discourse.*

Palavras-chave: *textual complexity, applied lingusitics, corpus linguistics, terminology, specialized.*

INTRODUÇÃO

O principal objetivo deste trabalho é refletir sobre modos para tratar do tema da complexidade textual (doravante CT) no âmbito dos estudos sobre textos e linguagens especializadas. A pergunta que guia o trabalho é a seguinte: haveria como avaliar em que medida textos científicos do tipo artigo seriam mais ou menos complexos em relação a textos científicos de outros perfis ou mesmo em relação a textos não-especializados[2]?

Ao ponderar sobre as contribuições da Terminologia de perspectiva textual (CIAPUSCIO, 2003), da Linguística Aplicada (LA), da Linguística de *Corpus* (LC), especialmente a Análise Multi-Dimensional da LC (proposta por BIBER, 1988), e da Linguística Computacional (LCOMP), representada aqui por um sistema para mensuração de graus de complexidade ou de inteligibilidade de textos para diferentes usuários (SCARTON, ALUÍSIO, 2010), discute-se como essas perspectivas poderiam cooperar em prol de um entendimento sobre fatores e fenômenos que perfazem a complexidade de um texto científico. Esse entendimento, tal como posto aqui, pode beneficiar sobretudo linguistas interessados em descrever e em analisar a constituição de textos especializados, seja do ponto de vista terminológico, seja do ponto de vista discursivo-textual ou mesmo gramatical.

Em síntese, pretende-se evidenciar como diferentes metodologias descritivas, de diferentes procedências teóricas, poderiam ser aplicáveis à observação de textos do tipo artigo científico.

---

[1] Este texto contém as bases da pesquisa de pós-doutoramento realizada entre fevereiro e julho de 2011I junto ao NILC-ICMC-USP (Núcleo Interinstitucional de Linguística Computacional do Instituto de Ciências Matemáticas e Computacionais da Universidade de São Paulo, *campus* de São Carlos – SP).

[*] Pesquisadora do Grupo TERMISUL, coordenadora do Projeto TEXTECC e TEXTQUIM, bolsista produtividade em pesquisa do CNPq, pós-doutoranda NILC-ICMC-USP.

[2] Não faremos aqui a ponderação de praxe sobre a diferença entre textos especializados e não especializados. Essa é uma discussão que replica a oposição termo/palavra "comum". Assumiremos apenas, tal como explica Maciel (2010, p.25), que "a realização Linguística do texto especializado, nela compreendida sua estruturação gramatical, textual e terminológica e ainda sua formatação gráfica, depende de fatores temáticos e pragmáticos. A influência desses fatores se faz sentir tanto na ativação do valor especializado das palavras que, no evento comunicativo, desempenham a função de vetor da transmissão da informação, da instrução, do mandamento, da sugestão e do conselho, como na seleção dos elementos lexicais que os articulam na estrutura sintática e na configuração discursiva."

Nesse sentido, tanto em termos de tratamento de unidades textuais quanto em termos de tratamento de *corpora,* sinaliza-se a importância do diálogo entre LA, LC e LCOMP a favor de incrementar-se o estudo de padrões de complexidade textual (CT) associados aos textos especializados em geral.

Como pano de fundo para ilustrar o potencial de sinergia dessas contribuições, é abordada a complexidade de um exemplar de artigo científico de Pediatria sob a ótica das diferentes perspectivas mencionadas. E, como um brevíssimo contraponto ilustrativo para a condição de especialização[3] do texto em foco, toma-se um texto de editoria geral de notícias de um jornal popular brasileiro[4]. É um jornal diário da cidade de Porto Alegre – RS dirigido a públicos leitores de menor poder aquisitivo, adultos com escolaridade média estimada correspondente ao Ensino Fundamental completo de oito anos[5].

O trabalho está organizado da seguinte maneira: na **primeira parte**, denominada *complexidade textual em revisão*, é feita uma varredura bibliográfica em busca de trabalhos ou de propostas, de diversas procedências, com destaque para os estudos de leitura, que possam ser associadas de alguma forma ao tratamento do tema da CT em textos especializados. Depois, **na segunda parte**, caracteriza-se a perspectiva da AMD para a descrição de textos, ilustrando-se essa proposta metodológica com a síntese de um trabalho brasileiro (SHERGUE, 2003) dedicado ao estudo de artigos médicos da área de Hematologia, tendo sido tais textos contrapostos a textos transcritos de comunicações orais em congresso na mesma especialidade.

Na **terceira parte**, traz-se o enfoque da LCOMP, em uma parte eminentemente experimental e exploratória deste texto, na qual observam-se diferentes medidas de complexidade textual geradas pela ferramenta computacional Coh-Metrix partindo-se de um trecho da Constituição do Brasil e de um artigo de Pediatria. Na **quarta parte**, relaciona-se a metodologia Coh-Metrix e os seus resultados com as considerações de níveis textuais de Ciapuscio (2003), conectando-se dimensões e fatores da AMD. Ao final dessa parte, concluindo o texto, o trabalho faz considerações sobre possibilidades para agregação do tema da CT aos estudos sobre textos especializados.

PRIMEIRA PARTE - COMPLEXIDADE TEXTUAL EM REVISÃO

No panorama da Linguística Aplicada (LA) nacional e internacional**,** o tema da *complexidade textual* (CT) integrou estudos sobre Leitura, incluindo pesquisas sobre compreensão e estratégias de leitura, sobre tipificação de leitores e sobre elementos linguísticos associados a dificuldades de compreensão de leitura. Embora esses estudos tenham gerado importantes contribuições, como a distinção entre *complexidade informativa* e *complexidade Linguística*, permanecem escassos, no Brasil, os trabalhos baseados em *corpora*, realizados com grandes extensões de dados e apoio informatizado, dedicados a reconhecer características estruturais globais de textos mais ou menos complexos em função das habilidades ou condições de determinados tipos de leitores.

Essa escassez, conforme se pode interpretar, está relacionada a dois fatores. Primeiro, ao relativamente recente enfrentamento do objeto texto, geralmente preterido em função de enfoques dedicados a frases, palavras ou expressões sintagmáticas. Segundo, a uma pouca experiência com a manipulação computacional de grandes *corpora*, algo recente no âmbito dos

[3] Essa condição de especialização, bem sabemos, tem sido muito discutida e debatida. Para um boa revisão a respeito, recomendo o trabalho de Zilio (2009).

[4] Textos do jornal popular porto-alegrense *Diário Gaúcho*, disponíveis para estudo no *site* www.ufrgs.br/textecc , projeto PorPopular.

[5] O nível de escolaridade do leitor do DG está aqui apenas grosseiramente estimado. O público leitor corresponde ao que se denomina públicos das classes C e D. Sua tiragem media diária é de 150 mil exemplares; é apenas vendido em bancas, não tem assinatura. Circula apenas na cidade de Porto Alegre e região metropolitana e cada exemplar exemplar tende a ser compartilhado por pelo menos 5 pessoas. O jornal circula há 11 anos e é publicado pela empresa RBS, que também publica jornais para os públicos das classes A e B.

Estudos da Linguagem no nosso país.

Por outro lado, há, na bibliografia estrangeira, registros de pesquisas sobre *readbility* ou legibilidade ou complexidade Linguística pelo menos desde os anos 1920, conforme já assinalaram Davison e Green (1988, p.1-4). Esses trabalhos trataram desde a compreensão de palavras até a compreensão de sentenças, chegando a textos de literatura, especialmente histórias curtas ou contos para crianças e jovens, tendo sido contemplada inclusive a compreensão de leitura de adultos com dificuldades cognitivas[6]. Sobre compreensão do texto científico ou técnico, entretanto, as referências são relativamente poucas[7].

Na obra fundamental das linguistas norte-americanas Davison e Green (*op.cit*., 1988) intitulada *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, por exemplo, há apenas dois trabalhos dedicados a problemas de compreensão ou de acessibilidade de textos científicos ou técnicos em um nível global. Há um trabalho dedicado a uma amostra de textos operativos da Marinha da OTAN (BAKER, ATWOOD E DUFFY, 1988). Esse trabalho tratou de trechos de manuais de instrução, os quais foram apresentados em versões originais e simplificadas para testes de compreensão com um grupo de leitores técnicos de formações diferenciadas. O outro trabalho que há nessa obra foi dedicado a cartas de *recall* de fabricantes de veículos[8] (CHARROW, 1988). Nele há interessantes propostas para a elaboração dessas cartas de um modo mais acessível para um consumidor leigo; entretanto, como a compreensão de um todo é o objeto privilegiado, a presença de terminologias como um fator de dificuldade é tratada apenas de modo incidental.

Enfim, desde muito tempo, buscaram-se fórmulas ou modelos – sempre muito discutidos e criticados – que fossem capazes de prever quais elementos textuais estariam mais associados à dificuldade de compreensão da escrita, de modo que pudessem ser gerados textos de acesso mais facilitado para uma grande fatia de população leitora. Essa população, cabe situar, correspondia a grupos sociais de escolarização recente. Entre esses estudos mais antigos de amplo espectro, não associados a uma perspectiva específica de Linguística, produzidos por volta dos anos 70, entretanto, não encontramos muitas referências sobre as condições de legibilidade de textos especializados.

No Brasil, um dos primeiros linguistas a se debruçar sobre o tema da leitura funcional e da maior ou menor habilidade de leitura foi Perini (1982), com o trabalho *Tópicos discursivos e legibilidade* (*apud* FULGÊNCIO, LIBERATO, 2004, p. 9). Propunha o autor, então, que os estudantes brasileiros tivessem acesso a materiais de leitura graduados de acordo com o seu nível de escolaridade e nível de dificuldade de compreensão.

A partir do legado de trabalhos fundadores tais como o de Perini, antes referido, Neis (1982) e Kleiman (1987, 1989, 1993, 1997), Kato (1982) e Averbuck, Appel e Hessel (1983), entre outros, produzidos especialmente ao longo dos anos 80 e 90, temos hoje no Brasil um vasto e multifacetado alicerce de estudos sobre o tema da Leitura. Esse corpo de conhecimento permitiu-nos hoje distinguir especificidades das noções de leitura, alfabetização, letramento, competência textual, competência lexical e competência leitora. Isso sem mencionarmos os inúmeros trabalhos sobre o tema da Leitura na área da Educação, Ensino de Língua Portuguesa e de Línguas Estrangeiras ou de Psicolinguística.

Nacionalmente, entre os vários trabalhos dedicados ao tema da compreensão de leitura, a partir dos anos 90, destacam-se as obras de Kleiman (1997) e de Leffa (1996). Leffa, por exemplo, já apontava que uma descrição completa do processo da compreensão deve levar em conta, no mínimo, três aspectos essenciais: o texto, o leitor e as circunstâncias em que se dá o

---

[6] Uma obra indicada pelas autoras é *What makes a book readable*?, publicada em 1935 (GRAY, LEARY,1935). Essa obra tentava predizer dificuldades de compreensão de leitura de adultos com algum tipo de déficit cognitivo considerando um universo de 350 livros.

[7] Naturalmente, há que se considerar que o texto científico se só coloca como tal, institucionalmente, a partir dos anos 1930, quando ocorre uma primeira reunião internacional de editores de textos científicos. Além disso, a leitura "técnica" ou científica só se distingue como tal à medida que haja também uma institucionalização da formação profissional, a qual gera e consome registros escritos sobre um saber e um saber-fazer.

[8] Interessante como esse tipo de texto atualmente tornou-se abundante no Brasil; cada vez compramos mais automóveis e já temos o anúncio de *recall* recorrentemente presente em jornais de circulação diária.

encontro entre ambos.

Ao tratar do papel do texto, Leffa (*op.cit*) observa que, nos estudos atuais, ainda persiste a preocupação centrada no léxico e na estrutura sintática das frases. Porém, conforme situa, diferentemente de estudos desenvolvidos durante as décadas de 50 e 60, a análise do objeto-texto evoluiu da micro para a macroestrutura. Assim, na sua interpretação sobre uma trajetória de investigações, a compreensão de um texto deixou de ser entendida apenas como um processo linear. Isso ocorreu à medida que se passou a valorizar a apreensão não-linear de segmentos selecionados.

Além da apresentação gráfica do texto (que o autor associa com *legibilidade*) e do uso de palavras freqüentes e estruturas sintáticas menos complexas (relacionada por ele com *inteligibilidade*), fatores tradicionalmente conhecidos como facilitadores da compreensão e a organização interna ou estrutural do texto também conquistaram destaques importantes em meio às investigações. Ainda que não tivessem o objetivo de tratar de um determinado tipo de texto, tampouco de textos especializados, os trabalhos de Kleiman e de Leffa, por sua amplitude e qualidade, têm sido muito referidos quando se trata de ensino de leitura em línguas estrangeiras, sobretudo no segmento denominado *Leitura Instrumental*.

Pois, justamente no âmbito dos estudos de Leitura Instrumental ou de LSP (*Language for Specific Purposes*), encontra-se uma significativa produção sobre leitura e escrita de textos científicos e técnicos. Ainda que o foco seja bastante centrado em uma escrita acadêmica associada ao ensino/aprendizagem de línguas estrangeiras, há muito que se pode aproveitar para a descrição de características desses textos, mesmo que a complexidade seja um assunto um pouco incidental. E, mais recentemente, pelo menos no Brasil, com a combinação dos estudos em *corpora* com ensino de línguas estrangeiras, tal como vemos em Viana e Tagnin (2010), há uma oferta de subsídios aproveitáveis para caracterizar diferentes LSPs, importantes também para o ensino de tradução científica e técnica.

Outras áreas de estudos que têm rendido boas considerações sobre a natureza e especificidades do texto científico ou técnico são a Análise Crítica do Discurso (ACD) e os estudos sobre Gêneros Textuais, cabendo destacar, no âmbito internacional, os trabalhos de Swales (1990) dedicados à escrita acadêmica, especialmente artigos científicos. No cenário brasileiro, a obra de Meurer e Mota-Roth (2005), por exemplo, apresenta a ACD e traz uma série de ensaios que visam, entre outros assuntos, identificar traços de gêneros textuais e discursivos tais como folhetos e relatórios de empresas.

Sob a perspectiva da Linguística de *Corpus* (LC), num âmbito global, pode-se considerar que o tema da CT (independentemente do tipo de texto envolvido, seja literatura ou texto técnico-científico) tenha sido parcialmente contemplado.

No âmbito brasileiro, por exemplo, não se pode deixar de citar o pioneiro Projeto DIRECT[9] – em direção à linguagem do trabalho. Esse grupo de pesquisa, desde 1991, junto ao LAEL da PUC-SP, trata de textos especializados da área de Administração de Empresas e Negócios. O DIRECT objetiva promover estudos sobre a linguagem das profissões, em português, inglês e espanhol. Descreve contextos originais de interação profissional, tais como reuniões de negócios, documentos empresariais de circulação restrita, e textos empresariais de domínio público em que o português (como língua materna) e o inglês (como língua estrangeira) são utilizados. Além disso, visa identificar as causas de possíveis problemas de comunicação no ambiente empresarial através da análise detalhada de situações bem-sucedidas. Volta-se, desse modo, para a descrição de gêneros linguísticos e de processos discursivos.

Quanto à LC em um âmbito internacional mais global, a qual dá suporte a vários estudos do DIRECT, é importante registrar a contribuição da Análise Multidimensional (AMD), proposta por Biber em 1988 (BIBER, 1988). Essa proposta para tratamento da variação ao longo de gêneros textuais e discursivos ou registro não trata diretamente do tema da CT. Entretanto, a metodologia descritiva da AMD associada a todo um conjunto de princípios teóricos da LC (tal

[9] Para mais detalhes, ver <http://www2.lael.pucsp.br/direct/projeto.htm >

como apresentados no Brasil por BERBER SARDINHA, 2004), conforme se pretende explicitar mais adiante, colocaria, desde suas bases nos anos 80, uma série de procedimentos aplicáveis à investigação de um fator como a CT.

O enfoque da AMD combinou análises de *corpus* de nível macro com análises de nível micro, em encaminhamentos da macrodimensão do *corpus* à microdimensão do texto e das sentenças que o integram. Nessa perspectiva, a microdescrição dos traços de cada texto deveria permitir a indução dos macro-agrupamentos textuais, tipificando-os por gêneros (cf. explica BERBER SARDINHA, 2000). Assim, a AMD, ao identificar tipos discursivos ou textuais, embora não tenha tratado diretamente de CT, propôs métodos descritivos da linguagem escrita úteis para a ponderação sobre características de determinados tipos de texto. Esses métodos, sem desconsiderar uma certa e inerente complexidade de aplicação para pessoas pouco afeitas a análises estatísticas multifatoriais, poderiam ser associados a medidas ou fatores de maior ou menor complexidade textual.

Na esteira da AMD, no cenário fora do Brasil, o trabalho de Atkinson (1992), por exemplo, tratou de artigos científicos sob uma perspectiva diacrônica. Seu diferencial foi justamente o de ter empregado uma metodologia de observação baseada em *corpus* para descrever o comportamento desse tipo de texto num intervalo de 1735 a 1985. No seu trabalho, não encontramos menção direta sobre complexidade textual, salvo o reconhecimento de uma certa prolixidade deliberada da retórica científica em inglês, recomendada por *Sir* Robert Boyle, precursor da Física e Química Modernas no século 18. Não será aprofundada aqui a descrição desse estudo visto que, mais adiante, dedicamos um segmento desta revisão para a AMD.

De outro lado, no âmbito da Linguística Computacional (LCOMP), pelo menos desde os anos 60, muito já foi e tem sido produzido sob forma de sistemas que geram versões mais simplificadas de textos, incluindo a produção de ferramentas capazes de indicar diferentes tipos de medidas de CT. Essas ferramentas também conseguem produzir diferentes tipos de representações esquemáticas do conteúdo de um texto ou de todo um *corpus*. Esses sistemas de LCOMP, de base fundamentalmente estatística, conseguem inclusive reconhecer tipologias textuais e graus de complexidade a eles associadas. Voltaremos mais adiante a esse tipo de enfoque computacional quando tratarmos do sistema Coh-Metrix.

Por sua vez, na perspectiva dos estudos de Terminologia, muito já se escreveu sobre o texto e/ou discurso científico-técnico, geralmente reconhecido como o "habitat das terminologias" e realização das linguagens especializadas. Essas linguagens, obviamente, serão realizadas sobretudo sob a forma de textos escritos. Assim, passou-se a reconhecer o texto do tipo científico, o qual, por força de sua institucionalização e da normatização terminológica, tende a seguir padrões mais ou menos fixos peculiares: padrões lexicais, terminológicos, retórico-argumentativos e de macroestruturação textual, entre outros. Além dos estudos de Terminologia, cabe também registrar o enfoque denominado *Linguística do Texto Especializado* (KALVERKÄMPER, 1983).

Por fim, mas não menos importante, resta ainda mencionar nesta breve revisão a linha dos estudos de Terminologia que se associaram aos estudos do texto especializado. Um trabalho que tratou, ainda que indiretamente do tema da CT, foi o de Ciapuscio (1998). Essa autora avaliou o grau de abstração conceitual em diferentes tipos de textos que tratavam de uma mesma temática, mas que eram dirigidos a diferentes perfis de leitores (cientistas, público semi-leigo e leigo). Considerou como fatores distintivos dos graus de especialização desses textos, produzidos por cientistas e por jornalistas que cobrem temas científicos coincidentes, o uso de terminologia específica e a presença de variação terminológica, realizada na forma de sinônimos, paráfrases e explanações.

Conforme explica Maciel (2010, p. 23-24), Ciapuscio examinou como a variação conceitual do termo se adaptava "à variação discursiva, a fim de modular o grau de densidade da informação a ser oferecida ao usuário, de maneira que o texto se tornasse mais ou menos transparente". Quando não havia variação da terminologia, o texto exibia um maior grau de densidade do conhecimento especializado.

Mais recentemente, em 2003, Ciapuscio desenvolveu essas idéias no livro *Textos*

especializados y terminologia (CIAPUSCIO, 2003). A partir do modo de apresentação de esquemas de conteúdo e das terminologias nos textos que tratam de temas científicos, propondo uma tipologização multinível. Para chegar a uma categorização dos textos, a autora propõe a consideração de quatro níveis:

a) o nível funcional do texto – que trata da sua função ou propósito;
b) o nível situacional – associado aos interlocutores e tipo de comunicação envolvidos;
c) o nível de conteúdo semântico, que inclui modos de tratamento e de apresentação do tema; e,
d) nível formal-gramatical, que inclui aspectos gramaticais, lexicais e terminológicos.

Cada um desses níveis receberá uma gradação, e a sua junção permitirá identificar tipos de textos em função de diferentes condições. Conforme é fácil perceber, há aqui, à semelhança da AMD, uma perspectiva multinível para a consideração de um todo de sentido que é naturalmente multifacetado. Como pretendemos voltar à proposta de Ciapuscio mais adiante, passamos agora a uma apresentação mais detalhada da AMD com vistas a identificar suas potencialidades para o estudo da CT de textos científicos.

SEGUNDA PARTE - ANÁLISE MULTI-DIMENSIONAL (AMD) NA LINGUÍSTICA DE CORPUS

Como já mencionado, a abordagem Multi-Dimensional, proposta por Douglas Biber a partir de 1988 (BIBER 1988 e 1995), propunha combinar análises de *corpus* de nível macro com análises de nível micro. A microdescrição dos traços de cada texto visa permitir a indução dos macro-agrupamentos textuais ou genéricos (BERBER SARDINHA, 2000, p.100). Assim, pode-se supor que essa seja uma metodologia do tipo *bottom-up*, pois, a partir do que se verificar nos textos, averiguando-se inúmeros traços, é que os textos serão categorizados em função de diferentes elementos.

De acordo com Berber Sardinha (2000),

> *a análise Multidimensional foi criada por Douglas Biber com o objetivo de permitir uma descrição rica e complexa de corpora inteiros de textos por meio estatísticos bem como a extração precisa de características textuais em comum entre corpora. Anteriormente à Análise Multimensional, a tendência era de que se estudasse a co-ocorrência de poucos traços e que se fizesse a interpretação de modo intuitivo. A variação entre registros era investigada comumente por meio de poucos parâmetros,*
>
> *Desse modo, é possível empreender-se uma análise de larga escala de um corpus fazendo-se descrições individuais ao longo do tempo, combinando-se posteriormente as análises para fins comparativos. Por isso, a abordagem Multidimensional presta-se perfeitamente a projetos de descrição de bancos de dados em crescimento, ou seja, aquelas bases de dados linguísticos que estão em processo de coleta.*

Conforme afirmava Berber Sardinha, já há dez anos (op.cit., 2000), trabalhos que incluíam análises multidimensionais de dados de *corpora* ainda não eram muito abundantes no Brasil, embora sua proposta tenha sido apresentada internacionalmente desde 1988. Isso leva-nos a imaginar que, independentemente de maior ou menor divulgação entre nós, esse tipo de investigação deve ter - e tem -, naturalmente, suas dificuldades operativas. Afinal, associar análises de nível geral – do *corpus* como um todo - com análises de nível textual – de um texto no *corpus* e dele com suas frases ou expressões em função de diferentes dimensões - é uma tarefa complexa. Há que considerar, também, algumas críticas importantes à proposta da AMD como elementos inibidores de sua disseminação, principalmente o fato de que o tipo de análise de texto

empreendida por Biber, originalmente, ter sido feita no nível da palavra em inúmeros contextos sentenciais e não no nível do texto.

A despeito de quaisquer limitações ou críticas, é preciso reconhecer o caráter inovador dessa proposta. O ideal, para se descrever os diferentes tipos de texto, conforme Biber propôs, seria combinar a descrição firmada em características situacionais da comunicação com a descrição baseada em traços linguísticos. E aqui já temos pelo menos duas dimensões.

A AMD se propõe justamente a isso, ou seja, a fornecer o instrumental para a identificação de padrões de co-ocorrências dos dois tipos de características, Linguísticas e situacionais. Visa caracterizar uma língua como um todo ou um conjunto de textos, de modo abrangente. Possui caráter essencialmente quantitativo e computacional, descrevendo seus objetos por meio de uma grande quantidade de características.

No Brasil, um dos trabalhos que justamente que associou AMD e textos científicos foi o de Shergue (2003). Seu estudo incidiu sobre dimensões de variação do discurso médico em inglês, tendo em vista auxiliar a produção oral e a compreensão de leitura em inglês de profissionais brasileiros. Seu *corpus* foi constituído por uns poucos artigos de pesquisa e textos transcritos de apresentações orais de trabalhos científicos em congressos. Em que pese a pequena dimensão de textos sob exame, conforme explica o autor, recuperando princípios do modelo de Biber, a qualidade da seleção do *corpus* é o fator mais preponderante nesse tipo de enfoque, em detrimento da quantidade.

Como nosso objeto para exploração do tema da CT é justamente o artigo de Pediatria em escrito português, utilizaremos esse trabalho de Shergue como um exemplo ilustrativo das metodologias e princípios da AMD.

Conforme seu autor, o trabalho procurou,

> *partindo da co-ocorrência de variáveis, buscar funções comunicativas subjacentemente compartilhadas nos corpora que, marcando o uso sistemático dessas características, podem determinar onde gêneros podem ser distribuídos em um espaço oral/escrito de variação contínua, ao invés de simples similaridades e diferenças."(SHERGUE, 2003, p.6)*

PASSOS DA AMD NO TRABALHO DE SHERGUE

Antes de mais nada, conforme é praxe na AMD, o que o autor fez foi revisar a bibliografia que tratou dos tipos de texto em questão para nela colher indicativos de características Linguísticas para compor as variáveis de estudo.

A partir de um conjunto de características linguísticas, foram a elas vinculadas algumas funções conforme exemplificado do quadro 1 a seguir. É importante notar que a co-relação característica-função ilustrada está associada a indicações da bibliografia e que não há, nelas, um recorte entre o que seria do texto oral ou do escrito.

CARACTERÍSTICA LINGUÍSTICA	FUNÇÕES
Conjunção coordenada	Conexão entre orações, fragmentar o texto (Pacheco, 1997: 95)
Conjunção subordinada	Conexão entre orações, complexidade estrutural (Pacheco, 1997: 95)
Pronomes pessoais de 1a. e 2a. pessoas	Interação e envolvimento (Biber, 1988:225)
Salvaguardas	marcar incerteza do autor ou apresentar o conteúdo de forma mais generalizada, distanciamento (Chafe e Danielewicz, 1986 em Biber, 1988:106, 240; Salager-Meyer (1994:154)
Passiva	Distanciamento e abstração (Biber, 1988:228)

Look e *See*	Interação com o ouvinte em chamadas de atenção para o tópico que está sendo apresentado (Serafini & Shergue, 2002)
Densidade e nominalizações	Organizar o texto não em função de nós mesmos mas em função de idéias, razões, causas, distanciamentos (Eggins, 1994:59)

Quadro 1 – Características Linguísticas e funções. Fonte: Adaptado de Shergue, 2003, p. 13

No que se refere à definição do que seja uma **dimensão**, vale reproduzir a indicação do autor, citando Berber Sardinha (2000, p.106): "Dimensão é o status que um fator assume assim que ele é interpretado do ponto de vista da sua função comunicativa." Para ficar mais claro, é preciso compreender que um fator é um conjunto de características linguísticas, tais como as elencadas no quadro 1, de modo que elas, as dimensões, não são um ponto de partida, mas um ponto de chegada da observação. Essas noções devem ficar mais claras para o leitor deste texto mais adiante.

Na Figura 1 a seguir, vemos três dimensões, que são correlações, entre a maior ou menor presença de algumas características linguísticas e um dado grupo de textos. As características (no caso, uso de passivas, pronomes, verbos no passado, nominalizações, contrações) são agrupadas de modos diferentes e graduadas para os textos em foco. Esses textos são um artigo científico, uma discussão sobre um pôster, uma conversa e um texto de ficção.

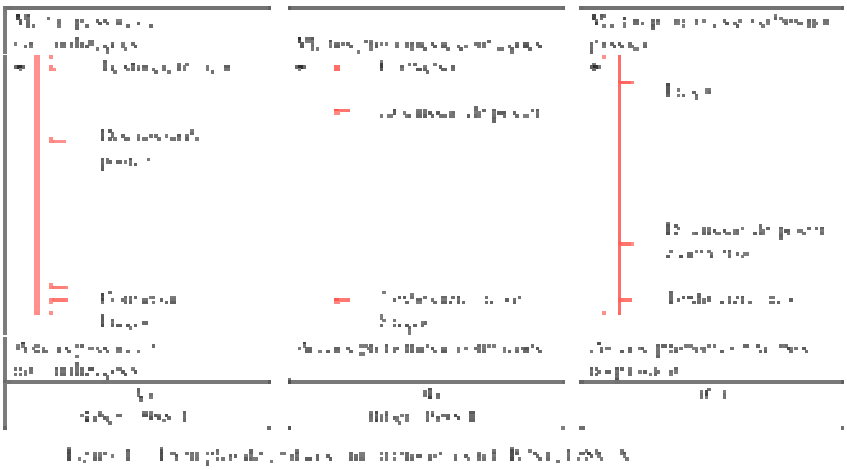


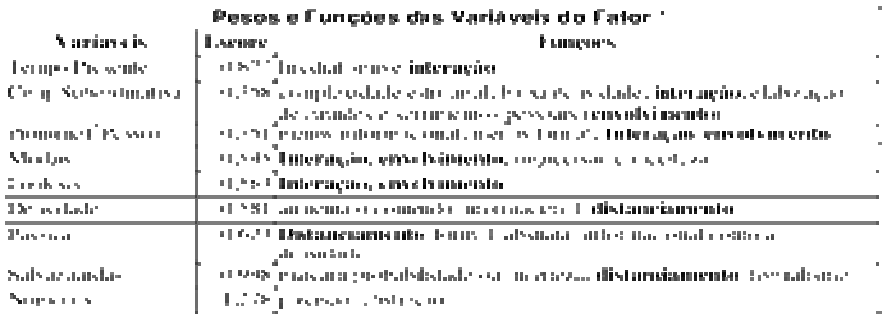
Figura 1- Reprodução de ilustração de *dimensão* de Shergue, citando Biber.

Esses agrupamentos, que são as dimensões, são vistos como "um conjunto de **características Linguísticas** que co-ocorrem em um texto porque **operam juntas** para marcar alguma função comum subjacente" (BIBER, 1988, p. 55, *apud* SHERGUE, grifos deste). Essa *alguma função subjacente* será depreendida pelo analista considerando-se a combinação entre a situação comunicativa e as funções (gramaticais e semântico-pragmáticas) dos elementos linguísticos levantados. Neste ponto da operação, salienta-se que a combinação entre uma característica X e uma Y (como, por exemplo, a característica *pronomes* e a característica *verbos no passado*, que fazem a dimensão *C* da Figura 1) não é aleatória, mas, sim, estatisticamente depreendida.

Para não estender demasiadamente esta parte dedicada a sintetizar o trabalho-exemplo com uso da AMD, passa-se agora a uma apresentação esquemática dos seus passos, das dimensões e das características identificadas para os artigos de Medicina. Os passos metodológicos do trabalho de Shergue (*op.cit*) foram os seguintes:

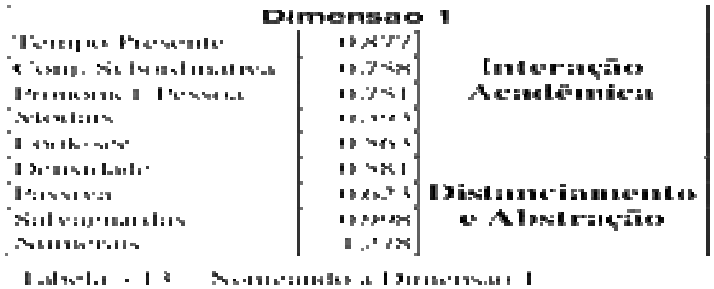
Primeiro passo: a) construção de *corpus* de modo que seja representativo em relação ao que se pretende observar (no caso, há uma distinção entre textos orais, artigos, e textos orais, as apresentações); b) etiquetagem do *corpus*;

Segundo passo: revisão da bibliografia sobre características e funções dos textos do *corpus*. Essa revisão serve para definir as variáveis presumidamente associadas aos textos;
Terceiro passo: contagens de freqüência de ocorrências das variáveis em cada texto. Nesta atividade entram diversas ferramentas computacionais (observam-se freqüências em geral, *clusters*, e palavras-chave – no sentido da LC);
Quarto passo: normalização das freqüências, objetivando um efeito de nivelamento da extensão irregular dos textos. Após a normalização, é feita uma seleção e descarte, restando as variáveis finais a serem submetidas à análise fatorial na próxima fase;
Quinto passo: análise microscópica e análise macroscópica. A análise macro chega nas dimensões globais da variação Linguística das variedades de elementos do *corpus* (tal como escrito vs. oral ou outra difereciação que se utilize, como, por exemplo, artigo de Pediatria e texto de jornal popular). Na parte micro, temos a identificação das funções comunicativas das características Linguísticas individuais;
Sexto passo: análise fatorial. É utilizado o pacote SPSS, programa de computador que faz uma série de testes estatísticos, cálculo de fatores, índices estatísticos de significância, média, desvio padrão, etc. Aqui diferentes características são correlacionadas em grupos. O grupo é denominado *fator*, conforme se vê no Quadro 2 a seguir. É importante salientar que as variáveis têm pesos positivos e negativos.



Quadro 2 – Pesos e funções de variáveis agrupados em um fator.

Sétimo passo: identificação e denominação de dimensões. Nesse momento, unem-se as funções e correlações acima identificadas e é depreendida uma globalidade. Isso é o que ilustra o Quadro 3 a seguir. No trabalho de Shergue, foram identificadas apenas duas dimensões.



Quadro 3 – Dimensão 1 - Interação acadêmica *vs* Distanciamento e abstração.

Conforme explica o autor,

> "A Dimensão 1, rotulada como ***Interação Acadêmica versus Distanciamento e Abstração,*** representa, de um lado, o discurso acadêmico oral com propósito interacional e envolvimento pessoal e, de outro lado, o discurso escrito de conteúdo formal e abstrato, marcado pela apassivização, pelos processos de salvaguardas, pela precisão numérica e densidade de conteúdo, promovendo o distanciamento entre o autor e o leitor."

Uma segunda dimensão identificada no trabalho de Shergue, denominada ***Dimensão 2 – Nominalização Técnica Específica versus Informalidade Textual Acadêmica***, incorporou, em diferentes níveis de variação nos mesmos textos, as funções comunicativas que promovem adensamento de conteúdos com nominalização específica e um certo grau de informalismo.

Finalizando esta seção, dedicada à AMD e ao seu *modus operandi*, aqui colocado em termos de passos, resta ainda dizer que muitas das características correlacionadas poderiam ser associadas a níveis ou a um dado nível de complexidade textual – o *adensamento de conteúdos via nominalizações* seria apenas um dos exemplos dessa condição. A complexidade do texto poderia, assim, corresponder a uma dimensão, isto é, a um dado grupo de características correlacionadas que operassem juntas para marcar alguma função comum.

De outro lado, é importante registrar que no próprio trabalho de Biber (1988, p.10), quando ele coloca as bases da sua idéia de *dimensão,* há, como exemplo, um trecho de um livro científico e um trecho de uma conversa entre duas pessoas sobre gostarem ou não de cerveja feita em casa. Ele mesmo aponta que, nesse exemplo, tem-se, entre outras, a dimensão *comum vs. especializado*. Conforme explica, à medida que esses dois textos fossem "ladeados" por textos de outros tipos, veríamos que a dimensão em questão mostraria-se como um *continuum* (tanto quanto as outras dimensões que traz para essa dupla de textos: *interativo vs não-interativo* e *planejado vs não-planejado*). Em cada dimensão, há diferentes elementos linguísticos relacionados que se graduam positiva e negativamente ao longo dos tipos de texto envolvidos.

Feita essa caracterização da AMD, passamos agora à apresentação do sistema Coh-Metrix, que será aqui tomado como um exemplo prototípico do enfoque da LCOMP.

TERCEIRA PARTE - SOBRE OS SISTEMAS COH-METRIX EM LCOMP

A Linguística Computacional (LCOMP) ou Processamento da Linguagem Natural (PLN) é a área de conhecimento que explora as relações entre Linguística e Informática, tornando possível a construção de sistemas com capacidade de reconhecer e de produzir informação apresentada em linguagem natural (LIMA; STRUBE, 2001). Seu objetivo é, assim, essencialmente aplicado, relacionado à produção de um sistema concreto.

Conforme Vieira e Lopes (2010), desde o surgimento das técnicas de PLN, muitos avanços foram obtidos, mas a compreensão plena de linguagem natural por métodos computacionais está ainda longe de ser resolvida. Ainda assim, o tratamento computacional da língua é um campo muito promissor[10].

Nascido do PLN, o sistema Coh-Metrix, que significa *cohesion metrics*, é uma ferramenta para análise de textos em inglês, disponível gratuitamente *on-line*. Elaborada por pesquisadores da Universidade de Memphis, nos Estados Unidos (GRAESSER; McNAMARA; LOUWERSE; CAI, 2004), tem como propósito calcular índices de coesão e de coerência textual num amplo espectro de medidas lexicais, sintáticas, semânticas e referenciais com o fim de indicar a adequação de um texto a seu público-alvo (a "demanda cognitiva" e a legibilidade do texto). Também tem a função de apontar dados para identificar problemas textuais de ordem estrutural.

[10] Um marco recente e concreto dessas promessas de desenvolvimento é o computador WATSON, da IBM. Em fevereiro de 2011, foi apresentado, com sucesso, em uma competição de perguntas e respostas de um programa de TV norte-americano. WATSON enfrentou humanos e foi capaz de reconhecer perguntas feitas oralmente e de produzir linguagem oral com padrão de naturalidade para as respostas que dava. O computador venceu os humanos na disputa. Há vídeos a respeito no YouTube.

Até o momento, mais de 500 métricas estão disponíveis um uma versão restrita do Coh-Metrix. Dessas 500, apenas 60 estão disponíveis na versão gratuita *on-line* no *site* do projeto. Para todas essas avaliações (chamadas de *métricas* na terminologia de Linguística Computacional) vários recursos e ferramentas de Processamento de Linguagem Natural são utilizados. A versão livre Coh-Metrix 2.0[11] opera com índices que vão desde *métricas* simples (como contagem de palavras) até medidas mais complexas, envolvendo algoritmos de resolução anafórica. Os 60 índices estão divididos em seis blocos que avaliam a complexidade de um texto a partir da mensuração dos seguintes elementos:

1) Identificação Geral e Informação de Referência, Índices de Inteligibilidade, Palavras Gerais e Informação do Texto, Índices Sintáticos, Índices Referenciais e Semânticos e Dimensões do Modelo de Situações. Essa primeira classe corresponde às informações que referenciam o texto, como título, gênero entre outros;

2) Índices de inteligibilidade calculados com as fórmulas *Flesch Reading Ease* e *Flesch Kincaid Grade Level*. Essas fórmulas consideram tamanho de sentença, número de palavras por sentença e número de palavras diferentes por sentença;

3) Verificação de quatro subclasses: Contagens Básicas, Freqüências, Concretude, Hiperônimos;

4) Verificação de cinco subclasses: Constituintes, Pronomes, Tipos e *Tokens,* Conectivos, Operadores Lógicos e Similaridade sintática de sentenças;

5) Verificação de três subclasses: Anáfora, Co-referência e Análise Semântica Latente;

6) Verificação de quatro subclasses: Dimensão Causal, Dimensão Intencional, Dimensão Temporal e Dimensão Espacial.

Em síntese, trata-se de uma ferramenta que calcula índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto em diferentes níveis. Esses níveis incluem os níveis lexical, sintático, discursivo e um nível denominado *conceitual*, observando-se fatores tais como número de sentenças, número de palavras por sentença, co-referências, anáforas, presença de conectores e de itens com ambigüidade semântica e número de pronomes por sintagma.

A partir do Coh-Metrix em inglês, uma iniciativa de adaptação para o português brasileiro das sessenta métricas oferecidas gratuitamente surgiu no âmbito do Projeto PorSimples[12], cujo objetivo era o de identificar índices de complexidade textual para simplificação de textos e facilitação do acesso à informação a analfabetos funcionais e pessoas com deficiências cognitivas. O nome da ferramenta correspondente em português é **Coh-Metrix-Port** e está disponível no *site* do PorSimples. Esse sistema foi desenvolvido pelo NILC (Núcleo Interinstitucional de Linguística Computacional da USP). Para mais detalhes sobre o NILC, veja-se Nunes, Aluísio e Pardo (2010).

É importante ressaltar que, até o momento, apenas **35** das **60** métricas originais do Coh-Metrix foram adaptadas para o português do Brasil. Para que se tenha uma idéia dos tipos de métrica, medidas ou índices em questão, reproduzimos a seguir, nas Figuras 2 e 3, respectivamente, uma amostra das métricas para o inglês e uma parte da tela de saída de análise para o português.

---

[11] (<http://cohmetrix.memphis.edu/cohmetrixpr/index.html>)

[12] O projeto **PorSimples** (http://caravelas.icmc.usp.br/wiki/index.php/Principal) iniciou em novembro de 2007. Tem apoio da **FAPESP** (Fundação de Amparo à Pesquisa de São Paulo) e da **MSR (Microsoft Research)**. Propõe o desenvolvimento de uma tecnologia para facilitar o acesso à informação dos analfabetos funcionais (AF) e, potencialmente, de pessoas com outras deficiências cognitivas, como afasia e dislexia. Essa tecnologia está oferecida em dois sistemas destinados a públicos alvos diferentes: a) um sistema de autoria para ajudar redatores a produzir textos simplificados destinados aos AFs, textos estes que serão validados pelos redatores e b) um sistema facilitador para ajudar AFs a lerem um dado conteúdo da Web. Este último inclui tarefas de sumarização textual e simplificação sintática (sistema FACILITA) e elaboração léxica, apresentação do texto salientando as relações retóricas entre as idéias do texto, explicitação das Entidades Mencionadas e dos argumentos dos verbos (sistema FACILITA EDUCATIVO).

No.	Description	Measure	Full description
1	Title	Title	Title
2	Genre	Genre	Genre
3	Source	Source	Source
4	JobCode	JobCode	JobCode
5	LSASpace	LSASpace	LSASpace
6	Date	Date	Date
7	Causal content	CAUSVP	Incidence of causal verbs, links, and particles
8	Causal cohesion	CAUSC	Ratio of causal particles to causal verbs (cp divided by cv+1)
9	Pos. additive connectives	CONADpi	Incidence of positive additive connectives
10	Pos. temporal connectives	CONTPpi	Incidence of positive temporal connectives
11	Pos. causal connectives	CONCSpi	Incidence of positive causal connectives
12	Neg. additive connectives	CONADni	Incidence of negative additive connectives
13	Neg. temporal connectives	CONTPni	Incidence of negative temporal connectives
14	Neg. causal connectives	CONCSni	Incidence of negative causal connectives
15	All connectives	CONi	Incidence of all connectives
16	Adjacent argument overlap	CREFA1u	Argument Overlap, adjacent, unweighted
17	Adjacent stem overlap	CREFS1u	Stem Overlap, adjacent, unweighted
18	Adjacent anaphor reference	CREFP1u	Anaphor reference, adjacent, unweighted
19	Argument overlap	CREFAau	Argument Overlap, all distances, unweighted
20	Stem overlap	CREFSau	Stem Overlap, all distances, unweighted
21	Anaphor reference	CREFPau	Anaphor reference, all distances, unweighted
22	NP incidence	DENSNP	Noun Phrase Incidence Score (per thousand words)
23	Pronoun ratio	DENSPR2	Ratio of pronouns to noun phrases
24	Conditional operators	DENCONDi	Number of conditional expressions, incidence score
25	Negations	DENNEGi	Number of negations, incidence score
26	Logic operators	DENLOGi	Logical operator incidence score (and + if + or + cond + neg)
27	LSA sentence adjacent	LSAassa	LSA, Sentence to Sentence, adjacent, mean
28	LSA sentence all	LSApssa	LSA, sentences, all combinations, mean
29	LSA paragraph	LSAppa	LSA, Paragraph to Paragraph, mean
30	Personal pronouns	DENPRPi	Personal pronoun incidence score

Figura 2 – Métricas do Coh-Metrix para o inglês (amostra).

Texto		
Título	Constituição - Dos direitos e deveres individuais	Título
Autor	Brasil	Autor
Fonte	http://www.senado.gov.br/legislacao/const/con1988/CON1988_05.10.1988/art_5_.shtm	Fonte
Data de Publicação		Data de Publicação
Gênero	Jurídico	Gênero
Contagens Básicas		
Índice Flesch	24.17043715847	Índice Flesch
Número de Palavras	549.0	Número de palavras do texto.
Número de Sentencas	27.0	Número de sentenças de um texto.
Número de Parágrafos	27.0	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
Palavras por Sentenças	20.3333333333333	Número de palavras dividido pelo número de sentenças.
Sentenças por Paragrafos	1.0	Número de sentenças dividido pelo número de parágrafos.
Sílabas por Palavras de Conteúdo	3.21621621621622	Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios).
Incidência de Verbos	127.504553734062	Incidência de verbos em um texto.
Incidência de Substantivos	326.047358834244	Incidência de substantivos em um texto.
Incidência de Adjetivos	132.969034608379	Incidência de adjetivos em um texto.
Incidência de Advérbios	20.0364298724954	Incidência de advérbios em um texto.
Incidência de Pronomes	30.9653916211293	Incidência de pronomes em um texto.
Incidência de Palavras de Conteúdo	606.55737704918	Incidência de Palavras de Conteúdo (substantivos, adjetivos, advérbios e verbos).
Incidência de Palavras Funcionais	377.049180327869	Incidência de Palavras Funcionais (artigos, preposições, pronomes, conjunções e interjeições).

Figura 3 – Saída do Coh-Metrix para o português – Texto: Constituição do Brasil - Título II: Dos direitos e deveres individuais - Capítulo I.

Um item de destaque, nesse sistema de medidas, é o índice Flesch[13]. É uma das diferentes medidas de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores. O resultado é um número de 0 a 100 que é assim mensurado (com a devida adaptação para o sistema escolar brasileiro feita pela equipe PorSimples):

- **muito fáceis** índice entre **75 - 100**, textos adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental
- **fáceis** índice entre **50 - 75**, textos adequados a alunos com escolaridade até a oitava série do ensino fundamental
- **difíceis** índice entre **25 - 50**, textos adequados para alunos cursando o ensino médio ou universitário
- **muitos difíceis** índice entre **0 - 25**, textos adequados apenas para áreas acadêmicas específicas

O segmento de texto examinado na Figura 3, trecho da Constituição do Brasil, teve índice Flesch de 24,17. Isso o coloca como um texto extremamente difícil. Naturalmente, essa é apenas **uma** das mais de trinta métricas oferecidas, tendo sido ilustrada apenas a parte denominada *Contagens Básicas*. Para mais detalhes sobre o sistema Coh-Metrix-Port, recomendamos a

[13] O nome *Flesch* deve-se a Rudolf Flesch (1911-1986). Esse autor foi um especialista em avaliações de índices de inteligibilidade de textos e defensor da idéia de se usar um *plain English* (inglês simplificado) em determinada situações de ensino/formação. Criou o Flesch Reading Ease Test e foi co-criador do Flesch-Kincaid Teste de Legibilidade.

leitura do manual produzido por Almeida e Aluisio (2009).

Como se pode perceber, no âmbito da LCOMP, a observação do texto está muito mais associada a medidas de complexidade, relativamente dispersas, de um modo diferente do que já vimos na AMD.

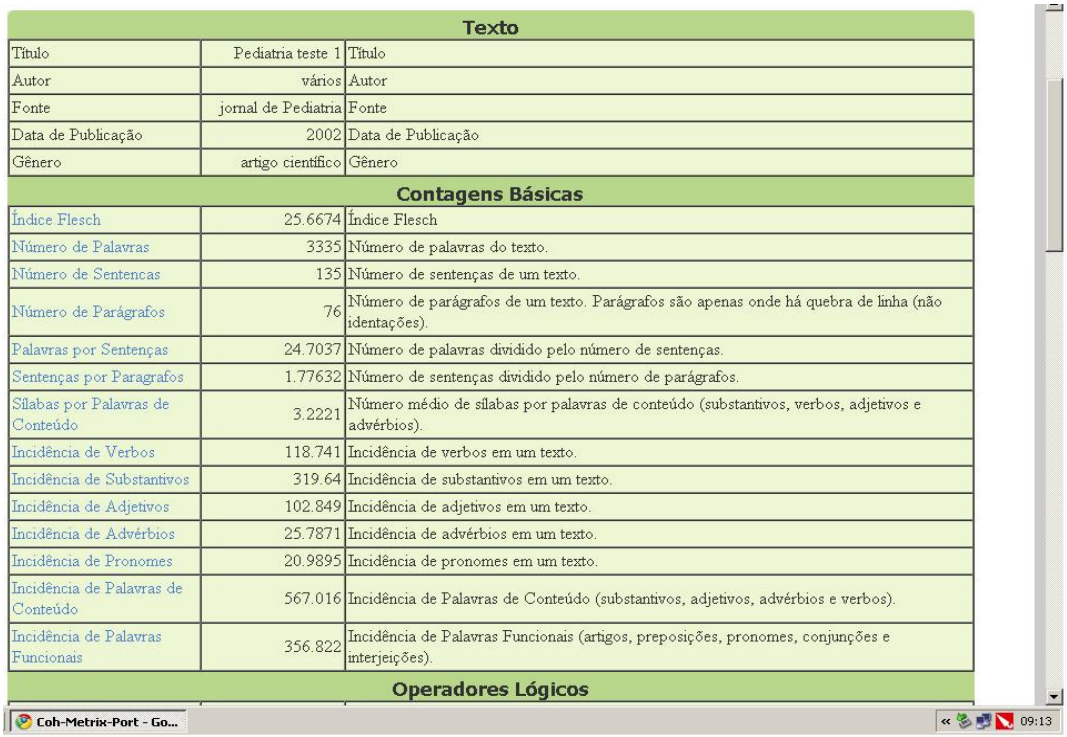
Tal como antes citado, é possível imaginar que o índice Flesh pudesse ser aproveitado pela AMD, assim como a maior inter-relação de características da AMD poderia ser aproveitada pelo sistema Coh-Metrix. Note-se, por exemplo, que o sistema Coh-Metrix para o português não contempla a presença de voz passiva, tampouco associa funções ou atribui pesos diferenciados por tipo ou gênero do texto avaliados automaticamente. Outros aspectos que poderiam ser apontados como peculiares – e até intrigantes, por exemplo, seriam a maior ou menor presença de adjetivos ou de advérbios, que integra o segmento *Contagens Básica*s no Coh-Metrix. Esses tipos de elementos, a adjetivação e a modalização adverbial[14], que papel teriam em um texto especializado? A maior diferença, naturalmente, entre AMD e Coh-Metrix também reside no processamento de um só texto por vez.

Vejamos agora o que a ferramenta Coh-Metrix mostra sobre a complexidade de um artigo de Pediatria coletado na revista brasileira Jornal de Pediatria

Observando um artigo científico de Pediatria

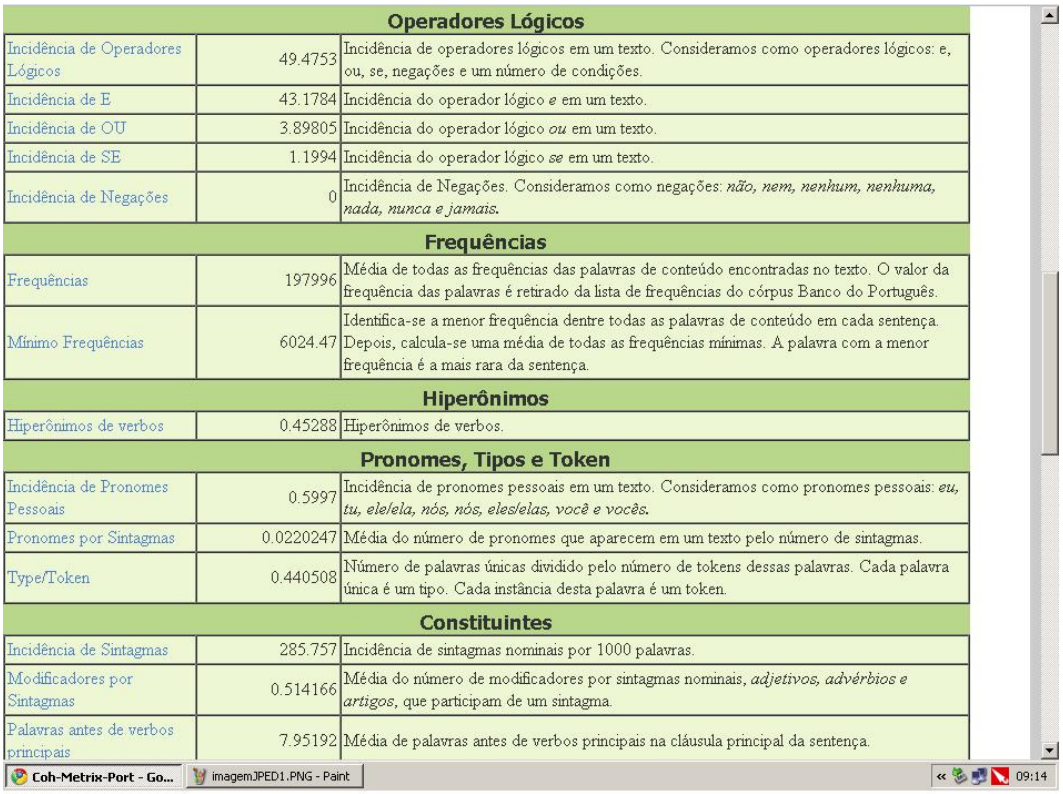
A seguir está um conjunto de figuras (Figura 4 até Figura 7) com algumas das avaliações do sistema Coh-Metrix-Port para um texto de Pediatria intitulado *Seguimento nutricional de pacientes com fibrose cística: papel do aconselhamento nutricional*, publicado na revista *Jornal de Pediatria* em 2004 (ADDE; RODRIGUES; CARDOSO, 2004). Não serão apresentadas todas as medidas, tampouco feitos maiores comentários, visto que os dados parecem auto-explicativos em função do que foi exposto na seção anterior.

[14] Um estudo exploratório sobre adjetivos e advérbios em Química e Medicina foi feito por Finatto e Huang (2005).



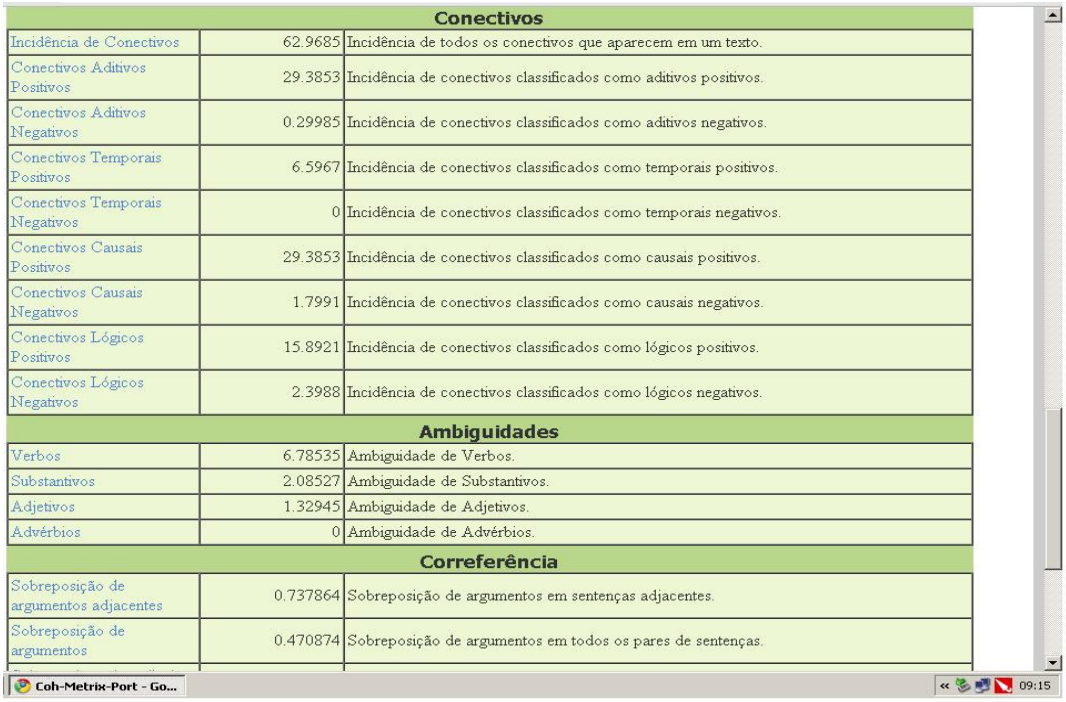
Texto		
Título	Pediatria teste 1	Título
Autor	vários	Autor
Fonte	jornal de Pediatria	Fonte
Data de Publicação	2002	Data de Publicação
Gênero	artigo científico	Gênero
Contagens Básicas		
Índice Flesch	25.6674	Índice Flesch
Número de Palavras	3335	Número de palavras do texto.
Número de Sentencas	135	Número de sentenças de um texto.
Número de Parágrafos	76	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
Palavras por Sentenças	24.7037	Número de palavras dividido pelo número de sentenças.
Sentenças por Paragrafos	1.77632	Número de sentenças dividido pelo número de parágrafos.
Sílabas por Palavras de Conteúdo	3.2221	Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios).
Incidência de Verbos	118.741	Incidência de verbos em um texto.
Incidência de Substantivos	319.64	Incidência de substantivos em um texto.
Incidência de Adjetivos	102.849	Incidência de adjetivos em um texto.
Incidência de Advérbios	25.7871	Incidência de advérbios em um texto.
Incidência de Pronomes	20.9895	Incidência de pronomes em um texto.
Incidência de Palavras de Conteúdo	567.016	Incidência de Palavras de Conteúdo (substantivos, adjetivos, advérbios e verbos).
Incidência de Palavras Funcionais	356.822	Incidência de Palavras Funcionais (artigos, preposições, pronomes, conjunções e interjeições).
Operadores Lógicos		

Figura 4 – Coh-Metrix- Port para um artigo de Pediatria - parte 1.



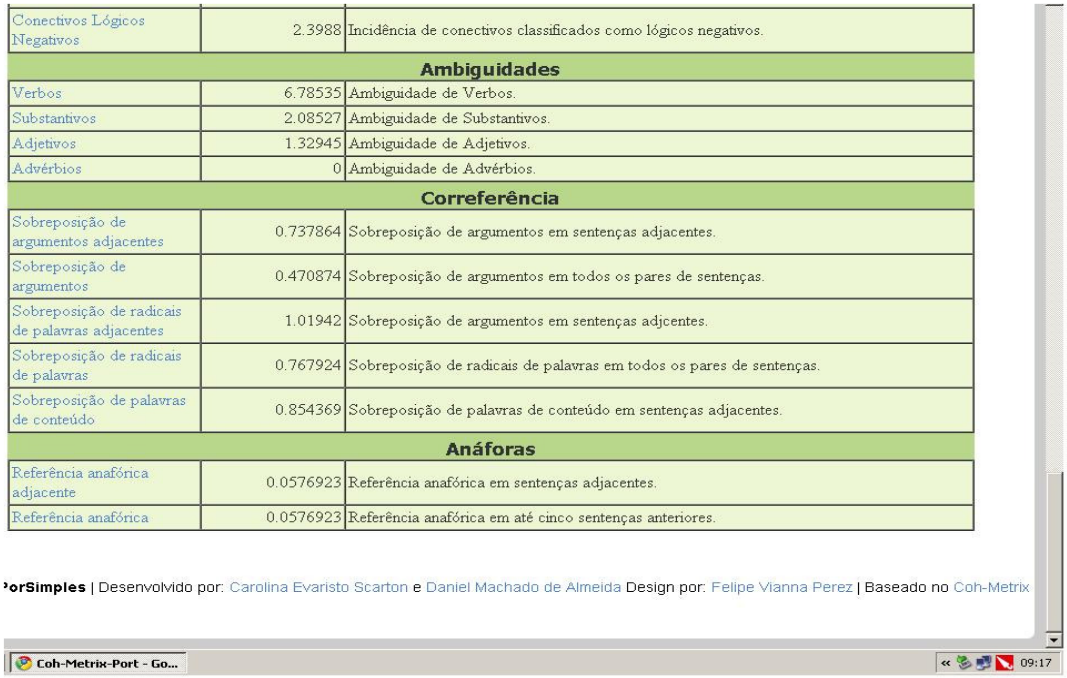
Operadores Lógicos		
Incidência de Operadores Lógicos	49.4753	Incidência de operadores lógicos em um texto. Consideramos como operadores lógicos: e, ou, se, negações e um número de condições.
Incidência de E	43.1784	Incidência do operador lógico *e* em um texto.
Incidência de OU	3.89805	Incidência do operador lógico *ou* em um texto.
Incidência de SE	1.1994	Incidência do operador lógico *se* em um texto.
Incidência de Negações	0	Incidência de Negações. Consideramos como negações: *não, nem, nenhum, nenhuma, nada, nunca e jamais.*
Frequências		
Frequências	197996	Média de todas as frequências das palavras de conteúdo encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do córpus Banco do Português.
Mínimo Frequências	6024.47	Identifica-se a menor frequência dentre todas as palavras de conteúdo em cada sentença. Depois, calcula-se uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença.
Hiperônimos		
Hiperônimos de verbos	0.45288	Hiperônimos de verbos.
Pronomes, Tipos e Token		
Incidência de Pronomes Pessoais	0.5997	Incidência de pronomes pessoais em um texto. Consideramos como pronomes pessoais: *eu, tu, ele/ela, nós, nós, eles/elas, você e vocês.*
Pronomes por Sintagmas	0.0220247	Média do número de pronomes que aparecem em um texto pelo número de sintagmas.
Type/Token	0.440508	Número de palavras únicas dividido pelo número de tokens dessas palavras. Cada palavra única é um tipo. Cada instância desta palavra é um token.
Constituintes		
Incidência de Sintagmas	285.757	Incidência de sintagmas nominais por 1000 palavras.
Modificadores por Sintagmas	0.514166	Média do número de modificadores por sintagmas nominais, *adjetivos, advérbios e artigos*, que participam de um sintagma.
Palavras antes de verbos principais	7.95192	Média de palavras antes de verbos principais na cláusula principal da sentença.

Figura 5 - Coh-Metrix- Port para um artigo de Pediatria - parte 2.



Conectivos		
Incidência de Conectivos	62.9685	Incidência de todos os conectivos que aparecem em um texto.
Conectivos Aditivos Positivos	29.3853	Incidência de conectivos classificados como aditivos positivos.
Conectivos Aditivos Negativos	0.29985	Incidência de conectivos classificados como aditivos negativos.
Conectivos Temporais Positivos	6.5967	Incidência de conectivos classificados como temporais positivos.
Conectivos Temporais Negativos	0	Incidência de conectivos classificados como temporais negativos.
Conectivos Causais Positivos	29.3853	Incidência de conectivos classificados como causais positivos.
Conectivos Causais Negativos	1.7991	Incidência de conectivos classificados como causais negativos.
Conectivos Lógicos Positivos	15.8921	Incidência de conectivos classificados como lógicos positivos.
Conectivos Lógicos Negativos	2.3988	Incidência de conectivos classificados como lógicos negativos.
Ambiguidades		
Verbos	6.78535	Ambiguidade de Verbos.
Substantivos	2.08527	Ambiguidade de Substantivos.
Adjetivos	1.32945	Ambiguidade de Adjetivos.
Advérbios	0	Ambiguidade de Advérbios.
Correferência		
Sobreposição de argumentos adjacentes	0.737864	Sobreposição de argumentos em sentenças adjacentes.
Sobreposição de argumentos	0.470874	Sobreposição de argumentos em todos os pares de sentenças.

Figura 6 – Coh-Metrix- Port para um artigo de Pediatria - parte 3.



Conectivos Lógicos Negativos	2.3988	Incidência de conectivos classificados como lógicos negativos.
Ambiguidades		
Verbos	6.78535	Ambiguidade de Verbos.
Substantivos	2.08527	Ambiguidade de Substantivos.
Adjetivos	1.32945	Ambiguidade de Adjetivos.
Advérbios	0	Ambiguidade de Advérbios.
Correferência		
Sobreposição de argumentos adjacentes	0.737864	Sobreposição de argumentos em sentenças adjacentes.
Sobreposição de argumentos	0.470874	Sobreposição de argumentos em todos os pares de sentenças.
Sobreposição de radicais de palavras adjacentes	1.01942	Sobreposição de argumentos em sentenças adjcentes.
Sobreposição de radicais de palavras	0.767924	Sobreposição de radicais de palavras em todos os pares de sentenças.
Sobreposição de palavras de conteúdo	0.854369	Sobreposição de palavras de conteúdo em sentenças adjacentes.
Anáforas		
Referência anafórica adjacente	0.0576923	Referência anafórica em sentenças adjacentes.
Referência anafórica	0.0576923	Referência anafórica em até cinco sentenças anteriores.

orSimples | Desenvolvido por: Carolina Evaristo Scarton e Daniel Machado de Almeida Design por: Felipe Vianna Perez | Baseado no Coh-Metrix

Coh-Metrix-Port - Go... 09:17

Figura 7 - Coh-Metrix- Port para um artigo de Pediatria - parte 4

Como é possível notar pelo o que está nas figuras, o sistema funciona perfeitamente bem também para um texto Pediatria, com o diferencial de nos revelar que seu índice Flesch fica em 25,66. O texto em questão é um artigo original do qual se extrai apenas o corpo do texto, incluindo apenas a seção de Agradecimentos. Esse escore Flesch o situa na categoria dos textos **difíceis,** categoria que fica entre as medidas **25 - 50**, sinalizando-se, assim, um texto adequado para alunos cursando o ensino médio ou universitário. Para que essa argumentação não careça da evidência, veja-se abaixo o primeiro trecho da introdução do todo do texto submetido ao Coh-Metrix Port:

> *A fibrose cística (FC) é uma desordem autossômica recessiva que afeta vários sistemas do corpo humano, em especial o trato respiratório. A importância do estado nutricional para aumento da sobrevida e bem-estar dos pacientes com FC é bem documentada na literatura (2). No entanto, a desnutrição continua sendo um sério problema em pacientes com FC. Nos Estados Unidos, o peso e a estatura de cerca de 20% das crianças e adolescentes com FC estão abaixo do percentil 5 (3). Dados a respeito da população com fibrose cística no Reino Unido (UK) também mostram déficits de peso e estatura, principalmente na faixa etária entre 1 e 10 anos de idade, embora tenha havido uma melhora no estado nutricional desses pacientes com relação às décadas anteriores (4). A magnitude desse problema pode ser ainda pior em países subdesenvolvidos, pois pode haver uma sobreposição de desnutrição primária e secundária na população com FC. (ADDE et al., 2004)*

Até esse ponto deste texto, o leitor que acompanha deve se perguntar o que há de novo nessa medição para esse tipo de texto, visto que, em tese, e pela situação comunicativa posta, há uma harmonia, também em tese, entre tipo de leitor e tipo de texto. Parece algo óbvio.

Entretanto, não é tão óbvia a condição do texto, tampouco o fato de tal consideração ter sido gerada automaticamente e de estar acompanhada por toda uma série de outras medidas. Pois é,

justamente, na expansão desse único ponto-medida do sistema Coh-Metrix-Port que reside um potencial de entrelaçamento muito novo com o modo de caracterização de gêneros textuais da AMD. Haveria uma inter-relação – ou co-relação – entre a medida *Pronomes por Sintagma* e medida *Índice Flesch*? Isso, essa co-relação, o sistema Coh-Metrix ainda não mostra, enquanto correlações são, justamente, um carro-chefe das dimensões de AMD.

Um diálogo entre as duas metodologias e seus princípios parece ser necessário para o mútuo enriquecimento de ambas. Naturalmente, mesmo que nenhuma das partes possa estar interessada nessa troca, do modo como é desenhada aqui, pode o linguista que se ocupa do tema do texto científico, como um terceiro envolvido, unir essas duas pontas e utilizar os elementos de contato entre AMD e Coh-Metrix em prol de seus interesses de pesquisa.

QUARTA PARTE - CONSIDERAÇÕES FINAIS

Conforme há pouco referido, entremeando-se esses dois campos, a AMD e a LCOMP, representada aqui pelo sistema Coh-Metrix, pode situar-se a Terminologia e os estudos do texto técnico-científico, associados como uma Linguística do Texto Especializado.

O texto de Pediatria, considerado especializado, pode, de certo modo, conforme se vê na bibliografia de Terminologia de perspectiva textual, ser distinguido do não-especializado no que se refere a esquemas de conteúdo e ao uso ou não uso de terminologias (CIAPUSCIO, 2003, p.71). Nos textos menos especializados, conforme a autora, as terminologias podem ser reescritas ou parafraseadas, dada uma situação de popularização para leigos ou semileigos.

Entretanto, como é fácil concluir, uma linguagem científica ou técnica não se faz assim apenas em função dos seus termos "técnicos" (FINATTO, AZEREDO, 2010, p.560), que cada vez mais parecem figurar também na linguagem cotidiana. Por isso, um outro traço de especificidade do texto científico, explorado por Ciapuscio em trabalho mais recente (CIAPUSCIO, 2005) é também a presença de metaforizações. Conforme a autora, as metáforas seriam un elemento extremamente interessante nos distintos estágios do *continuum* da comunicacão da ciência, desde a criação de conhecimentos no âmbito mais especializado até sua divulgacão para o público leigo. Assim, ela propõe também as metáforas como elementos que pontuam e constituem esse *continuum* que se desenha do mais ao menos especializado (condição que se pergunta aqui se poderia ser associada a um texto mais ou menos complexo).

Considerando toda uma diversidade de fatores que poderiam ser evocados para indicar prováveis condicionantes da CT em textos especializados de diferentes perfis e as características dos gêneros ou dos registros envolvidos, pela conjunção de referenciais vistos até aqui, parece ser possível realizar um movimento de reavaliação sobre a complexidade de textos que tenham mais ou menos terminologias – além de outros elementos, naturalmente.

Conforme vimos, a partir do modo de apresentação de esquemas de conteúdo e das terminologias nos textos que tratam de temas científicos, Ciapuscio propôs uma tipologização multinível. Vale a pena relembrar os níveis:

a) o nível funcional do texto – que trata da sua função ou propósito;
b) o nível situacional – associado aos interlocutores e tipo de comunicação envolvidos;
c) o nível de conteúdo semântico, que inclui modos de tratamento e de apresentação do tema; e,
d) nível formal-gramatical, que inclui aspectos gramaticais, lexicais e terminológicos.

Para chegar a uma categorização dos textos, a autora indica a consideração desses quatro níveis simultaneamente. Esses planos, como parece fácil concluir, assemelham-se a uma base que impregna a proposta da AMD, a qual defende a conjugação da dimensão Linguística com a dimensão funcional do texto para que se possa tratar das variações entre gêneros ou registros.

De outro lado, ao examinar um *output* do sistema Coh-Metrix, cujo resultado é gerado em segundos, um linguista experimentará vários questionamentos. Entres esses questionamentos, vejamos alguns:

- O que significam tantos índices ou métricas postos lado a lado de uma única vez? Ou melhor, o que se entende por complexidade do texto a partir desses diferentes escores?
- Por que esse sistema foi construído desse modo? No que ele poderia ser melhorado?
- A complexidade do texto, entendida globalmente, pode ser considerada apenas como uma média dos diferentes fatores/métricas?
- Que pesos diferentes poderiam ter diferentes medidas em diferentes situações de texto?
- Que elementos desse sistema podem ser melhor aproveitados em um estudo de Linguística Aplicada?
- Como se pode juntar 34 ou 54 métricas em torno de alguma condição do texto, para além do índice de complexidade da medida Flesch?

Essas perguntas, pontuais, somam-se, naturalmente, à pergunta que abre este trabalho: haveria como avaliar em que medida textos científicos do tipo artigo seriam mais ou menos complexos em relação a textos científicos de outros perfis ou mesmo em relação a textos não-especializados? Ao que parece, a resposta é sim, há como avaliar, mas é preciso definir antes, algum parâmetro ou uma série deles em função do objetivo que venha cumprir tal avaliação. Além disso, pelo visto até agora, pelo menos no território da LCOMP e dos estudos de Leitura, já estão disponíveis vários recursos que poderiam nos ajudar na empreitada.

Para terminar este texto, já demasiadamente longo, vejamos a seguir, o que o sistema Coh-Metrix mostra a respeito de um pequeno texto extraído de um jornal popular, publicação que é dirigida a público de menor poder aquisitivo e que, em geral, tem também menor nível de escolaridade ou letramento. O texto vem reproduzido antes da apresentação das medidas Coh-Metrix. A saída do sistema está exemplificada nas Figuras 8 e 9 a seguir:

Por Adriana Franciosi – Editoria Geral – Jornal Diário Gaúcho, 2008.

Pacote do trânsito Rigor nas multas

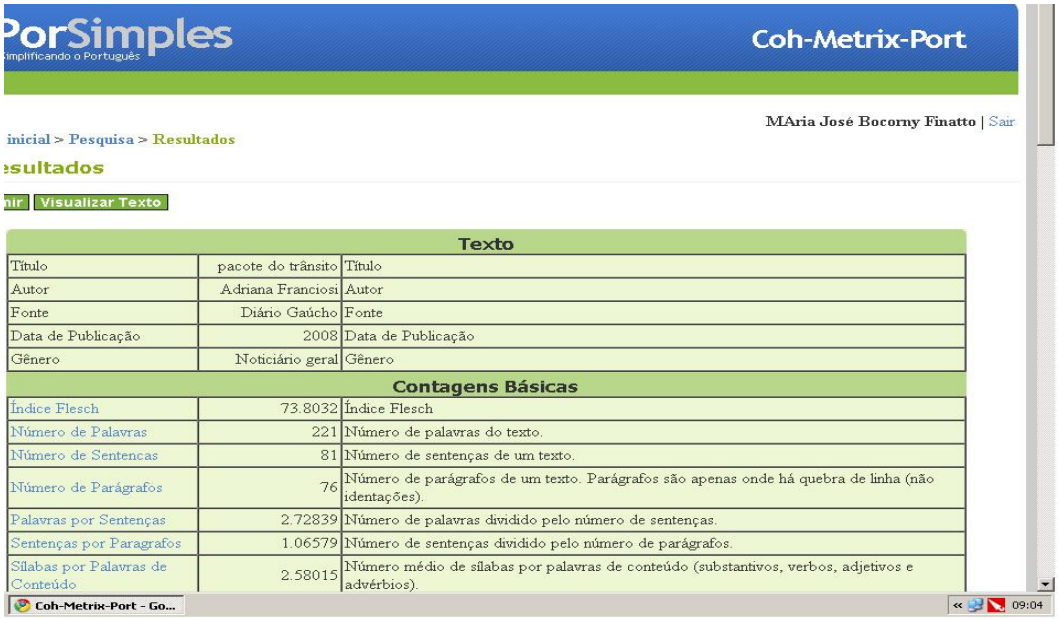
O ministro da Justiça, Tarso Genro, anunciou ontem um pacote de medidas para tornar leis de trânsito mais rígidas. São 28 as alterações, que agora precisam ser aprovadas pelo Congresso Nacional. Se as mudançasforemconfirmadas, a multa da infração gravíssima, dependendo do caso, poderá passar de R$ 1,5 mil. Atualmente, o valor mais alto é de R$ 572,40. Outra medida é a redução, pela metade, do nível de álcool tolerado no sangue. Dirigir embriagado passaria a ser crime, assim como ser flagrado duas vezes em um ano trafegando em uma rodovia a mais de 50km/h acima da velocidade permitida.

Mudanças

Multa mais alta passa de R$ 572,40 para R$ 1,5 mil. Motoristas multados mais de duas vezes em um mesmo ano, por dirigirem com velocidade mais de 50% acima da permitida, responderão por crime.

Carteira

A carteira de motorista vai ficar R$ 60 mais barata no Rio Grande do Sul. A redução foi anunciada ontem pela governadora Yeda Crusius. O valor cai de R$ 805 para R$ 744. A diminuição se deve ao corte nas taxas cobradas pelo Detrane na redução no preço dos serviços oferecidos pelos centros de formação de condutores. A medida foi anunciada na véspera do aumento do preço da carteira, que devido ao reajuste anual da Unidade Padrão Fiscal passaria para R$ 840.



PorSimples
Simplificando o Português
Coh-Metrix-Port

MAria José Bocorny Finatto | Sair

inicial > Pesquisa > Resultados

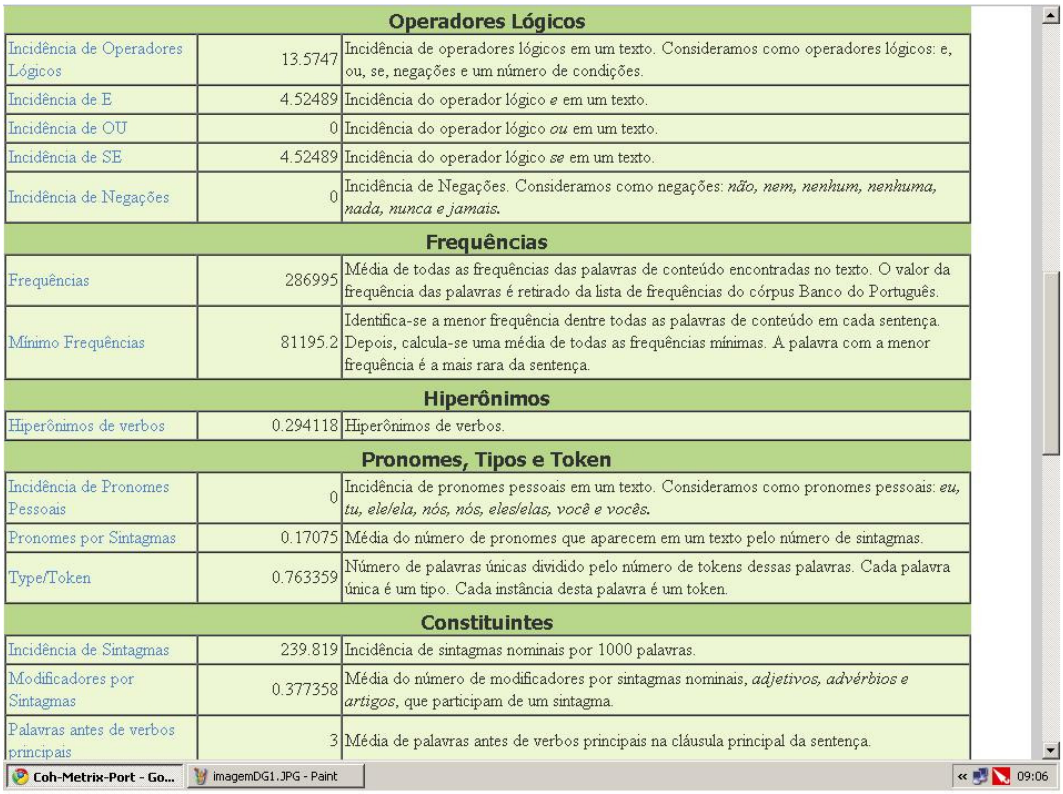
esultados

nir | Visualizar Texto

Texto		
Título	pacote do trânsito	Título
Autor	Adriana Franciosi	Autor
Fonte	Diário Gaúcho	Fonte
Data de Publicação	2008	Data de Publicação
Gênero	Noticiário geral	Gênero
Contagens Básicas		
Índice Flesch	73.8032	Índice Flesch
Número de Palavras	221	Número de palavras do texto.
Número de Sentencas	81	Número de sentenças de um texto.
Número de Parágrafos	76	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
Palavras por Sentenças	2.72839	Número de palavras dividido pelo número de sentenças.
Sentenças por Paragrafos	1.06579	Número de sentenças dividido pelo número de parágrafos.
Sílabas por Palavras de Conteúdo	2.58015	Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios).

Coh-Metrix-Port - Go... 09:04

Figura 8 - Saída do Coh-Metrix-Port – Contagens Básicas, texto de jornal popular- parte 1.



Operadores Lógicos		
Incidência de Operadores Lógicos	13.5747	Incidência de operadores lógicos em um texto. Consideramos como operadores lógicos: e, ou, se, negações e um número de condições.
Incidência de E	4.52489	Incidência do operador lógico *e* em um texto.
Incidência de OU	0	Incidência do operador lógico *ou* em um texto.
Incidência de SE	4.52489	Incidência do operador lógico *se* em um texto.
Incidência de Negações	0	Incidência de Negações. Consideramos como negações: *não, nem, nenhum, nenhuma, nada, nunca e jamais.*
Frequências		
Frequências	286995	Média de todas as frequências das palavras de conteúdo encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do córpus Banco do Português.
Mínimo Frequências	81195.2	Identifica-se a menor frequência dentre todas as palavras de conteúdo em cada sentença. Depois, calcula-se uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença.
Hiperônimos		
Hiperônimos de verbos	0.294118	Hiperônimos de verbos.
Pronomes, Tipos e Token		
Incidência de Pronomes Pessoais	0	Incidência de pronomes pessoais em um texto. Consideramos como pronomes pessoais: *eu, tu, ele/ela, nós, nós, eles/elas, você e vocês.*
Pronomes por Sintagmas	0.17075	Média do número de pronomes que aparecem em um texto pelo número de sintagmas.
Type/Token	0.763359	Número de palavras únicas dividido pelo número de tokens dessas palavras. Cada palavra única é um tipo. Cada instância desta palavra é um token.
Constituintes		
Incidência de Sintagmas	239.819	Incidência de sintagmas nominais por 1000 palavras.
Modificadores por Sintagmas	0.377358	Média do número de modificadores por sintagmas nominais, *adjetivos, advérbios e artigos*, que participam de um sintagma.
Palavras antes de verbos principais	3	Média de palavras antes de verbos principais na cláusula principal da sentença.

Coh-Metrix-Port - Go... | imagemDG1.JPG - Paint | 09:06

Figura 9 - Saída do Coh-Metrix-Port – Contagens Básicas, texto de jornal popular- parte 2.

Como se pode perceber por esses resultados, pelo menos no que se refere ao Índice Flesch, temos um escore de 73.80, o que corresponde a um texto do tipo **fácil**, enquadrado no parâmetro dos índices entre **50 – 75**. Isso dá uma classificação de textos adequados a alunos com escolaridade até a oitava série do ensino fundamental.

Tal como ocorreu com o texto de Pediatria, a classificação parece bem justa se considerar-se o perfil do jornal e de seu público-alvo. Mas, o que mais há além disso? Se detivermos nossa atenção na comparação entre o artigo científico de Pediatria e a notícia do jornal popular, veremos que a presença de pronomes parece ser um diferencial e que há a terminologia, naturalmente (que não constou dos excertos de quadros, mas aparecerá na parte das contagens lexicais nominalizadas num caso e noutro, não).

De outro lado, importa mencionar aqui também que há padrões de texto associados – e cultivados - no jornalismo, independentemente do caráter popular. A escrita de jornal se pretende objetiva e sem repetições. A propósito, vale mencionar que um famoso jornalista[15] já disse que se a língua fosse mais rica em substantivos e verbos, não precisaríamos usar tantos adjetivos e advérbios em um bom texto de jornal. Segundo entende, essas palavras embaçam a exatidão e fazem o texto parecer chumbo em lugar de cristal. Essa seria uma indicação sobre o papel de adjetivos e de advérbios na CT do jornal? De todo modo, contagens de adjetivos e de advérbios associadas a graus de intelegibilidade perfazem um padrão nas métricas do Coh-Metrix-Port.

Assim, com 34 medidas diferentes associadas em torno de índices de inteligibilidade de um texto, não há como não pensar em diferentes níveis ou dimensões das distintas e variadas complexidades mobilizadas. Nesse ponto, mais uma vez, a cooperação com a AMD e as tipologias multiníveis de Ciapuscio (2003) parece ser um objetivo a ser seguido quando pensamos no texto científico em contraste com o texto do jornal popular. Por fim, cabe dizer que o propósito deste texto será cumprido se o leitor que o seguiu até aqui também tiver experimentado essas e outras suspeitas. De nossa parte, seguiremos em busca das inter-relações mencionadas acima, tratando tanto de investigar tanto as já postas quanto as presumidas.

BIBLIOGRAFIA

.

ADDE, Fabíola V.; RODRIGUES, Joaquim C.; CARDOSO, Ary L. Seguimento nutricional de pacientes com fibrose cística: papel do aconselhamento nutricional. *J. Pediatr.* Porto Alegre, v. 80, n. 6, p.475-482, 2004. [recurso eletrônico] Acesso em < http://www.jped.com.br/conteudo/04-80-06-475/port.asp?cod=1261> .

ALMEIDA, D.M de; ALUISIO, S.M. Manual de Uso do Coh-Metrix-Port 1.0. Agosto de 2009. NILC-TR-09-05. Disponível em< caravelas.icmc.usp.br/wiki/images/f/fc/NILC-TR-09-05.pdf> Arquivo acessado em 02/03/2011.

AVERBUCK, L. M.; APPEL, M. B.; SILVEIRA, R. M. H. Leitura: fatores que interferem na compreensao de textos no ensino de primeiro grau. *Leitura*: *Teoria & Prática*. Campinas, v. 1, p. 26-39, 1983.

BAKER, Eva L.; ATWOOD, Nancy K.; DUFFY, Thomas M. Cognitive Approaches to Assessing the Readability. In: DAVISON, Alice; GREEN, Georgia M. (eds.) *Linguistic complexity and text comprehension.* Readability issues reconsidered. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1988.

[15] Otavio Frias Filho, no Antimanual de jornalismo. **Folha de S.Paulo**, 18.nov.1984. Caderno Folhetim, p. 7. Citado por Carlos Kaufmann (KAUFMANN, 2005)

BERBER SARDINHA, Tony. Análise multidimensional. *DELTA*, São Paulo, v. 16, n. 1, p. 99-127, 2000.
____. *Linguística de corpus*. Barueri: Manole, 2004.
BIBER, Douglas. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988.
____. *Dimensions of Register Variation* – A Cross-Linguistic Comparison. Cambridge: Cambridge University Press, 1995.
CHARROW, Veda. Readability vs. comprehensibility: a case study in improving a real document. In: DAVISON, Alice; GREEN, Georgia M. (eds.) *Linguistic Complexity and text comprehension. Readability Issues Reconsidered*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1988, p.85-114.
CIAPUSCIO, Guiomar. La terminología desde el punto de vista textual: selecció, tratamiento y variación. Porto Alegre, *Organon*, v.12, n.26, p.43-65, 1998
____. *Textos especializados y terminología*.Barcelona: IULA, 2003.
____ . Las metáforas en la comunicación de ciencia. In: HARVEY, Anamaría (org.) *En torno al discurso:* Estudios y perspectivas. Santiago: Universidad Católica de Chile, 2005, p. 81-93.
DAVISON, Alice; GREEN, Georgia M. (eds.) *Linguistic complexity and text comprehension:* readability issues reconsidered. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1988.
FINATTO, M. J. B.; HUANG, C. Da adjetivação em Química e Medicina: algumas implicações para os estudos do léxico e de textos técnico-científicos. *Revista Língua & Literatura*, Frederico Westphalen-RS, v. 6 e 7, , p. 45-56, 2005**.**
FINATTO, M.J.B; AZEREDO, Susana de. Observações da tessitura do texto especializado são observações de/em Terminologia? In: *As Ciências do Léxico*. Lexicologia, Lexicografia, Terminologia. v.4. Campo Grande, MS: Editora da UFMS/ Porto Alegre: Editora da UFRGS, 2010. p. 557-578.
FULGÊNCIO, Lúcia, LIBERATO, Yara. *Como facilitar a leitura: como se processa a leitura; orientação para textos didáticos; aspectos discursivos*. São Paulo: Contexto, 1992.
GRAESSER, A.C., MCNAMARA, D.S., LOUWERSE, M., & CAI, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, v. 36, p.193-202. [recurso eletrônico] Acesso em http://www.memphis.edu/psychology/graesser/publications/documents/BSC505.pdf .
GRAY, Willian. S,; LEARY, Bernice E. *What makes a book readable?* With special reference to adults of limited reading ability an initial study. Chicago: The University of Chicago Press, 1935.
KALVERKÄMPER, H. Textuelle Fachsprachen-Linguistik als Aufgabe. *Zeitschrift für Literaturwissenschaft und Linguistik*, Siegen, v. 51/52, n. 13, p. 124-166, 1983.
KATO, Mary. *Reconhecimento instantâneo e processamento* em leitura. Uberaba, MG: 1982 (*Série Estudos*, 8).
KAUFMANN, Carlos. *O corpus do jornal:* variação linguística, gêneros e dimensões da imprensa diária escrita. Dissertação (Mestrado) Programa de Estudos Pós-Graduados em Linguística Aplicada, Pontifícia Universidade Católica de São Paulo, São Paulo, SP, 2005.
KLEIMAN, A. Aprendendo palavras, fazendo sentido: o ensino de vocabulário nas primeiras séries. In: *Trabalhos em Linguística Aplicada 9*. Campinas, SP: Universidade Estadual de Campinas, 1987. p. 47-81.
KLEIMAN, A. *Leitura:* Ensino e Pesquisa. Campinas, SP: Pontes, 1989.
KLEIMAN, A. *Oficina de Leitura teoria e prática.* Campinas, SP: Pontes, 1993.
KLEIMAN, A. *Texto e leitor:* aspectos cognitivos da leitura. Campinas: Pontes, 1997. 5.ed.
LEFFA, V. J. *Fatores da compreensão na leitura.* Projeto ELO, Ensino de línguas online: 1996. Disponível em: www.leffa.pro.br Arquivo acessado em 30/08/2007.
LEFFA, V. J. O conceito de leitura. In: LEFFA, V. J. *Aspectos da leitura.* Porto Alegre: Sagra-Luzzato, 1996. p. 9-24.
MACIEL, Anna Maria B. Linguagens especializadas e terminologia: o passado projetando o futuro. In: PERNA, C.; DELGADO, H.K.; FINATTO, M.J.B. *Linguagens especializadas em*

corpora: modos de dizer e interfaces de pesquisa [recurso eletrônico]. Porto Alegre: EDIPUCRS, 2010. Modo de Acesso: < http://www.pucrs.br/edipucrs >

MEURER, J. L.; MOTA-ROTH, D. *Gêneros textuais e práticas discursivas*. Florianópolis: EDUSC, 2005.

NEIS, Ignacio A. A competência de leitura. Porto Alegre: *Letras de Hoje*, 15 (2), 1982, p.43-57.

NUNES, M. G. V.; ALUÍSIO, S. M.; PARDO, T. A. S.. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. *Linguamática.* Revista para o Processamento Automático das Línguas Ibéricas, v. 2, n.2 , p. 13-27, 2010. Disponível em < http://www.linguamatica.com/linguamatica-v2n2.pdf > Arquivo acessado em 08/08/2010).

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática* (Revista para o Processamento Automático das Línguas Ibéricas), v. 2, n.1, p. 45-61, 2010. Disponível em < http://linguamatica.com/index.php/linguamatica/article/viewFile/44/59 > Arquivo acessado em 08/08/2010.

SHERGUE, Orlando. *Dimensão de Variação no Discurso Médico- Acadêmico:* o artigo de pesquisa e a apresentação de trabalhos científicos em congressos. Dissertação (Mestrado) Programa de Estudos Pós-Graduados em Linguística Aplicada, Pontifícia Universidade Católica de São Paulo, São Paulo, SP, 2003,

SWALES, J.M. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press, 1990.

VIANA, Vander; TAGNIN, Stella E. O. (orgs.). *Corpora no ensino de línguas estrangeiras*. São Paulo: HUB Editorial, 2010.

VIEIRA, Renata; LIMA,Vera Lúcia Strube. "JAIA/Linguística Computacional: p rincípios e aplicações". In: MARTINS, Ana Teresa; BORGES, Díbio Leandro (eds.),*:* XXI CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO *As Tecnologias da informação e a questão social.* Anais ... Fortaleza, 2001. Porto Alegre: Sociedade Brasileira de Computação, 2001.

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas**.** In: PERNA, C.; DELGADO, H.K.; FINATTO, M.J.B. Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa [recurso eletrônico]. Porto Alegre: EDIPUCRS, 2010, p. 184-201. Modo de Acesso: < http://www.pucrs.br/edipucrs >

ZILIO, L. *Colocações especializadas e Komposita: um estudo contrastivo alemão-português na área de cardiologia*. Dissertação (Mestrado) Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.