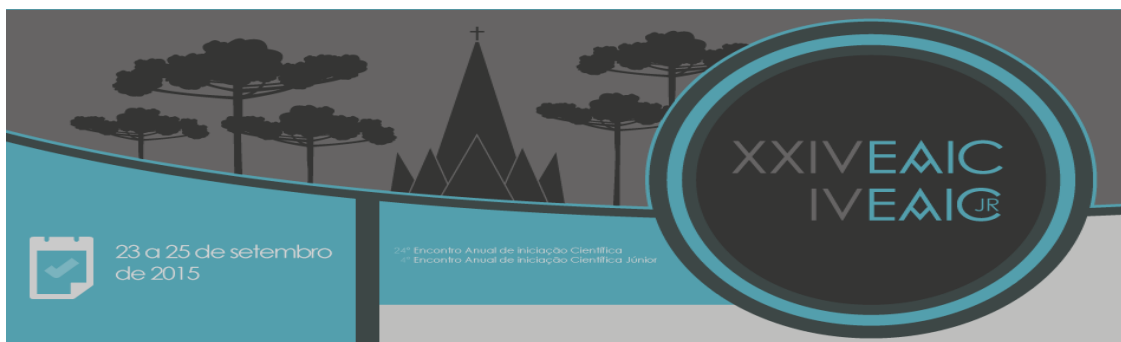




IMPLEMENTAÇÃO DE UM CORPUS ANOTADO DISCURSIVAMENTE PARA IDENTIFICAÇÃO DA UNIDADE CENTRAL DE TEXTOS ARGUMENTATIVOS

Kátia Roseane Cortez dos Santos (PIC/Uem), Juliano Desiderato Antonio
(Orientador), e-mail: katiacortez_@hotmail.com.

Universidade Estadual de Maringá / Centro de Ciências Humanas/Maringá

Área: Linguística, Letras e Artes. **Subárea:** Teoria e Análise Linguística

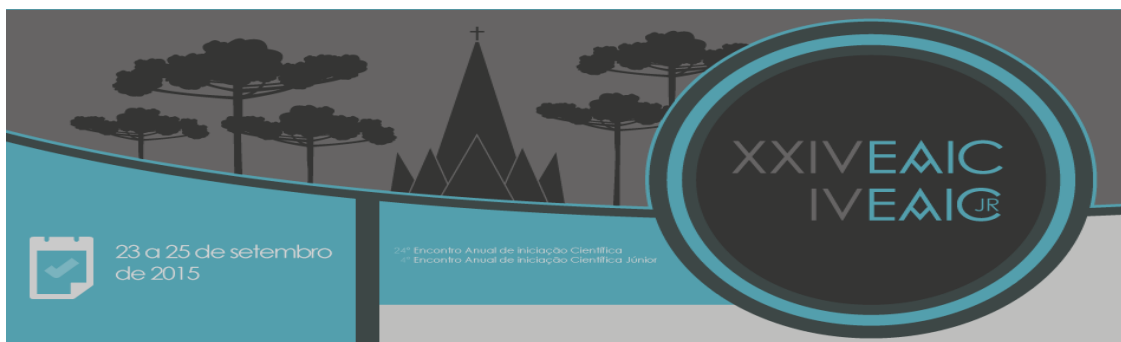
Palavras-chave: RST, Unidade central, Gênero resposta argumentativa.

Resumo: Este trabalho tem como objetivo investigar o acordo dos anotadores na detecção da unidade central em um corpus composto por cem textos do gênero resposta argumentativa. A teoria subjacente à investigação é a *Rhetorical Structure Theory* (RST), proposta por Mann e Thompson (1988), que tem como objetivo investigar a coerência do texto no que diz respeito às relações mantidas entre suas partes, tanto na macro quanto na microestrutura. A pesquisa revelou o padrão mais comum das unidades centrais: a unidade central é a declaração inicial do texto. As discrepâncias foram motivadas por erros dos anotadores ou por erros dos escritores. Os erros dos anotadores foram os seguintes: identificar a conclusão como a unidade central. Os erros dos escritores foram os seguintes: escolher um fator e mais à frente no texto escolher outro(s) fator(es) como o segredo do Vestibular e apresentar a mesma resposta duas vezes por paráfrase. Os resultados deste estudo podem ser úteis para o desenvolvimento de aplicações de avaliação automática.

Introdução

De acordo com Irukieta, Ilaraza e Lersundi (2015), detectar a unidade central do texto é um passo fundamental para anotar relações retóricas, uma vez que "um maior grau de concordância em relação à unidade central leva a um maior grau de concordância nas relações retóricas ligadas à unidade central".

Mann e Thompson (1988) afirmam que a nuclearidade é o principal organizador central da coerência do discurso. Assim, a detecção da unidade central foca a mais relevante e essencial parte da informação, o núcleo a que todas as outras porções de texto estão relacionadas.



Materiais e métodos

Os 100 textos do *corpus* foram escritos por candidatos para o vestibular de verão de 2013 da UEM, em resposta à pergunta:

Como vestibulando, redija, em até 15 linhas, uma resposta argumentativa à pergunta: “Qual o segredo do vestibular: inteligência, esforço ou sorte?”. Você pode basear-se nas informações dos textos de apoio, mas não deve apresentar cópia deles.

Após o período de avaliação, os pesquisadores solicitaram à UEM o acesso a um lote de cem textos. O pedido não mencionou quaisquer critérios para a seleção dos textos, porque o objetivo da pesquisa não foi apenas descrever um certo padrão para a unidade central, mas também investigar os motivos para discrepâncias entre os anotadores.

Os textos foram digitalizados e segmentado em unidades discursivas elementares manualmente. Quatro anotadores com formação básica em RST que participam na pesquisa foram convidados a indicar a unidade central.

Depois que o acordo entre os anotadores foi calculado, as unidades centrais identificadas pelos quatro anotadores sem discrepância foram investigadas a fim de encontrar um padrão de frequência de substantivos, advérbios, verbos e também a posição da unidade central no texto e se havia restabelecimento da pergunta no início da resposta.

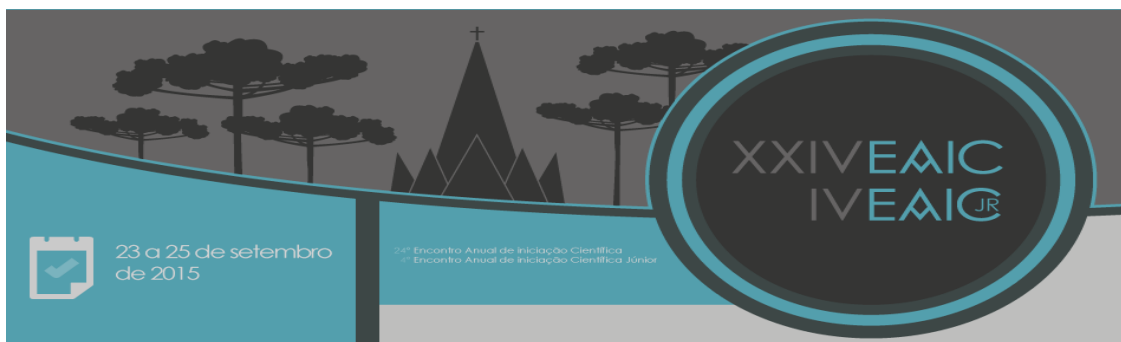
Para os textos em que houve discrepância a metodologia foi diferente. Um super anotador (Hovy, 2010) também identificou a unidade central de cada texto.

Resultados e Discussão

Os quatro anotadores identificaram a mesma unidade central em 74% dos textos (acordo total). Em 24% dos textos, duas unidades centrais diferentes foram identificados pelos anotadores. Em 2% dos textos, anotadores identificados três unidades centrais diferentes.

No caso de um acordo total, para o gênero resposta argumentativa, as características relevantes investigadas foram:

- a) a posição da unidade central: estima-se que os textos comecem com a resposta, que é a unidade central;
- b) retomada da pergunta: espera-se que a resposta retome a pergunta;
- c) frequência de substantivos, advérbios, verbos: espera-se que a alta frequência esteja associada a algum padrão de resposta.



O padrão mais frequente para as unidades centrais em que houve total concordância por todos os anotadores pode ser resumido da seguinte forma:

- (a) A unidade central é a declaração inicial do texto;
- (b) A resposta retoma a pergunta e a unidade central apresenta a estrutura "O segredo do Vestibular + cópula + fator(es)" ou uma estrutura que leva em conta mais de um fator: "O segredo do Vestibular + cópula + [junção de] fator(es)";
- (c) A declaração inicial pode ser incorporada a verbos evidenciais de atitude proposicional, como "pensar", "crer" e "acreditar": "[Eu + evidencial + que] o segredo do Vestibular + cópula + [junção de] fator(es)";
- (d) Advérbios também podem ser utilizados, especialmente os advérbios asseverativos epistêmicos "[Eu + verbo de evidência + que] o segredo da do Vestibular + cópula + [advérbio] [junção de] fator(es)".

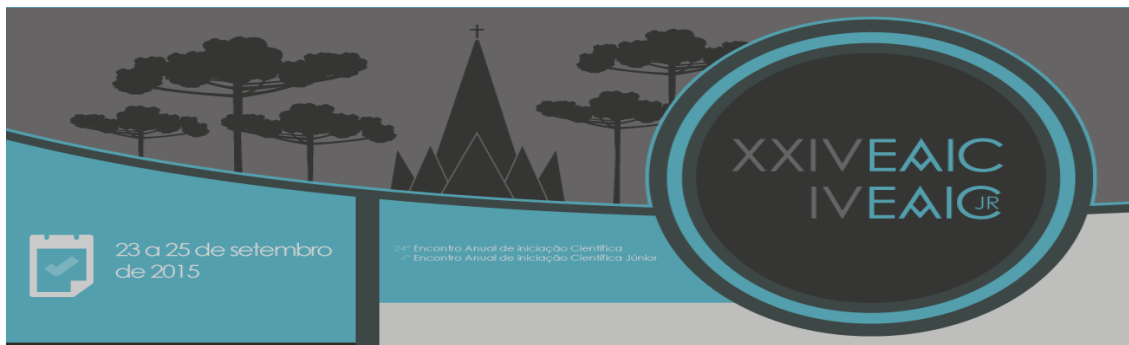
Em relação às discrepâncias, foram encontradas vinte e seis, sendo que doze podem ser consideradas erros dos anotadores, já que a unidade central pode ser facilmente identificada usando o padrão descrito anteriormente. Cinco dos doze erros aconteceram porque os anotadores identificaram a conclusão do texto como unidade central. Os restantes sete erros foram motivados por alguma palavra no texto que parece sinalizar a unidade central mais fortemente do que o padrão apresentado anteriormente.

Dois tipos de problemas na estrutura dos textos foram responsáveis pelas outras catorze discrepâncias: (i) duas respostas diferentes; (ii) a mesma resposta em duas posições diferentes no texto.

Conclusões

Os resultados deste estudo podem ser úteis para o desenvolvimento de aplicações de avaliação automática e também para a formação da comissão de professores de língua portuguesa que avaliam os exames vestibulares. No caso da avaliação automática, é necessária a implementação de aplicações para segmentação automática e detecção da unidade central. No caso da avaliação manual pelos professores, treinamentos devem ser fornecidos por duas razões: uma delas é distinguir respostas argumentativas bem escritas, em que a unidade central apresenta o padrão esperado e o restante do texto desenvolve a unidade central, de textos com problemas de estrutura e de coerência, que apresentam mais do que uma unidade central; outra razão é evitar erros como os cometidos pelos anotadores que tomaram parte na pesquisa.

Agradecimentos



Camila Christiane Moreschi (PIBIC/CNPq-FA-UEM), que contribuiu com esta pesquisa.

Referências

Hovy, E. 2010. Annotation: a tutorial. En **48th Annual Meeting of the Association for Computational Linguistics**. Uppsala.

Iruskieta, M., A. D. Ilaraza y M. Lersundi. 2015. The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations. En **Proceedings of COLING 2014: Technical Papers**, Dublin.

Mann, W. C. y S. A. Thompson. 1988. **Rhetorical Structure Theory: toward a functional theory of text organization**. Text, 8(3):243-281.