Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero

Maria José B. Finatto[1], Carolina E. Scarton[2], Amanda Rocha[2], Sandra Aluísio[2]

[1]Instituto de Letras – Universidade Federal do Rio Grande do Sul (UFRGS), Pós-Doutoranda USP-ICMC-NILC - 91540-000 – Porto Alegre – RS – Brazil

[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

{maria.finatto,carol.scarton}@gmail.com,{amandarc, sandra}@icmc.usp.br

Abstract. *This paper compares the readability of popular Brazilian newspapers to traditional ones using cohesion, syntax and vocabulary metrics, including ellipsis. The newspapers used are the popular Diário Gaúcho (DG) and the traditional Zero Hora (ZH); both published in Porto Alegre, RS. ZH aims at more educated readers, whereas DG is dedicated for a public with lower education. The aim of this study was to set the first steps in the analysis of popular newspapers as an emerging journalistic Brazilian genre. We concluded that the most discriminative features between both newspapers are a set of 14 features extracted using Coh-Metrix-Port, but ellipsis does not have a distintictive role.*

Resumo. *Este trabalho contrasta a inteligibilidade de jornais populares e de jornais tradicionais brasileiros a partir de atributos coesivos, vocabulares e sintáticos, incluindo elipses. Os jornais são o popular Diário Gaúcho (DG) e o tradicional Zero Hora (ZH), publicados em Porto Alegre, RS. ZH dirige-se a leitores de maior escolaridade, enquanto DG visa público de renda e escolaridade inferiores. Pretende-se auxiliar a análise do jornal popular como um gênero textual jornalístico emergente. Conclui-se que os atributos mais distintivos são um conjunto de 14 extraídos com ajuda da ferramenta Coh-Metrix-Port e que as elipses não contribuem para a diferenciação.*

1. Introdução

O mercado de jornais populares brasileiros (JPBs) cresceu e mudou muito nos últimos anos, obrigando quem só conheça o lugar-comum do tablóide sensacionalista ou da "imprensa marrom" a rever ideias e preconceitos sobre esse tipo de publicação [Amaral, 2006, p. 9-11]. Trata-se de um segmento de jornais comerciais bastante recente, cuja história não tem mais de 15 anos e que visa uma aproximação com as amplas camadas da população urbana de menor poder econômico, consumidores com nível de escolaridade relativamente baixo e pouco hábito de leitura. Para atingir esse público, os JPBs oferecem preço baixo, usam textos curtos, fartura de imagens, linguagem simples e didatismo, tendendo a explorar bastante o tema de prestação de serviços para comunidades de baixa renda [Oliveira, 2009]. Diferente da linha essencialmente sensacionalista, "marca registrada" desse tipo de jornal vinte anos atrás, hoje os JPBs mostram um gradativo novo posicionamento. Tratam prioritariamente de temas relacionados ao cotidiano do seu público (saúde, mercado de trabalho, transporte e

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 49–58,
Cuiabá, MT, Brazil, October 24–26, 2011.

educação), reservando boa parte de seu conteúdo para temas de entretenimento e notícias sobre esportes e celebridades. Apesar dessas mudanças, conforme estudos de Jornalismo [Amaral, 2006], permanece bom espaço para casos policiais, histórias de interesse humano e feitos extraordinários, tal qual ocorria nas antigas publicações sensacionalistas.

A despeito da importância que assumem os JPBs, pois têm incrementado o índice de leitura de grandes parcelas da população urbana de baixa renda e oferecido um mercado de trabalho expressivo [Amaral, 2006, p.80], são poucas as pesquisas linguísticas e de Jornalismo/Comunicação dedicadas a observar a constituição do seu texto e a descrever esse novo gênero textual jornalístico brasileiro [Amaral, 2004 e 2006; Bernardes, 2004; Oliveira, 2009; Silva e Finatto, 2009]. Assim, são necessários mais estudos para subsidiar a verificação de seus traços constitutivos e diferenciais, quer em relação a jornais realmente sensacionalistas ou apelativos, que seguem existindo, quer em relação àqueles jornais tradicionais que visam atingir a população mais escolarizada, também denominados *jornais de referência* [Amaral, 2006, p.55].

Nsse cenário, faz-se uma análise contrastiva da inteligibilidade de textos de jornais populares e textos de jornais tradicionais a partir da verificação de um conjunto de características coesivas, vocabulares, sintáticas e discursivas. Os jornais sob exame são o popular *Diário Gaúcho* (DG) e o tradicional *Zero Hora* (ZH), ambos publicados em versão impressa na cidade de Porto Alegre, RS, pela empresa jornalística RBS. Considerando o padrão atual para classes sócio-econômicas no Brasil, ZH é dirigido para as classes de maior poder aquisitivo, classes A, B e C; enquanto o DG visa atingir leitores das classes C, D e E, cuja faixa de renda é inferior. É interessante observar que ambos veículos compartilham leitores da classe C.

O foco maior, obviamente, recai sobre o DG, visto que integra gênero pouco estudado, além de ser um dos mais antigos no Brasil no seu segmento (fundado no ano de 2000). Além disso, possui alta tiragem – cerca de 160 mil exemplares/dia, sendo que cada exemplar tende a ser compartilhado por cinco pessoas. Após o seu lançamento, uma pesquisa do IBOPE revelou que o DG foi responsável pela elevação do índice de leitores da região metropolitana de Porto Alegre para o maior de todo o Brasil [Amaral, 2006, p.38-39]. O grau de instrução dos leitores do DG está assim distribuído: 60% com Ensino Fundamental; 34% com Ensino Médio e 6% com Ensino Superior. Sua renda média fica na faixa de até 5 salários mínimos. A missão do jornal é mostrar maneiras de melhorar a vida desse segmento social [Amaral, 2006, p. 80-85]. Conforme ratifica o *Manual de Redação do DG* (2005), seus pilares editoriais são: enfoque no leitor, interatividade, utilidade e serviço, vida real e otimismo.

A análise aqui apresentada visa auxiliar a verificar em que medida o DG se colocaria como um gênero textual jornalístico emergente e a questão de pesquisa deste estudo é: considerando sua conformação lexical e sintática (incluindo a sintaxe frasal e textual), por meio de quais características o texto do jornal popular se diferenciaria do texto do jornal tradicional? Como instrumento de auxílio para responder a questão, são explorados algoritmos de aprendizado de máquina supervisionado que usam como *features* medidas oriundas da Linguística Computacional e da Psicolinguística para avaliação da inteligibilidade. Essas *features* são as métricas do sistema Coh-Metrix-Port [Scarton e Aluísio, 2010]. Esse sistema é a versão brasileira do sistema Coh-Metrix original para o inglês [Graesser; Mcnamara; Louwerse; Cai, 2004]. O propósito desses

sistemas é 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. Como inovação ao sistema brasileiro atualmente disponível, que opera com 48 métricas, foi testado aqui um conjunto de cinco métricas que dão conta da presença de elipses de sintagmas nominais, tendo em vista mensurar seu papel para a avaliação da inteligibilidade.

Para Lima (2001), elipses tratam de um recurso textual condensador caracterizado por ser um elemento de coesão referencial em que o item elíptico só pode ser interpretado semanticamente graças aos pressupostos (os itens explicitados no texto, também denominados antecedentes) ou por um conhecimento do leitor sobre a situação ou contexto comunicativo. Como são elementos "economizadores" de palavras, tenderiam a ser utilizados em textos jornalísticos também para evitar a repetição, salientando-se que baixa repetitividade é índice da boa escrita jornalística conforme atestam diferentes manuais de redação de grandes jornais brasileiros [Villar Belmonte, 2010]. Contudo, supondo que a omissão de um termo na estrutura frasal possa dificultar a compreensão de leitores de menor proficiência, como é o caso de 60% dos leitores do DG, pretendeu-se então medir também a ocorrência de elipses. A inserção de *features* relacionadas a elipses, todavia, é experimental e apóia-se em um trabalho de anotação manual realizado por anotador único.

Na Seção 2, são apresentados os trabalhos relacionados da área de inteligibilidade textual. Na Seção 3, é apresentada a abordagem proposta, descrevendo-se os corpora e as *features* utilizadas nos experimentos, bem como os experimentos e resultados. Por fim, a Seção 4 fica reservada para as conclusões e trabalhos futuros.

2. Trabalhos relacionados em Avaliação da Inteligibilidade

A avaliação da inteligibilidade quantifica a dificuldade de compreensão de leitura ou a complexidade de um texto. Dubay (2004) relata que surgiram 200 fórmulas entre 1920 e 1980 para avaliar a complexidade de textos em inglês, as quais geralmente recorrem ao uso de propriedades superficiais como tamanho de sentenças e de palavras. O índice *Flesch Reading Ease,* p.ex., foi adaptado para o português do Brasil, apresentando a complexidade em quatro níveis, correspondentes a séries escolares [Martins et al., 1996]. Entretanto, essas fórmulas mostram-se insuficientes quando não conseguem, por exemplo, abranger a importância de marcadores discursivos [Williams, 2004].

A maioria dos trabalhos mais atuais sobre avaliação da inteligibilidade de textos usa métodos de aprendizado de máquina (AM) e avalia suas abordagens para textos em inglês. Em outra abordagem, um trabalho para a língua alemã [Glöckner et al., 2006] propõe avaliar a inteligibilidade via uso de várias *features* e de uma medida global de inteligibilidade similar ao índice *Flesch*. A língua portuguesa também foi contemplada em trabalhos recentes [Aluísio et al., 2010; Scarton e Aluísio, 2010; Scarton et al., 2010]. Esses trabalhos empregam a abordagem de avaliação da inteligibilidade para apoiar a simplificação de textos destinados a leitores com níveis de letramento baixos. Os trabalhos atuais, geralmente, tratam de cinco aspectos principais:

a) Avaliam o conjunto de *features* usado para capturar os vários aspectos da inteligibilidade e a contribuição das *features* de vários níveis linguísticos. É o caso do trabalho de Pitler and Nenkova (2008) que, com base nos trabalhos de criação de rubricas para avaliação de redações de alunos (*essay scoring*) [Burstein et al., 2003],

propõem um *framework* unificado composto de *features* relacionadas a vocabulário, sintaxe, elementos de coesão lexical e relações discursivas para medir a qualidade de um texto. Feng et al. (2010), seguindo os estudos de Pitler e Nenkova (2008), propõem o uso de várias *features,* que são comparadas e avaliadas em termos de seu impacto para prever uma série de livros de leitura adequados para estudantes do nível fundamental.

b) Focam uma dada audiência para a qual a avaliação da inteligibilidade é destinada. É o caso de trabalhos focando em aprendizes do inglês como língua estrangeira [Schwarm and Ostendorf, 2005], pessoas com capacidade intelectual reduzida [Feng et al., 2009], pessoas com problemas cognitivos causados por Alzheimer [Roark at al., 2007], textos para adultos ou para crianças [Scarton e Aluísio, 2010] e textos para um determinado nível de letramento [Aluísio et al., 2010].

c) Tratam dos efeitos do gênero textual no cálculo do índice de inteligibilidade. É o caso dos trabalhos de Sheehan et al. (2007), que estudam modelos para textos expositivos e literários, considerando que o uso de índices simples, como *Flesch-Kincaid Level*, tendem a subestimar a dificuldade dos primeiros e sobrestimar a dos últimos e Scarton et al. (2010) que apresentam os primeiros resultados na proposta de um avaliador de inteligibilidade global, usando as *features* do Coh-Metrix-Port.

d) Avaliam qual o modelo estatístico é mais apropriado para os índices (escalas nominais, ordinais ou intervalares) e para a precisão de métodos de aprendizado de máquina. Heilman et al. (2008) investigaram as escalas de medida para dificuldade de leitura – nominal, ordinal e intervalar – via comparação da efetividade de modelos estatísticos para estes tipos de dados, enquanto que Petersen and Ostendorf (2009) defenderam o uso de *Support Vector Machines* tanto como modelo de regressão quanto de classificação para predizer níveis de inteligibilidade. Para o português, Aluísio et al. (2010) avaliaram modelos nominal, ordinal e intervalar para textos em português (originais e simplificados) obtendo resultados similares.

e) Focam-se em aplicações computacionais que utilizam métodos de avaliação de inteligibilidade. Heilman et al. (2007) usaram um avaliador de inteligibilidade em sistemas tutores inteligentes para indicar textos de leitura com o nível adequado de dificuldade para aprendizes do inglês como segunda língua. Miltsakali e Troutt (2007 e 2008) propuseram uma ferramenta automática para avaliar textos da *Web* que fossem adequados para adolescentes e adultos com níveis baixos de letramento, enquanto Aluísio et al. (2010) utilizaram um avaliador de inteligibilidade em um editor de simplificação para medir o nível de inteligibilidade com relação a três padrões de letramento (rudimentar, básico e pleno).

Considerando os trabalhos citados, este estudo pretende contribuir com a área de pesquisa em inteligibilidade textual ao tratar de uma audiência ainda não estudada – pessoas das diferentes classes sócio-econômicas brasileiras, relacionando-as ao consumo de jornais populares e tradicionais (aspecto b). Além disso, propõe o cômputo de uma nova classe de *features* (aspecto c) em meio às 48 pré-existentes do sistema Coh-Metrix-Port, o uso de elipses, a fim de verificar se a sua ocorrência, tipos e localização em relação a seu antecedente contribuiriam na classificação de um texto jornalístico como sendo destinado a um ou outro grupo sócio-econômico.

3. Abordagem Proposta

3.1. *Corpora* e *Features*

Na Tabela 1, estão algumas estatísticas dos *corpora* utilizados neste trabalho. Foram 80 textos do ZH dos anos de 2006 e 2007 e 80 textos do DG do ano de 2008. Os textos foram selecionados buscando-se a maior variedade possível de temas e editorias, tendo em comum o fato de serem antecedidos por um parágrafo destacado que funciona como uma pequena síntese-guia de seu conteúdo, denominada tecnicamente *lead*.

Tabela 1: Estatísticas dos *corpora* ZH e DG

Corpus	Número de textos	Número de palavras	Número de palavras por texto	Público-alvo
ZH	80	55.528	694,100	Classes A, B, C
DG	80	32.706	408,825	Classes C, D, E

Como features para o aprendizado de máquina, foram utilizadas as 48 métricas, geradas automaticamente pela ferramenta Coh-Metrix-Port (http://caravelas.icmc.usp.br:3000), mais cinco novas features, derivadas de uma anotação manual de co-referência de elipses. A anotação do *corpus* envolveu identificar elipses de três tipos: nominais, verbais e sentenciais. Para esta análise, decidiu-se destacar somente a anotação das elipses nominais, pois muitas delas se apresentam distantes de seus antecedentes, o que poderia dificultar a sua recuperação por parte do leitor. Já as elipses verbais ocorrem, normalmente, a uma distância de três *tokens*, ou seja, são, em tese, facilmente recuperáveis, e elipses sentenciais foram pouco frequentes. A seguir, listam-se todas as *features*:

- Contagens básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças, sílabas por palavras, incidência de verbos, incidência de substantivos, incidência de adjetivos, incidência de advérbios, incidência de pronomes, incidência de palavras de conteúdo (verbos, substantivos, adjetivos e advérbios) e incidência palavras funcionais (artigos, preposições, pronomes, conjunções e interjeições). *Incidência* corresponde a uma medida de "densidade". Por exemplo, a incidência de verbos é calculada pela fórmula (nº de verbos/nº de palavras)*1000.
- O índice Flesch para o português [Martins et al., 1996].
- Constituintes: ocorrência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Frequências: frequência de palavras de conteúdo e mínimo da frequência de palavras de conteúdo.
- Conectivos: incidência dos seguintes elementos: de todos os conectivos, de conectivos aditivos positivos, de conectivos aditivos negativos, de conectivos temporais positivos, de conectivos temporais negativos, de conectivos causais positivos, de conectivos causais negativos, de conectivos lógicos positivos e incidência de conectivos lógicos negativos.
- Operadores lógicos: incidência de: operadores lógicos, número de *e*, número de *ou*, número de *se* e número de negações.
- Pronomes, tipos e *tokens*: incidência de pronomes pessoais, pronomes por sintagmas nominais e relação tipo/*token*.
- Hiperônimos: hiperônimos de verbos
- Ambiguidades: ambiguidade de verbos, ambiguidade de substantivos, ambiguidade de adjetivos e ambiguidade de advérbios.

- Correferências: sobreposição do argumento em sentenças adjacente, sobreposição de argumento, sobreposição do radical de palavras em sentenças adjacente, sobreposição do radical de palavras, sobreposição de palavras de conteúdo em sentenças adjacentes.
- Anáforas: referência anafórica em sentenças adjacentes e referência anafórica.
- Distância média de *tokens* entre antecedente e elipse: quantidade média de palavras presente entre a elipse e seu antecedente.
- Distância sentencial média entre antecedente e elipse: distância média de afastamento entre a elipse e seu antecedente. A distância inicial começa em zero, caso em que antecedente e elipse se encontram na mesma sentença.
- Total de elipses cujos antecedentes são sintagmas nominais.
- Total de elipses extratextuais.
- Total de elipses indeterminadas, cujo antecedente, geralmente, remete a um padrão sintático de indeterminação de sujeito gerada por um verbo.

3.2 Experimentos e Resultados

Foram definidos quatro cenários para execução dos algoritmos de AM: (a) todas as *features*; (b) sem as *features* de elipses; (c) *features* selecionadas pelo algoritmo InfoGainAttributeEval; e (d) *features* selecionadas pelo algoritmo SVMAttributeEval. Com ambos os algoritmos de seleção de *features*, foram selecionadas 21, apresentando 14 *features* em comum (Tabela 2). Os experimentos com seleção de *features* foram realizados para identificar quais *features* melhor representavam o domínio proposto. Foram escolhidos dois algoritmos de seleção (com implementações distintas) para verificar se ambos selecionavam *features* em comum.

Tabela 2: 14 *features* em comum entre InfoGainAttributeEval e SVMAttributeEval

1. Número de Palavras	2. Sobreposição de palavras de conteúdo em sentenças adjacentes
3. Palavras por sentenças	4. Número de sentenças
5. Pronomes por sintagmas	6. Modificadores por sintagmas
7. Frequência de palavras de conteúdo	8. Palavras antes de verbos principais
9. Sobreposição do radical de palavras	10. Sílabas por palavras
11. Incidência de adjetivos	12. Incidência de verbos
13. Sobreposição do argumento	14. Número de parágrafos

Além das *features* apresentadas na Tabela 2, cada algoritmo selecionou sete *features* distintas. No caso do InfoGainAttributeEval, foram selecionadas: mínimo da freqüência de palavras de conteúdo, Flesch, sobreposição do argumento em sentenças adjacente, sobreposição de radical de palavras em sentenças adjacentes, incidência de palavras funcionais, relação tipo/token e ambiguidade de verbos. Já com o SVMAttributeEval foram selecionadas: incidência de substantivos, incidência de negações, sentenças por parágrafos, conectivos causais negativos, total de elipses indeterminadas, incidência de *ou*, ambiguidade de substantivos.

Para todos os experimentos (seleção de *features* e treinamento), foi utilizado o pacote Weka (Witten e Frank, 2005). Os algoritmos de AM utilizados foram: SMO (implementação do algoritmo SVM no Weka) com PolyKernel e expoente 1.0, 2.0 e 3.0, MultilayerPerceptron, SimpleLogistic (algoritmo de máxima entropia) NaiveBayes, IBk (implementação do algoritmo K-NN no Weka) com o melhor k escolhido em cada caso (foram realizados experimentos variando o valor de k de 1 até 40), J48 (implementação do algoritmo de árvores de decisão C4.5) e JRIP (implementação do *Repeated*

Incremental Pruning to Produce Error Reduction (RIPPER) no Weka). Com estes cinco algoritmos, representam-se os cinco grandes conjuntos de algoritmos de AM: baseados em funções matemáticas, probabilísticos, *lazy*, baseados em árvores e baseados em regras. Na Tabela 3 são apresentados os resultados obtidos para *F-measure* para cada algoritmo em cada cenário proposto.

Tabela 3: Resultados de *F-measure* dos experimentos com métodos de AM

	Cenários			
	(a) k= 15	**(b) k= 28**	**(c) k= 11**	**(d) k= 24**
SMO – PolyKernel p = 1.0	0.825	0.837	0.844	0.837
SMO – PolyKernel p = 2.0	0.769	0.825	0.869	0.825
SMO – PolyKernel p = 3.0	0.769	0.769	**0.877**	0.812
MultilayerPerceptron	0.769	0.819	0.819	0.756
SimpleLogistic	**0.831**	**0.85**	0.837	0.837
NaiveBayes	**0.831**	0.838	0.843	0.85
JRIP	0.799	0.843	0.844	0.819
J48	0.793	0.794	0.831	0.818
IBk	0.779	0.756	0.806	**0.838**

Considerando somente o cenário (a) (coluna 2 da Tabela 3) o melhor resultado foi de 0.831 de *F-measure* em dois algoritmos: SimpleLogistic e NaiveBayes. Em (b) (coluna 3 da Tabela 3) o melhor resultado obtido foi de 0.85 de *F-measure* com o algoritmo SimpleLogistic. Já em (c) (coluna 4 da Tabela 3) o melhor resultado foi de 0.877 de *F-measure* para o algoritmo SMO com PolyKernel cúbico (p=3.0) - este resultado também foi o melhor resultado global. Por fim, em (d) (coluna 5 da Tabela 3), o melhor resultado foi de 0.838 de *F-measure* para o algoritmo IBk com k = 24.

Aparentemente, as métricas de elipses acrescentam ruído à maioria dos classificadores, pois os resultados de *F-measure* pioram do cenário (a) (todas as métricas – coluna 2 da Tabela 3) para o cenário (b) (sem as métricas de elipses – coluna 3 da Tabela 3) nos classificadores SMO com p = 1.0 e p = 2.0, MultilayerPerceptron, SimpleLogistic, NaiveBayes, JRIP e J48. Outro resultado que merece ser comentado é que, utilizando o InfoGainAttributeEval para seleção dos atributos, nenhuma métrica de elipse foi selecionada, enquanto que, utilizando SVMAttributeEval, somente uma métrica de elipse foi selecionada (total de elipses indeterminadas). Este é outro indício de que as *features* de elipses não acrescentam um diferencial na distinção entre os dois jornais quando somadas às demais. Além disso, acredita-se que as melhores *features* para a tarefa de distinguir o jornal popular do jornal tradicional são as apresentadas na Tabela 2, pois são as *features* comuns a todos os cenários. Em outras palavras, essas *features* foram selecionadas por ambos os algoritmos de seleção de atributos utilizados. Entretanto, vários outros algoritmos de seleção de atributos podem ser explorados em trabalhos futuros. Outro fator que vale destacar é que os resultados nos cenários em que foram utilizadas as *features* selecionadas foram melhores do que os resultados obtidos com todas as *features*. Isto pode indicar que as *features* não selecionadas inserem um ruído na classificação.

4. Conclusões e Trabalhos Futuros

No início deste estudo, perguntou-se em que medida o texto do jornal popular se diferenciaria do texto do jornal tradicional. Pelos dados obtidos, há pelo menos 14 medidas ou *features* que auxiliam a distinguir um jornal do outro. Entre elas, não estão as medidas associadas a elipses. Isso revela que o DG, popular, exibe, comparado ao tradicional ZH, similar uso de elipses. Portanto, se elipses equivalessem a texto mais

complexo e até sofisticado em termos de elaboração, poder-se-ia detectar justamente aí um dos traços de um gênero novo, popular e ao mesmo tempo complexo. Chama atenção também a incidência de adjetivos como fator distintivo entre ambos, visto que, assim como a repetitividade, a adjetivação também é recurso desaconselhado para que haja um "bom" texto de jornal [Oliveira, 2009], pois deve ser o mais neutro possível. O DG, novamente, inverte as expectativas pré-concebidas e mostra, neste *corpus*, adjetivação bem menos freqüente que a do ZH, atestada nos dados comparativos do Weka, mostrando-se, em tese, mais objetivo. Naturalmente, seria possível fazer longas considerações sobre o papel de cada uma das *features* identificadas como distintivas, o que não é objetivo deste trabalho. Entretanto, percebe-se o quanto podem contribuir com o trabalho de descrição e caracterização de um novo estilo de texto de jornal, o JPB. Além disso, este estudo tem implicações para várias tarefas e áreas de pesquisa relacionadas ao PLN, como a mineração de textos, pois, ao distinguir tipos ou gêneros próximos, aumenta a precisão da recuperação e identificação de textos. E, distinguindo textos que possuem uma dificuldade de leitura menor, pode-se facilitar também o *parsing* destes. Embora mais pesquisa seja necessária, este estudo gerou uma contribuição na análise estatística de textos, mostrando que a ferramenta Coh-Metrix-Port é capaz de distinguir tipos de jornais, separando com uma boa *F-measure* o jornal popular do tradicional.

Um trabalho futuro é avaliar, com um *corpus* pareado ZH-DG (tratando do mesmo assunto publicado em ambos), diferenças de valores nos grupos de medidas fornecidas pela ferramenta Coh-Metrix-Port. Quanto a ferramentas, pode-se imaginar a criação de uma de suporte à escrita, por exemplo, para jornalistas novatos elaborarem e identificarem textos mais ou menos adequados para uma ou outra classe de leitores.

Agradecimentos

Ao CNPq e à FAPESP pela concessão de bolsas e de auxílios à pesquisa. Ao NILC-ICMC- USP pelo apoio institucional.

Referências

Aluísio, S. M.; Specia, L.; Gasperin, C. andScarton, C. E. (2010) "Readability Assessment for Text Simplification", In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA2010) in conjunction with NAACL HLT 2010*, Los Angeles, CA, p. 1-9.

Amaral, M. F. (2004). "Lugares de fala do leitor no Diário Gaúcho", Tese (Doutorado em Comunicação e Informação), Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, 273f.

Amaral, M. F. (2006) "Jornalismo Popular", São Paulo: Contexto, 144p.

Bernardes, C. B. (2004) "As Condições de produção do jornalismo popular massivo: o caso do Diário Gaúcho", Dissertação (Mestrado em Comunicação e Informação), Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, 258f.

Burstein, J.; Chodorow, M. and Leacock, C. (2003) "CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays", In *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial*

Intelligence (AAAI2003), Acapulco, Mexico, p. 3-10.

DuBay, W. H. (2004) "The principles of readability", Costa Mesa, CA: Impact Information, 74p.

Feng, L.; Jansche, M.; Huenerfauth, M. and Elhadad, N. (2009) "Cognitively Motivated Features for Readability Assessment", In *Proceedings of the 12th Conference of European Chapter of the Association for Computational Linguistics (EACL2009)*, Athens, Greece, p. 229-237.

Feng, L.; Jansche, M.; Huenerfauth, M. and Elhadad, N. (2010) "A Comparison of Features for Automatic Readability Assessment", In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Poster Volume, Beijing, China, p. 276-284.

Graesser, A. C.; McNamara, D. S.; Louwerse, M., and Cai, Z. (2004) "Coh-Metrix: Analysis of text on cohesion and language". BehaviorResearchMethods, Instruments, &Computers, v.36, n.2, p. 193-202.

Glöckner, I.; Hartrumpf, S.; Helbig, H.; Leveling, J. and Osswald, R. (2006) "An architecture for rating and controlling text readability". In *Proceedings of the 8th Conference on Natural Language Processing (KONVENS2006)*, Konstanz, Germany, p. 32-35.

Heilman, M.; Collins-Thompson, K.; Callan, J. and Eskenazi, M. (2007) "Combining lexical and grammatical features to improve readability measures for first and second language texts". In the *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-2007)*, Rochester, NY, p. 460-467.

Heilman, M.; Collins-Thompson, K.; Callan, J. and Eskenazi, M. (2008) "An Analysis of Statistical Models and Features for Reading Difficulty Prediction". In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA2008) in conjunction with ACL 2008*, Columbus, OH, p. 71-79.

Lima, R. L. de M. (2001) "Um mecanismo de coesão: a elipse". Todas as Letras, n.3, n.1, p. 25-35.

Manual de Redação do Diário Gaúcho (2005). Porto Alegre, RS: Rede Brasil Sul de Comunicação, 45p.

Martins, T. B. F.; Ghiraldelo, C. M.; Nunes, M. G. V. and Oliveira Jr, O. N. (1996) "Readability formulas applied to textbooks in Brazilian Portuguese". São Carlos: Notas do ICMC-USP, Série Computação, n. 28, 11p.

Miltsakaki, E. and Troutt, A. (2007) "Read-X: Automatic Evaluation of Reading Difficulty of Web Text", In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2007)*, Quebec, Canada, p. 7280-7286.

Miltsakaki, E. and Troutt, A. (2008) "Real Time Web Text Classification and Analysis of Reading Difficulty". In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA2008) in conjunction with ACL 2008*, Columbus, OH, p. 89-97.

Oliveira, M. R. A. R. (2009) "Jornal Popular X Jornal Tradicional: Análise léxico-gramatical da notícia a partir da Linguística de Corpus Um estudo de casos dos jornais cariocas "O Globo" e "O Dia"". Veredas – Revista de Estudos Linguísticos, v.13, n.2, p. 07-19.

Petersen, S. E. and Otendorf, M. (2009) "A machine learning approach to reading level assessment". Computer Speech and Language, v.23, n.1, p. 89-106.

Pitler, E. and Nenkova, A. (2008) "Revisiting readability: A unified framework for predicting text quality". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, Waikiki, Honolulu, Hawaii, p. 186-195.

Roark, B.; Mitchell, M. and Hollingshead, K. (2007) "Syntactic complexity measures for detecting mild cognitive impairment". In *Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing (BioNLP 2007)*, Prague, Czech Republic, p. 1-8.

Scarton, C. E. e Aluísio, S. M. (2010) "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.

Scarton, C. E.; Gasperin, C. and Aluísio, S. M. (2010) "Revisiting the Readability Assessment of Texts in Portuguese". In *Proceedings of the 12th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2010)*, Bahia Blanca, Argertina, p. 306-315.

Silva, B. R. da e Finatto, M. J. B. (2009) "Português Popular Escrito: o Vocabulário do Jornal Diário Gaúcho". In *Anais do X Salão de Iniciação Científica da PUCRS*, Porto Alegre: EDIPUCRS, p. 3332-3334.

Schwarm, S. E. and Ostendorf, M. (2005) "Reading Level Assessment Using Support Vector Machines and Statistical Language Models". In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL2005)*, University of Michigan, Ann Arbor, p. 523–530.

Sheehan, K. M.; Kostin, I. and Futagi, Y. (2007) "Reading Level Assessment for Literary and Expository Texts". In *Proceedings of the 29th Annual Cognitive Science Society*, Austin, TX, p. 1853.

Villar Belmonte, R. (2010) "A coesão textual frente à regra jornalística da não-repetição de palavras". Porto Alegre, RS: Curso de Especialização em Estudos Linguísticos do Texto, Universidade Federal do Rio Grande do Sul.

Williams, S. (2004) "Natural Language Generation (NLG) of discourse relations for different reading levels". PhD Thesis, University of Aberdeen.

Witten, I. H. and Frank, E. (2005) "Data Mining: Practical machine learning tools and techniques". San Francisco: Morgan Kaufmann, 2nd Edition.