Relações de coerência do espanhol peninsular: Um estudo bibliográfico-documental da *Rhetorical Structure Theory*

Ewerson Dantas[1], Roana Rodrigues[2], Jackson Wilke da Cruz Souza[3]

[1]Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

[2]Programa de Pós-Graduação em Letras – Universidade Federal de Sergipe (UFS) – São Cristóvão/SE

[3]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA) – Salvador/BA

roana@academico.ufs.br; {ewersonndantad, jackcruzsouza}@gmail.com

Abstract. *This work aims to develop a bibliographical study [Gil,2002] of academic works on RST* (*Rhetorical Sistematic Theory*)*, an analytical discourse theory focused on the organization and analysis of texts in peninsular Spanish (ESP). After collecting texts from international journals and databases, we proposed an organization and classification of the works, based on how RST is used.*

Resumo. *Esse trabalho visa o desenvolvimento de um estudo bibliográfico [Gil 2002] de trabalhos acadêmicos em RST* (*Rhetorical Sistematic Theory*)*, teoria discursivo-analítica focada na organização e análise de textos, no espanhol peninsular (ESP). Após a coleta de textos, a partir de revistas internacionais e bases de dados, foi proposta uma organização e classificação dos trabalhos, tendo como base a forma como a RST é utilizada.*

1. Introdução

A RST (*Rhetorical Sistematic Theory*) é uma teoria discursivo-análitica desenvolvida, a príncipio, para análise de textos por Mann e Thompson (1987), sendo muito utilizada em pesquisas de Processamento de Língua Natural (PLN), no campo da Linguística Computacional. O principal ponto dessa teoria é o estudo da organização do texto, tendo em vista a coerência estabelecida nele, utilizando dos marcadores discursivos para esse estudo. Nesse sentido a RST é vista em diversos tipos de pesquisas e trabalhos dentro da PLN, para exemplificação, temos a criação do DiSeg [Da Cunha *et al*. 2010], um segmentador discursivo para língua espanhola, como também dos diversos trabalhos comparatistas entre espanhol e basco feitos por Da Cunha e Iruskieta.

Neste trabalho, o RST é o principal tema para o levantamento bibliográfico [Gil 2002] realizado com pesquisas dentro do espanhol peninsular (ESP), ou seja, aquele falado na Espanha. Após a coleta dos textos, a partir de revistas internacionais e ferramentas de bases de dados, foi feita uma organização e classificação destes a partir da observação da forma como a teoria é trabalhada.

A organização do artigo se dará da seguinte forma, além desta Introdução: na Seção 2 será abordada a metodologia utilizada para a pesquisa dos trabalhos em RST no espanhol peninsular; na Seção 3 os resultados obtidos serão relatados com foco na

exposição da classificação feita a partir do levantamento bibliográfico; na Seção 4 constam as considerações finais do trabalho.

2. Metodologia

O primeiro passo para o levantamento bibliográfico foi a procura por trabalhos de pesquisa que incluíssem a RST dentro do espanhol peninsular, para isso foram utilizadas duas plataformas de pesquisa: o Google Acadêmico e o *ACL Anthology*. Os termos de buscas utilizados foram "RST", "spanish" e "español". Como resultado foram encontrados 2330 resultados no Google acadêmico, e 601 resultados no *ACL Anthology*, totalizando 2931 trabalhos. A partir dos resultados encontrados houve uma separação entre aqueles que realmente se enquadram como objeto de estudo deste trabalho, seguindo os seguintes critérios; (i) a forma como a RST era citada no resumo e nas palavras-base do trabalho, (ii) se os criadores da teoria, Mann e Thompson, estavam citados na referência, (iii) se o trabalho realmente focava na língua espanhola, utilizando dos resumos e introduções dos textos como base. Ao final desta separação, 30 trabalhos se enquadraram nos requisitos citados.

Após essa separação, uma segunda fase foi feita focando em coletar somente textos que trabalhassem com o espanhol peninsular, para isso duas coisas foram observadas: (i) a instituição a qual o trabalho está ligado, e (ii) os pesquisadores responsáveis pela sua execução. Assim, ao final, somente textos que trabalham com a RST dentro da área da PLN, e tem como objeto de estudo o espanhol peninsular, permaneceram para análise.

3. Resultados e discussão

Partindo da pesquisa realizada, chegou-se a sete trabalhos que cumpriam todos os requisitos apresentados. Em seguida, foi proposta uma classificação em função da maneira como a RST é abordada nos trabalhos. Chegou-se, então, em duas categorias, a saber: "Estudos comparativos", que são pesquisas que compararam duas ou mais línguas, ou ainda gêneros textuais, observando estruturas e relações de coerência mais recorrentes; e "Estudos em interface com PLN", que são trabalhos mono e multilíngue centrados na criação de ferramentas e recursos em Linguística computacional.

3.1. Estudos comparativos

a) Trabalho de Da Cunha e Iruskieta (2010)

Neste estudo os autores utilizaram a RST para uma análise comparativa entre o Espanhol e o Basco, tendo como foco as ocorrências de relações retóricas, como também seus respectivos marcadores. Para tanto, realizaram a análise das relações RST e seus marcadores a partir de 20 resumos retirados do periódico científico *Gaceta Médica de Bilbao*, que tem como domínio a medicina. Após a coleta dos textos, houve mais duas fases metodológicas: a fase *quantitativa*, feita a partir da contabilidade dos aspectos discursivos do *corpus*; e a fase *qualitativa*, focada na análise da ambiguidade dos marcadores e na forma como eles refletem cada relação. Os autores observaram um número maior de marcadores discursivos no Basco, além de notarem similaridade na quantidade de relações retóricas entre as duas línguas analisadas.

b) Trabalho de Iruskieta e Da Cunha (2010)

Neste trabalho os autores usam a RST em um estudo comparatista entre domínios distintos (Medicina e a Terminologia) de textos produzidos em Espanhol e Basco, e publicados na *Gaceta Médica de Bilbao* e no Congresso Internacional de Terminologia de 1997. O objetivo dos autores foi utilizar as relações retóricas como um meio para caracterizar os domínios. A análise dos textos foi feita a partir do processo de anotação do *corpus*, seguido pela análise discursiva, realizada para delimitar as diferenças entre as anotações entre os dois idiomas. Por fim, foi feita a análise quantitativa, focada na quantidade de relações ocorridas em cada área. Os autores notaram pontos interessantes tanto na comparação *entre áreas* (como o maior aparecimento da relação *Result* no *corpus* de Medicina, e de *Interpretation*, no *corpus* de Terminologia) e *entre as línguas* (diferenças na pontuação e sintáticas, principalmente pelo Basco ser uma língua aglutinante, onde a maioria dos morfemas está junto a palavra).

c) Trabalho de Iruskieta, Da Cunha e Taboada (2014)

Neste trabalho os autores fizeram uso de três diferentes idiomas (Inglês, Espanhol e Basco), buscando comparar a estrutura retórica das línguas a partir da anotação e análise de *corpus* dos três autores/anotadores. O *corpus* trabalhado continha textos da Conferência Internacional de Terminologia, realizada no ano de 1997. Os autores analisaram 15 resumos que continham as três línguas. A anotação semi-automática foi feita a partir da ferramenta RSTTool [O'Donnell 2000]. Ao final, os autores chegaram à conclusão de que o par Inglês-Espanhol continha o maior grau de concordância entre os anotadores, seguido pelo Espanhol-Basco, e por último Inglês-Basco, sendo estas últimas as línguas com menor concordância entre si porque estão tipologicamente mais distantes, além de não ter contato próximo como o par Espanhol-Basco.

3.2. Estudos em interface com PLN

a) Trabalho Da Cunha et al. (2010)

Os autores propuseram um segmentador automático de unidades mínimas de análise em RST, tidas como *Elementary Discourse Units* (EDU). Tais unidades podem ser frases ou orações, a partir das quais se constroem as árvores discursivas [Tofiloski, Brook e Taboada 2009]. Como resultado, desenvolveram o DiSeg, testando-o em um *corpus gold standart*, composto por artigos retirados da *Gaceta Médica de Bilbao*. Após testes e análises realizados, os autores destacam que a ferramenta DiSeg teve um bom desempenho ao segmentar EDU, especialmente quando comparado a outros segmentadores encontrados fora da língua espanhola. Por fim, os autores apontaram que os erros de segmentação se concentraram diante da partícula "*y*" quando aparecia antes da partícula "*que*".

b) Trabalho de Da Cunha, Torres-Moreno e Sierra (2011)

Nesta pesquisa o objetivo é desenvolver um *corpus* anotado em RST para o Espanhol (*RST Spanish Treebank*) visando ao desenvolvimento de um analisador discursivo. Para a formação do *corpus*, os autores objetivaram a diversidade de gêneros textuais (como artigos científicos, teses de doutorado e livros didáticos, por exemplo) e domínios (como Astrofísica, Engenharia, Economia e Direito, por exemplo). A anotação RST foi feita com o RSTTool, por uma equipe de 10 anotadores com diferentes graus de formação

acadêmica. Os autores evidenciaram que o *RST Spanish Treebank* é um recurso gratuito, desenvolvido para melhoria dos estudos da teoria, contando inclusive com uma interface *online* [Da Cunha *et al.* 2011]. Ao final, os autores também ressaltaram a necessidade de melhorias no processo de anotação e no *corpus*, como a necessidade de se incluir mais anotadores, a utilização de diferentes medidas de concordância, além do aumento do número de texto.

c) Trabalho de Cao, Da Cunha e Iruskieta (2018)

Os autores objetivaram a criação do primeiro *corpus* paralelo entre o Espanhol e o Chinês anotado em RST, o *Spanish-Chinese Treebank*. Os textos que compunham o *corpus* são de diferentes gêneros textuais (como resumos de trabalhos acadêmicos, notícias e anúncios), com diferentes quantidades de palavras. A ferramenta utilizada para anotação dos textos foi a RSTTool, e o grupo de anotadores era bilíngue ou falantes nativos de somente uma das línguas. Ao fim, o *Spanish-Chinese Treebank* obteve uma anotação com alta concordância, podendo ser um recurso a ser utilizado no desenvolvimento de ferramentas (como tradutores automáticos) ou abordagens (como Aprendizado de Máquina) em PLN. No âmbito do Espanhol e Chinês.

4. Considerações finais

Neste trabalho foi desenvolvido um levantamento bibliográfico e documental acerca de pesquisas acadêmicas que abordassem a teoria RST tendo como escopo o Espanhol peninsular. A partir do levantamento realizado foram encontrados seis trabalhos que se enquadraram em todos os requisitos impostos nesta pesquisa, sendo a pesquisadora Iria Da Cunha colaboradora em todos eles.

As análises feitas a partir destes textos mostraram um interesse na evolução da teoria RST no Espanhol, destacado pelos trabalhos com foco no desenvolvimento de ferramentas e recursos em PLN para o idioma. Demais estudos demonstram uma tendência aos estudos comparatistas em conjunto com o ESP, tanto com línguas também faladas na Espanha (Basco), como com línguas mais distantes geograficamente (Inglês e Chinês).

Tais observações corroboram os apontamentos feitos por Rodrigues, Souza e Cardoso (2023). Os autores destacaram a importância dos estudos centrados em segmentação, anotação e análise comparativa entre línguas, reforçando inclusive a ideia entre estudos descritivos e comparativos utilizando o Português e o Espanhol, em suas mais diversas variações, aspecto que é fortalecido pela dinâmica político-linguística na América do Sul.

Importante ressaltar que é possível encontrar outros resultados a partir de diferentes palavras chaves utilizadas no momento das pesquisas e buscas por textos. Sendo assim, trabalhos futuros poderão utilizar dessa lacuna para novas revisitações à bibliografia da RST no Espanhol peninsular.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este

projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento e suporte.

5. Referências

Cao, Shuyuan, Da Cunha, Iria, e Iruskieta, Mikel. (2018) The rst spanish-chinese treebank. In: *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions*. New Mexico/USA: Association for Computational Linguistics. p. 156-166. Disponível em: https://aclanthology.org/W18-49

Da Cunha, Iria *et a*l. (2011) The RST Spanish treebank on-line interface. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar/Bulgaria: Association for Computational Linguistics. p. 698-703. Disponível em: https://aclanthology.org/R11-1000

Da Cunha, Iria *et a*l. (2012) DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, v. 39, n. 2, p. 1671-1678. Disponível em: https://hal.science/hal-01314824

Da Cunha, Iria, e Torres-Moreno, Juan-Manuel; e Sierra, Gerardo. (2011) On the development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop*. Oregon/USA: Association for Computational Linguistics. p. 1-10. Disponível em: https://aclanthology.org/W11-0400

Gil, Antônio Carlos. (2002) *Como elaborar projetos de pesquisa*. Editora Atlas SA.

Iruskieta, Mikel; e Da Cunha, Iria (2010) *Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera*. Vigo: Universidade de Vigo. p. 146-159. Disponível em: http://ixa.si.ehu.es/sites/default/files/dokumentuak/3965/AESLA_marcadores.pdf

Iruskieta, Mikel; e Da Cunha, Iria. (2010) El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español. *Calidoscópio*, v. 8, n. 3, p. 181-202. Disponível em: https://www.redalyc.org/articulo.oa?id=571561875003

Iruskieta, Mikel; e Da Cunha, Iria; e Taboada, Maite. (2015) A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual *corpora*. *Language resources and evaluation*, v. 49, p. 263-309. DOI: https://doi.org/10.1007/s10579-014-9271-6

Mann, William C., e Thompson, Sandra (1987) A. *Rhetorical structure theory: Description and construction of text structures*. In: Natural language generation: New results in artificial intelligence, psychology and linguistics. Dordrecht: Springer Netherlands. p. 85-95.

O'Donnell, Michael. (2000) RSTTOOL 2.4-A markup tool for rhetorical structure theory. In: *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*. Mitzpe Ramon/Israel: Association for Computational Linguistics. p. 253-256. Disponível em: https://aclanthology.org/W00-14

Rodrigues, R., Souza, J., e Cardoso, P. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC. p. 249-257. DOI: https://doi.org/doi:10.5753/stil.2023.234120

Tofiloski, Milan; e Brooke, Julian; e Taboada, Maite. (2009) A syntactic and lexical-based discourse segmenter. In: *Proceedings of the ACL-IJCNLP 2009* conference short papers. Suntec/Singapure: Association for Computational Linguistics. p. 77-80. Disponível em: https://aclanthology.org/P09-2020