AAACT – Avaliador e Analisador Automático de Coesão Textual

Thiago Gaglietti de Cândido[1], **Carine Geltrudes Webber**[1]

[1]Área do Conhecimento de Ciências Exatas e Engenharias
Universidade de Caxias do Sul (UCS)
Rua Francisco Getúlio Vargas, 1130 - CEP 95070-560 Caxias do Sul - RS - Brasil

{tgcandido,cgwebber}@ucs.br

Abstract. *This article presents the process of analysis and evaluation of textual cohesion through the implementation of an online software. The method uses the lexical cohesion mechanism and is based on Focusing Theory [Sidner 1979, Sidner 1981] and Centering Theory[Grosz et al. 1983], and the evaluation is performed by using an adaptation of the method used by Nobre (2011).*

Resumo. *Esse artigo apresenta o processo de análise e avaliação da coesão textual através da implementação de uma ferramenta on-line. O método apresentado utiliza o mecanismo de coesão lexical como elemento principal da análise, e é baseado na Teoria do Foco [Sidner 1979, Sidner 1981] e Teoria da Centragem [Grosz et al. 1983]. A avaliação é realizada por meio de uma adaptação do método utilizado por Nobre (2011).*

1. Introdução

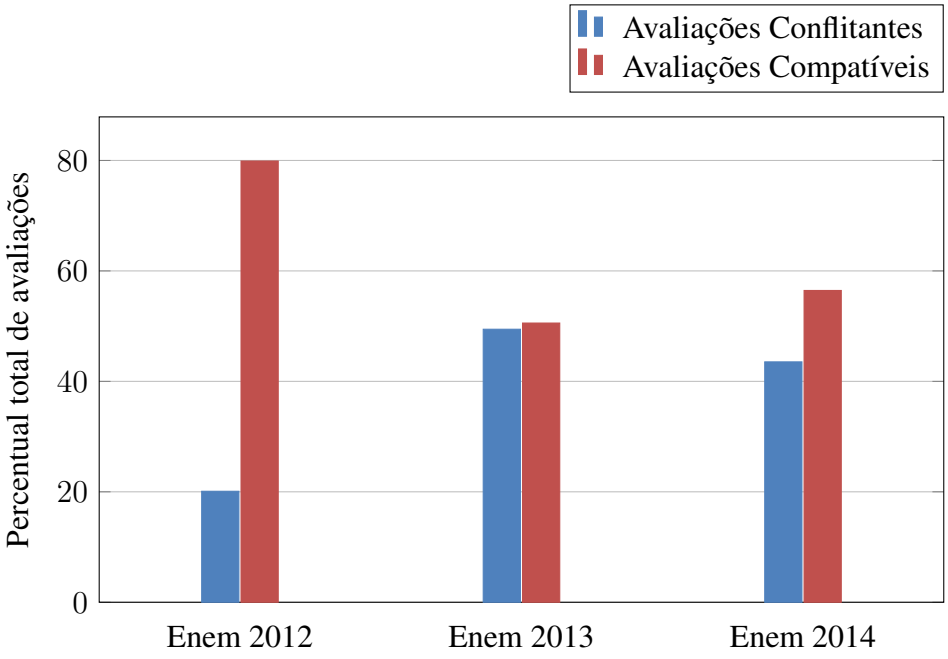
A avaliação de uma redação é uma tarefa complexa, dada sua subjetividade. Mesmo quando efetuada por especialistas, a avaliação pode resultar em diferentes pontuações dentro de uma banca avaliadora, o que causa problemas quando se trata de trabalhos acadêmicos ou de redações de vestibulares. Para tratar desse problema, é possível empregar a tecnologia para auxiliar no processo avaliativo de redações.

A área de *Análise Discursiva Automática* (ADA) compreende técnicas para automatizar pela via computacional tarefas como a avaliação de redações, atividade antes apenas realizada por especialistas objeto principal de estudo da subárea de *Automated Essay Scoring* (AES). Ao avaliar uma mesma redação, dois peritos podem obter resultados diferentes, demonstrando que existem obstáculos no processo efetuado por humanos. Pode-se observar isso analisando dados da correção de redações de provas de vestibular, como o *ENEM - Exame Nacional do Ensino Médio*. Na Figura 1, demonstra-se o percentual de compatibilidade entre as notas atribuídas por avaliadores nos exames de 2012, 2013 e 2014.

A dificuldade para obter um consenso de notas entre avaliadores é ocasionada pelo extenso conjunto de metodologias utilizado para a realização da análise textual, muitas das quais são mutuamente contraditórias e incompatíveis [McKee et al. 2001]. Também pode-se levar em consideração o fator subjetivo da análise textual, que pode ser justificado pelas diferentes trajetórias acadêmicas dos avaliadores.

Ao avaliar redações acadêmicas, podemos elencar três fatores determinantes à pontuação final, todos pertencentes ao nível semântico da língua. São eles: a

Figura 1. Compatibilidade dos resultados atribuídos pelas bancas avaliadoras nas redações do *ENEM* dos anos de 2012, 2013 e 2014 [Pereira 2014, Pereira 2013, Gei 2015]



Fonte: Elaborada pelo autor

coerência, a coesão e a aderência do conteúdo da redação ao tema proposto [Nobre 2011]. O presente artigo possui como foco a análise e a avaliação da coesão textual, inserida no nível linguístico da semântica, que trata do significado do que é escrito [Hasan and Halliday 1976]. A análise possui como foco o mecanismo da coesão lexical.

O objetivo principal deste trabalho é a implementação de uma ferramenta gratuita para o uso acadêmico e *on-line* que realize a análise e a avaliação automática da coesão textual através de técnicas baseadas na *Teoria do Foco* [Sidner 1979, Sidner 1981], na *Teoria da Centragem* [Grosz et al. 1983] e no método utilizado por Nobre (2011) no sistema *Avaliador Automático de Redação – AVAR*.

Esse artigo é estruturado da seguinte forma: trabalhos correlatos são apresentados na Seção 2. Na Seção 3, existe a revisão bibliográfica, onde conceitua-se a coesão (3.1), anáforas e o mecanismo coesivo de coesão lexical (3.2), sintagmas (3.3), a *Teoria do Foco* e a *Teoria da Centragem* (3.4) e explana-se o cálculo do índice coesivo (3.5). Na Seção (4), detalha-se arquitetura da ferramenta, o processo de análise e avaliação da coesão textual em um texto dissertativo, e os experimentos realizados. Na Seção (5), apresenta-se os resultados dos experimentos. Na Seção (6) conclui-se com observações sobre a pesquisa realizada e apresenta-se possibilidades para trabalhos futuros.

2. Trabalhos Correlatos

O sistema *Avaliador Automático de Redação – AVAR* [Nobre 2011] implementa a avaliação e a valoração da coesão, coerência e adequação ao tema de uma redação de vestibular. O *AVAR* utiliza a *Inteligência Artificial* (IA), a partir da captação de elementos relevantes para o processo de atribuição de nota à redação, por meio do *Sistema de In-*

ferência Fuzzy (SIF) para valorar cada quesito com base na *Teoria de Conjuntos Fuzzy* e na *Lógica Fuzzy*. Os resultados apresentados equipararam-se aos dos especialistas, indicando a viabilidade da aplicação do modelo e dos métodos em contextos reais.

Pardo apresenta o *DiZer – an Automatic Discourse Analyzer for Brazilian Portuguese* [Pardo et al. 2004]. Trata-se de um analisador discursivo automático para a língua português brasileiro, que segue como teoria discursiva a *Rhetorical Structure Theory* (RST). Um corpus com anotações das relações retóricas entre as preposições expressas no texto, chamado de *Rhetalho*, foi utilizado para realizar o treinamento da ferramenta. Os resultados obtidos por meio do software foram satisfatórios para textos científicos, e utilizando textos jornalísticos os resultados apresentados foram aceitáveis [Pardo et al. 2004].

Para a língua inglesa, pode-se citar o *Intelligent Essay Assessor* (IEA). Trata-se de um conjunto de funcionalidades para avaliar a qualidade do conteúdo da redação [Foltz et al. 1999]. O IEA utiliza uma abordagem matemática por meio do método *Latent Semantic Analysis (LSA)*, que é empregado para extrair e representar a semântica contextual das palavras mediante cálculos estatísticos aplicados em uma grande coleção de textos [Landauer and Dumais 1997].

Também para a língua inglesa, o *Criterion Online Essay Evaluation Service* é um sistema *on-line* que realiza a avaliação automática de redações de alunos, por meio de duas aplicações complementares: o sistema de avaliação *E-rater*® que identifica e analisa elementos relacionados a proficiência da escrita; e uma suíte de programas que detectam erros de gramática e da utilização de mecanismos linguísticos chamado de *Critique Writing Analysis Tools* [Burstein et al. 2003].

Levando em consideração a quantidade de ferramentas disponíveis para a realização de análises discursivas automáticas em outros idiomas, pode-se observar que o número de softwares existentes para a língua português brasileiro é limitada [Fonseca et al. 2017]. Isso é mais evidente quando se leva em consideração na comparação apenas ferramentas gratuitas ou de código aberto.

3. Compreendendo os Elementos da Análise Coesiva

Dentre as várias maneiras de analisar um texto, pode-se verificar sua apreensibilidade, frequência de palavras, facilidade de leitura, coesão, coerência, entre outras características. Tratar a coesão de um texto de forma automática é um desafio para a área de *Processamento de Linguagem Natural* (PLN). Elementos linguísticos são utilizados por técnicas computacionais para realizar a avaliação de uma redação. Nessa seção, apresenta-se as principais técnicas, teorias e os elementos para a realização da análise coesiva.

3.1. Coesão

A coesão aborda as articulações gramaticais entre as palavras, orações e frases para garantir uma boa sequenciação de eventos, ou seja, é aspecto fundamental do discurso, por tratar-se da ideia de ordem entre seus elementos [Crystal 2011]. Halliday e Hasan (1976) definem coesão como um conceito semântico que se refere às relações de sentido existentes no interior do texto e que o definem como um texto. Os mesmos autores propõem a distinção dos mecanismos coesivos em cinco categorias. Cada categoria deve respeitar

o modo como os itens lexicais e gramaticais relacionam-se com o texto e no texto. As categorias são: referência, substituição, elipse, conjunção e coesão lexical.

A primeira categoria é a *referência*, um dos principais mecanismos para evitar repetições desnecessárias. Para elucidação das categorias, foram utilizados exemplos de Perez (2016). Observe o primeiro exemplo:

"**As crianças** foram passear no parque. **Elas** foram acompanhadas de seus pais."

Na segunda frase, "elas" refere-se a "as crianças". Já neste segundo exemplo, observa-se a repetição do sujeito:

"**As crianças** foram passear no parque. **As crianças** foram acompanhadas de seus pais."

Embora as duas versões estejam corretas, a primeira forma utiliza uma anáfora pronominal para retomar o elemento referente.

A *substituição* é outra categoria de mecanismo coesivo; por meio dela, palavras e expressões que retomam termos já enunciados são utilizados. O exemplo a seguir demonstra a utilização da substituição:

" **Os alunos** foram advertidos pelo **mau comportamento**. Caso **isso** volte a acontecer, **eles** serão suspensos."

A frase também poderia ser escrita da seguinte maneira:

"**Os alunos** foram advertidos pelo **mau comportamento**. Caso o **mau comportamento** volte a acontecer, **os alunos** serão suspensos."

Porém, não empregaria elementos anafóricos para realizar a substituição.

A *elipse*, terceira categoria de mecanismo coesivo, ocorre por meio da omissão de uma ou mais palavras sem comprometer a clareza da oração, como podemos ver no seguinte exemplo:

"**Mariana** faz o dever de casa e simultaneamente conversa com as amigas pelo bate-papo."

Sem utilizar o mecanismo da elipse, obtemos a seguinte frase:

"**Mariana** faz o dever de casa e simultaneamente **Mariana** conversa com as amigas pelo bate-papo."

A categoria de mecanismo coesivo chamada *conjunção* possibilita relações entre os termos do texto através do emprego de conjunções, como podemos ver no exemplo:

"**Como** estava doente, não fui à escola, **embora** tivesse provas nesse dia."

E, por fim, a quinta categoria dentre os mecanismos coesivos é a *coesão lexical*, que ocorre por meio da reiteração termos, utilizando sinônimos, pronomes, hipônimos e hiperônimos, e da colocação de itens lexicais.

"**Carlos Drummond de Andrade** é considerado o maior poeta brasileiro. **O itabirano** nasceu no dia 31 de outubro de 1902 e faleceu no dia 17 de agosto de 1987. **Gênio das letras**, deixou imortalizada em seus diversos livros sua contribuição para a Literatura Brasileira.".

A classificação de mecanismos coesivos não deve ser vista como uma divisão rígida com uma clara separação entre eles. Por exemplo, os mecanismos de referência e substituição são diferenciados apenas pela escolha das palavras, e não por uma diferença de significado [Hasan and Halliday 1976].

McNamara (1996) afirma que existem dois tipos de coesão, a *Coesão Local* (CL) e a *Coesão Global* (CG). Textos que possuem apenas CL tendem a inibir a compreensão e, decorrentemente, a recuperação de informações. Outros textos que apenas possuem CG são geralmente difíceis de ler e compreender [McNamara et al. 1996].

3.2. Mecanismo da Coesão Lexical

O conceito de anáfora é essencial para o entendimento do mecanismo da coesão lexical. Pode-se definir anáfora como uma unidade linguística que tem sua interpretação definida por uma entidade previamente expressada, chamada de antecedente [Crystal 2011].

A coesão lexical é um mecanismo coesivo empregado por meio da escolha de vocabulário, e é substancialmente obtida por meio da *reiteração* e da *colocação* [Hasan and Halliday 1976].

Por meio de sinônimos e de entidades genéricas, como membros superordenados (hiperônimos) e membros subordinados (hipônimos), ou por itens lexicais idênticos, obtém-se a reiteração. Assim, nomes como *o ser humano* e *o lugar* são itens de referência anafóricos e lexicalmente genéricos. A colocação se dá por meio da ligação de itens lexicais que ocorrem frequentemente no texto [Hasan and Halliday 1976].

3.3. Sintagma, Classificações e Exemplos

Pode-se definir sintagma como um conjunto de constituintes contíguos de uma sentença [Crystal 2011]. Os sintagmas desempenham diferentes funções na sentença, e essa função é definida pelo seu núcleo [Fonseca et al. 2017]. Os tipos de sintagmas são os seguintes [Duarte 2017]: nominal, verbal, adverbial e proposicional.

Um sintagma com o núcleo constituído por um nome ou pronome caracteriza um *sintagma nominal*, e geralmente são representados pelo sujeito e pelos complementos verbais da oração. Observe no exemplo:

”**Os alunos** entregaram **os livros**.”

Um sintagma com o núcleo composto por um verbo caracteriza um *sintagma verbal*:

”As visitas **chegaram**.”

Quando constituído por um complemento nominal e seus respectivos advérbios modificadores e pelo predicativo, é denominado de *sintagma adjetivo*:

”Aquela garota é **gentil**.”

O *sintagma adverbial* é formado pelo advérbio e seus respectivos modificadores. Vejamos:

”**Jamais** acreditarei em você.

E por último, tem-se o *sintagma preposicional* que é constituído de uma locução e serve como modificador de qualquer outro sintagma:

”Somos todos muito responsáveis **pelo cumprimento de todos os deveres**.”

3.4. Teoria do Foco e Teoria da Centragem

As teorias explicadas nessa seção foram utilizadas para realizar a análise do grau de coesão entre as unidades textuais das redações avaliadas.

A *Teoria do Foco* (TF) tem como objetivo a análise de elementos anafóricos [Sidner 1979, Sidner 1981]. Ela sugere que as entidades mais salientes de uma frase devem ser os antecedentes preferenciais para a resolução de uma anáfora numa frase seguinte [Nobre and Pellegrino 2010]. O algoritmo da TF tem como função reduzir o conjunto de possíveis antecedentes introduzidos no universo do receptor durante a interpretação de novas frases proferidas em um dado contexto e propor um caminho mais eficiente para percorrer este universo, já reduzido, em busca de um antecedente [Nobre and Pellegrino 2010]

A *Teoria da Centragem* (TC) tem como proposta medir como a coesão do discurso é influenciada pela compatibilidade entre os centros de atenção e a escolha das expressões de referenciação [Grosz et al. 1995]. O centro de atenção é utilizado para designar o objeto mais relevante do discurso [Grosz et al. 1977]. Essa teoria é uma proposta diretamente relacionada com a *Grosz and Sidner Discourse Theory* (GSDT) [Grosz and Sidner 1986].

O foco é a entidade que o emissor toma como centro de sua atenção em determinado ponto do texto [Nobre 2011]. Para a resolução de anáforas, dois centros de atenção são definidos [Sidner 1981]: o *Foco do Ator* (FA) e o *Foco do Discurso* (FD), os quais são determinados pelo agente e pelo tema de cada frase, utilizando a informação temática [Gruber 1976], a informação gramatical (sujeito, objeto direto, objeto indireto, etc.) e a informação sobre quais são as entidades mais salientes da fase anterior, ou seja, o *Foco Local* (FL) [Grosz et al. 1977].

Seguindo a abordagem utilizada por Nobre (2010), a classificação de *Foco Explícito* (FE) e *Foco Implícito* (FI) é feita da seguinte maneira:

a) FE: É a lista de entidades explicitamente contidas em cada frase do texto. São elementos anafóricos e sintagmas nominais existentes na sentença [Sidner 1979, Sidner 1981]; e
b) FI: É a lista de rótulos semânticos das entidades em FE. No caso de um nome próprio, o rótulo semântico será um identificador único para o termo.

As possíveis leituras do relacionamento em uma frase F entre o FE e FI são [Nobre and Pellegrino 2010]:

a) se existe um elemento E de FE_i em FE_{i+1} ($E \in FE_{i+1}$) e também existe um elemento I de FI_i em FI_{i+1} ($I \in FI_{i+1}$), então as frases F_i e F_{i+1} estão em *processo de elaboração*, visto que compartilham as mesmas entidades;
b) se existe um elemento E de FE_i em FE_{i+1} ($E \in FE_{i+1}$), mas não existe um elemento I de FI_i em FI_{i+1}, então as frases F_i e F_{i+1}($I \notin FI_{i+1}$) estão num processo de *manutenção de tópico*, visto que compartilham alguns elementos explícitos;
c) se não existe um elemento E de FE_i em FE_{i+1} ($E \notin FE_{i+1}$), mas existe um elemento I de FI_i em FI_{i+1} ($I \in FI_{i+1}$), então as frases F_i e F_{i+1} estão num

processo de *mudança de tópico*, haja vista compartilhar as mesmas entidades semânticas; e

d) se não existe elemento E de FE_i em FE_{i+1} ($E \notin FE_{i+1}$), nem um elemento I de FI_i em FI_{i+1} ($I \notin FI_{i+1}$), então as frases F_i e F_{i+1} estão num processo de *mudança de assunto*, pois não compartilham as mesmas entidades explícitas e nem semânticas.

Através do conhecimento dessas teorias, os algoritmos 1 e 2 foram elaborados para a realização da análise da utilização do mecanismo coesivo da substituição no texto [Nobre 2011]. Tem-se F como a representação de uma frase de um discurso representado por D, e P um parágrafo de um discurso, assim $D = \{P_i = \{F_1, F_2\}, ..., P_n = \{F_1, F_2, ..., F_n\}\}$. Um conjunto de elementos de FE contido em F_i é representado por FE_i. Um conjunto de elementos de FI contido em F_i é representado por FI_i. Para os parágrafos, da mesma maneira, um conjunto de elementos de FE contido em P_i é representado por FEP_i. Um conjunto de elementos de FI contido em F_i é representado por FIP_i.

Algoritmo 1: AVALIAÇÃO DE COESÃO LOCAL CL DE UM DISCURSO D

Entrada: *Tokens* obtidos após a análise morfossintática do Discurso D
Saída: Classificação coesiva dos pares (F_i, F_{i+1}) do discurso D
início
  para *cada* $F \in D$ **faça**
    FE_i = Obter substantivos, pronomes, e outros elementos anafóricos e a que termo referem-se de F_i
    FI_i = Obter rótulos semânticos de FE_i
    LI_i = Obter sinônimos de FE_i
  fim
  para *cada par* $(F_i, F_{i+1}) \in D$ **faça**
    $RFE_{i,i+1} = FE_i \cap FE_{i+1}$
    $FI_i = FI_i - (FI_i \cap RFE_{i,i+1})$
    $FI_{i+1} = FI_{i+1} - (FI_{i+1} \cap RFE_{i,i+1})$
    $RFI_{i,i+1} = FI_i \cap FI_{i+1}$
    $PSF_{F_i,F_{i+1}}$ = Classificação de (F_i, F_{i+1}) utilizando a tabela de coesão e os conjuntos $RFE_{i,i+1}$ e $RFI_{i,i+1}$
  fim
fim
retorna C

Por meio da aplicação dos Algoritmos 1 e 2, obtém-se dados que possibilitam a realização do cálculo do *Índice Coesivo*.

3.5. Índice Coesivo

O *Índice Coesivo* (IC) é utilizado para determinar a força coesiva do texto através das relações constituídas entre suas frases e parágrafos [Nobre 2011]. IC é definido pela Equação 1:

Algoritmo 2: AVALIAÇÃO DE COESÃO GLOBAL CG DE UM DISCURSO D

Entrada: *Tokens* obtidos após a análise morfossintática do Discurso D
Saída: Classificação coesiva dos pares de parágrafos (P_i, P_{i+j}) do discurso D
início
  para *cada* $P \in D$ **faça**
    $RFEP_i = (FEP_iF_1 \cap FEP_iF_2) \cup (FEP_iF_1 \cap FEP_iF_n) \cup ... \cup (FEP_iF_{n-1} \cap FEP_iF_n)$
    $RFIP_i = (FIP_iF_1 \cap FIP_iF_2) \cup (FIP_iF_1 \cap FIP_iF_n) \cup ... \cup (FIP_iF_{n-1} \cap FIP_iF_n)$
  fim
  para *cada par* $(P_i, P_{i+j}) \in D$ *com adjacência máxima* **faça**
    $RFEP_{i,i+j} = FEP_i \cap FE_{i+j}$
    $FIP_i = FIP_i - (FIP_i \cap RFEP_{i,i+j})$
    $FIP_{i+j} = FIP_{i+j} - (FIP_{i+j} \cap RFEP_{i,i+j})$
    $RFIP_{i,i+j} = FIP_i \cap FIP_{i+j}$
    $PSP_{F_i,F_{i+j}}$ = Classificação de (P_i, P_{i+j}) utilizando a tabela de coesão e os conjuntos $RFEP_{i,i+j}$ e $RFIP_{i,i+j}$
  fim
fim
retorna C

Tabela 1. Relação de FE e FI para o estabelecimento de coesão, sendo E um elemento de FE_i e I um elemento de FI_i

	$E \in FE_{i+1}$	Pontuação	$E \notin FE_{i+1}$	Pontuação
$I \in FI_{i+1}$	Elaboração	1	Mudança de tópico	0,5
$I \notin FI_{i+1}$	Manutenção do tópico	0,75	Mudança de assunto	0,0

Fonte: [Nobre and Pellegrino 2010]

$$IC = \left(\frac{\sum_{i=1}^{s-1} PSF_{i,i+1}}{s-1} + \frac{\sum_{i=1}^{p-1} \sum_{j=1}^{p-1} PSP_{i,i+j}}{\sum_{j=1}^{p-1} j}\right)/2 \quad (1)$$

Onde s representa o total de sentenças, p o total de parágrafos, $PSF_{i,i+1}$ a nota da relação entre as frases adjacentes, $PSP_{i,i+j}$ a nota da relação entre parágrafos adjacentes, de acordo com a Tabela 1.

4. Materiais e Métodos

A análise realizada pelo software *AAACT – Avaliador e Analisador Automático de Coesão Textual* produz uma pontuação de $0,0$ a $2,0$ pontos que valora o nível coesivo de uma redação levando em consideração o mecanismo de coesão lexical. O processo de análise de ser separada em duas fases: a fase de obtenção de informações e a fase de aplicação das teorias. Na primeira fase o software utiliza recursos externos para a obtenção de informações e, na segunda fase, implementa os algoritmos que realizam a análise e a avaliação da coesão.

4.1. Processo de Análise

Na fase de obtenção de informações é realizada uma análise utilizando o *CORP: Coreference Resolution for Portuguese* concebido por Fonseca et al. (2016). Esse analisador realiza a resolução de correferências em língua portuguesa e possui como objeto principal a resolução das correferências pertencentes às categorias de entidades nomeadas de Pessoa, Local e Organização. O treinamento desse software é realizado por meio dos corpus *Summ-it* que é composto por cinquenta textos jornalísticos do caderno de Ciências da Folha de São Paulo [Collovini et al. 2007], e pelo corpus do *HAREM*, que foi avaliado conjuntamente com o intuito de incentivar as pesquisas na área de PLN com foco na língua portuguesa [Freitas et al. 2010, Santos and Cardoso 2007].

Durante o processo de análise textual, o *CORP* realiza a *tokenização*, cria *chunks* a partir dos *tokens* reconhecidos, e efetua a classificação sintática e a atribuição de rótulos semânticos [Fonseca et al. 2017]. Como saída, além desses elementos, apresenta as cadeias de sintagmas nominais e as menções únicas no texto analisado. A ferramenta é de uso gratuito para trabalhos acadêmicos. Ainda durante a fase de obtenção de informações utilizou-se a base *Tep2* [Maziero et al. 2008] para obter sinônimos do núcleo dos sintagmas nominais.

Na segunda fase, as teorias elencadas são aplicadas utilizando o conhecimento obtido na fase de análise. A TF e a TC foram utilizadas para identificar o grau de coesão entre as unidades do texto e, a partir da obtenção desses dados, é aplicado o cálculo do IC para obter o índice coesivo da redação. O resultado do índice coesivo é utilizado como pontuação da coesão textual.

Os principais componentes da ferramenta são representados no formato de diagrama na Figura 2. A análise e a avaliação da coesão textual é realizada por uma aplicação *Java*, disponibilizada em um servidor e acessível por meio de métodos *HTTP*, que implementa a lógica do analisador e realiza a interação com o analisador *CORP* e a base *Tep2*. Para submeter redações e visualizar os resultados da análise criou-se uma página *web* com *HTML*, *CSS* e *JavaScript*, além das bibliotecas *D3.js* e *Angular.js*.

4.2. Análise Exemplificada

Como exemplo, possui-se o discurso D contendo as frases F_1 e F_2 para demonstrar o processo de análise e avaliação.

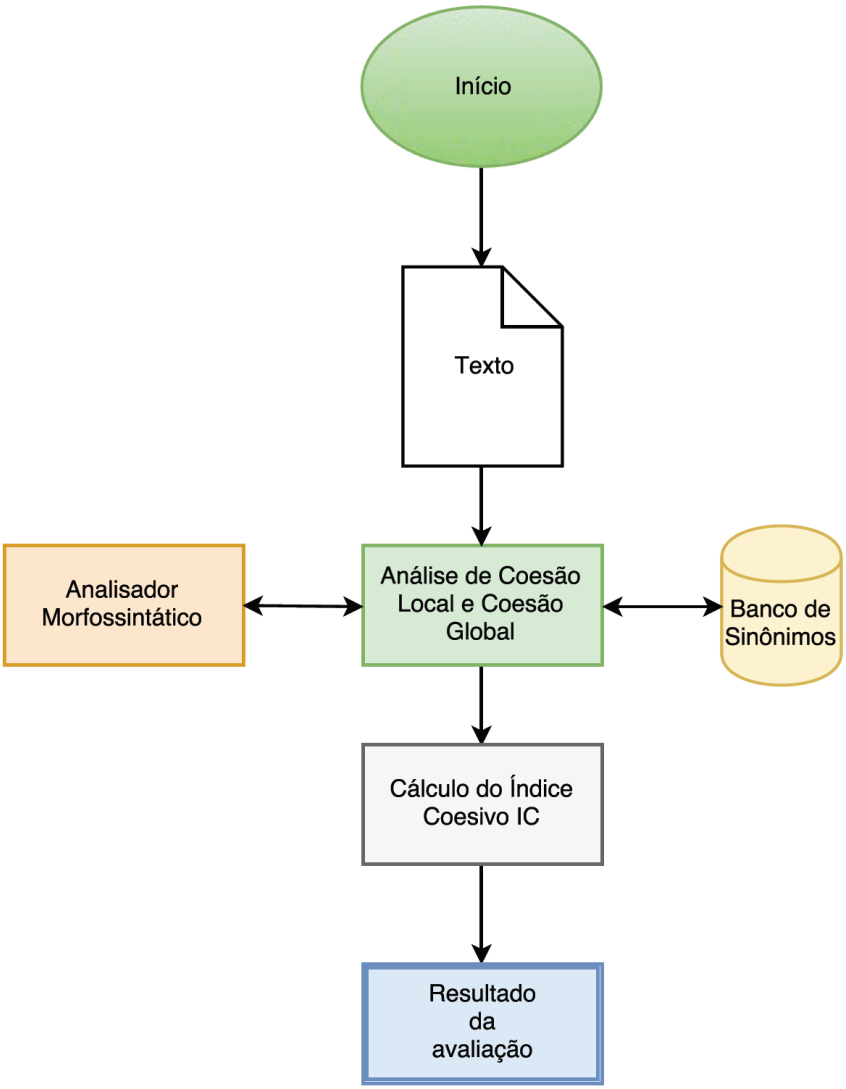
$$F_1 = \text{“}O\ Brasil\ passa\ por\ um\ momento\ dif\acute{\imath}cil\ devido\ aos \tag{2}$$

$$esc\hat{a}ndalos\ pol\acute{\imath}ticos.\text{”} \tag{3}$$

$$F_2 = \text{“}O\ pa\acute{\imath}s\ sente\ os\ efeitos\ disso\ na\ economia.\text{”} \tag{4}$$

Logo, o discurso pode ser representado por $D = \{F1, F2\}$. Para cada F_i de D, calculam-se os conjuntos FE_i e FI_i, ou seja, o conjunto de elementos do Foco Explícito e o conjunto de elementos do Foco Implícito. Realiza-se esse processo por meio do analisador *CORP* [Fonseca et al. 2017]. Para F_1, obtêm-se os seguintes conjuntos:

Figura 2. Componentes da ferramenta



Fonte: Elaborada pelo autor

$$FE_1 = \{\text{“}O\ Brasil\text{”}, \text{“}um\ momento\ dif\acute{\imath}cil\text{”}, \text{“}a\ os\ esc\hat{a}ndalos\ pol\acute{\imath}ticos\text{”}\} \quad (5)$$

$$FI_1 = \{\text{“}ORGANIZA\text{Ç}\tilde{A}O|LOCAL\text{”}, \text{“}OUTRO\text{”}, \text{“}OUTRO\text{”}\} \quad (6)$$

E para F_2, os conjuntos são os seguintes:

$$FE_2 = \{\text{“}O\ país\text{”}, \text{“}os\ efeitos\ de\ isso\text{”}, \text{“}isso\text{”}, \text{“}a\ economia\text{”}\} \quad (7)$$

$$FI_2 = \{\text{“}OUTRO\text{”}, \text{“}OUTRO\text{”}, \text{“}ORGANIZA\text{Ç}\tilde{A}O|LOCAL\text{”}, \quad (8)$$

$$\text{“}COMUNICA\text{Ç}\tilde{A}O|PREDIC\text{”}\} \quad (9)$$

Durante a definição dos conjuntos, buscam-se sinônimos do núcleo de cada elemento de foco explícito, utilizando como fonte a base de sinônimos ***Tep2*** [Maziero et al. 2008]. Por exemplo, como sinônimos de um elemento E, obtém-se o conjunto Syn_e. Com $E =$ “$O\ país$”, obtém-se $Syn_E = \{\text{“}na\tilde{c}\tilde{a}o\text{”}, \text{“}territ\acute{o}rio\text{”}\}$.

Com os conjuntos definidos, aplicam-se as teorias para avaliar o grau de coesão textual. Utiliza-se os Algoritmos 1 e 2 para a análise. A partir do Algoritmo 1 (Coesão Local), o primeiro passo é obter a intersecção, entre sentenças adjacentes, dos elementos de foco explícito $RFE_{1,2}$. O analisador ***CORP*** identifica que o sintagma “$O\ país$” é um hiperônimo do sintagma “$O\ Brasil$”, e assim obtém-se:

$$RFE_{1,2} = \{\text{“}O\ Brasil\text{”}, \text{“}O\ país\text{”}\} \quad (10)$$

Continua-se removendo os elementos de foco implícito dos conjuntos FI_1 e FI_2 equivalentes aos elementos pertencentes a $RFE_{1,2}$. Após isso, calcula-se a intersecção dos conjuntos FI_1 e FI_2 que resulta no conjunto $RFI_{1,2}$. Para a realização desse cálculo foi feita uma adaptação devido as capacidades do analisador ***CORP***.

Como a análise semântica do analisador em questão possui foco nas categorias de Pessoa, Organização e Local, as categorias desconhecidas são rotuladas com a etiqueta ”OUTRO”. Pelo fato de que a análise do foco implícito é realizada utilizando o rótulo semântico, essa análise se demonstrou imprecisa quando se compara elementos que possuem a etiqueta semântica ”OUTRO”. Nesse caso, o sistema realiza a comparação por meio dos sinônimos dos elementos, para obter maior precisão.

No exemplo proposto, a partir da intersecção dos de FI_1 e FI_2 obtém-se o conjunto $RFI_{1,2} = \{\}$. O conjunto é vazio pois nenhum par de elementos possui o mesmo rótulo semântico, e dentre os pares que possuem o rótulo ”OUTRO”, não foram encontrados sinônimos equivalentes. Após isso possui-se os conjuntos necessários para identificar a relação do estabelecimento de coesão, utilizando a Tabela 1, e para realizar o cálculo do Índice Coesivo (IC). Para realizar o cálculo do IC, aplica-se a Equação 1 e é feita a proporção para o resultado enquadrar-se nos valores da Tabela 2. A avaliação da coesão obtém nesse caso uma pontuação de $1,5$, e conclui-se que o discurso D utiliza, sem refinamento, os mecanismos coesivos para o desenvolvimento do texto.

Tabela 2. Relacionamento entre classe e nota atribuída por banca

Classes	Formas de utilização de elos coesivos	Nota
1	Utiliza recursos coesivos da língua que afetam a coerência global do texto.	0,0 - 0,4
2	Utiliza recursos coesivos da língua que não afetam a coerência global do texto.	0,5 - 0,8
3	Utiliza, ainda que com alguns problemas, recursos coesivos da língua, adequadamente.	0,9 - 1,2
4	Utiliza, sem demonstrar grande refinamento, os recursos da língua para o desenvolvimento do tipo de texto.	1,3 - 1,6
5	Utiliza com proficiência os recursos coesivos da língua para o desenvolvimento do tipo de texto.	1,7 - 2,0

Fonte: [Nobre and Pellegrino 2010]

Nesse exemplo não aplicamos o Algoritmo 2 (Coesão Global) pois o discurso D possui apenas um parágrafo. O Algoritmo 2 é semelhante ao Algoritmo 1, e possui como principal diferença a análise entre os parágrafos ser realizada utilizando adjacência máxima, ou seja, compara-se os elementos de foco explícitos, implícitos e seus sinônimos de um parágrafo com todos os outros parágrafos do discurso.

4.3. Experimentos

Para a realização do experimento, foram utilizadas trinta e cinco redações de alunos dos cursos de engenharia da Universidade de Caxias do Sul (UCS). Após as redações serem avaliadas pela ferramenta, os valores obtidos por meio do cálculo do IC foram comparados com a média da pontuação obtida pela banca avaliadora, que contava como membros especialistas graduados no curso de Licenciatura em Letras. Apenas as redações que apresentaram uma diferença inferior a $0,4$ pontos foram comparadas.

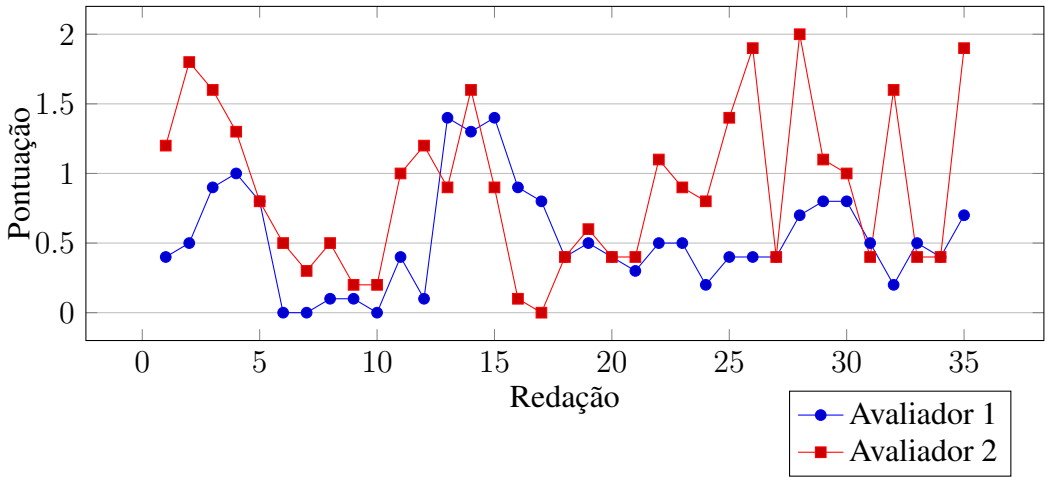
Com a finalidade de obter resultados padronizados, os peritos utilizaram a Tabela 2 como critério de avaliação. É importante destacar que a avaliação foi realizada apenas levando em consideração a coesão textual.

5. Análise dos Resultados

Ao analisar os resultados, é possível verificar que o *AAACT* obteve um desempenho satisfatório. A compatibilidade das notas atribuídas pelos especialistas é apresentada na Figura 3, totalizando um percentual de $51,42\%$. Ao remover as redações que obtiveram uma diferença de pontuação superior a $0,4$ pontos, os avaliadores apresentaram uma média de divergência de resultados de $0,16$ pontos, representada na Figura 4.

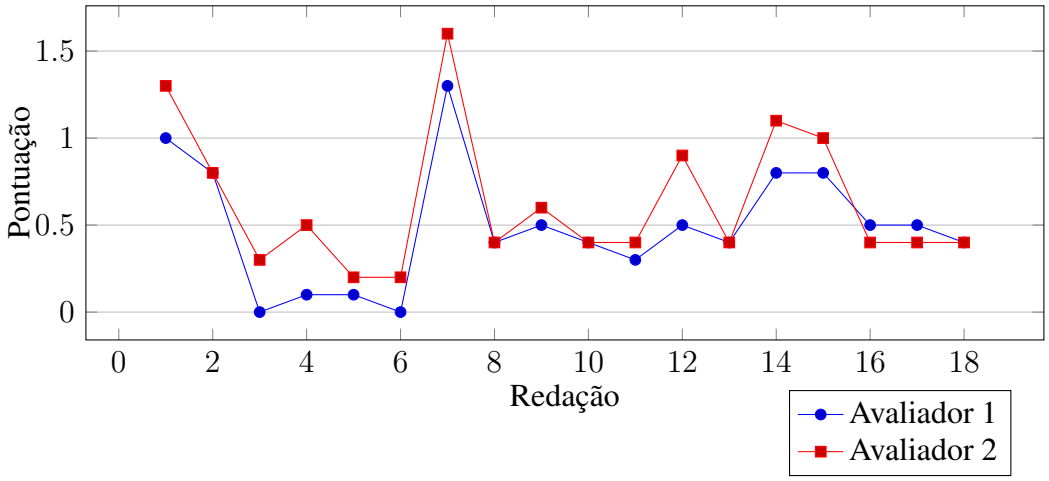
Aplicando a avaliação automática nas redações em que os peritos alcançaram pontuações compatíveis, ou seja, nos textos em que a discrepância entre as notas foi menor que $0,4$ pontos, os resultados atribuídos pelo sistema atingiram em média $0,46$ pontos de diferença, demonstrados na Figura 5. Dentro desse conjunto de redações, a ferramenta obteve em 50% das análises uma discrepância inferior a $0,4$ pontos e uma diferença média de $0,23$ pontos, representada na Figura 6.

Figura 3. Pontuação atribuída por avaliadores



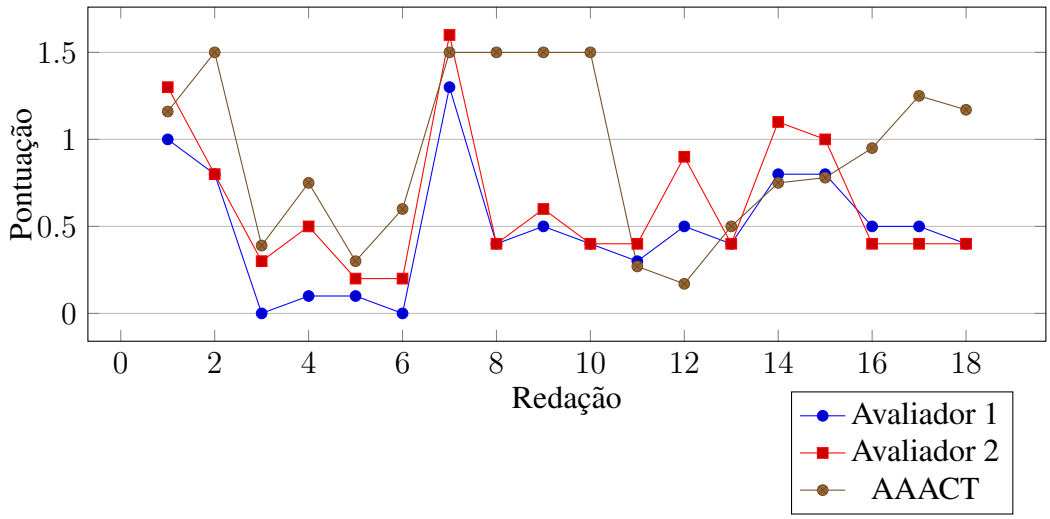
Fonte: Elaborada pelo autor

Figura 4. Resultados compatíveis entre avaliadores



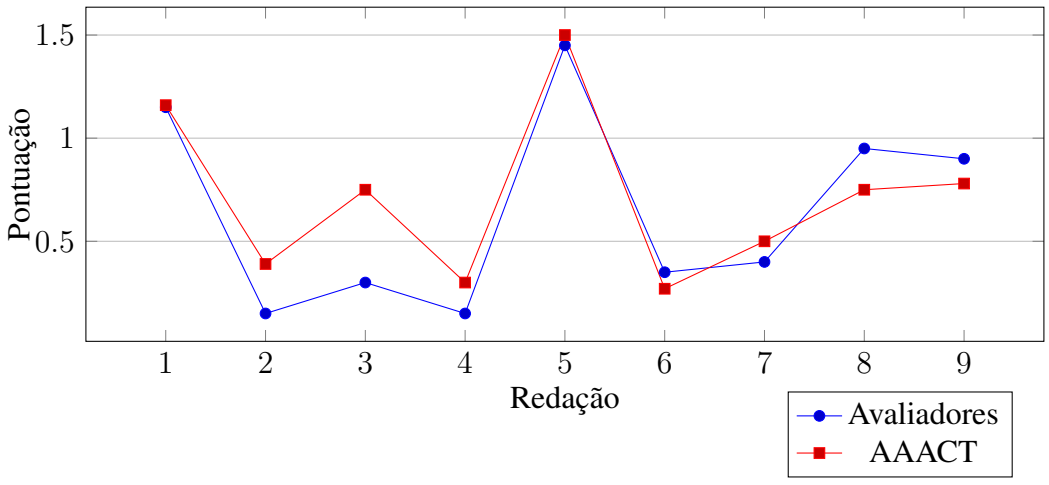
Fonte: Elaborada pelo autor

Figura 5. Resultados das avaliações que obtiveram compatibilidade entre os especialistas, e a pontuação do *AAACT* para essas avaliações



Fonte: Elaborada pelo autor

Figura 6. Média dos resultados de redações compatíveis entre avaliadores e o o resultado obtido pelo AAACT para os mesmos textos



Fonte: Elaborada pelo autor

6. Conclusão

O presente trabalho dedicou-se à implementação de um analisador e avaliador de coesão textual por meio de uma ferramenta *on-line*. Por ser de uso gratuito, apenas softwares de uso livre foram elencados como componentes. As principais teorias utilizadas baseiam-se no algoritmo apresentado por Nobre (2011) e foram obtidas por meio da *Teoria do Foco* [Sidner 1979, Sidner 1981] e a *Teoria da Centragem* [Grosz et al. 1983]. A metodologia para pontuação coesão é baseada no sistema *Avaliador Automático de Redação – AVAR* [Nobre 2011].

A avaliação automática obteve uma compatibilidade de $50,00\%$ com as notas dadas pelos avaliadores nas redações que apresentaram diferença de até $0,4$ pontos. Apesar de não estar no nível do estado da arte, é um resultado satisfatório, dado que o sistema realiza a avaliação apenas utilizando um dos cinco mecanismos coesivos. Ao se limitar à redações que aplicam predominantemente o mecanismo da coesão lexical, o *AAACT* obteve em média uma discrepância de $0,23$ pontos, que reforça a possibilidade de utilizar o método proposto, com algumas melhorias, como parte da realização de um avaliador de redações automático que é capaz de analisar textos de forma regular e imparcial. Conclui-se também que avaliações de redações irregulares realizadas por especialistas podem apresentar um alto índice de divergência, fato que motiva a criação de um avaliador automático para redações.

A pesquisa apresenta contribuições à área de AES para a língua português brasileiro, implementando o primeiro avaliador automático de coesão textual que utiliza apenas recursos gratuitos.

Para pesquisas futuras, tem-se como prioridade a integração de um analisador morfossintático para obter os referentes das anáforas que utilizam outros mecanismos coesivos, como o da referência e da substituição, assim complementando a entrada de dados para a aplicação dos métodos de análise de coesão textual e obtendo maior compatibilidade com avaliadores humanos. Também tem-se em vista a integração desse trabalho com outros relacionados a AES, para a realização de um sistema de avaliação de redações automático, utilizando apenas ferramentas gratuitas.

Referências

Burstein, J., Chodorow, M., and Leacock, C. (2003). Criterionsm online essay evaluation: An application for automated evaluation of student essays. In *IAAI*, pages 3–10.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summ-it: Um corpus anotado com informaç oes discursivas visandoa sumarizaç ao automática. *Proceedings of TIL*.

Crystal, D. (2011). *Dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.

Duarte, V. (2017). Tipos de sintagmas. Disponível em: <http://portugues.uol.com.br/gramatica/tipos-sintagmas.html/>. Acesso em: 21 jun. 2017.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Fonseca, E., Sesti, V., Antonitsch, A., Vanin, A., and Vieira, R. (2017). Corp: Uma abordagem baseada em regras e conhecimento semântico para a resoluçao de correferências. *Linguamática*, 9(1):3–18.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., and Carvalho, P. (2010). Second harem: Advancing the state of the art of named entity recognition in portuguese. In *LREC*.

Gei, A. (2015). Em 2014 mais de 200 mil redações do enem foram canceladas por fuga ao tema. Disponível em: <http://universitario.net/enem/em-2014-mais-de-200-mil-redacoes-do-enem-foram-canceladas-por-fuga-ao-tema/>. Acesso em: 21 jun. 2017.

Grosz, B. J. et al. (1977). The representation and use of focus in a system for understanding dialogs. In *IJCAI*, volume 67, page 76.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 44–50. Association for Computational Linguistics.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Gruber, J. (1976). *Lexical structures in syntax and semantics*, volume 25. North-Holland.

Hasan, R. and Halliday, M. A. (1976). *Cohesion in English*. Longman London.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Maziero, E. G., Pardo, T. A., Di Felippo, A., and Dias-da Silva, B. C. (2008). A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, pages 390–392. ACM.

McKee, A. et al. (2001). A beginner's guide to textual analysis. *Metro Magazine: Media & Education Magazine*, (127/128):138.

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1):1–43.

Nobre, J. C. S. (2011). *Modelo Computacional para Valoração e Avaliação de Redações Baseado em Lógica Fuzzy*. PhD thesis, Área de Informática - Instituto Tecnológico de Aeronáutica, São José dos Campos.

Nobre, J. C. S. and Pellegrino, S. R. M. (2010). Anac: um analisador automático de coesão textual em redação. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1.

Pardo, T. A. S., Nunes, M. d. G. V., and Rino, L. H. M. (2004). Dizer: An automatic discourse analyzer for brazilian portuguese. In *Brazilian Symposium on Artificial Intelligence*, pages 224–234. Springer.

Pereira, C. D. P. (2013). Dados sobre a correção da redação do enem 2012. Disponível em: <https://www.infoenem.com.br/dados-sobre-a-correcao-da-redacao-do-enem-2012/>. Acesso em: 21 jun. 2017.

Pereira, C. D. P. (2014). Dados da correção da redação do enem 2013. Disponível em: <https://www.infoenem.com.br/dados-da-correcao-da-redacao-do-enem-2013/>. Acesso em: 21 jun. 2017.

Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português. *Linguateca, Portugal*, 7(7):1.

Sidner, C. L. (1979). Towards a computational theory of definite anaphora comprehension in english discourse. Technical report, Massachusetts Inst of Tech Cambridge Artificial Intelligence lab.

Sidner, C. L. (1981). Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231.