Criação e Anotação do *corpus* de resumos científicos de Ciências Sociais Aplicadas

Sabrina de Fátima Barbosa Taniwaki, Jackson Wilke da Cruz Souza

Instituto de Ciências Sociais Aplicadas – Universidade Federal de Alfenas (UNIFAL-MG)
Varginha-MG – Brasil

{sabrina.tanikaki, jackson.souza}@unifal-mg.edu.br

Abstract. *With the growing interest and need in the creation of (semi)automatic tools that help the literacy process in academic textual genres, we present in this work the corpus of scientific abstracts in applied social sciences. Our objective was to study, at the sentential level, the rhetorical structure of the abstracts in the areas of Public Administration, Accounting and Economics, based on the typology proposed in the literature, using the WebAnno corpus annotation tool. As a result, (i) the organization of a set of 200 texts, (ii) the preliminary study of the rhetorical structure of those areas and (iii) the production of an annotation manual with specific guidelines on the identification of the rhetorical structure of the abstracts scientific.*

Resumo. *Com o crescente interesse e necessidade na criação de ferramentas (semi)automáticas que auxiliem o processo de letramento em gêneros textuais acadêmicos, apresentamos neste trabalho o corpus de resumos científicos em ciências sociais aplicadas. Nosso objetivo foi estudar, a nível sentencial, a estrutura retórica dos resumos das áreas de Administração Pública, Contabilidade e Economia, com base na tipologia proposta na literatura, por meio da ferramenta de anotação de corpus WebAnno. Tivemos como resultado, (i) a organização de um conjunto de 200 textos, (ii) o estudo preliminar da estrutura retórica das referidas áreas e (iii) a produção de um manual de anotação com diretrizes específicas sobre a identificação da estrutura retórica dos resumos científicos.*

1. Introdução

Os Gêneros Textuais (doravante, GTs) são "textos materializados que encontramos em nossa vida diária e que apresentam características sociocomunicativas definidas por conteúdos, propriedades funcionais, estilo e composição característica" [Marcuschi 2002, p.4]. Segundo Marcuschi [2002], só conseguimos nos comunicar por GTs, que surgem emparelhados às nossas necessidades comunicativas. Nesse sentido, o ensino-aprendizagem dos gêneros estaria condicionado à experiência linguística com que o falante tem com cada um deles.

Essas concepções norteiam, há alguns anos, o ensino de GTs tanto no ensino básico, bem como no ensino superior. Por estarem conectados a situações comunicativas, os gêneros que ficam à disposição dos alunos variam de acordo com o nível de letramento de cada um deles. Vieira e Faraco [2019] classificam os gêneros em função da formalidade (informal, semiformal, formal e ultraformal) inerente ao contexto sociocomunicativo em que emergem, como os gêneros acadêmicos.

Partindo desse princípio, as dificuldades enfrentadas pelos alunos no processo de letramento de GTs, especificamente o acadêmico, são caracterizadas por conta de os alunos não terem participado anteriormente de situações comunicativas que exigissem

determinados gêneros. Isso não quer dizer que o aluno-aprendiz não domine a língua e suas regras de funcionalidade [Bakhtin e Volochinov 2006].

Para mitigar as dificuldades de letramento de GTs acadêmicos, os ambientes computacionais de auxílio à escrita disponibilizam ao usuário modelos de estruturas retóricas de GTs, assistindo-o na organização e produção textual, como o SciPo [Antiqueira *et al.* 2003]. Apesar de abordarem gêneros estabilizados e que sofrem pouca variação (como os resumos acadêmicos), o fato de esses ambientes terem sido construídos dependentes de domínio (Computação, Física e Farmácia, por exemplo), faz com que a aprendizagem do gênero possa ser insuficiente.

Visando contribuir em ampliar a disposição de subsídios linguísticos às áreas de Letramento acadêmico e de Processamento de Línguas Naturais (doravante, PLN), apresentamos neste trabalho a construção de um *corpus* de resumos científicos das áreas de Administração Pública, Contabilidade e Economia. Para tanto, os textos foram anotados semi automaticamente a nível sentencial de acordo com a tipologia da estrutura retórica proposta por Feltrim [2004] e Feltrim *et al.* [2004].

Este artigo está organizado em cinco seções, além desta Introdução. Na Seção 2, apresentamos trabalhos que se relacionam a esta proposta de pesquisa, os quais investigaram estruturas retóricas de GTs. Na Seção 3 apresentamos a metodologia deste trabalho, bem como a caracterização do *corpus* organizado. Na Seção 4 demonstramos os resultados e a discussão, além de considerações finais, na Seção 5.

2. Trabalhos relacionados

Os trabalhos mais recentes que investigaram a estrutura retórica (como Iriguti e Feltrim [2019] e Teufel e Moens [2002]), baseiam-se em métodos de aprendizado supervisionado, portanto, dependentes de conjuntos de textos dos quais se possam extrair informações e conhecimentos linguísticos.

Para o Português do Brasil (PB), destacamos o trabalho de Feltrim *et al.* [2004], em que foi proposto o *Argumentative Zoning for Portuguese* (AZPort). Após o estudo no CorpusDT, composto por resumos científicos da área de Ciências da Computação, os autores apresentaram o conjunto de seis macrocategorias que caracterizam a estrutura retórica de resumos científicos nessa área. Tais macrocategorias são *Contexto* (55,7% de frequência), *Lacuna* (42,3%), *Propósito* (100%), *Metodologia* (63,4%), *Resultado* (67,3%) e *Conclusão* (30,7%); cada uma dessas categorias subdividem-se em mais três microcategorias.

Feltrim *et al.* [2004] anotaram manualmente todas as sentenças no conjunto de 52 textos (que somam 366 sentenças), e concluíram que nem todas as categorias são necessárias para compor um resumo científico. Ademais, apesar de o modelo retórico prever uma ordem com que essas categorias ocorrem, ela deve ser compreendida como diretriz e não como regra; isso quer dizer, que a sequência de ocorrência não é fixa, corroborando o posicionamento de Vieira e Faraco [2019] quanto à plasticidade dos GTs.

Tomando o trabalho de Feltrim *et al.* [2004] como base para a descrição da organização retórica de resumos em PB, é necessário ressaltar a importância da tarefa específica de anotação de *corpus*. Pustejovsky e Stubbs [2012] apontam que um dos papeis do PLN é capturar propriedades das estruturas linguísticas que possam ser

aprendidas por sistemas computacionais. Para que os algoritmos aprendam de maneira eficiente e eficaz, a anotação e identificação de tais propriedades devem ser precisa e relevante.

3. Metodologia

Metodologicamente, este trabalho foi realizado segundo a síntese das etapas de anotação de *corpus* propostas por Hovy e Lavid [2010], a saber: (i) Elaboração do *corpus*, (ii) Criação do manual de anotação e (iii) Anotação e avaliação dos *subcorpora* (estudo e treinamento). Tais tarefas são apresentadas respectivamente nas subseções, a seguir.

3.1. Elaboração do *corpus*

Partindo do pressuposto teórico de Sardinha [2004], decidimos que o *corpus* deveria ter o seguinte *design*: *Modo-Escrito* (composto por textos escritos); *Tempo-Diacrônico* (composto por resumos publicados entre 2009 e 2019); *Seleção-Equilibrado* (composto por resumos de periódicos de Qualis CAPES elevada quando possível); *Conteúdo-Especializado* (resumos científicos/acadêmicos); *Autoria-De língua nativa* (autores falantes de português); Finalidade (*subcorpora* de estudo e de treinamento); *Tamanho-Médio-grande* (200 textos, 1.118 sentenças e 35.904 palavras); *Anotação-Corpus anotado* (nível sentencial).

As áreas do conhecimento que compõem o *corpus* são Contabilidade/Economia (73 textos), Administração Pública (97 textos) e Economia (30 textos). A coleta e armazenagem dos textos durou cerca de um mês.

3.2. Criação do manual de anotação

A partir da elaboração do *corpus*, desenvolvemos o manual de anotação[1], o qual contém as diretrizes que nortearam este trabalho. Para tanto, estudamos em 1/4 do *corpus* (a saber, 50 resumos) o *tagset* proposto por Feltrim *et al.* [2004]. Esse estudo inicial foi feito por três anotadores, após quatro meses de análise da proposta dos autores, resultando numa anotação preliminar.

Utilizamos as macro e microcategorias propostas pelos autores, pois, ao longo dos estudos iniciais, percebemos que poderíamos identificar características relativas às próprias (sub)áreas de Ciências Sociais Aplicadas, e essas em relação às Ciências da Computação. Assim, na versão final do manual apresentamos as etiquetas utilizadas em nosso trabalho, a definição de cada uma delas, exemplos de sentenças anotadas, e instruções técnicas para utilização do ambiente virtual de anotação.

3.2. Anotação e avaliação dos *subcorpora*

Nesta tarefa, anotamos o *subcorpus* de treinamento, que corresponde a 3/4 do conjunto completo, totalizando 150 resumos. A anotação durou quatro meses, com uma equipe de dois anotadores. Para tanto, utilizamos o ambiente virtual WebAnno [CASTILHO *et al.* 2016]. Trata-se de um ambiente que reúne um conjunto de ferramentas relativas à tarefa de anotação de *corpus*, como a possibilidade de escolha da granularidade da anotação (palavra, sintagma, sentença etc.) e a disponibilidade de ferramentas de estatística textual (para medir concordância, tokens e types etc.), por exemplo.

[1] Disponível em: https://github.com/jackcruzsouza/EstruturaRetorica

Além disso, outra vantagem oferecida pelo referido ambiente de anotação é o fato de ser colaborativo, fazendo com que haja diferentes papeis entre a equipe do projeto (como moderador, anotador, gerente etc.), sem que haja necessidade de instalar alguma ferramenta e/ou ter conhecimento prévio de programação. O formato de saída do arquivo é em XML, e é compatível com a ferramenta Brat [Stenetorp *et al.* 2012].

Quanto à avaliação, obtivemos 60% e 72% de concordância nos *subcorpora* de estudo e de treinamento, respectivamente, segundo a medida Kappa, calculada automaticamente no próprio ambiente WebAnno.

4. Resultados e discussão

A partir da observação da anotação do *corpus*, foi possível tecer algumas considerações acerca das categorias retóricas.

Com relação ao *Contexto*, os resumos introduzem mais brevemente o assunto abordado no trabalho, além de explicação de alguns termos técnicos para melhor entendimento do leitor. Sobre a categoria *Propósito*, os resumos da área de Ciências Sociais Aplicadas apresentam, na maioria das vezes, apenas um objetivo principal do trabalho, não apresentando objetivos específicos da pesquisa realizada. A *Lacuna* ocorreu mais extensivamente em textos da área de Contabilidade. Quanto à *Metodologia*, os resumos apresentaram-se bastante explicativos com relação aos métodos e material utilizados, contendo ainda, por vezes, a justificativa de uso pelo de autores/referencial teórico. Sobre os *Resultados*, observamos que ocorrem nos resumos utilizando o tipo textual de descrição. Por fim, a *Conclusão* ocorreu em dois textos, sendo constituídas de breves explicações ou apresentações da importância do trabalho à área de pesquisa.

Diante disso, é possível propor que a estrutura retórica genérica e preliminar dos resumos científicos da área de Ciências Sociais Aplicadas é composta por *Contexto*, *Propósito*, *Metodologia* e *Resultado*, tendo a *Lacuna* e a *Conclusão* como macrocategorias pouco exploradas na grande área. Destaca-se também que no decorrer do trabalho e das anotações realizadas não houve sentenças que não se enquadrassem na proposta do conjunto de etiquetas utilizado.

5. Considerações finais

O estudo da estrutura retórica dos resumos das (sub)áreas de Ciências Sociais Aplicadas demonstrou-nos que há particularidades que a diferencia e a aproxima do que está proposto na literatura. Os dados obtidos a partir da anotação de cada sentença ou trecho dos textos com a identificação sintática delas constituem uma importante base para o treinamento, desenvolvimento e/ou atualização de um ambiente de auxílio à escrita acadêmica.

Em trabalhos futuros, esperamos aprofundar as análises com relação à estrutura retórica dos resumos e, em especial, apresentarmos caracterizações linguísticas de cada uma das categorias (macro e micro). Ademais, com o objetivo de propiciar uma melhor experiência de letramento acadêmico com o ambiente de auxílio à escrita, objetivamos construir um *corpus* a partir de ingressantes no Ensino Superior para detectarmos suas maiores dificuldades em aprender o GT resumo científico/acadêmico. Assim, pretendermos ter um *corpus* paralelo, em que para possíveis desvios de ordem retórica e linguística tenhamos correspondentes *gold standart*.

Referências

Antiqueira, L., Feltrim, V. D., & Nunes, M. D. G. V. (2003*). Projeto e implementação do sistema SciPo*. São Carlos, Brasil. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação (nº 223).

Bakhtin, M., & Volochinov, V. N. (2006). Marxismo e filosofia da linguagem (Vol. 7). São Paulo: Hucitec.

Castilho, R.E., Mujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. Em *Proceedings of the workshop on language technology resources and tools for digital humanities* (LT4DH) (pp. 76-84).

Feltrim, V. D., Pelizzoni, J. M., Teufel, S., Nunes, M. D. G. V., & Aluísio, S. M. (2004). Applying argumentative zoning in an automatic critiquer of academic writing. Em *Brazilian Symposium on Artificial Intelligence* (pp. 214-223). Springer, Berlin, Heidelberg.

Feltrim, V.D. (2004). *Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes web de auxílio à escrita acadêmica em português*. Universidade de São Paulo, São Carlos, Brasil. Tese de Doutorado.

Hovy, E., & Lavid, J. (2010). Towards a 'science'of *corpus* annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1), 13-36.

Iriguti, A. H., & Feltrim, V. D. (2019). Avaliando atributos para a classificação de estrutura retórica em resumos científicos*. Linguamática, 11*(1), pp.41-53.

Marcuschi, L. A. (2002). Gêneros textuais: definição e funcionalidade. *Gêneros textuais e ensino*, 2, pp.19-36.

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media.

Sardinha, T. B. (2004). *Linguística de corpus*. Barueri/SP: Manole Ltda.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). BRAT: a web-based tool for NLP-assisted text annotation. Em *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107).

Teufel, S. & Marc, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics 28*(4). 409–445.

Vieira, F. E., & Faraco, C. A. (2019). *Escrever na universidade: fundamentos*. São Paulo: Parábola.