COMUNICAÇÃO E TECNOLOGIA: DELINEAMENTO ESTATÍSTICO DE UMA ANÁLISE SEMÂNTICA AUTOMÁTICA DE TEXTOS CIENTÍFICOS

Gabriel Evaristo Santana da Silva[1]

Rosana Ferrareto Lourenço Rodrigues[2]

Gustavo Aurelio Prieto[3]

RESUMO

A comunicação científica ocorre entre os processos de pesquisa e comunicação, a partir dos quais cognição e língua estão conectadas para comunicar a narrativa da pesquisa, um processo complexo que pode ser gerenciado a partir da semântica de frames. Este trabalho teve por objetivo analisar como os frames semânticos, detectados em resumos científicos, estão associados com a narrativa da pesquisa, utilizando-se um modelo de delineamento estatístico. Dessa forma, foi proposto um esforço entre a linguística aplicada, a informática e a estatística, com foco na semântica de frames. As atividades do projeto são pautadas na análise dos dados dos frames semânticos em um *corpus* de 1.800 resumos científicos, buscando-se identificar os padrões presentes no *corpus* e as associações necessárias, a partir de uma análise baseada em tabelas e gráficos, propiciando mais clareza no entendimento dos dados e dos resultados provenientes da análise do *corpus*. Concluiu-se, portanto, que o projeto cumpriu seu propósito, pois entregou como resultado uma análise estatística e semântica do *corpus*, permitindo reconhecer o viés da área de estudo e servindo como orientação para o ensino de redação científica no processo de educação de cientistas.

Palavras-chave: Estatística; Semântica; Frames; Resumos Científicos; Comunicação Científica.

COMMUNICATION AND TECHNOLOGY: STATISTICAL DESIGN OF AN AUTOMATIC SEMANTIC ANALYSIS OF SCIENTIFIC TEXTS

ABSTRACT

Scientific communication occurs between the research and communication processes, from which cognition and language are connected to communicate the research, a complex process that can be managed from the semantic frames. This project aims to analyze how the semantic frames, detected in scientific abstracts, are associated with the research narrative, using a statistical design model. Thus, an effort between applied linguistics, information technology and statistics is proposed, with a focus on frame semantics. The project activities are based on the analysis of data from the semantic frames in a corpus of 1800 scientific abstracts, seeking to identify the patterns present in the corpus and the necessary associations, based on an analysis upon tables and graphs, providing more clarity in the understanding of the data and results from the corpus analysis. It is concluded, therefore, that the project fulfilled its purpose, as it delivered as a result a statistical and semantic analysis of the corpus, allowing for the recognition of the bias of the study area and serving as a guideline for teaching scientific writing in the process of educating scientists.

Keywords: Statistic; Semantics; Frames; Abstracts; Scientific Communication.

[1] evaristo-gabriel@hotmail.com
[2] rosanaferrareto@ifsp.edu.br
[3] gaprieto@ifsp.edu.br

INTRODUÇÃO

A comunicação científica ocorre entre os processos de pesquisa e comunicação, a partir dos quais cognição e língua estão conectadas para comunicar a narrativa da pesquisa. O papel da ciência na relação entre a comunicação e a tecnologia é crucial para transformar informação em conhecimento. Nas discussões sobre a mediação da comunicação pela tecnologia, seja na esfera social, corporativa ou educacional, desde o advento de seu uso, os dispositivos tecnológicos são ferramentas de extensão da comunicação humana, porque viabilizam a ubiquidade das interações e da troca de informações. Na comunicação, o dado significado torna-se informação, que, sistematizada cientificamente, torna-se conhecimento. A tecnologia nesse processo é ferramenta – é meio e não fim em si mesma. A Semântica de Frames, nesse contexto, é um modelo do significado linguístico, no escopo da abordagem teórica linguístico-cognitivista, aplicada para comprimir o todo de qualquer processo complexo, como o científico e o da escrita, em partes gerenciáveis (RODRIGUES, 2020)

Dessa forma, nos estudos linguísticos, a Ciência de Dados tem propiciado a visualização de padrões em textos, a partir dos quais encontram-se as evidências necessárias para extrair conclusões que direcionam o uso da língua de maneira mais eficaz na comunicação. Os modelos científicos em cada área do conhecimento contribuem para o delineamento dessas generalizações extraídas dos dados. Nesse processo de investigação, as ferramentas computacionais capacitam o cientista da linguagem a olhar os textos e enxergar elementos concretos da criatividade humana, da invenção e da inovação no processo de comunicação.

Os padrões que são investigados são os narrativos e os figurativos dos textos científicos, por meio da modelagem estatística de dados, obtida a partir anotação automática de frames, constructos semânticos evocados pelas palavras dos textos de língua natural (RODRIGUES, 2019).

A semântica de frames é uma área de estudo da Linguística Cognitiva que postula que “o significado das línguas naturais é relativizado a cenas” (FILLMORE, 1982), chamadas de frames. Ao descrever sistematicamente o significado das línguas naturais, essa abordagem entende o frame como um constructo cognitivo indexado por palavras associadas a ele, composto de componentes provindos da cultura, da experiência e da imaginação do falante. “A ciência é a mãe da tecnologia” é um fato frequentemente comunicado nos ambientes em que se produz ciência com o uso de

tecnologia. Essa construção linguística – O X é o Y de Z – em que X = ciência; Y = mãe e Z = tecnologia é uma analogia estudada em detalhes por Turner (2008). A ciência está para a mãe como a tecnologia está para a filha (o elemento W = filha é inferido a partir do valor identificado em mãe). Esse tipo de construção linguística tem potencial didático persuasivo à medida em que ancora a nova informação (tecnologia) em conhecimento prévio (ciência). Resumindo, organizadas em uma só sentença, "empacota-se" elementos de cenas/domínios distintos para propor o novo através do lógico. Essas cenas, na linguística cognitiva, são denominadas frames semânticos. Em geral, na língua escrita e falada, estruturas pertencentes a várias classes gramaticais são capazes de evocar frames. Essas estruturas são chamadas de predicadores. Segundo Abreu (2018), os predicadores são itens lexicais que têm uma estrutura argumental em torno da qual se constrói orações, que são projeções dessa estrutura. Eles podem ser verbos, substantivos, preposições, determinantes, etc.

Dessa forma, para analisar como os frames semânticos estão associados com a narrativa da pesquisa científica, utilizou-se a ferramenta anotadora automática, *Semafor*. Para fazer a anotação automática, o *Semafor* utiliza os dados provenientes da *FrameNet*. A *FrameNet* é um projeto de lexicografia computacional que extrai informações sobre as propriedades semânticas e sintáticas de grandes *corpora* de texto eletrônico. Utiliza procedimentos manuais e automáticos, apresentando as informações com variedade de relatórios. O nome '*FrameNet'* é utilizado por basear-se na teoria da Semântica de Frames, se preocupando em analisar as redes de significado das quais as palavras participam. (FILLMORE; JOHNSON; PETRUCK, 2003, p. 1)

OBJETIVOS

O objetivo geral do projeto foi analisar como os frames semânticos detectados em resumos científicos estão associados com a narrativa da pesquisa em dada área do conhecimento e disciplina e identificar os fatores envolvidos nessa associação, a partir de um modelo de delineamento estatístico[4].

O objetivo específico deste trabalho foi caracterizar as variáveis independentes (áreas do conhecimento e disciplinas) e as variáveis dependentes (frames, unidades lexicais e elementos de frames) a partir do corpus de 1.800 *abstracts*, escolhendo o

[4] Este projeto faz parte de um programa de pesquisa em Comunicação Científica do Grupo de Pesquisas do IFSP-SBV – o aCOMTECe <https://acomtece.sbv.ifsp.edu.br>

modelo de delineamento estatístico com base na literatura de Volpato e Barreto (2016). Foram realizados testes de hipótese de comparação e verificou-se se as variáveis se afetam ou não e como, além de detectar se a abordagem estatística requer um teste paramétrico ou não paramétrico para a compreensão da distribuição de dados no corpus.

MATERIAL E MÉTODOS

Foi utilizado um *corpus* de 1.800 resumos científicos coletados da *Web of Science*. Para realizar a coleta do corpus, definiu-se que seriam coletados abstracts de 18 disciplinas (*disciplines*), sendo 6 de cada grande área do conhecimento (*fields*). No Quadro 1, estão indicadas as disciplinas separadas por área do conhecimento.

Quadro 1. Disciplinas separadas por área do conhecimento.

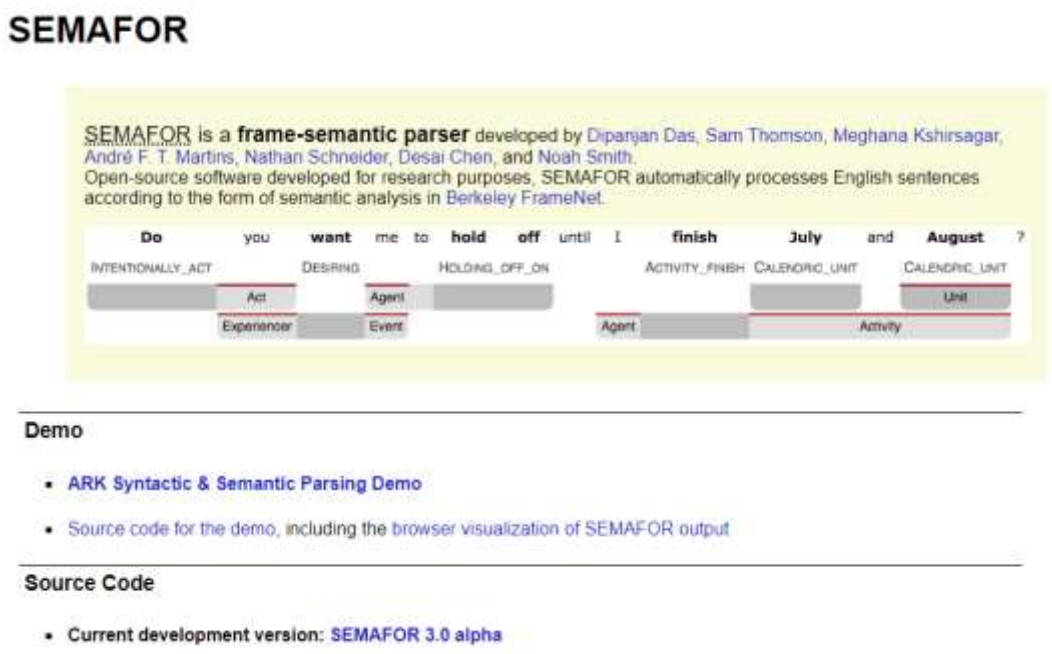
Área (*field*)	**Disciplina (*discipline*)**
CIÊNCIAS BIOLÓGICAS	BIOLOGIA
	CIÊNCIAS DA SAÚDE
	MEDICINA, PESQUISA EXPERIMENTAL
	NEUROCIÊNCIAS
	DIETÉTICA NUTRICIONAL
	PSICOLOGIA MULTIDISCIPLINAR
CIÊNCIAS EXATAS	AGRONOMIA
	QUÍMICA
	CIÊNCIA DA COMPUTAÇÃO
	ENGENHARIA
	FÍSICA
	TELECOMUNICAÇÕES
CIÊNCIAS HUMANAS E SOCIAIS	*BUSINESS*
	LÍNGUISTICA COGNITIVA
	COMUNICAÇÃO
	EDUCAÇÃO, DISCIPLINAS CIENTÍFICAS
	CIÊNCIA DA INFORMAÇÃO E BIBLIOTECA

	LINGUÍSTICA

Fonte: Autores.

Com o corpus adequadamente selecionado, utilizou-se a ferramenta anotadora automática, *Semafor* (Figura 1), para produzir uma saída estatística, com listas de frequências, do processamento do *corpus*.

Figura 1. Interface do Semafor.



Fonte: http://www.cs.cmu.edu/~ark/SEMAFOR/

A partir das listas de frequência geradas automaticamente pela ferramenta, utilizou-se modelos de delineamento estatístico aplicados à pesquisa descritiva com técnicas de associação entre variáveis (VOLPATO; BARRETO, 2016; VOLPATO, 2019).

RESULTADOS E DISCUSSÃO

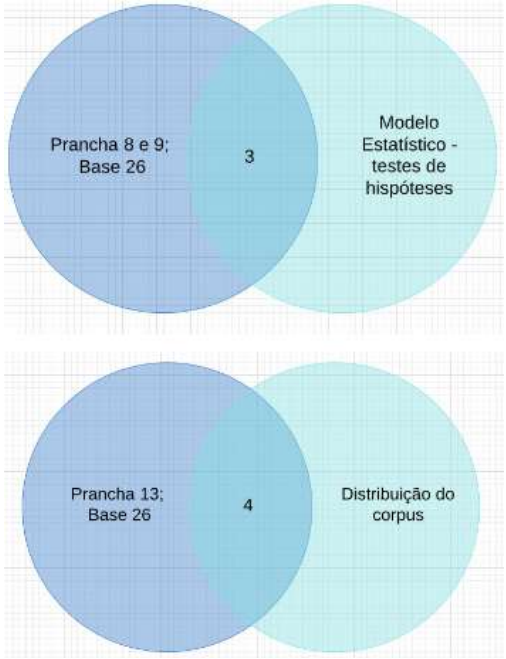
Gerou-se, para basear o estudo estatístico, um modelo de triangulação visual a partir de Rodrigues (2020a); Rodrigues (2020b); e Volpato e Barreto (2016). Esse modelo que integra os dois artigos e o livro, que podem ser observados nas Figuras 2 e 3, onde a esfera à esquerda indica as bases e pranchas do livro que podem ser utilizadas, a esfera à direita indica quais análises e resultados serão obtidos e a intersecção indica a etapa do trabalho. Desta forma, a primeira etapa corresponde ao início da análise estatística, onde será feita uma análise qualitativa do corpus, utilizando-se de gráficos e

listas de frequências para facilitar a visualização. A segunda etapa refere-se à classificação do corpus a partir de uma análise comparativa. Na terceira etapa tem-se os testes de hipóteses a partir de um modelo estatístico. A quarta e última etapa está em concordância com o final do estudo estatístico e com a classificação da distribuição dos dados do corpus.

Figura 2 – Triangulação para as etapas 1 e 2

Fonte: Autores.

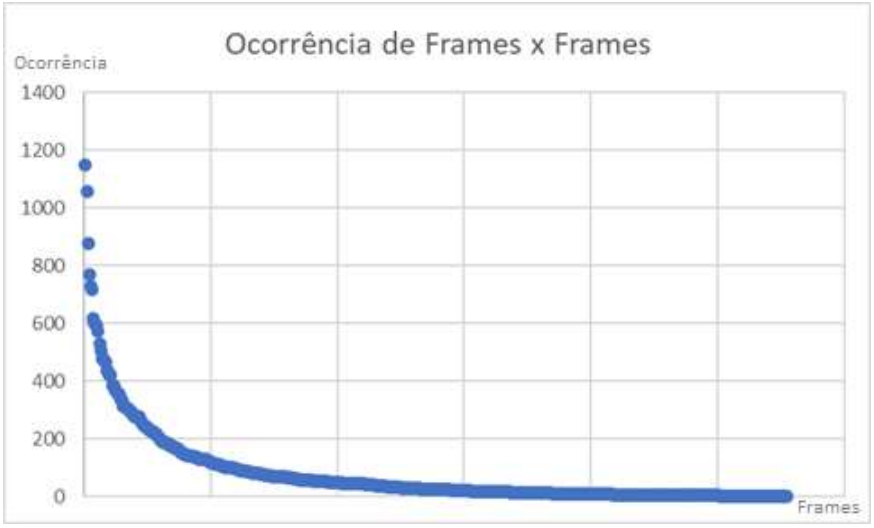
Figura 3 – Triangulação para as etapas 3 e 4



Fonte: Autores.

Para a caracterização do corpus, utilizou-se a lista de frequência dos frames mais recorrentes no corpus com base em Volpato e Barreto (2016). Como resultados, obteve-se que as variáveis do corpus são de natureza quantitativa discreta. As variáveis independentes são as áreas do conhecimento e disciplinas e as variáveis dependentes são os frames, unidades lexicais e elementos de frames. A distribuição dos dados é não normal e heterocedástica, portanto testes não paramétricos serão necessários. Ademais, plotou-se um gráfico de dispersão (Figura 4) dos dados para comprovar, de acordo Volpato e Barreto (2016), qual é a distribuição dos dados e qual a melhor medida para este tipo de gráfico. O gráfico gerado se assemelha ao gráfico apresentado na Prancha 3, Base 14, Caso 3, (VOLPATO; BARRETO, 2016, p. 34, 84, 85) que comprova que que a distribuição de dados é não padrão e que a melhor medida para esse estudo é a mediana.

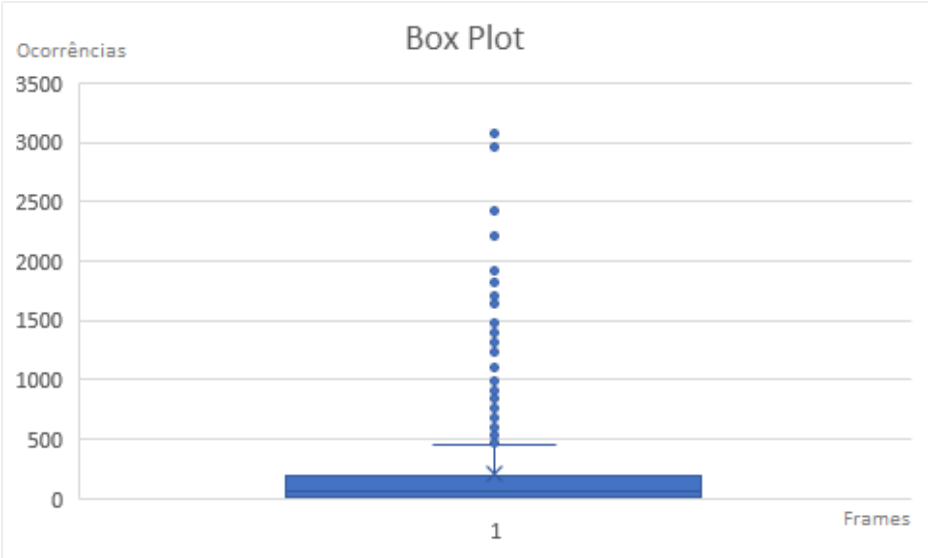
Figura 4 – Gráfico de dispersão de dados



Fonte: Autores.

Buscando-se uma análise mais apurada do corpus, fez-se a plotagem de um gráfico do tipo *box-plot* para os dados das três grandes áreas do conhecimento (*fields*) - *Hard Sciences*, *Social & Human Sciences* e *Biological Sciences* - para identificação do comportamento geral do corpus (Figura 5).

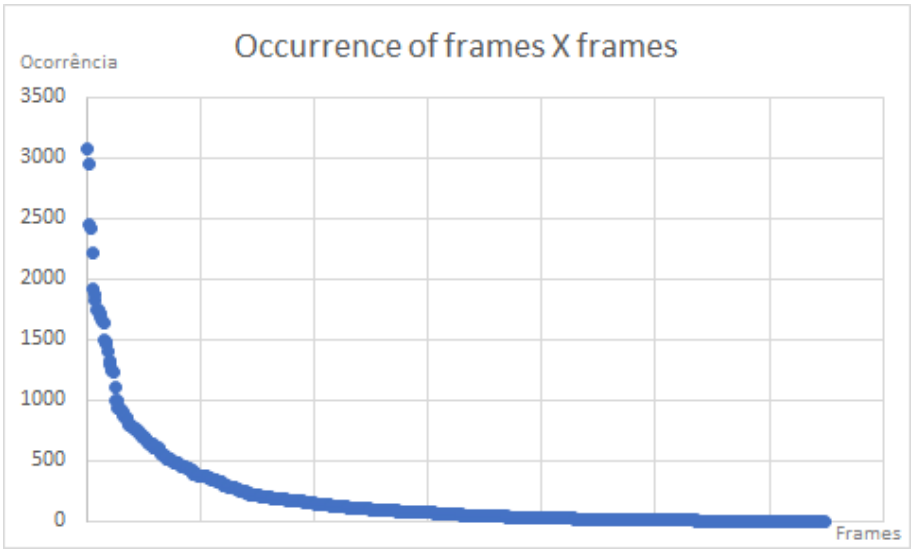
Figura 5 – Box-plot para os dados gerais do corpus.



Fonte: Autores.

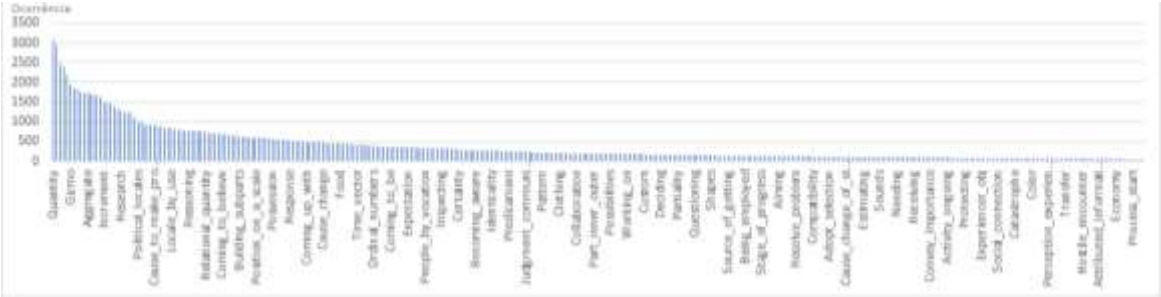
Após, foi plotado o gráfico de dispersão (Figura 6) e de barras (Figura 7) também para os dados do corpus geral. Com esses gráficos, pode-se concluir algumas informações acerca do comportamento do corpus geral e sua caracterização.

Figura 6 – Gráfico de dispersão para o corpus.



Fonte: Autores. DEFINIR UNIDADE NOS DOIS EIXOS

Figura 7 – Gráfico de barras para o corpus.



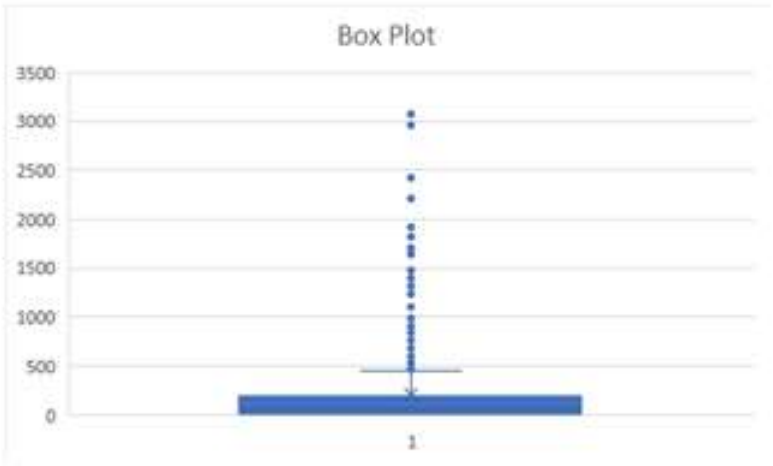
Fonte: Autores.

Utilizou-se a lista de frequência dos frames mais recorrentes no corpus, para a caracterização com base em Volpato e Barreto (2016). Pode-se observar que as variáveis do corpus (frames) são de natureza discreta quantitativa. Nesse caso, as variáveis independentes são as áreas do conhecimento e disciplinas e as variáveis dependentes são os frames, unidades lexicais e elementos de frames. A distribuição dos dados é não normal e heterocedástica, sendo necessários testes não paramétricos.

O gráfico de dispersão (Figura 6) se assemelha ao gráfico apresentado na Prancha 3, Base 14, Caso 3, (VOLPATO; BARRETO, 2016, p. 34, 84, 85) que comprova que nossa distribuição de dados é não padrão e que a melhor medida para esse estudo é a mediana.

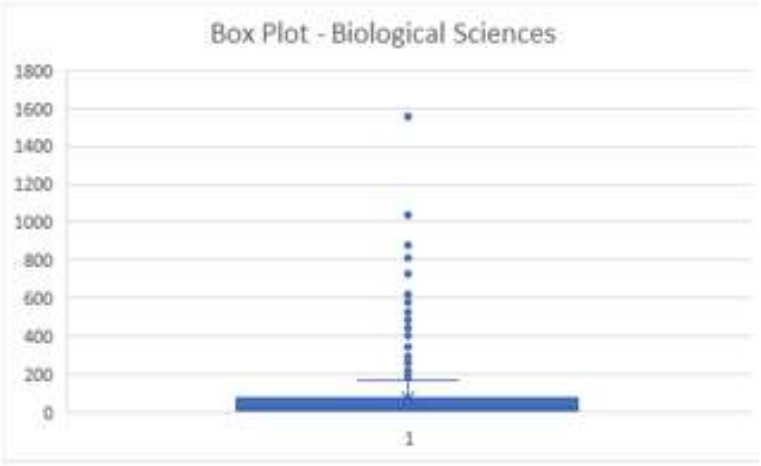
Posteriormente, fez-se a plotagem do box-plot de ocorrência de frames por frames para os *fields* geral e para os *fields Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* respectivamente (Figuras 8a, 8b, 8c e 8d).

Figura 8a – Box-Plot para *fields* geral dentro de *fields*.



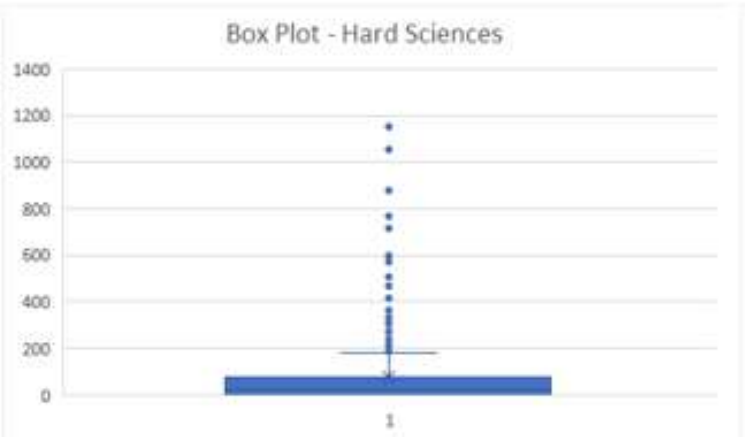
Fonte: Autores.

Figura 8b – Box-Plot para *fields Biological Sciences* dentro de *fields*.



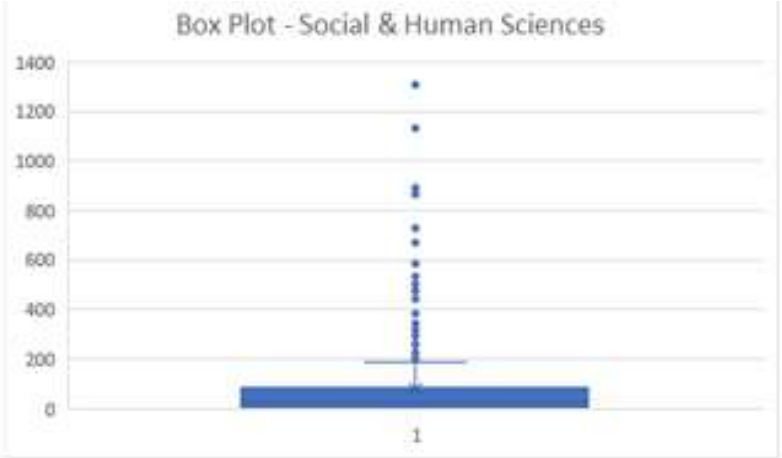
Fonte: Autores.

Figura 8c – Box-Plot para *fields Hard Sciences* dentro de *fields*.



Fonte: Autores.

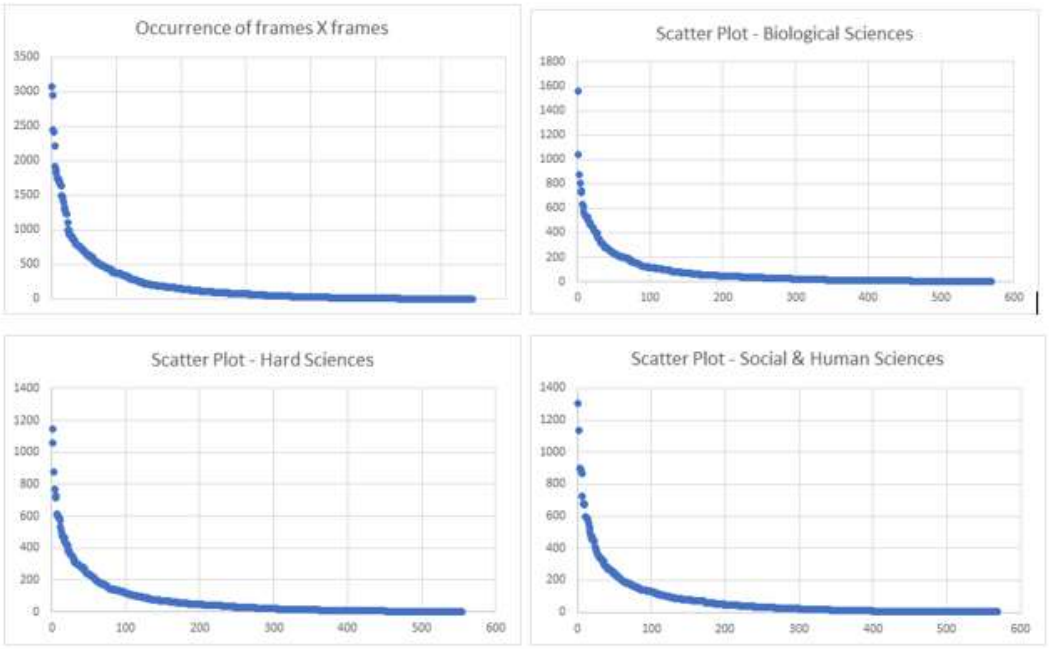
Figura 8d – Box-Plot para *fields Social & Human Sciences* dentro de *fields*.



Fonte: Autores.

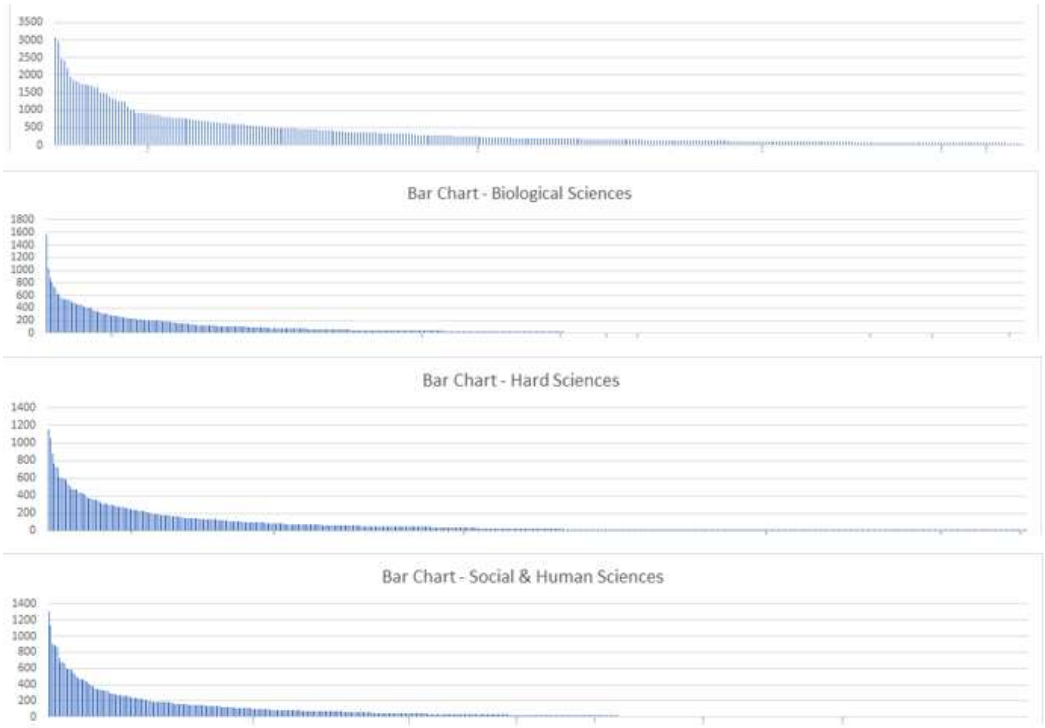
Ademais, fez a plotagem de gráficos de dispersão (Figura 9) e de barras (Figura 10) para os *fields* anteriormente citados.

Figura 9 – Gráficos de dispersão para fields geral, *Biological Sciences*, *Hard Sciences e Social & Human Sciences* dentro de *fields*



Fonte: Autores.

Figura 10 – Gráficos de dispersão para *fields* geral, *Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* dentro de *fields*.

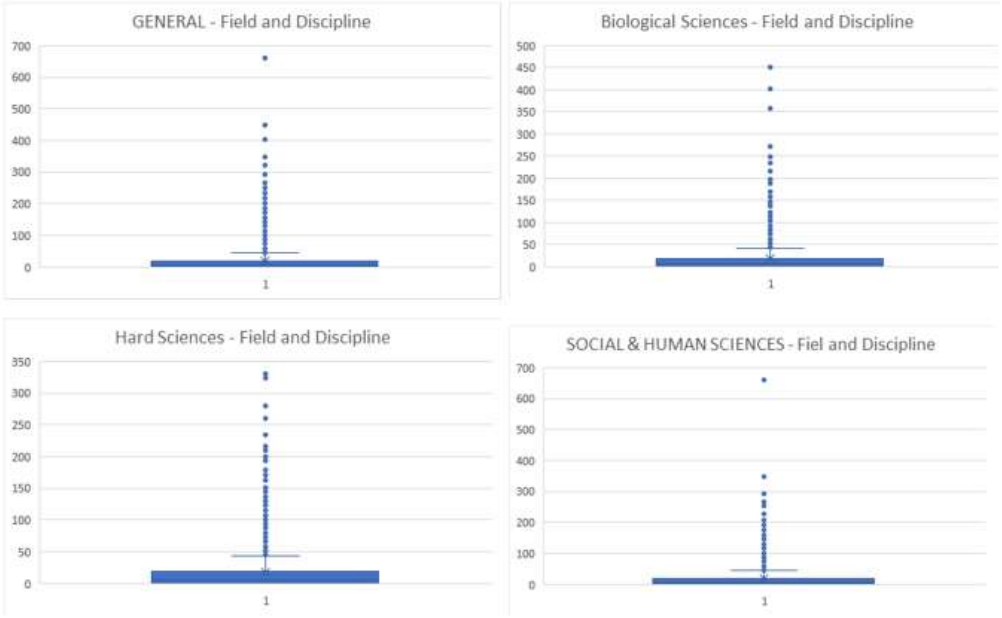


Fonte: Autores.

Como um comportamento esperado, todos os gráficos seguiram o mesmo padrão. Esse comportamento presente em todos os gráficos se deve ao comportamento do corpus, que uma grande quantidade de frames se repete apenas uma vez e outra grande quantidade repete em grandes quantidades. Isso pode ser observado em todos os gráficos acima. Numa forma mais visual de análise, observa-se que nos gráficos de dispersão e de barras vemos é verificado uma cursa muito acentuada que representa a grande quantidade de frames que se repetem poucas vezes no corpus. No box-plot isso pode ser vista pela caixa, que está na parte inferior do gráfico, o que represente o grande número de frames com poucas ocorrências.

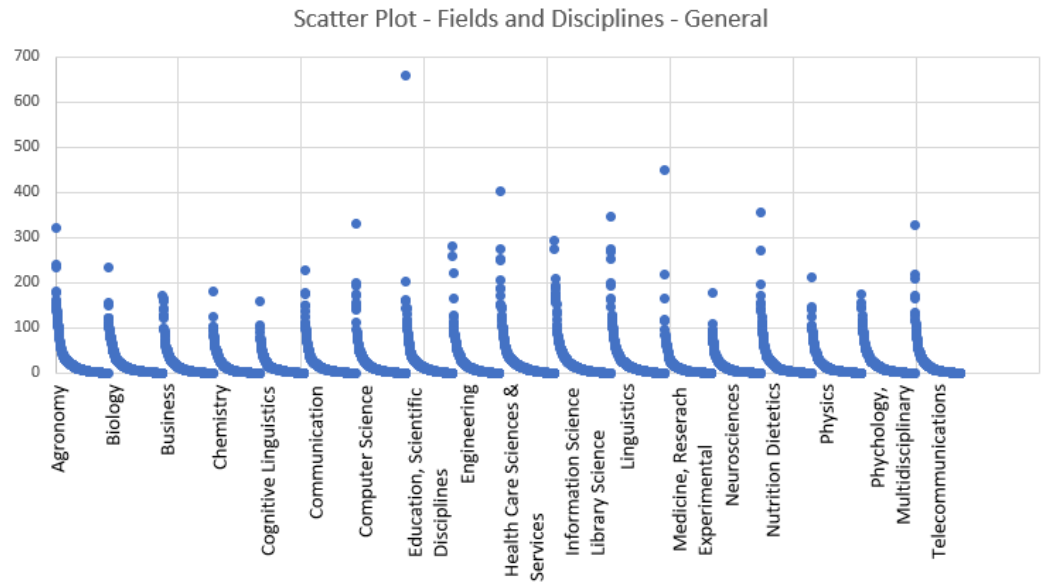
Após isso, fez-se a plotagem dos gráficos box-plot (Figura 11) para o corpus geral e para os 3 *fields* dentro de *fields* e *disciplines*. Além disso, plotou-se os gráficos de dispersão (Figura 12) e de barras (Figura 13) para o corpus geral dentro de *fields* com *disciplines*.

Figura 11 – Box-Plot para *fields* geral, *Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* dentro de *fields* com *disciplines*.



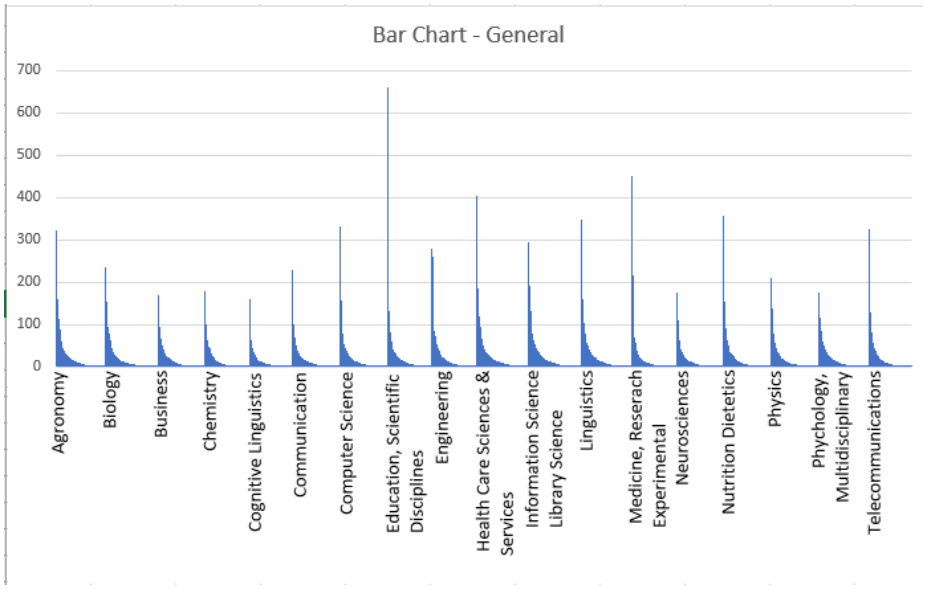
Fonte: Autores.

Figura 12 – Gráfico de dispersão para *fields* geral dentro de *fields* com *disciplines*.



Fonte: Autores.

Figura 13 – Gráfico de barras para *fields* geral dentro de *fields* com *disciplines*.

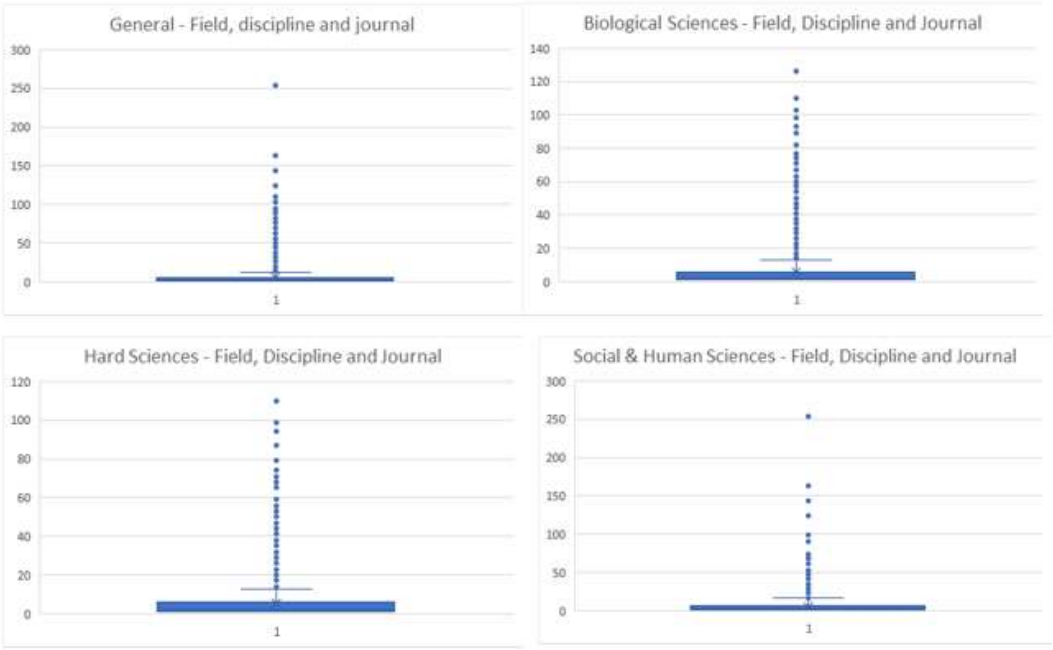


Fonte: Autores.

Com a plotagem desses gráficos dentro de *fields* com *disciplines*, podemos perceber uma repetição do comportamento padrão anteriormente observado apenas para *fields*. Um comportamento que pode ser observado é que dez *disciplines* tem o top frame com menos de 300 recorrências, comprovando assim o comportamento do corpus, onde uma grande quantidade de frames recorrerem poucas vezes.

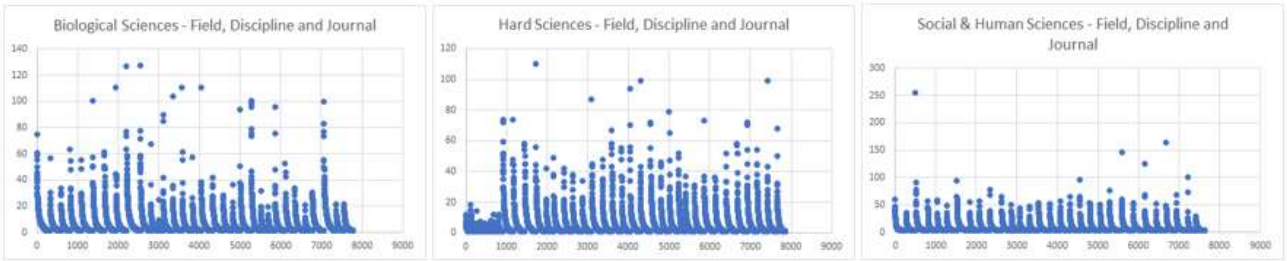
Ademais, por vias de conclusão das análises e confirmações dos resultados, fez-se a plotagem dos gráficos box-plot (Figura 14) para o corpus geral e para os 3 *fields* dentro de *journal* com *field* e *discipline*. Além disso, plotou-se os gráficos de dispersão (Figura 15) e de barras (Figura 16) para os 3 *fields* dentro de *journal* com *field* e *discipline*.

Figura 14 – Box-Plot para *fields* geral, *Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* dentro de *journals* com *fields* e *disciplines*.



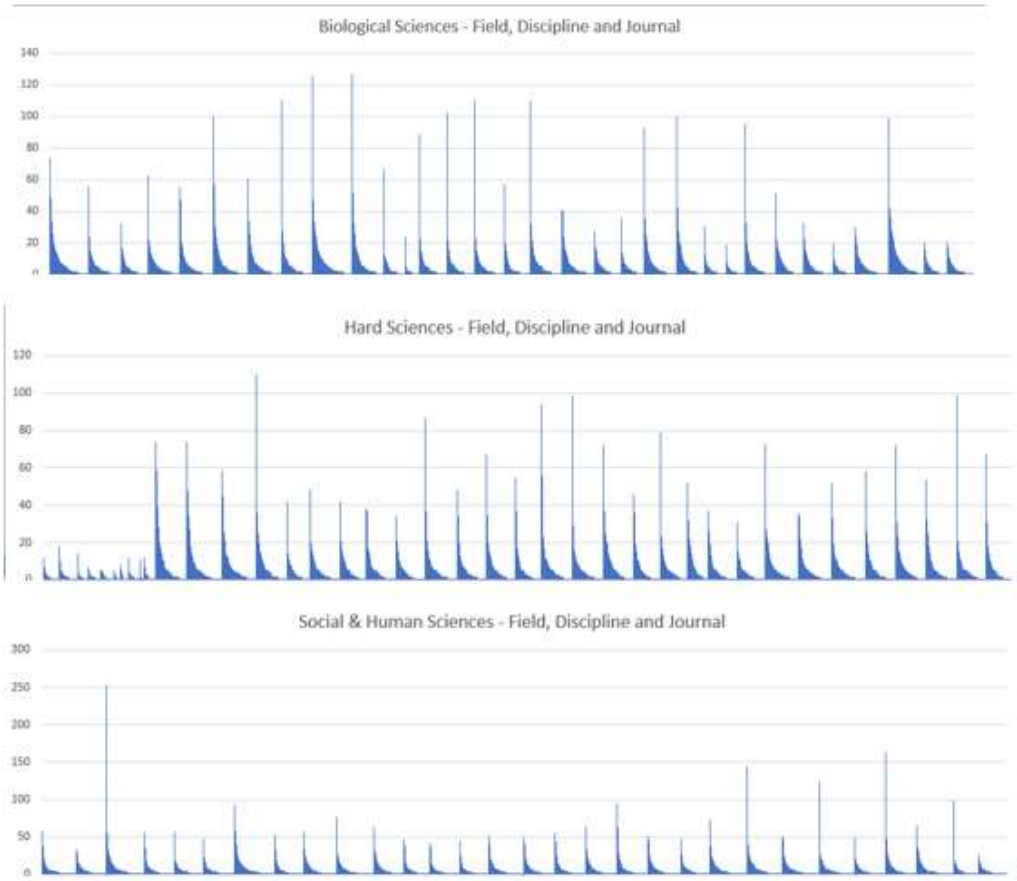
Fonte: Autores.

Figura 15 – Gráfico de dispersão para *Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* dentro de *journals* com *fields* e *disciplines*.



Fonte: Autores.

Figura 16 – Gráfico de barras para *Biological Sciences*, *Hard Sciences* e *Social & Human Sciences* dentro de *journals* com *fields* e *disciplines*.



Fonte: Autores.

CONCLUSÕES

Conclui-se que todas as curvas dos gráficos seguem o mesmo padrão. No entanto, conforme visto na análise, cada periódico de cada disciplina seguirá seu próprio padrão. Portanto, algumas curvas acabam tendo uma altura maior e outras, menor. O comportamento individual da curva será baseado no padrão do periódico e no padrão de publicação que o periódico segue. Analisando o comportamento de todos os gráficos, percebemos que, independentemente da área do conhecimento, o comportamento dos dados permanece o mesmo. Assim, podemos concluir que os dados do corpus sempre seguirão um padrão.

Em termos semânticos, os dados resultantes da análise fornecem ao linguista o padrão de conceitos (frames) comunicados nos textos científicos em cada disciplina, em consonância com a área do conhecimento e em atendimento ao periódico de publicação. Esse padrão permite reconhecer o viés da área de estudo e serve como orientação para o ensino de redação científica no processo de educação de cientistas. É possível também investigar o viés de cada periódico, a partir de uma análise contrastiva mais refinada da relação entre as disciplinas e os periódicos em que os textos são publicados. Como perspectivas futuras, pode-se estudar a relação entre os itens lexicais (texto) e os frames (conceitos), contribuição que possibilitará ao linguista que atua na educação científica desenvolver atividades de ensino de escrita científica não só relativas ao pareamento conceito : área de estudo, mas também relativa à escolha lexical e fraseológica : conceito : área de estudo.

Este estudo interdisciplinar entre Ciência Linguística e Ciência de Dados, com aplicação de modelagem estatística a partir da compilação de um *corpus* eletrônico anotado por *parser* semântico, comprova que a articulação entre saberes é não apenas viável, mas substancial para que o conhecimento semântico sobre a comunicação científica, sistematizado cientificamente, seja produzido e utilizado para a educação, para a ciência e para a tecnologia.

REFERÊNCIAS

ABREU, A. S. Gramática Integral da Língua Portuguesa: uma visão prática e funcional. Cotia/SP: Ateliê Editorial, 2018.

FILLMORE, C. Frame Semantics. In: Linguistics in the morning calm. Seoul: Hanshin, 1982. p.111-138.

FILLMORE, C.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to FrameNet. International Journal of Lexicography, Vol. 16, n. 3, p. 235-250. Oxford University Press, 2003.

RODRIGUES, R. F. L. *A ciência é uma jornada*: um projeto remodelado como programa de Pesquisa Linguística em Comunicação Científica com uso de Data Science. Sinergia (IFSP), ISSN:2177-451X, v. 20, Edição Especial - Comunicação Científica, Cognição e Persuasão, 2019. p. 60-81. Disponível em: <https://ojs.ifsp.edu.br/index.php/sinergia/article/view/1112> Acesso em: 11 Maio 2020.

RODRIGUES, R. F. L. *Frame narratives for the semantics of Science Communication*: a first qualitative sample analysis of the most prevalent frames in a corpus of abstracts tagged in Semafor. [S. l.: s. n.], 2020a.

RODRIGUES, R. F. L. *The semantics of Science Communication*: the research design of a study on automatic frame tagging of scientific abstracts. 2020b.

RODRIGUES, R. F. L.; BAPTISTA, A. E. O. B. Design thinking tools for scientific storytelling: a didactic innovation. *Proceedings of the 13th Annual International Technology, Education and Development Conference*, INTED 2019. 11th-13th March, 2019.

TAGLIACOLLO, Victor Alberto; VOLPATO, Gilson Luiz; JUNIOR, Alfredo Pereira. *Association of student position in classroom and school performance*. 2010. Disponível em: https://www.researchgate.net/publication/256502547_Association_of_Student_Position_in_Classroom_and_School_Performance#:~:text=We%20suggest%20that%20school%20performance,likely%20to%20improve%20school%20performance. Acesso em: 1 ago. 2020.

TURNER, M. Frame Blending. In: Frames, Corpora, and Knowledge Representation, edited by Rema Rossini Favretti. Bologna: Bononia University Press, 2008.

VOLPATO, G. L.; BARRETO, R. E. *Estatística sem dor*!!! 2. ed. Botucatu: Best Writing, 2016.