UNIVERSIDADE ESTADUAL PAULISTA
Faculdade de Filosofia e Ciências de Marília
Departamento de Ciência da Informação
Curso de Biblioteconomia

GIOVANNA TEODORO ROSA

A PRAGMÁTICA-LINGUÍSTICA NO PROCESSAMENTO DA LINGUAGEM NATURAL: Contribuições ao tratamento temático da informação

MARÍLIA
2025

GIOVANNA TEODORO ROSA

A PRAGMÁTICA-LINGUÍSTICA NO PROCESSAMENTO DA LINGUAGEM NATURAL: Contribuições ao tratamento temático da informação

Trabalho de Conclusão de Curso apresentado Conselho de Curso de Biblioteconomia da Faculdade de Filosofia e Ciências, da Universidade Estadual Paulista – UNESP – Campus de Marília, para a obtenção do título Bacharel em Biblioteconomia

Financiadora: CNPq – Proc. 122252/2023-3 e Proc.128763/2024-8

Linha de Pesquisa: Produção e organização da informação

Orientador: Prof. Carlos Cândido de Almeida

MARÍLIA
2025

R788p

Rosa, Giovanna Teodoro

A PRAGMÁTICA-LINGUÍSTICA NO PROCESSAMENTO DA LINGUAGEM NATURAL : Contribuições ao tratamento temático da informação / Giovanna Teodoro Rosa. -- Marília, 2025

157 p. : tabs.

Trabalho de conclusão de curso (Bacharelado - Biblioteconomia) - Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, Marília

Orientador: Carlos Cândido de Almeida

1. Organização da informação. 2. Tratamento temático da informação. 3. Pragmática. 4. Processamento da linguagem natural. 5. Indexação automática. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

GIOVANNA TEODORO ROSA

A PRAGMÁTICA-LINGUÍSTICA NO PROCESSAMENTO DA LINGUAGEM NATURAL: Contribuições ao tratamento temático da informação

Trabalho de Conclusão de Curso apresentado Conselho de Curso de Biblioteconomia da Faculdade de Filosofia e Ciências, da Universidade Estadual Paulista – UNESP – Campus de Marília, para a obtenção do título Bacharel em Biblioteconomia

Financiadora: CNPq – Proc.. 122252/2023-3 e Proc.128763/2024-8

BANCA EXAMINADORA

Orientador: Prof. Dr. Carlos Cândido de Almeida

Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP) – Faculdade de Filosofia e Ciências (FFC), Campus de Marília

2º Examinador: Prof. Dr. Edberto Ferneda

Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP) – Faculdade de Filosofia e Ciências (FFC), Campus de Marília

3º Examinador: Profª. Dra. Franciele Marques Redigolo

Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP) – Faculdade de Filosofia e Ciências (FFC), Campus de Marília

Marília, 02 de dezembro de 2025

AGRADECIMENTOS

Este trabalho não seria possível sem o apoio incessante dos meus pais, Áurea e Kleber, cujos vários anos trabalhando embaixo de muito sol me possibilitaram viver aqui, à sombra. O amor e sacrifícios de vocês me moldaram. Obrigada.

Agradeço minhas avós, Vera e Isaura, que tiveram a vida marcada pelo amor e entrega aos filhos e netos. Meus exemplos de resiliência, que foram rede de apoio e fonte de amor, carinho e companheirismo desde o dia que nasci.

Agradeço minhas (mães)drastas, Silvana e Érika, por terem me amado como filha e me acolhido como amiga.

Agradeço meus irmãos, Marcelo, Carolina, Lorena e Igor, por me mostrarem que a vida compartilhada é mais gostosa. Não existem limites para o amor.

Agradeço minha sobrinha, Heloísa, que todos os dias me motiva a ser um pouquinho melhor. Que eu possa ser seu exemplo.

Agradeço minha tia, Maira, por ter sido colo, mãe e amiga, e aos meus tios, Fábio, Fernando e Hailton, por terem preenchido minha vida familiar de afeto e bons momentos.

Agradeço meus primos, Bruno, Matheus, Lorena, Raphaella e Ravi, por serem também irmãos. Que eu possa inspirá-los a buscar um futuro melhor através do conhecimento.

Não poderia deixar de agradecer meu orientador, Carlos, por ter visto naquela caloura, ainda perdida na imensidão da vida universitária, a pesquisadora que me tornei. Suas contribuições foram muito além da vida acadêmica.

Agradeço minha banca avaliadora, Edberto e Franciele, por terem ajudado a lapidar este trabalho até sua melhor versão, além de me auxiliar durante todo meu processo de formação acadêmica.

Agradeço minha turma, especialmente Nicole e Isabela, por todos os momentos que dividimos nos últimos quatro anos. Foram vocês que fizeram minhas manhãs valerem a pena.

Agradeço a todos aqueles que passaram pela minha vida, independente de quando e como, pois o que somos nós se não uma miscelânea de todos aqueles que amamos?

“Tenho para mim que as palavras carregam mais do que seus significados. É preciso cavar e não apenas ouvir. [...] Em minha arqueologia das palavras, busco, incessante, seus vestígios.”

Carla Madeira (2022).

RESUMO

As atividades do tratamento temático da informação são mediadas pela linguagem natural, cuja variação linguística e ambiguidade representam desafios para os processos de representação e recuperação da informação. A automatização destes processos exige a compreensão e processamento de contextos informacionais, evidenciando a importância de diálogos entre ciência da informação, linguística e tecnologias de processamento da linguagem natural. Este estudo investiga em que medida *softwares* de análise textual incorporam fundamentos pragmático-linguísticos capazes de apoiar atividades de tratamento temático da informação em linguagem natural. Para isso, realizou-se pesquisa bibliográfica e análise comparativa de treze ferramentas: Yake!, MAUI, Annif, Medical Text Indexer, KEA, WEKA, Finto AI, Voyant Tools, Cortext, IRaMuTeQ, AntConc, Orange Data Mining e Yoshikoder. Os resultados indicam que nenhuma das aplicações examinadas se fundamenta explicitamente em teorias pragmáticas; contudo, observa-se a presença de métodos oriundos da linguística de *corpus*, que aproximam informação, representação e contexto. Conclui-se que, embora ofereçam recursos úteis para a automatização parcial do tratamento temático, os *softwares* analisados não contemplam dimensões pragmáticas essenciais à significação contextual, indicando uma lacuna e uma oportunidade para o desenvolvimento de abordagens interdisciplinares que integrem pragmática, semiótica informacional e processamento da linguagem natural.

Palavras-chave: Organização da informação; Tratamento temático da informação; Pragmática; Processamento da linguagem natural; Indexação automática.

ABSTRACT

The subject representation activities are inherently guided by natural language, whose linguistic variation and ambiguity represent challenges for information representation and retrieval processes. The automation of these processes requires the understanding and processing of informational contexts, highlighting the importance of dialogues between information science, linguistics, and natural language processing technologies. This study investigates to what extent text analysis software incorporates pragmatic-linguistic foundations capable of supporting subject representation activities in natural language. To this end, bibliographic research and a comparative analysis of thirteen tools were conducted: Yake!, MAUI, Annif, Medical Text Indexer, KEA, WEKA, Finto AI, Voyant Tools, Cortext, IRaMuTeQ, AntConc, Orange Data Mining, and Yoshikoder. The results indicate that none of the examined applications is explicitly based on pragmatic theories; however, the presence of methods originating from corpus linguistics is observed, bringing information, representation, and context closer together. It is concluded that, although they offer useful resources for the partial automation of subject representation, the analyzed software does not contemplate pragmatic dimensions essential to contextual meaning, indicating a gap and an opportunity for the development of interdisciplinary approaches that integrate pragmatics, informational semiotics, and natural language processing.

KEYWORDS: Knowledge organization; Subject representation; Pragmatics; Natural language processing; Automatic indexing.

LISTA DE ILUSTRAÇÕES

Figura 1 - Processos de representação documental 18
Figura 2 - Mapa do tratamento temático da informação 20
Figura 3 - Semântica fregeana 40
Figura 4 - Simplificação redutora de Morris.....42
Figura 5 - Ocorrência dos termos descritores destacada pelo sistema Yake! ..83
Figura 6 - Termos sugeridos como descritores pelo sistema Annif.....87
Figura 7 - Interface de entrada de dados do sistema Finto AI.....93
Figura 8 - Ferramentas e técnicas do Cortext *Manager*.....98
Gráfico 9 - Métodos de análise textual aplicados pelos *softwares*.....107
Gráfico 10 - Presença de funções/bases nos *softwares*.....110
Figura Apêndice 1 - Fatores que influenciam a qualidade da indexação.....139

LISTA DE QUADROS

Quadro 1 - Modelo de Leitura para indexação de textos científicos 25
Quadro 2 - Modelo de resumo de texto canônico 28
Quadro 3 - Modelo de resumo de texto dissertativo argumentativo 29
Quadro 4 - Modelo de resumo de texto expositivo 29
Quadro 5 - Correspondência entre as partes dos tipos de texto 29
Quadro 6 – Tipos de contextos associados à organização e recuperação da informação 65
Quadro 7 - Modelo de ficha de informações para a avaliação de *softwares* 81
Quadro 8 - Ferramentas do *Voyant Tools* 95
Quadro 9 - Ferramentas do Cortext 98
Quadro 10 - Síntese dos resultados 110
Quadro Apêndice 1 - Síntese das convergências entre critérios, indicadores e fatores dos autores 144
Quadro Apêndice 2 - Síntese das divergências entre critérios, indicadores e fatores dos autores 146
Quadro Apêndice 3 - Síntese das convergências entre critérios de qualidade de indexação e critérios de qualidade de *softwares* 148
Quadro Apêndice 4 - Síntese das convergências entre critérios de qualidade de indexação e critérios de qualidade de *softwares* 150
Quadro Apêndice 5 - Síntese das aplicações dos critérios da ISO/IEC 25010:2023 na avaliação de *softwares* para a automatização indexação 152

LISTA DE ABREVIATURAS E SIGLAS

AACR2R Código de Catalogação Anglo Americano 2° edição revisada

ABNT Associação Brasileira de Normas Técnicas

AFC Análise Fatorial de Correspondência

CI Ciência da Informação

CT *CheckTags*

DOC Documento

DOCX Documento *Word*

EPUB *Electronic Publication*

HTML *HyperText Markup Language*

IA Inteligência Artificial

IDF *Inverse Document Frequency*

IFRIS *Institut Francilien Recherche Innovation Société*

INESC TEC *Institute for Systems and Computer Engineering, Technology and Science*

INRAE Instituto Nacional de Pesquisa para a Agricultura, Alimentação e Meio Ambiente

IRaMuTeQ *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*

KEA *Keyphrase Extraction Algorithm*

KWIC *Key Word in Context*

L**ABEX SITES** Laboratório de Excelência em Ciências, Inovações e Tecnologias para o Meio Ambiente e a Sociedade

LISIS *Laboratoire Interdisciplinaire Sciences Innovations Sociétés*

MAUI *Multi-purpose Automatic Topic Indexing*

MeSH *Medical Subject Headings*

MH Lista de Cabeçalhos Principais do MeSH

MMI *MetaMap Indexing*

MS *Microsoft Word*

MTI *Medical Text Indexer*

NBR Norma Brasileira Regulamentadora

NLM	Biblioteca Nacional de Medicina dos Estados Unidos
OC	Organização do Conhecimento
ODT	*Open Document Text*
OI	Organização da Informação
PDF	*Portable Document Format*
PLN	Processamento da Linguagem Natural
PPTX	*PowerPoint Open XML Presentation*
RI	Recuperação da Informação
RNA	Redes Neurais Artificiais
RTF	*Rich Text Format*
SH	Subcabeçalhos MeSH
SKOS	*Simple Knowledge Organization System*
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TTI	Tratamento Temático da Informação
TXT	Arquivo de Texto
URL	*Uniform Resource Locator*
WEKA	*Waikato Environment for Knowledge Analysis*
XLS	*Extensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO 11
2. REFERENCIAL TEÓRICO 16
2.1. Organização da informação 16
2.2. Pragmática 30
2.2.1. Frege 33
2.2.2. Morris 41
2.2.3. Austin e Searle 53
2.3. Pragmática e ciência da informação 60
2.4. Processamento da linguagem natural 71
3. PROCEDIMENTOS METODOLÓGICOS 79
3.1. Tipo de pesquisa 79
3.2. Universo e amostra 79
3.3. Técnicas de coleta 80
3.4. Técnicas de análise 81
4. RESULTADOS 82
4.1. Análise dos *softwares* 82
4.1.1. Yake! 82
4.1.2. MAUI 84
4.1.3. Annif 85
4.1.4. *Medical Text Indexer* 87
4.1.5. KEA 89
4.1.6. WEKA 91
4.1.7. Finto AI 92
4.1.8. *Voyant Tools* 94
4.1.9. Cortext 97
4.1.10. IRaMuTeQ 100
4.1.11. AntConc 101
4.1.12. *Orange Data Mining* 103
4.1.13. Yoshikoder 105
4.2. Bases pragmático-linguísticas dos *softwares* 106
5. CONSIDERAÇÕES FINAIS 115
REFERÊNCIAS 118

APÊNDICE A – CONTRIBUIÇÕES DE CRITÉRIOS DE QUALIDADE DE *SOFTWARES* DE PROCESSAMENTO DA LINGUAGEM NATURAL À INDEXAÇÃO.....132

1. INTRODUÇÃO

Questionamentos relativos às condições de recuperação da informação (RI) mediada por instrumentos de busca constituem grande parte dos estudos empreendidos na ciência da informação (CI), estabelecendo relações multi e interdisciplinares com áreas como ciência da computação, inteligência artificial (IA), linguística, filosofia da linguagem, entre outras.

A recuperação efetiva da informação tem como base imprescindível a representação de seu conteúdo, expresso em signos linguísticos, como forma de mediação entre a linguagem utilizada pelos usuários no processo de busca e a linguagem expressa no sistema e objetos informacionais nele armazenados.

Investigações acerca da representação da informação como fundamento para a recuperação em sistemas informacionais encontram-se dispostas nos estudos sobre organização da informação e do conhecimento.

Segundo Bräscher e Café (2008), a organização da informação compromete-se com os estudos relacionados aos registros informacionais de existência física, favorecendo sua organização através de atividades como a catalogação, classificação, indexação e elaboração de resumos, enquanto a organização do conhecimento encarrega-se de conceitos a partir de abstrações da realidade com vistas à descrição e explicação de fenômenos observáveis.

Assim, as ações de representação no âmbito da organização da informação dividem-se em tratamento descritivo, com foco nos aspectos físicos documentais da informação, como a catalogação, e o tratamento temático, encarregado do conteúdo documental, como a indexação.

Independente de categorizações, as atividades representativas se propõem a condensar as informações de um documento criando uma versão sintetizada para a "substituição" por produtos informacionais, facilitando sua recuperação e gerando um ciclo de informações elaboradas a partir de outras informações (Kobashi; Fernandes, 2009).

Para isto, a representação temática busca identificar os termos mais relevantes e representativos dos significados de um documento, tornando possível o estabelecimento de uma essência temática ao documento, a partir de conceitos capazes de expressar a significação documental (Saracevic, 1996).

Todo este processo ocorre a partir da linguagem natural, que além de mediadora é também matéria prima fundamental para todo o ciclo informacional, desde sua produção até sua disseminação.

Embora sejam detentoras de grande capacidade de comunicação das mais diversas formas e sobre os mais diversos assuntos, assegurando sua versatilidade, a utilização de linguagens naturais também pode tornar-se um empecilho em situações em que se preza principalmente pela precisão, como nas atividades de representação informacional

Pensando no contexto no qual esta linguagem é utilizada, a consideração dos contextos os quais envolvem a informação e o ouvinte - ou receptor - torna-se relevante às ações de tratamento e comunicações da informação.

Neste sentido, a pragmática surge no campo informacional como possibilidade para a análise da relação entre locutores de uma situação comunicativa e os signos escolhidos e utilizados por eles, o que lhes atribui significado pela interação entre o contexto da situação comunicativa e a forma de expressão.

A partir disso surge uma nova perspectiva na qual a informação passa a ser considerada integralmente, atribuindo relevância também aos contextos de produção e recepção documental e informacional (Novellino, 1998; Azevedo Netto, 2002).

Sem estes aspectos considerados e delimitados, mesmo que representadas, as informações armazenadas em sistemas informacionais tornam-se ambíguas e sofrem com processos de ressignificação (Armengaud, 2006).

A utilização de sistemas informacionais, como explorada por Coneglian e Santarém Segundo (2022), enfrenta complicações no processamento dos contextos de informações e também nas buscas realizadas no sistema pelos usuários a partir da linguagem natural, além dos fenômenos ambiguidade e variação linguística, típicos na utilização da linguagem humana.

Pensando nisso, o processamento da linguagem natural (PLN) surge como alternativa para compreender o desenvolvimento de sistemas e aplicações computacionais para a análise e compreensão de textos produzidos em linguagem natural (Coneglian; Santarém Segundo, 2022).

Para Bräscher (2002), ferramentas de busca de sistemas de informação devem reconhecer os usos, expressões, relações e significados da linguagem natural, e serem também capacitadas para o tratamento das mais diversas manifestações

linguísticas e fenômenos, como a ambiguidade, que podem vir a afetar a qualidade da representação e recuperação de informações.

Pensando no contexto concebido, o estudo se propôs responder aos seguintes questionamentos: Quais os atributos da pragmática estão presentes nos *softwares* de representação da informação? Quais *softwares* oferecem um processamento pragmático da linguagem, atendendo as demandas dos processos de representação da informação?

De modo geral, buscou-se identificar os *softwares* de análise textual que se concentram no processamento técnico da linguagem natural e suas bases pragmáticas.

Em busca de atingir o objetivo geral, adotou-se os seguintes objetivos específicos: a) revisão da literatura produzida sobre a pragmática-linguística e processamento da linguagem natural; b) identificação das aplicações de *softwares* de processamento da linguagem natural para o tratamento temático da informação; c) identificar quais dessas aplicações se fundamentam nos estudos realizados no campo pragmático; d) relacionar os fundamentos pragmáticos dos *softwares* como forma de apoio aos estudos da semiótica informacional (semiótica aplicada aos estudos da ciência da informação).

Com base no exposto, são evidentes as possíveis contribuições dos estudos linguísticos e pragmáticos para a compreensão e tratamento da linguagem natural na representação e recuperação da informação. O reconhecimento e utilização dos elementos teóricos da pragmática aplicados à ciência da informação levaria à consolidação do vínculo estabelecido entre a organização da informação e os estudos da linguagem.

Em relação à sociedade, reconhece-se o papel mediador das atividades, processos e serviços informacionais entre a informação e seus potenciais usuários, como afirma Ribeiro (2010). Entre estes processos, o tratamento temático da informação se destaca por mediar linguagem do usuário e linguagem do sistema da informação.

Em relação às contribuições na esfera científica, a investigação identifica e aprofunda laços interdisciplinares entre ciência da informação e outros campos de investigação, como linguística, semiótica e processamento da linguagem natural.

Por conseguinte, são reafirmadas contribuições dos estudos da linguagem e das tecnologias na ciência da informação, como feito também por Borko (1968),

Baranow (1983), Saracevic (1996), Novellino (1998) e Kobashi (2007), validando a transversalidade dos estudos informacionais.

Assim, citam-se autores como Kobashi (2007), Kobashi e Fernandes (2009), Baranow (1983), Leal e Costa (2020), Novellino (1998), Mari (1996), Martines (2023), Saldanha (2008), que retratam contribuições dos estudos da linguagem - linguísticos e semióticos - para as atividades e processos informacionais.

Também se tem autores como Coneglian e Segundo (2022), Kamp e Stokhof (2008), Baeza-Yates (2004), Falcão, Lopes e Souza (2021), Ferrández, Rojas e Peral (2007), Gamallo e García (2012), Hirschberg e Manning (2015), Vallez e Jimenez-Pedraza (2007) que destacam a atuação dos *softwares* e do processamento da linguagem natural na ciência e para a informação.

Indo além e partindo do ponto de vista do crescente volume de produção informacional, os resultados desta investigação pôde apresentar novos instrumentos a serem utilizados na organização e representação da informação, dinamizando o tratamento temático para o armazenamento, busca e potencial uso da informação e do conhecimento.

Na ciência da informação os estudos de organização da informação favorecem outras atividades e fases do ciclo informacional, como o armazenamento, a busca dos usuários e a recuperação da informação, todas elas fundamentadas na organização temática da informação. Assim, esta investigação propõem contribuições para a dinamização e aprimorando de todas as fases do ciclo informacional, desde seu tratamento até sua utilização.

Valentim (2000, p. 8) salienta a busca da ciência da informação por um “[...] profissional dinâmico e competitivo que de fato atenda os anseios da sociedade brasileira”. Para isso, é necessário o empreendimento de estudos e pesquisas que colaborem com o desenvolvimento de novos métodos, técnicas e instrumentos, favorecendo novas competências e abordagens para a formação inicial, para novos profissionais, e continuada, para profissionais atuantes.

Este estudo também possui potencial de aprimoramento da formação e atuação do profissional da ciência da informação através da apresentação de novos aportes teóricos e a identificação de novos instrumentos e métodos para a dinamização das atividades profissionais, estimulando a atualização e desenvolvimento das competências do profissional já inserido no mercado de trabalho.

Para tanto realizou-se, nos moldes quali-quantitativos, a análise de 13 *softwares* de processamento da linguagem (Yake!, MAUI, Annif, *Medical Text Indexer*, KEA, WEKA, Finto AI, *Voyant Tools*, Cortext, IRaMuTeQ, AntConc, *Orange Data Mining* e Yoshikoder) como forma de identificar os fundamentos pragmático-linguísticos utilizados para o processamento da linguagem natural tendo em vista uma possível aplicação nas atividades de tratamento temático da informação.

A construção do Apêndice A teve como objetivo propor critérios de qualidade de sistemas de processamento da linguagem natural para a seleção de sistemas para a automatização da indexação, de acordo com a literatura, ramificando-se nos seguintes objetivos específicos: a) Revisão de literatura acerca da indexação, seu processo de automação e seus critérios de qualidade; b) Revisão de literatura referente aos *softwares* de análise textual, especialmente de processamento da linguagem natural, e seus critérios de qualidade; c) Sistematização dos critérios para a avaliação dos *softwares* de análise textual, tendo em vista o seu uso na avaliação e seleção de sistemas para a automatização da indexação.

Finalmente, a pesquisa é apresentada da seguinte maneira: 1) Introdução do problema de pesquisa, objetivos e justificativa; 2) referencial teórico fundamentais à investigação, abordando a organização da informação, a pragmática, as relações entre pragmática e ciência da informação, e o processamento da linguagem natural; 3) procedimentos metodológicos, que inclui o tipo de pesquisa, universo e amostra de análise, técnicas de coleta de dados e técnicas de análise dos resultados; 4) resultados, englobando a apresentação dos *softwares* e análise de suas funções, suas bases pragmático-linguísticas e a discussão dos resultados; 5) considerações finais; 6) referências da bibliografia utilizada; 7) apêndice de pesquisa relacionada que investiga as contribuições de critérios de qualidade de *softwares* de processamento da linguagem natural à automação da indexação.

2. REFERENCIAL TEÓRICO

Nesta seção, apresentam-se os principais conceitos e autores que fundamentam teoricamente a pesquisa. Inicialmente, apresenta-se um apanhado sobre a área de organização da informação e suas atividades, com destaque ao tratamento temático da informação. Logo após, apresentam-se definições de pragmática, junto a uma síntese de três de suas principais linhas de estudo, e uma subseção para as contribuições da pragmática identificadas no arcabouço teórico-metodológico da ciência da informação. Por fim, tem-se uma seção que trata exclusivamente do processamento da linguagem natural.

2.1. Organização da informação

A organização do conhecimento surge como um dos marcos conceituais característicos da ciência da informação que se dedica à estruturação, sistematização e representação do conhecimento produzido e registrado visando sua acessibilidade (Oliveira; Martínez-Ávila, 2019). Fundamentalmente, as atividades da organização do conhecimento, juntamente com sua representação, objetivam o acesso do conhecimento materializado em documentos em centros informacionais, e a recuperação de seu conteúdo por meio de catálogos e índices de assunto.

De acordo com Guimarães *et al*. (2015), a organização do conhecimento proporciona as bases para que a informação seja organizada a partir dos processos de representação empreendidos na organização da informação, mais especificamente o tratamento temático da informação.

Bräscher e Café (2008) se dedicaram à distinção entre os processos de organização do conhecimento (OC) e a organização da informação (OI), que se caracteriza como a descrição dos aspectos físicos e de conteúdo da informação através das atividades próprias da representação da informação, que buscam, por meio de atributos descritivos, gerar produtos informacionais - e representacionais - de determinado recurso informacional. Assim, a organização da informação conceitua-se como processo, enquanto a representação da informação se apresenta como produto desta operação.

Por ser direcionada às descrições materiais e de conteúdo de objetos informacionais, pode-se dizer que a organização da informação se dedica exclusivamente ao “mundo dos objetos físicos”, isto é, tem como objeto a informação registrada, o qual procura representar e organizar, enquanto a organização do

conhecimento busca sistematizar ideias e conceitos para que sua essência informacional perdure (Bräscher; Café, 2008).

Dessa forma, pode-se dizer que a organização da informação, como descrita por Oliveira (2021, p. 89), é a "[…] ação prática orquestrada com o fim de constituir arranjos, categorizar e individualizar conjuntos informacionais, com vistas a viabilizar o acesso ao conhecimento neles contido", realizada a partir de um "[…] olhar analítico sobre os documentos para gerar representações contextualizadas", apoiando a busca, recuperação e uso dessas informações (Oliveira, 2021, p. 89; Kobashi, 2017).

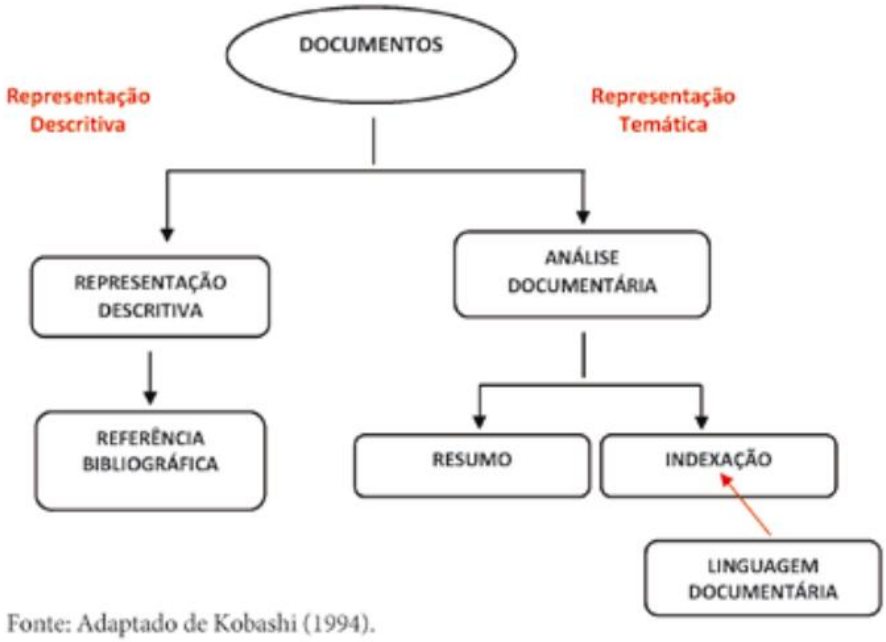
Assim, afirma-se que a organização da informação busca tratar as informações presentes nos conteúdos documentais, através do tratamento da informação, por representações de forma e conteúdo que viabilizem a encontrabilidade informacional. Esse campo se responsabiliza pelas atividades e processos de tratamento e organização da informação registrada nos mais diversos suportes, seja de sua forma física, como a catalogação descritiva, ou de seu conteúdo, como a indexação e classificação (Café; Sales, 2010).

Dias e Naves (2007) entendem o tratamento da informação como o conjunto de disciplinas, processos, métodos e técnicas de descrição física e de conteúdo de documentos e itens bibliográficos em centros informacionais, como bibliotecas e sistemas de informação, e desenvolve e se utiliza de instrumentos, como normas, códigos, padrões e linguagens, em suas descrições e estruturas bases de dados para o armazenamento de documentos e representações (p. 17).

Assim, Dias e Naves (2007 *apud.* Redigolo, 2014, p. 31) determinam que o tratamento da informação "[...] compreende as disciplinas de classificação, catalogação, indexação, bem como especialidades delas derivadas, ou terminologias novas nelas aplicadas, tais como metadados, e ontologias, entre outras".

O tratamento da informação se desdobra entre tratamento descritivo de forma e o tratamento temático da informação (TTI) que cumprem diferentes funções na cadeia informacional, como representado na figura 1.

Figura 1 - Processos de representação documental



Fonte: Maimone, Kobashi e Mota (2016), adaptado de Kobashi (1994).

Enquanto o tratamento descritivo descreve o aspecto de forma dos documentos de maneira a individualizá-los para sua identificação, localização e recuperação, o tratamento temático ou de assunto, segundo Guimarães (2008, 2009), constitui-se como um dos domínios formadores da organização da informação que procura proporcionar possibilidades de acesso à informação, e diz respeito às técnicas e processo de representação do conteúdo informacional em assuntos, sendo então o complemento da descrição física da informação materializada. Portanto, caracteriza-se como processos, instrumentos e produtos de tratamento do conteúdo em dois aspectos principais: a determinação do assunto e sua especificação (Oliveira; Martínez-Ávila, 2019; Bräscher; Guimarães, 2018).

Metodologicamente o tratamento descritivo é um processo mais estruturado, justamente por possuir códigos de catalogação, como o Código de Catalogação Anglo Americano 2° edição revisada (AACR2R), que dão bases e norteiam o trabalho do catalogador de maneira normalizada, e os padrões de metadados para a descrição eletrônica de objetos informacionais, como o *Dublin Core* e o Marc 21 (Dias; Naves, 2007; Redigolo, 2014).

Enquanto isso, os instrumentos do tratamento temático de conteúdo, como as linguagens documentárias, não são padronizadas e muito menos utilizadas, sendo instrumentos “acessórios” para o tratamento da informação, e muitas vezes desconsideradas durante o processo, embora sejam instrumentos de normalização

terminológica e cumprem um papel imprescindível na representação temática de documentos. Além destas, o tratamento temático, segundo Redigolo (2010, 2014), tem como instrumento normatizador os manuais de indexação, embora não sejam desenvolvidos e utilizados por todas as instituições que realizam o tratamento temático.

Para Kochani, Boccato e Rubi (2012), o tratamento temático da informação é a representação do conteúdo de documentos informacionais que se realiza a partir da análise documental, enquanto para Guimarães, Ferreira e Freitas (2012), é uma sequência formada pela análise, descrição e representação de conteúdos documentais com fins de recuperação. Já Oliveira (2021, p. 109) a define do seguinte modo:

> […] o TTI é desenvolvido de forma corriqueira por meio de ações biblioteconômicas pontuais, conjugando esforços da organização da informação e do conhecimento em prol desse intento. Estas, se voltam à representação do assunto, empregando, para tanto, instrumentos específicos e, gerando a partir disso, produtos característicos.

Pinto Molina, Garcia Marco e Agustín Lacruz (2002), destacam que, uma das principais dificuldades para a extração do conteúdo para o tratamento temático é “[...] compreender primeiro a mensagem, é dizer, o significado e o sentido da mensagem, e esse é um conceito que raramente se define” (p. 106).

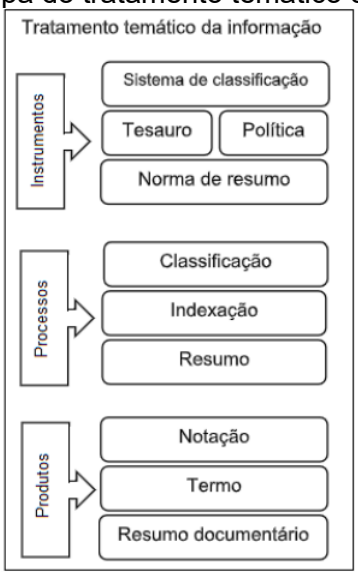
Esta atividade de processamento dos conteúdos informacionais, de acordo com Varela e Barbosa (2007), direciona a informação ao usuário, conferindo ao TTI um papel mediador, conectando a produção e o uso da informação, propiciando acessibilidade temática à informação (Guimarães; Ferreira; Freitas, 2012). Para possibilitar essa “acessibilidade temática”, o TTI conta com as atividades de classificação, elaboração de resumos, a catalogação de assuntos e a indexação. Além destes processos, o TTI ainda se apoia em instrumentos que viabilizam a criação de seus produtos, isto é, as notações classificatórias, os termos descritores e os resumos documentários.

Entre seus instrumentos auxiliares, pode-se destacar as linguagens documentárias, como os tesauros e vocabulários controlados, para a extração de termos de indexação; os sistemas de classificação bibliográfica para o desenvolvimento das notações classificatória; e a política de indexação e normas de resumo para a orientação da produção dos produtos documentais. Como base para suas atividades e para produção de seus produtos, o tratamento temático faz uso das

linguagens de indexação, normas para a elaboração de resumos, e manuais de indexação, de acordo com Dias e Naves (2013) e Oliveira (2021).

A relação entre processos, instrumentos e produtos no tratamento temático da informação foi sistematizado por Oliveira, Grácio e Martínez-Ávila (2020) e replicado por Oliveira (2021) na figura 2.

Figura 2 - Mapa do tratamento temático da informação



Fonte: Oliveira, Grácio e Martínez-Ávila (2020).

Café e Sales (2010) citam que, na literatura, o tratamento temático recebe denominações diversas, como análise temática, análise de assunto, análise documentária e descrição de conteúdo. A variação terminológica, aponta Guimarães e Sales (2010), é decorrente da evolução diversa das correntes teóricas do tratamento temático. Nesse sentido, Oliveira (2021) salienta que a variação terminológica identificada ao falar de TTI diz respeito, sobretudo, ao desenvolvimento epistemológico plural que este campo teve, especialmente a partir de teorias inglesas, francesas e norte-americanas. Assim, a expressão tratamento temático da informação também é designada como análise documentária, análise de assunto, análise temática, descrição temática, descrição de conteúdo, representação temática e organização temática.

A classificação, segundo Vieira (2014, p. 68) é o "[...] ato mental que visa entender ou relacionar coisas e ideias", podendo ser também a divisão em grupos e classes, ou dividir com base em semelhanças e diferenças. Ranganathan, ainda de acordo com Vieira (2014), define a classificação como "[...] a tradução dos nomes dados aos assuntos dos documentos da linguagem natural (título, autor, etc.) para a

linguagem artificial (letras, números e sinais) usadas nos sistemas de classificação bibliográfica".

Norte (2010) delimita a classificação como processo que reúne documentos com base em suas semelhanças ou diferenças a partir de um sistema, ou ainda a organização física dos livros e materiais bibliográficos de forma que as áreas, temas e assuntos que dizem respeito fiquem reunidos nas estantes. Cunha e Cavalcanti (2008, p. 84), a classificação, no contexto das atividades biblioteconômicas, define-se como o conjunto de processos que "[...] levam à colocação de um documento em uma determinada ordem" apoiando-se no uso de sistemas de classificação bibliográfica predeterminados.

Considerando que a atividade classificadora está "[...] fundamentada em semelhanças e contrastes", como declara Albuquerque (2015, p. 4), esta pode ser caracterizada por "reunir a síntese e análise", pois ao mesmo tempo separa e agrupa elementos em conjuntos (Gil, 2001).

De acordo com Kédrov (1974), a classificação das ciências diz respeito a sistematização e organização do conhecimento humano, reafirmando os vínculos entre as ciências, seus objetivos e fenômenos, unindo todo o conhecimento em um único sistema geral. Para isso, o desenvolvimento de propostas de classificações tornou-se um "hobby", como afirma Dahlberg (1979), influenciando também bibliotecários e documentalistas para a elaboração de sistemas como tentativas de organização do conhecimento.

Partindo de um ponto de vista biblioteconômico, San segundo Manuel (1996) entende a classificação como o agrupamento de documentos segundo seu conteúdo, o que resulta em esquemas de grupos nos campos do conhecimento para o compartilhamento conceitual entre eles.

Os sistemas de classificação bibliográficas em si surgem no final do século XIX e complementam as atividades de organização do conhecimento através de representações taxonômicas numeradas em ramos e subsidiários, e as atividades de organização da informação através do agrupamento de documentos com base em seu assunto ou conteúdo temático (Batley, 2005).

Burke (2003) aponta as bibliotecas, currículos e enciclopédias como os principais disseminadores dos modelos de classificação, embora a popularização dos sistemas de informação e das tecnologias da informação, surgem instrumentos que

complementam os sistemas de classificação, como as taxonomias, tesauros, ontologias e listas de cabeçalhos de assunto, por exemplo.

Com base em seus objetivos, os sistemas de classificação bibliográficas podem ser gerais, abrangendo todos os ramos do conhecimento, como a Classificação Decimal de Dewey e a Classificação Decimal Universal, ou especializados, abrangendo um ramo do conhecimento ou um tipo de material (Barbosa, 1969).

De toda forma, todas classificação bibliográfica têm como finalidade "[...] promover uma classificação sistemática, lógica, que reflita crítica e filosoficamente sobre os elementos de ligação que servem para a reunião de conceitos", como afirma Araújo (2006) e seu objetivo é a ordenação temática de unidades informacionais, promovendo a elaboração de bibliografias e catálogos para a recuperação informação (San Segundo Manuel, 1996).

Entre os principais sistemas de classificação bibliográficas estão a Classificação Decimal de Dewey, uma classificação enumerativa com indicações de assuntos por meio de símbolos, a Classificação Decimal Universal de Otlet e La Fontaine, que possibilita a criação de subsistemas secundários não previstos no sistema e a representação por meio de tabelas e símbolos, a Classificação Multifacetada de Ranganathan, que organiza o mesmo assunto em subclasses que caracterizam seus diversos aspectos, e a Classificação da Biblioteca do Congresso, no qual os assuntos são representados alfabeticamente (Albuquerque, 2015).

Já o processo de indexação, segundo Norte (2010, p. 24), é formado pelas atividades de identificação e descrição do conteúdo de um documento por meio de descritores - termos representativos do assunto - de uma linguagem de indexação. Para isso, o indexador faz uma leitura - leitura documentária - para a determinação de conteúdo do texto, que dá base para a seleção ou atribuição de termos descritores como forma de facilitar a recuperação do material (Reitz, 2004, p. 354).

De acordo com Maimone, Kobashi e Mota, caracteriza-se a indexação como

> [...] um conjunto de ações que visam identificar e descrever conteúdos de documentos, de acordo com seus assuntos, por meio da "atribuição" de termos de uma linguagem de indexação, também chamada de linguagem documentária (2016, p. 75).

Também denominada como análise de conteúdo e análise documentária, a indexação, de acordo com Cunha e Cavalcanti (2008, p. 193), diz respeito à representação do conteúdo temático de documentos através de termos próprios do documento - por extração - ou por meio de uma linguagem documentária - por

atribuição - predeterminada. De acordo com a Associação Brasileira de Normas Técnicas (ABNT), na NBR 12676 de 1992, a indexação diz respeito à identificação e descrição do conteúdo documental por meio de termos de representação de seu assunto que constituem uma linguagem de indexação.

De acordo com Fujita (2003), a indexação tem bases teóricas e metodológicas na análise de assunto de linha inglesa e na catalogação de assunto de linha norte-americana, assim, é compreendida como processo realizado nos serviços de informação, enquanto a catalogação de assunto se limitou à organização das bibliotecas.

É possível citar três marcos conceituais que justificam essa pluralidade terminológica: uma linha norte-americana (*subject cataloguing*) de abordagem pragmática, uma linha inglesa (*indexing*) de abordagem teórica, e uma linha francesa (*analyse documentaire*) de abordagem processual. Essa sistematização também pode identificar duas correntes principais: corrente anglo-saxônica, focada nos instrumentos e produtos (catalogação de assunto e indexação), e uma corrente francesa, focada nos elementos teóricos e metodológicos (análise documental), como aponta Oliveira (2021).

Segundo Redigolo (2010), a análise de assunto sofre com a carência profissional e metodológica que leva à irregularidade de seus processos, causada principalmente pela falta de metodologias desenvolvidas e atualizadas para a análise e a dificuldade na identificação de conceitos para a posterior seleção dos termos de representação. Além disso, a autora ainda destaca os principais obstáculos para a análise de assunto, sendo eles: a falta de tempo para a realização da leitura e da análise dos textos, a diversidade de temas e assuntos dos itens analisados, a falta de parâmetros metodológicos e o não uso de linguagens controladas.

Nesse sentido, a indexação e catalogação de assunto podem ser entendidas como processos análogos e correspondentes, por estarem ambas relacionadas à determinação de assunto de documentos e itens bibliográficos, embora sejam realizadas em ambientes diferentes, com a indexação em centros de informação mais especializados e a catalogação de assunto em bibliotecas públicas, escolares e universitárias (Milstead, 1983; Redigolo, 2014).

A indexação como processo, segundo Lancaster (2004a), entre outros autores, realiza-se a partir das etapas de análise conceitual e tradução. A análise conceitual

busca determinar ao que se refere o conteúdo do documento para a representação do assunto de um documento (Dias; Naves, 2007). A tradução, segundo Dias e Naves (2007), faz uso de instrumentos, como as linguagens documentárias, para a representação do assunto determinado na análise conceitual a partir de termos de indexação.

Enquanto isso, autores como ABNT (1992), Dias e Naves (2007) e Redigolo (2010) definem três etapas para a análise de assunto, sendo elas a leitura documental, a identificação de conceitos e a seleção/tradução de conceitos.

A leitura documental, diferente da leitura sequencial e integral de um documento, constitui uma leitura profissional ou documental, que analisa partes estratégicas do texto para a identificação dos conceitos principais do conteúdo documental.

No campo profissional, a leitura pode ser compreendida a partir das variáveis leitor, texto e contexto, conforme Giasson (1993)**.** A variável leitor diz respeito ao conhecimento prévio do leitor profissional em relação ao assunto do documento, seu conhecimento linguístico e dos gêneros textuais, e se relaciona diretamente com seu contexto social, psicológico e ambiental, com a estrutura do texto e suas possibilidades de interpretação.

Fujita, Nardi e Santos (1998), assim como Fujita (2004), declaram que a leitura, além de ser influenciável pelo contexto cultural e ideológico que a cerca, também é capaz de influenciá-lo, uma vez que [...] a interpretação de um texto depende do processamento das informações, do próprio texto de partida e do contexto de ambos (Maiome; Kobashi; Mota, 2016, p. 76).

Já a identificação de conceitos se realiza inicialmente com a leitura realizada e a compreensão dos elementos analisados, que leva a identificação e seleção dos conceitos principais do conteúdo.

Como guia para este processo a norma NBR 12676 (1992) desenvolve questões que facilitam a identificação dos conceitos de documentos, sendo eles: a) qual o assunto de que trata o documento?; b) como se define o assunto em termos de teorias, hipóteses etc.?; c) o assunto contém uma ação, uma operação, um processo?; d) o documento trata do agente dessa ação, operação, processo etc.?; e) o documento se refere a métodos, técnicas e instrumentos especiais?; f) esses aspectos foram considerados no contexto de um local ou ambiente especial?; g) foram identificadas

variáveis dependentes ou independentes?; h) o assunto foi considerado sob um ponto de vista interdisciplinar?

Com a mesma finalidade, Fujita (2003) desenvolve o modelo de leitura documentária para a indexação de textos científicos, que indica, além dos questionamentos de identificação dos contextos, as partes da estrutura textual que podem respondê-los, descrito no quadro 1.

Quadro 1 - Modelo de Leitura para indexação de textos científicos

CONCEITO	QUESTIONAMENTO	PARTES DA ESTRUTURA TEXTUAL	TERMOS IDENTIFICADOS
OBJETO (algo ou alguém que está sob estudo do pesquisador)	O documento possui em seu contexto um objeto sob efeito desta ação?	INTRODUÇÃO (OBJETIVOS) (Confirmar tema em referências)	
AÇÃO (processo sofrido por algo ou alguém)	O assunto contém uma ação (podendo significar uma operação, um processo etc.)?	INTRODUÇÃO (OBJETIVOS) (Confirmar tema em referências)	
AGENTE (aquele ou algo que realizou a ação)	O documento possui um agente que praticou esta ação?	INTRODUÇÃO (OBJETIVOS) (Confirmar tema em referências)	
MÉTODOS DO AGENTE (métodos utilizados para realização da pesquisa)	Para estudo do objeto ou implementação da ação, o documento cita e/ou descreve modos específicos, por exemplo: instrumentos especiais, técnicas, métodos, materiais e equipamentos?	METODOLOGIA	
LOCAL OU AMBIÊNCIA (local físico onde foi realizada a pesquisa)	Todos estes fatores são considerados no contexto de um lugar específico ou ambiente?	METODOLOGIA	
CAUSA E EFEITO Causa (ação+objeto)/Efeito	Considerando que a ação e o objeto identificam uma causa, qual é o efeito desta causa?	RESULTADOS; DISCUSSÃO DE RESULTADOS; CONCLUSÕES	

Fonte: Fujita, 2004.

Por fim, a seleção/tradução de conceitos tem em vista o objetivo da indexação, o público-alvo da indexação e de acordo com a política de indexação da instituição.

Assim, nem todos os conceitos identificados serão necessariamente selecionados (Fujita, 2003).

Para a tradução dos conceitos em termos de indexação utilizam-se linguagens documentárias como forma de normalização e padronização da representação, cuja principal função é “[...] compatibilizar a linguagem utilizada por uma comunidade de usuários e entre várias instituições” (Maimone; Kobashi; Mota, 2016, p. 77; Chaumier, 1988).

Hjørland (2003), assim como Redigolo (2010 e 2014), destaca que determinar o assunto de um documento não constitui uma atividade objetiva ou neutra, sendo influenciado por diferentes visões, interesses e fatores psicológicos e contextuais. Assim, uma análise de assunto ideal depende de condições específicas de conhecimentos prévios, questões psicológicas, ambiente de trabalho, instrumentos e estratégias.

Assim, o processo de indexação, realizada manualmente por indexadores profissionais, tem como uma das principais limitações a subjetividade de suas etapas, afetando diretamente a uniformidade da atividade, principalmente durante a análise do conteúdo, e a consistência na atribuição dos termos de indexação (Pinheiro, 1978).

A partir desta perspectiva, a partir da década de 1950 surge a primeira tentativa de automatização da indexação a partir da proposta de Luhn (1957 e 1958) que a recuperação de um documento, para ser efetiva, deveria basear-se em uma análise de conteúdo, contemplando seu vocabulário a partir do índice KWIC (*KeyWord in Context*), que seleciona termos representativos para um documento a partir de palavras do título.

Correa e Lapa (2013) caracterizam a face automatizada da indexação como processo de natureza estatística e/ou linguística realizado por máquinas e computadores para a seleção e extração de termos representativos descritivos de um documento a partir do processamento de seu conteúdo.

Assim, a indexação automatizada surge como uma alternativa ou recurso tecnológico que objetiva auxiliar indexadores e profissionais da informação para o tratamento temático de documentos de forma mais uniforme e prática (Silva; Correa; Gil Leiva, 2020). Para Lancaster (2004b), a indexação automática e/ou semiautomática serve de auxílio aos profissionais indexadores, automatizando partes repetitivas e altamente subjetivas em um trabalho mais dinâmico e sistemático,

tornando-se uma alternativa para lidar com as inconsistências e subjetividades que cercam a análise de assunto.

Além disso, torna-se importante ressaltar a necessidade das intervenções de profissionais da informação e indexadores para o uso mais adequado e ético de máquinas no processo de análise de assunto, como afirma Narukawa, Gil Leiva e Fujita (2009), permitindo a validação profissional dos termos selecionados pelo sistema de indexação.

Já a atividade de elaboração de resumos diz respeito à representação condensada de um documento que possibilita a recuperação da informação de forma facilitada. Segundo Kobashi (1994), textos completos são formados por repetições e redundâncias, possibilitando sua sintetização sem a descaracterização da informação central.

Os resumos documentários são os produtos deste processo de síntese e previnem a leitura integral de textos não pertinentes à busca do usuário. Assim, a função principal dos resumos documentários - diferente dos resumos produzidos pela mídia e por alunos da educação básica - é disponibilizar as informações principais sobre o conteúdo documental, seja ele um resumo informativo ou indicativo (Kobashi, 1994).

Como processo, a elaboração de resumos tem como produto de seus processos o resumo documentário, caracterizado como representação condensada de um texto e/ou documento com fins de organização e recuperação. Segundo Kobashi (1994), o fato de os textos integrais possuírem redundâncias, reiterações e informações acessórias, isto é, não essenciais, permite que o resumo documentário seja uma sintetização paralela, e até mesmo substituta, do documento original, sem comprometer seu sentido e informações essenciais.

Como produto, os resumos documentais são essenciais nos processos de divulgação, disseminação e comunicação técnico-científica, pois permitem buscas mais rápidas e eficazes, servindo, muitas vezes, como substituto do texto integral, evitando assim a leitura integral de documentos irrelevantes ao motivo da busca inicial do usuário da informação. Para isso, Medeiros (2018) declara a necessidade de o resumo ser portador de significado e sentido equivalentes ao texto integral a ser representado, principalmente ao adotar os aspectos de organização textual do texto base (Medeiros, 2015).

A elaboração de resumos, segundo Leitão e Simões (2017), deve fundamentar-se no seu público-alvo, ou seja, os usuários a quem se destina, já que os resumos documentários cumprem um papel intermediário entre o documento original e integral e um determinado público, que podem ser estudantes, pesquisadores, professores, até mesmo usuários em conhecimentos prévios de pesquisa e do assunto. Sendo assim, são pontes que conectam o conhecimento científico à sociedade.

Kobashi (1994) propõe um modelo para a análise textual na elaboração de resumos de acordo com sua natureza (informativo ou indicativo), compreendendo a seleção de ideais representativas do conteúdo do texto e sua combinação.

De acordo com a NBR 6028 (ABNT, 2021), norma de elaboração de resumos, resenhas e recensões, o resumo indicativa apenas indica pontos principais de um texto sem muito detalhamento, não dispensando a consulta do documento original, enquanto o resumo informativo informa o tema, objetivo, metodologia, resultado e conclusões sobre o documento, servindo até mesmo como substituto ao texto original.

Assim, a elaboração de resumos se inicia com a identificação do tema do texto, a identificação de suas informações, a seleção das informações essenciais ou relevantes, e a representação das informações.

Para textos canônicos (quadro 2) de ciências experimentais foca-se no tema, ou assunto, o problema ou lacuna estudado, as hipóteses ou conjectura, a metodologia utilizada no estudo, os resultados, com os fatos descobertos e suas interpretações, e as conclusões finais.

Quadro 2 - Modelo de resumo de texto canônico

Texto Canônico		Resumo indicativo	Resumo informativo
Tema		X	
Problema	Indagação	X	X
Hipótese	Conjectura		X
Metodologia	Observação		X
Resultado	Interpretação		X
Conclusão	Comentário final	X	X

Fonte: Kobashi (1997).

Para o texto dissertativo argumentativo (quadro 3) extraem-se a tese, ou ponto de vista apresentado no texto, os argumentos que comprovam a tese do autor, e a conclusão, confirmando o ponto de vista inicial.

Quadro 3 - Modelo de resumo de texto dissertativo argumentativo

Texto dissertativo argumentativo		Resumo indicativo	Resumo informativo
Tema		X	X
Tese	Ponto de vista		X
Argumentos	Provas		X
Conclusão	Confirmação	X	X

Fonte: Kobashi (1997).

No texto expositivo (quadro 4) são três categorias essenciais, sendo elas o problema ou questão que guia a pesquisa, as causas do problema estudado, e a solução, ou resposta, para o fenômeno pesquisado.

Quadro 4 - Modelo de resumo de texto expositivo

Texto expositivo		Resumo indicativo	Resumo informativo
Tema		X	X
Problema	Questão		X
Causas	Razões		X
Solução	Resposta	X	X

Fonte: Kobashi (1997).

Com base na metodologia apresentada, Kobashi (1997) sistematiza as correspondências entre os tipos de texto e suas partes essenciais que devem estar em seus resumos, representadas no quadro 5.

Quadro 5 - Correspondência entre as partes dos tipos de texto

Texto canônico	Texto dissertativo	Texto expositivo
Tema	Tema	Tema
Problema Hipótese	Tese	Problema
Metodologia Resultados	Argumento	Causas
Conclusão	Conclusão	Solução

Fonte: Kobashi (1997).

Em relação à qualidade dos resumos, Lancaster (2004a) caracteriza um bom resumo a partir de três pontos principais: brevidade, exatidão e a clareza. Medeiros (2018), por sua vez, faz algumas recomendações quanto a produção de resumos documentários, sendo elas: 1) identificar e apresentar o tema no início do texto; 2) identificar e apresentar também outras informações essenciais, como metodologia, resultados e considerações finais do texto base; 3) manter a objetividade, clareza, coesão e coerência na redação do resumo; 4) evitar transcrições e fragmentação do texto original.

Já Kobashi (1995), objetivamente, estabelece princípios como garantia da qualidade de resumos tendo em vista as máximas conversacionais de Greice. Dessa forma, um resumo documentário que cumpre com suas funções primordiais, sendo necessário: concisão, com informações suficientes; pertinência, sendo fiel ao texto original representação; precisão, suprimindo ambiguidades; objetividade, sem julgamentos de valor.

Assim, pode-se dizer que a organização da informação é a disciplina da ciência da informação que estuda e busca avanços teóricos, metodológicos e conceituais que visem a melhor eficiência nos processos de representação, sistematização e recuperação da informação. Entre suas atividades tem-se o tratamento temático da informação, que abrange operações de análise e representação de conteúdo informacional para a criação de produtos informacionais com o apoio de instrumentos, como as linguagens documentárias, proporcionando maior eficiência aos processos do ciclo informacional.

2.2. Pragmática

A linguística, área responsável pelo estudo do “[...] estado científico da linguagem e das línguas naturais”, como sua “[...] natureza, funcionamento e procedimentos de descrição das línguas naturais” (Greimas; Courtés, 2008, p. 291), investiga aspectos da linguagem como meio de comunicação ou expressão individual, e até mesmo a linguagem como “herança” de uma comunidade discursiva (Hartmann; Stork, 1972).

Tendo como objeto de estudo uma língua, uma família de línguas ou as línguas de modo geral - necessariamente humanas, de acordo com Camara Jr. (2009) -, a linguística busca compreender os “[...] princípios fundamentais que regem a

organização e o funcionamento da faculdade da linguagem entre os homens” (p. 197-198).

Considerando os diversos aspectos formadores de uma língua, os estudos linguísticos de ramificam fundamentalmente em sintaxe, semântica e pragmática. A sintaxe é determinada como o estudo da estrutura de sentenças (Trask, 2004, p. 272-273), ou uma das partes que compõem os estudos da gramática, sendo a outra a morfologia, que estuda as unidades formadoras das sentenças, a sintaxe estabelece suas relações e regras de construção (Greimas e Courtés, 2008, p. 471).

Já a semântica investiga o significado (Trask, 2004, p. 261) ou domínio que tem como objeto “[...] o sentido e as interpretações das unidades significativas da língua e de sua combinação no discurso” (Neveu, 2008, p. 269).

O campo da pragmática analisa a relação forjada entre os locutores de uma situação comunicativa e os signos escolhidos e utilizados por eles, o que lhes atribui significado que, diferente do significado semântico, resultará da interação entre o contexto da situação comunicativa e a escolha das formas de expressão escolhida pelo enunciador.

Na linguística, a pragmática foi aplicada para o estudo da língua a partir da perspectiva dos indivíduos que dela fazem uso, abordando questões como escolhas linguísticas realizadas pelos falantes, o efeito destas escolhas sobre os outros sujeitos participantes da comunicação, e até as restrições do uso da língua em interações determinadas (Crystal, 1985, p. 205-206).

Sob uma lente semiótica, a pragmática, segundo Morris (1938) é uma das três relações entre as dimensões da semiose: relação entre os signos, que forma a dimensão sintática, relação entre os signos e os objetos a que se referem, que forma a dimensão semântica; e a relação entre os signos e seus intérpretes, que forma a dimensão pragmática.

Nessa linha, também se define pragmática como a parte da semiótica que estuda as relações entre os signos e seus usuários, o que ultrapassa os limites do campo linguístico e atinge o campo real do ser humano e sua relação e interações com elementos da realidade, ou seja, como os objetos, seres e fenômenos afetam os usuários de uma linguagem (Armengaud, 2006).

O domínio pragmático das ciências da linguagem, para Neveu (2008, p.2 44), é constituído a partir da convergência de disciplinas, “[...] sobretudo as linguísticas enunciativas, a sociolinguística, a psicolinguística, a semântica de textos, a semiótica,

a análise da conversação, as ciências da comunicação, ou as ciências cognitivas", e estuda o uso da língua, não seu sistema constitutivo, e o sentido.

Assim, a pragmática, o terceiro elemento constitutivo dos estudos da linguagem, potencializa e complementa os trabalhos já realizados na sintaxe e na semântica a partir de investigações sobre a forma de comunicação dos significados de um enunciado em um contexto determinado (Trask, 2004, p. 232).

Diferente da semântica, os estudos dos significados empreendidos na pragmática têm como ponto diferencial a consideração de um contexto extralinguístico que condiciona o enunciado, isto é, "[...] toda vez que o contexto muda, também muda o que é comunicado" (Trask, 2004, p. 233).

Por existir uma pluralidade de tentativas de definição da pragmática a partir das mais diversas áreas e abordagens - filosóficas, linguísticas, semióticas, lógicas etc. -, como sistematiza Armengaud (2006), é possível informar a não existência de uma visão unificada do que é a pragmática, já que alguns a compreendem como "práxis", ou uma integração entre teoria da ação e comportamento linguístico, ou até mesmo como o uso de signos em contextos (p. 17).

Assim como Armengaud (2006, p. 20), esta seção buscou introduzir as investigações pragmáticas, suas teorias, teóricos e conceitos "sem apresentar uma teoria original da pragmática, nem tentar uma impossível síntese", mas sim "oferecer algumas referências úteis" para a compreensão geral e inicial de um tema tão amplo e facetado.

Assim, apresenta-se respectivamente a semântica desenvolvida por Gottlob Frege com foco em seus conceitos de sentido e referência, a tripartição semiótica proposta por Charles Morris e sua dimensão pragmática, e, por fim, a teoria dos atos de fala introduzida por Austin e Searle.

A escolha destes autores para o desenvolvimento da seção foi realizada de forma a sintetizar algumas das principais perspectivas de estudos pragmáticos, apresentando também uma visão geral de um campo de estudos vasto e diversificado.

O trabalho de Frege dá bases ao aprofundamento do aspecto pragmático a partir de uma visão filosófica, desenvolvendo também as noções de sentido e referência que contribuem também para a relevância dos contextos de uso como

fatores determinantes na atribuição de significado e sentido de expressões e proposições lógicas.

Charles Morris se destaca pelo desenvolvimento de uma teoria geral dos signos e por sua tripartição semiótica, realçando a irredutibilidade da pragmática e suas relações indissociáveis com as esferas sintática e semântica da semiose, o que confere um papel fundamental aos contextos no que diz respeito à interpretação sígnica e significação.

Por fim, a introdução dos atos de fala de Austin e Searle parte do pressuposto do dizer como fazer, permitindo entender os atos realizados pelos intérpretes ao se comunicarem através de signos, relacionando assim uso, contexto, ação, signos e significação à comunicação de sentido.

2.2.1. Frege

Friedrich Gottlob Frege, matemático, filósofo e lógico alemão, define como objeto primeiro da lógica o estudo da verdade a partir do reconhecimento das leis do pensamento - que determinam o verdadeiro naturalmente - que a constituem. Seu trabalho influenciou o trabalho de autores como Ludwig Wittgenstein, Bertrand Russell e Edmund Husserl (Martínez Vásquez, 2020, p. 3).

Em sua análise lógica, Frege substitui a divisão tradicional da oração em sujeito e predicado pela sua dissolução em função e argumento - que depois passa a denominar-se conceito e objeto -, pois “[...] distinções como ativo ou passivo, sujeito e predicado são chamadas por Fregue de não-pertinentes quando se levam em conta a verdade e a falsidade” (Armengaud, 2006, p. 31-32).

Assim, a verdade ou falsidade do pensamento que será julgado deverá ser definida, pois “[...] apenas a combinação de função e argumento pode nos fornece um conteúdo judiciável”, que mais tarde foi denominado Pensamento (Wischin, 2012, p. 95, tradução nossa; Pitt, 2018).

A função seria, em uma definição ainda vaga, o sentido insaturado do pensamento, enquanto o argumento, que pode sofrer substituições, atribui os diferentes valores de verdade a uma oração, e o objeto engloba as entidades que podem ser um argumento e que não é uma função.

Em seu trabalho *Conceitografia: uma linguagem formular do pensamento puro decalcada sobre a aritmética* (1879), Frege deixa de utilizar-se da linguagem natural para a expressão matemática e descreve uma linguagem com bases lógicas para que

a forma de expressar as relações presentes entre o conteúdo judiciável, e que preservasse a verdade nas inferência matemáticas “[...] sem deixar lacunas e sem apelo à intuição kantiana” (Pitt, 2018, p. 331).

Enquanto a linguagem comum se caracteriza se preocupa com o êxito da comunicação, a linguagem lógica, ou científica, concentra seus esforços na determinação da verdade (Armengaud, 2006).

A linguagem comum também se define a partir da sua ambiguidade, sua multiplicidade de expressões e sua possibilidade de deixar brechas, enquanto a linguagem científica deve contar com sua univocidade, sua precisão nas expressões e clareza lógica (Armengaud, 2006).

Por fim, a linguagem lógica de Frege também conta com a impessoalidade de suas construções, enquanto na linguagem natural seu foco está na emoção, no convencer, impressionar e interessar (Armengaud, 2006, p. 31).

A partir do método de provas axiomático e o uso de axiomas lógicos para a eliminação de lacunas, essa nova linguagem seria capaz de atender aos objetivos de Frege, contando também com regras sintáticas para o uso do alfabeto e a expressão a partir de regras de inferências - para a derivação de novas expressões - e axiomas, de forma que “[...] tudo que é necessário para uma inferência correta será totalmente expressado, [...], nada será deixado à adivinhação” (Frege, 1997, p. 54 *apud* Pitt, 2018, p. 331).

Em sua obra *Os fundamentos da aritmética* (1884), Frege apresenta um dos aspectos mais importantes de seu trabalho lógico. Neste trabalho será apresentado um número, que para ser um argumento com conteúdo passível de julgamento, antes deve ser um objeto.

Sem a percepção ou imagem mental desse número, será através do sentido da oração, que depende diretamente de seu contexto, que o numeral poderá ser conhecido (Frege, 1988 *apud* Wischin, 2012).

Outra conclusão que impactou em seu programa lógico foi a necessidade de separação do conteúdo judiciável, dando origem às noções de sentido e referência[1], tornando-se termos distintos com papeis próprios dentro do sistema lógico.

[1] Aqui adotou-se o termo “referência” como tradução do original *Bedeutung*, mas Wischin (2012, p. 97-98), considera a abrangência de “significado”, que se refere tanto a *meaning* quanto a *significance*, tal qual ocorre com *Bedeutung*, no alemão. Além do mais, Wischin também aponta a necessária distinção entre o significado atribuído a um signo e o objeto que este denomina. Dessa forma, Wischin opta por “significado” em seu trabalho. Porém,

Assim, referência é compreendida como o objeto denotado pelo termo, enquanto sentido, que pode ser captado por qualquer um com conhecimento o suficiente do idioma, diz respeito a forma que a referência é apresentada.

> as noções fregeanas de sentido e referência estipularam as bases para o aprofundamento na linguagem em seus níveis mais importantes e em suas relações com outras disciplinas filosóficas, de modo que, a partir de sua crítica, tem presenciado um aprofundamento crescente na história do pensamento contemporâneo (Martínez Vásquez, 2020, p. 15).

Wischin (2012, p. 97) sintetiza a ideia mencionando que "[...] o sentido de uma oração inteira determina a referência dos nomes que nela aparecem". Em *Princípios Fundamentais da Aritmética*, Frege aponta que o sentido de uma oração seria a condição necessária para que a referência seja verdadeira, contribuindo para o pensamento que está sendo expresso pela construção linguística.

Abbagnano (1962, p. 858) sintetiza o pensamento de Frege, reforçando a necessidade de vincular o signo a duas coisas, o significado (nesse caso, a referência), e o sentido (do original alemão *Sinn*), responsável pela sua denotação. O termo "sentido" seria atribuído às indicações que funcionassem como nome próprio de um objeto determinado.

Brugger (1962, p. 473) também apresenta uma definição análoga, caracterizando o sentido, como definido por Frege, ao sentido semântico, "[...] a referência indicadora própria de um → sinal, relativamente ao significado ou à sua significação".

A referência se definiria como o objeto de que se fala através de uma expressão/construção linguística, ou seja, de natureza extralinguística, e a sua determinação "[...] frequentemente requer considerações extra literais ou extralinguísticas, isto é, se refere à forma que a referência foi disposta, como apresenta Armengaud (2006, p. 33).

Uma mesma referência pode ser apresentada em duas ou mais expressões diferentes com sentidos diferentes. Um exemplo utilizado por Armengaud (2006)

Mora (1993) aponta que o signo se caracterizaria como qualquer designação que seja um nome próprio e que a referência seja um objeto, isto é, a referência do signo - ou de nomes próprios - é o objeto que por ele é designado. "Um nome próprio (uma palavra, um signo, uma combinação de signos, uma expressão) *exprime* o seu sentido e designa ou reporta-se à sua referência" (Simpson, 1973 *apud* Mora, 1993, p. 624).

seriam as expressões “estrela da tarde” e “estrela da manhã” que, mesmo possuindo sentidos distintos, a referência continua sendo equivalente, Vênus, nesse caso.

Sobre a referência, Mora (1993, p. 623) traz a importante observação de que a obra *Über Sinn und Bedeutung*, de Frege recebeu as mais diversas traduções, como “Sobre sentido e *denotatum*”, “Sobre sentido e denotação”, “Sobre significação e *denotatum*”, “Sobre significação e denotação”, “Sobre sentido e referência” - sendo esse o mais utilizado - e “Sobre significação e referência”. Embora *Bedeutung* seja corretamente traduzida para “significação” ou “significado”, buscando evitar confusões terminológicas, adota-se o termo “referência”[2].

> Na semântica de Frege, além da substituição, a noção de referência desempenha um papel fundamental. Isso ocorre porque o valor de verdade das declarações complexas, construídas por meio de operadores proposicionais e quantificadores, é derivado do valor de verdade das declarações atômicas. A determinação do valor de verdade dessas declarações atômicas depende dos referentes dos termos singulares, já que requer a inserção de termos singulares nos lugares argumentais da declaração atômica (Armero, 1993, p. 143).

As representações, formadas a partir de experiências sensíveis, impressões e ações, tanto internas quanto externas, são variáveis e dependem de quando e quem a idealiza, além de suas experiências vivenciadas. Por se tratar de uma construção mental e puramente humana, as representações, tal qual a referência e suas possibilidades interpretativas, possui, intrínseca a sua existência, a subjetividade, “[...] assim, diante da busca natural do ser humano pela realidade, integra-se a busca pela representação” (Martínez Vásquez, 2020, p. 6, tradução nossa).

Martínez Vásquez (2020) apresenta os seguintes aspectos como características que fazem a distinção entre as representações e os objetos:

> [...] essas representações não podem ser vistas, tocadas, cheiradas, saboreadas ou ouvidas. São impressões que temos, mas que não podemos apreender com os sentidos. Nunca temos uma apreensão sensorial de uma cor, apenas de um objeto com essa cor (p. 6)

Representações também são o conteúdo de nossa consciência, isto é, só podem ser determinadas e compreendidas integralmente pelo indivíduo que a constituiu, já que está sujeita a subjetividade de cada ser, suas percepções do mundo e vivências.

---

[2] Mora (1993) ainda aponta que, embora o uso dos termos “denotação” e “*denotatum*” não esteja incorreto, há a preferência por “referência”, também por ser o termo mais utilizado, da mesma forma que *Sinn*, mesmo podendo ser traduzido como “significação” e “significado”, normalmente é utilizado como “sentido”.

Ainda, as representações necessariamente precisam de um “portador”, enquanto os objetos possuem existência independente. Em continuidade com essa característica, as representações são próprias de seu portador, ou seja, cada representação é única de seu portador.

Em suma, o objeto pode ser apreendido através dos sentidos, enquanto a representação só serão captadas caso materializadas como objetos (Martínez Vásquez, 2020).

De acordo com Frege, existem duas formas das palavras “obterem” significado, seja por meio da definição e de sua elucidação. Em uma definição, de acordo com Rott (2000), o termo é apresentado como tendo o mesmo significado ou sentido da combinação de termos cujos significado e sentidos já são conhecidos. Para ser definido, deve se tratar de um termo novo na linguagem em questão, ou, caso seja um termo já conhecido, seu significado se renovará.

Nesse sentido, as definições são caracterizadas como convenções, estipulações ou abreviações. Já na elucidação o termo será caracterizado vagamente a partir de uma linguagem natural e expressões instáveis, ou a partir de outros termos da linguagem do objeto que não possuem definições precisas.

Dessa forma, elucidações podem caracterizar o significado de termos já definidos quando usados de novas formas, ou o significado de termos novos obtidos através de análises lógicas. Outro fator que as distingue das definições é que a elucidação não precisa, necessariamente, ser concebida verbalmente em sua completude.

> Elas são semelhantes às definições por também se preocuparem em estabelecer o significado de um signo (ou palavra). Portanto, também possuem algo cujo significado não pode ser presumido como conhecido antecipadamente, pelo menos não completamente e além da dúvida, talvez porque são usados na linguagem da vida, de forma flutuante ou em vários sentidos (Frege, 1899 *apud* Rott, 2000, p. 624, tradução nossa).

Cabe mencionar que, para Frege, segundo Rott (2000), a única maneira admissível de determinar o significado de palavras e signos em teorias formalizadas é por meio de definições explícitas dos termos. As elucidações, nesse sentido, apenas elucidam as palavras, mas elas não vão, e nem mesmo podem, defini-las, ou seja, sua tarefa é puramente pragmática.

Pensando em significado, Rott (2000, p. 626) apresenta o princípio do contexto de Frege a partir da simplificação: “o significado de pequenas unidades linguísticas e

o significado de operações gramaticais elementares só pode ser determinado no contexto de grandes complexos linguísticos". Levando esse princípio em consideração, até mesmo termos fundamentais do programa lógico de Frege, como "objeto", "argumentação", "conceito", "verdade" e "função" só poderiam se caracterizar a partir de seu papel - seu *uso* - dentro da teoria.

Mesmo que esse princípio seja de grande importância dentro do trabalho de Frege, ele aponta uma oposição entre seus princípios de contexto e composição, gerando o questionamento, também considerado em Rott (2000), do que viria em primeiro lugar? As partes ou o todo? As palavras ou a oração?

O princípio de composição, também conhecido por sua relevância no trabalho de Frege, diz respeito à decomposição de construções linguísticas completas em partes, como as palavras e suas regras combinatórias como forma de torná-las manuseáveis e mais compreensíveis.

Nesse ponto, Rott (2000, p. 627) considerou que, no princípio do contexto, Frege afirma que o significado das palavras - das partes - apenas será determinado a partir do sentido das orações onde estas ocorrem, enquanto no princípio de composição, o significado das orações podem ser determinados a partir do significado de cada uma das palavras que ocorrem em sua construção.

> [...] nós escolhemos interpretar a totalidade dos produtos linguísticos do enunciador ou escritor, de forma que a verdade e a adequação - julgada pela perspectiva do intérprete - sejam maximizados. A tarefa de um intérprete é essa: Atribuir significados para as palavras de forma que o *interpretandum*, como um todo, atenda, a um nível máximo, aos critérios de perfeição que podemos imaginar (Rott, 2000, p. 628).

Nesse processo interpretativo nosso objetivo é a atribuição de sentido ao produto linguístico em sua completude. Esse processo interpretativo, ainda citando Rott (2000, p. 634, tradução nossa), se inicia com as unidades primárias do significado e da significação, que são "[...] produtos da abrangente tarefa de interpretar os falantes e escritores de uma determinada língua".

Requena (1974) salienta a contradição gerada pelos próprios escritos de Frege em relação ao significado e contexto em Frege. Em um de seus escritos, Frege afirma que o significado das palavras deve ser buscado através de seu contexto proposicional, deixando de lado sua consideração isolada. Porém, há uma contradição ao considerarmos seu trabalho sobre o Sentido e a Referência, onde expõe que toda

expressão que possua uma totalidade de signos corresponde a um sentido determinado.

A explicação apresentada por Requena (1974, p. 583) se baseia na possibilidade de que em cada uma das afirmações Frege se referia a um tipo de linguagem. Portanto, na linguagem lógica proposta em *Conceptografia* o princípio contextual não poderá se sobressair, já que essa linguagem se fundamenta na consideração de que todo signo deve possuir um sentido unívoco, isto é, não dependerá da construção linguística de que faz parte. Já a linguagem natural, por terem os sentidos dos elementos que constituem de suas expressões subjugados a seu contexto preposicional, não compartilha da mesma prerrogativa.

Tendo essa consideração em vista, torna-se possível visualizar que ambos os princípios se encontram constantemente em colaboração durante a atividade interpretativa realizada a partir de construções linguísticas abrangentes decompostas em partes menores. Dessa forma, a interpretação tem início com a compreensão ainda geral de produtos linguísticos mais abrangentes e, a partir do princípio do contexto, atribuímos significado às construções menores e unidades (Rott, 2000).

O princípio de vericondicionalidade, também postulado por Frege, parte da consideração de que o sentido de uma oração está fundado na noção de sua condição de verdade, logo “captar o sentido de uma frase é saber as condições que devem ser preenchidas para que ela seja verdadeira” (Armengaud, 2006, p. 34).

A semântica fregeana pôde ser sistematizada em duas fases de sua teoria. Entre 1879 e 1891, segundo Coffa (1991), tratava-se de uma semântica monista - um item gramatical se relaciona unicamente a uma entidade semântica que se encontra no mundo - baseada nas ideias de conceito, objeto e, com mais destaque, o conteúdo judicável. Esse conteúdo seria o objetivo da lógica de Frege, ou seja, “o legítimo candidato a ser o portador da Verdade” (Pitt, 2018, p. 331).

Pitt (2018) afirma que, nesse momento, já era subentendida a divisão entre os três reinos dos estudos fregeanos: o reino da subjetividade (estados mentais), o reino da efetividade (objetos presentes no espaço tempo em relações causais) e o reino da objetividade (entidades válidas, como pensamentos, sentidos, os valores da verdade, fatos, conceitos, números, funções e relações.

A transição de uma semântica monista a uma semântica dualista, entre 1891 e 1925, a partir da definição e caracterização da identidade entre conteúdos. Inicialmente é apresentado que a identidade considera o conteúdo informacional de cada juízo, porém, mais tarde, afirma-se que a identidade tem em vista o objeto que foi designado, assim, a identidade seria a relação entre os signos - sinais - dos objetos, e não entre eles em si.

Essa dualidade causa um problema na definição do critério de identidade, pois Frege aponta que dois sinais de objetos podem ser substituídos com um mesmo conteúdo, isto é, dois sinais com um mesmo conteúdo não sofrerão alterações com essa mudança.

Pitt (2018) representa esse pensamento apontando que se “a” = “b” é uma identidade, na qual “a” designa o mesmo objeto que “b”, expressando um mesmo conteúdo, sendo possível realizar a troca de “a” por “b” - “a” = “a” ou “b” = “b” - ainda mantendo seu conteúdo informacional, mesmo não sendo o que acontece.

Como solução para o problema apresentado, Frege assume a divisão de sentido e referência e sentido, dessa forma as expressões “a” = “a” e “b” = “b”, considerando que a identidade “a” = “b” é verdadeira, possuem diferenças no modo de apresentação da designação, e não em seu conteúdo.

Essa nova consideração elucida que Frege declara “[...] os símbolos linguísticos sinais, os objetos designados por eles a referência desses sinais e o modo com que designamos os objetos o sentido dos sinais” (Pitt, 2018, p. 333). Segue um esquema da semântica de Frege em sua segunda fase na figura 3.

Figura 3 - Semântica fregeana

Sentença	Nome Próprio	Termo Conceitual		
↓	↓	↓		
Sentido da Sentença (Pensamento)	Sentido do Nome Próprio	Sentido do Termo Conceitual		
↓	↓	↓		
Referência da Sentença (Valor de Verdade)	Referência do Nome Próprio (Objeto)	Referência do Termo Conceitual (Conceito)	→	objetos caindo sobre o conceito

Fonte: Esquema de Frege (1997, p. 149) demonstrado por Pitt (2018, p. 333).

Nessa segunda fase de sua semântica, a noção de pensamento como o sentido de uma asserção, se caracteriza pela sua relevância. Nesse momento o pensamento

passa a ser entendido como o conteúdo objetivo das ações de pensar, não pertencentes a atos individuais e subjetivos de pensar, mas independentes e objetivos em relação àqueles que o pensam. O pensamento não é produzido durante o pensar, não se relaciona com esse como um produto, mas sim apreendido através deste (Pitt, 2018).

Conclui-se que o projeto lógico e a semântica de Frege, juntamente com sua distinção entre sentido e referência, deram bases conceituais, semânticas e filosóficas para as análises de proposições lógicas, além de contribuir com o campo pragmático ao destacar o contexto e o uso como fatores que contribuem para a determinação do significado de expressões e proposições. Assim, a diferenciação de referência, como o objeto denotado, e o sentido, como a forma de expressão do objeto, permite a verificação da condição de verdade das proposições e justifica as diferentes formas de expressões - a significação - de um mesmo objeto - palavra - de forma dependente do contexto e da forma de uso das expressões linguísticas.

2.2.2. Morris

Semioticista e filósofo estadunidense, Charles William Morris se baseou nos estudos pragmáticos de Peirce e em Frege e a questão da representação simbólica nas ciências, para o desenvolvimento de uma teoria geral dos signos.

Em um emaranhado de abordagens e teorias, Morris buscava aquela que seria unificada e que fosse capaz de ampliar seu alcance aos campos de estudos naturais, antropológicos e filosóficos (Armengaud, 2006).

Seus esforços seguem sua concepção de que a ciência e os signos se encontram associados de forma inerente, já que o papel da ciência seria oferecer “[...] signos mais fiáveis e, por outro, inserir os resultados obtidos em sistemas de signos” (Morris, 1938, p. 1). Nesse mesmo sentido, a relação da semiótica com as ciências possui duas vertentes, sendo ela uma das ciências e um instrumento delas, já que fazem uso de seus signos para a expressão de seus resultados.

Morris define o processo em que algo funciona como um signo como semiose, processo formado por três aspectos: aquilo que age como um signo, o veículo sígnico; aquilo que o signo se refere, o *designatum*; efeito sobre o intérprete, o interpretante. Nesse caso, o intérprete também pode ser considerado nessa relação como o quarto aspecto.

Assim, o signo pode ser caracterizado da seguinte forma:

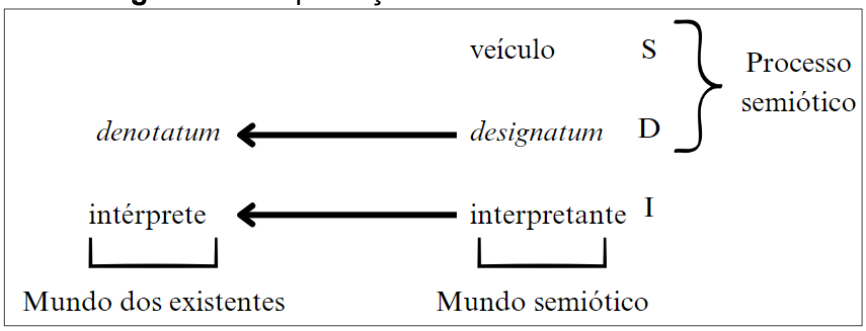
> *S* é um signo de *D* para *I* na medida em que *I* se dá conta de *D* em virtude da presença de *S*. Assim, na semiose, alguma coisa dá-se conta de uma outra coisa mediatamente, isto é, mediante uma terceira. Consequentemente, a semiose é um dar-se-conta-mediato. Os mediadores são os veículos sígnicos; os dar-se-conta-de são os interpretantes; os agentes do processo são os intérpretes; aquilo de que se dá conta são os *designata* (Morris, 1938, p. 4).

Uma importante consideração a ser feita é a de que termos como "signo" "interpretante" e "*designatum*" não representam objetos em si, mas referem-se a fatores que compõem o processo da semiose. Morris destaca que a semiótica não se ocupa do estudo de um tipo específico de um objeto, e sim de qualquer objeto que, em alguma circunstância, faça parte do processo da semiose.

Morris (1938, p. 45) apresenta que "[...] nada é intrinsecamente um signo ou um veículo sígnico, mas converte-se não somente na medida em que permite a alguma coisa dar-se conta de alguma coisa através da sua mediação", logo o significado não possui uma existência localizada em alguma parte da semiose, mas são determinados de acordo com o processo como um todo. Não contamos com a existência de uma classe de objetos que são significados, mas sim que cada objeto obtenha sua função nas condições do processo da semiose.

Em consideração dos três aspectos formadores da semiose - veículo sígnico, *designatum* e intérprete - Morris deriva três relações diádicas, gerando as três dimensões da semiose, representadas na figura 4.

Figura 4 - Simplificação redutora de Morris



Fonte: Baseado em Armengaud (2006, p. 43).

A relação entre os signos forma a dimensão sintática da semiose, ou *Dsin*; a relação entre signos e os objetos a que se referem forma a dimensão semântica da semiose, ou *Dsem*. A relação entre signos e seus intérpretes forma a dimensão pragmática da semiose, ou *Dp*.

Assim, “[...] será conveniente ter termos especiais para designar algumas das relações dos signos a signos, a objectos, e a intérpretes. 'Implica' restringir-se-á a *Dsin*, 'designa' e 'denota' a *Dsem*, e 'expressa' a *Dp*” (Morris, 1938, p. 10).

Com a caracterização das dimensões do processo de semiose, Morris transforma a relação triádica do signo de Peirce - signo, *designatum* e interpretante - em três relações diádicas - signo e objeto, signo e signo, signo e intérprete - com a adição do interpretante e consequente supressão do interpretante (Armengaud, 2006).

Em sua caracterização do termo “língua”, Morris destaca que essa caracterização também interage com as dimensões do processo da semiose, podendo ser: formalista, que considera todo sistema axiomático, sendo ou não denotativa, independentemente de ser utilizado ou não por um grupo de intérpretes; empirista, com destaque para a necessidade de interação entre signo e objeto denotado; pragmatista, tendo em vista uma atividade social e comunicativa e uma necessidade, seja ela individual ou comum.

Tornar-se-á também claro que, justamente como um signo individual é completamente caracterizado ao fornecer-se a sua relação a outros signos, objetos e utilizadores, uma língua também é completamente caracterizada ao fornecer-se aquilo que posteriormente se chamará as regras sintáticas, semânticas e pragmáticas que governam os veículos sígnicos (Morris, 1938, p. 11).

Morris inicia a explanação das dimensões da semiose a partir da dimensão sintática, responsável pelas relações dos signos entre eles. A partir dos esforços de Leibniz, nasce a elaboração de uma ciência formal geral, a partir de considerações da linguística, lógica e matemática, com uma ciência característica geral é uma ciência combinatória geral (Morris, 1938).

A sintaxe lógica apoia-se na desconsideração dos fatores das dimensões semânticas e pragmáticas da semiose, mantendo em vista apenas sua dimensão sintática e sua estrutura lógica.

Com isso em vista, “língua” - *Lsin*, nesse caso - passa a ser caracterizada como, segundo Morris (1938, p. 17):

> conjunto de coisas relacionadas consoante dois tipos de regras: regras de formação, que determinam combinações permissivamente independentes de membros do conjunto (chamando-se essas combinações proposições), e regras de

> transformação, que determinam as proposições que podem ser obtidas de outras proposições.

Morris ainda revela que os estudos desenvolvidos pela sintaxe lógica (*syntax*), integram a sintaxe (*syntactics*) em seu aspecto científico, sendo ela a parte mais desenvolvida da sintaxe - e, consequentemente da semiótica -, sendo responsável também por contribuições para os campos da semântica e da pragmática.

A estrutura sintática de uma linguagem é a inter-relação de signos causada pela inter-relação das respostas de que os veículos sígnicos são produtos ou partes.

Sobre a natureza sígnica, um signo singular, isto é, que denota um único objeto - como a ação de apontar - se caracteriza como um signo indexical. Já um signo que denota uma diversidade de objetos, dá-se o nome de signos caracterizantes, podendo passar por combinações com signos que restringem sua aplicação. Por fim, nomeia-se como signos universais aqueles que podem denotar qualquer objeto.

Porém, o destaque vai para a combinação de signos indexicais e signos caracterizantes, que define a referência do signo indexical em adição da determinação de uma expectativa a partir do signo caracterizante. Essas expectativas podem sofrer um aumento a partir da utilização de signos complementares, gerando graus de especificação e explicitando as relações existentes entre os signos (Morris, 1938, p. 18-19):

> [...] pode dizer-se que qualquer proposição contém um signo dominante e certos especificadores, sendo estes termos relativos uns aos outros, pois o que é um signo dominante relativamente a certos especificadores pode ele mesmo ser um especificador relativamente a um signo dominante mais geral [...] uma proposição capaz de verdade e falsidade envolve signos indexicais, um signo caracterizante dominante com especificadores possivelmente caracterizantes, e alguns signos para mostrar a relação dos signos indexicais e caracterizantes entre si e para com os membros da sua própria classe.

Essa consideração aponta que são as condições do uso que irão constatar quando e qual signo será um signo dominante, já que “[...] condições pragmáticas determinam aquilo que, de fato, é o signo dominante” (Morris, 1938, p. 19).

É também a partir dessa explanação que surge a chamada fórmula geral da proposição:

Signo caracterizante dominante [especificadores caracterizantes (signos indexicais)].

Para que as combinações entre signos sejam compreensíveis, são utilizados signos especiais responsáveis por indicar as relações de outros signos, porém mantendo-se diferenciados dos signos que designam as relações na sintaxe. Esses signos especiais indicam qual signo é especificado e qual o especifica, quais são os signos indexicais ou caracterizantes, e até mesmo qual é designado pelo signo indexical.

Assim, na linguagem escrita, esse papel é realizado pelos sinais de pontuação, tamanho da letra, itálicos e negritos, parênteses, entre outros, enquanto na linguagem falada, são as entonações, as pausas e as ênfases. Nesse sentido, "[...] esses signos dentro da linguagem exercem primeiramente uma função pragmática" (Morris, 1938, p. 20).

Responsável pelos aspectos da relação entre os signos e seus *designata* e, consequentemente, aos objetos que estes denotam e que podem vir a denotar, a dimensão semântica da semiose é dotada de uma ligação profunda com a determinação do valor de verdade de proposições, e com o desenvolvimento sintático, já que, baseando-se em Morris (1938), a explanação das relações entre signo e objeto devem apoiar-se em uma teoria da estrutura formal das construções linguísticas, a linguagem da dimensão da sintaxe.

Sobre essa associação entre sintaxe e semântica, Morris (1938, p. 23) salienta que "a semântica pressupõe a sintaxe, mas abstrai da pragmática; seja lidando com signos simples ou complexos (como seria um sistema matemático inteiro), a semântica restringe-se à dimensão semântica da semiose". Porém, Morris (1938, p. 29) destaca:

> Se fatores pragmáticos apareceram frequentemente em páginas pertencentes à semântica, é porque a intelecção geral de que a sintaxe deve ser complementada pela semântica ainda não se estendeu de um modo tão geral à intelecção de que a semântica deve por seu lado ser complementada com a pragmática. [...] as regras sintácticas e semânticas são apenas as formulações verbais dentro da semiótica do que em qualquer caso concreto da semiose são hábitos da utilização sígnica pelos utilizadores efetivos dos signos.

Tal qual a dimensão sintática com suas regras de formação e transformação, essa dimensão conta com sua regra semântica, que irá determinar as condições que possibilitarão a aplicação de um signo a um objeto.

Um signo denota o quer que se conforme às condições estabelecidas na regra semântica, enquanto a própria regra estabelece as condições de designação e, desse modo, determina o *designatum* (a classe ou a espécie de *denotata*) (Morris, 1938, p. 23).

O destaque de Morris (1938, p. 26) também recai sobre o fato de que as regras que determinam as circunstâncias de aplicação de um signo não são estabelecidas, ou pelo menos não completamente, pelos usuários dos signos, e sim pelos comportamentos e hábitos, "[...] de modo que só certas combinações de signos ocorrem de fato, só certas combinações de signos se derivam de outras, e só certos signos se aplicam a certas situações".

Assim, a presença da dimensão da semântica em relação a um signo se concretiza a partir da existência das regras semânticas que irão determinar suas condições de uso. Considerando que o emprego dos signos dependerá, também, de outros signos, uma fórmula geral foi concebida: "O veículo sígnico 'x' designa as condições a, b, c... em que é aplicável. O estabelecimento dessas condições fornece as regras semânticas de 'x'. Se um qualquer objeto ou situação preenche as condições requeridas então é denotado por 'x'".

A regra semântica dos signos indexicais estabelece que esse signo designará aquilo que chama sua atenção - no ato de apontar, tomado como um signo indexical, o signo designa aquilo que está sendo apontado - sem caracterizá-lo e também não possui semelhanças ou propriedades em comum com aquilo que está sendo indicado (índice).

No caso de um signo característico, seu *designatum* também é caracterizado por ele, seja através da demonstração de suas propriedades pelo próprio signo (ícone) ou não (símbolo).

A regra semântica de uma proposição considera a referência e as regras semânticas de seus signos, assim, as proposições designam tanto aquilo que é designado pelo signo indexical quanto o que é designado pelo signo característico.

Relativamente a suas regras, a regra da dimensão semântica encontra um análogo na pragmática, onde o hábito/comportamento do intérprete durante a utilização do signo em certas condições enquanto possui expectativas de respostas ao uso desse signo. Além do mais, as regras do campo sintático - regras de formação

e de transformação -, na dimensão pragmática, dizem respeito às combinações utilizadas pelos intérpretes durante o uso dos veículos sígnicos e, de acordo com Morris (1938, p. 32), às regras do uso de certos signos estabelecidas pelos próprios usuários, de modo a “[...] controlar deliberadamente outros modos de comportamento em relação a pessoas e a coisas”.

Com isso em vista, no âmbito pragmático, a estrutura linguística converte-se no sistema de comportamentos onde, em proposições analíticas, as relações cultivadas entre signos são respostas dadas a outras respostas sígnicas mais amplas de que são partes, enquanto, em proposições sintéticas, as relações entre signos são respostas sígnicas que não fazem parte de um todo como sua parte.

Desse modo, um signo indexical, como o ato de apontar, desloca a atenção de um intérprete para algo que faz parte de um contexto - uma parte do todo -, enquanto o signo característico dominante refere-se a uma resposta geral às partes, e os especificadores delimitam esses signos característicos e as expectativas que estes geram. Caso a expectativa ocorra como o esperado, o signo será confirmado, o que, em geral, ocorre apenas de forma parcial (Morris, 1938).

As regras da dimensão pragmática têm como propósito fundamental definir as condições relativas aos intérpretes para que um veículo sígnico seja um signo. Como forma de exemplificação temos as interjeições e expressões como “bom dia” ou “venha cá”, que ocorrerão apenas sob condições favoráveis à sua compreensão.

A partir da determinação desta última regra, Morris (1938, p. 35) fórmula a caracterização da língua: “uma língua no pleno sentido semiótico do termo é qualquer conjunto intersubjetivo de veículos sígnicos cujo uso é determinado por regras sintácticas, semânticas e pragmáticas”.

Sob um sentido pragmático, um signo é utilizado em combinação com outros signos por aqueles que são parte de um grupo social, assim como a língua é um sistema social de signos que irão fazer a mediação entre os integrantes de uma comunidade e sua relação com o mundo, assim como foi colocado por Morris:

> Compreender uma língua é empregar apenas aquelas combinações e transformações de signos não proibidas pelos usos do grupo social em questão; é denotar objetos e situações tal como o fazem os membros desse grupo, é ter as expectativas que os outros têm quando se empregam certos veículos sígnicos, e é expressar os seus próprios estados tal como os outros fazem − em

> suma, compreender uma língua ou usá-la corretamente é seguir as regras do seu uso (sintácticas, semânticas, e pragmáticas) numa dada comunidade social (1938, p. 36).

Morris (1938, p. 40) considera legítimo a utilização dos signos com o objetivo único de gerar interpretações e respostas específicas, mesmo que esses signos não denotem algum objeto ou que não sejam combinados de acordo com as regras dispostas na dimensão sintática da semiose.

Atos como ordens, perguntas etc., podem ser utilizados como forma de controle comportamental e para a produção específica de certos interpretantes, assim como seu uso nas artes e literatura:

> [...] a sintaxe, a semântica, e a pragmática dos signos usados em literatura, nas artes, na moral, na religião, e em juízos de valor, são em geral tanto do seu foro como dos estudos dos signos usados na ciência. Tanto num caso como no outro, o uso do veículo sígnico varia com o fim a que se presta (Morris, 1938, p. 42).

Ainda nessa linha, Morris (1938) delineia a questão da universalidade dos signos. Para isso, a questão da universalidade foi “desconstruída” de acordo com as dimensões da semiose, e foi abordada a partir da distinção de sinsigno e legisigno, feita por Peirce. Nesse sentido, um sinsigno é uma coisa - particular - que, em certas condições, assume a função de signo, enquanto o legisigno pode ser entendido como uma “lei” ou regra que assume a função de signo.

Assim, temos como exemplo que um conjunto de “marcas” em um contexto específico, como “casa”, é um sinsigno, enquanto esse mesmo conjunto não caracteriza a palavra “casa”, sendo essa um hábito de uso “universal” em oposição ao uso particular.

Para essa compreensão, é necessário entender a universalidade - ou generalidade - em todas as dimensões do processo da semiose. A indiscriminação entre a universalidade de cada uma das dimensões pode gerar confusões. Ao afirmar que um signo é “universal”, afirma-se que ele integra uma classe de objetos que são capazes de suprir expectativas, combinar-se com determinados signos, denotar certos objetos e sujeitar-se às mesmas regras de utilização.

Os signos com diferentes regras de uso serão, consequentemente, diferentes, não importa sua semelhança formal ou fonética, já que “[...] nenhuma das disciplinas relacionadas aos signos está interessada na descrição física completa do veículo

sígnico, mas preocupa-se com o veículo sígnico na medida que este esteja em conformidade com as regras de utilização" (Morris, 1938, p. 50).

Em usos específicos, o signo caracteriza-se como sinsigno, sua universalidade, isto é, sua caracterização como legisigno dependerá de este ser parte de uma classe de objetos que possam atuar em uma mesma função na semiose.

A questão posta sobre a universalidade sígnica também irá se relacionar com a dimensão semântica. Seguindo a regra semântica, o *designatum* de certo signo será a classe de objetos que podem ser denotados por esse signo, seja ela formada por apenas um objeto, múltiplos objetos, nenhum ou todos os objetos.

Assim, a universalidade do signo na dimensão semântica diz respeito à possibilidade desse signo denotar mais de um objeto, e a universalidade dos objetos será verdadeira quando estes podem são denotados pelo mesmo signo. Tal fato também se faz presente na dimensão sintática, onde mesmo em uma combinação particular de signos - regidas pelas regras de formação e transformação -, sua forma pode ser compartilhada com outras combinações sígnicas, gerando uma universalidade formal, ou sintática (Morris, 1938).

A partir da dimensão pragmática, dois fatos se relacionam com a possibilidade da universalidade de um signo: o fato de que um mesmo signo pode denotar mais que um objeto ou situação também se refere às diferentes expectativas determinadas, que também podem ser arrematadas - mesmo que não completamente, por diferentes objetos, trazendo à tona a universalidade do interpretante, mesmo que este seja específico à situação; e o outro fato que determinada a universalidade do signo corresponde à potencialidade do signo de ser partilhado entre uma diversidade de intérpretes, originando a generalidade social.

Assim distinguem-se os cinco tipos de universalidade (generalidade) sígnica, respectivamente: universalidade do veículo sígnico, universalidade da forma, universalidade da denotação, universalidade do interpretante e universalidade social.

Mesmo se dedicando para caracterizar cada uma das dimensões da semiose individualmente, Morris (1938, p. 52) reafirma a inter-relação inerente entre essas dimensões, mesmo que elas sejam consideradas e estudadas separadamente, isto é, "[...] a semiótica [...] está principalmente preocupada com suas inter-relações e,

portanto, com o caráter unitário da semiose que essas disciplinas individualmente ignoram".

A inter-relação entre dimensões torna-se evidente ao considerarmos que, mesmo que cada uma dessas dimensões lide com um aspecto sígnico, seja ele sobre signos entre si, signos e objetos ou signos e intérpretes, nenhuma delas é capaz de definir o signo ou sua própria dimensão.

A sintática, ao considerar as regras de formação e transformação, estará considerando aspectos pragmáticos da semiose, já que regras determinam comportamentos e intérpretes; a semântica concentra seus esforços na relação dos signos com seus objetos e situações designadas, porém também leva em conta as regras semânticas de uso e seus intérpretes; a pragmática analisa a os signos como são interpretados em condições determinadas por intérpretes, que só poderão ser definidos considerando os veículos sígnicos e seu *designatum*.

As regras que regem os veículos sígnicos da língua em estudo devem ser compreendidas, e "compreensão" é um termo pragmático. As regras para combinação e transformação de possíveis veículos sígnicos não podem ser meramente compostas de possíveis veículos sígnicos, mas devem, na verdade, funcionar como signos.

Na sintaxe descritiva devem existir signos para denotar os veículos sígnicos sendo estudados, e o objetivo deve ser fazer afirmações verdadeiras sobre esses veículos sígnicos - mas "denotar" e "verdadeira" não são termos sintáticos.

A semântica vai estudar a relação de uma combinação sígnica com o que ela denota ou pode denotar, mas isso envolve o conhecimento da estrutura da combinação sígnica e das regras semânticas [...] A pragmática não pode ir longe sem levar em conta as estruturas formais para as quais deve procurar o correlato pragmático, e a relação dos signos com os objetos que procura explicar através da noção de hábito de uso (Morris, 1938, p. 53).

Até mesmo as linguagens utilizadas em cada uma das dimensões da sintaxe, semântica e pragmática possuem três dimensões: elas irão designar fases - ou aspectos - do processo da semiose, são dotadas de estruturas formais e contam com o fator pragmático ao serem usadas e compreendidas. Mesmo que sejam partes de uma mesma ciência, a sintaxe, a semântica, e a pragmática são irredutíveis e intrinsecamente relacionadas.

Em busca de articular sobre os problemas e possíveis campos de aplicação da semiótica, Morris (1938) os agrupa em unificação das ciências semióticas, semiótica como *organon* das ciências e implicações humanísticas da semiótica.

Falando sobre a unificação das ciências semióticas, antes faz-se necessário considerar que, mesmo que possua uma delineação bem feita, a semiótica ainda não seria uma ciência avançada. Para isso, seriam necessários dois tipos de profissionais para isso, os chamados "*fact-finders*" - ou descobridores - que esclarecerá as condições de ocorrência da semiose e os acontecimentos do processo, e os sistematizadores, que fazem uso desses fatos para o desenvolvimento de uma estrutura teórica sistematizada que será utilizada pelos próprios descobridores.

A história da semiótica também se mostra como um campo para os semioticistas, como a revisitação, esclarecimento de antigas doutrinas, como as categorias, os transcendentais e as previsões, assim como debates entre signos repressivos e indicativos, as doutrinas de intenção, imposição e suposição e até a história da linguística, da retórica, da lógica, do empirismo e da ciência experimental.

Por darem bases à semiótica, ciências como a matemática, lógica e linguística se apresentam como campos amplos para os estudos semióticos: a matemática fornece conhecimentos sobre a estrutura das construções linguísticas; a lógica com seus paradoxos, sua teoria dos tipos, suas leis, teoria da probabilidade, os processos de dedução, indução e hipóteses; a linguística, por fim, que concentra esforços nas estruturas sígnicas complexas que constituem as línguas (Morris, 1938).

Morris (1938) aponta até mesmo aos problemas metodológicos e epistemológicos como partes da esfera semiótica, como o empirismo e racionalismo que colaboram com as relações de notação, questões de validade e de conhecimento apoiando estudos semânticos e pragmáticos, os procedimentos científicos e estudos estéticos do funcionamento sígnico.

O desenvolvimento da semiótica é, em si, um estágio da unificação das ciências lidam total ou parcialmente com signos; também pode desempenhar um papel importante preenchendo a lacuna entre ciências biológicas, por um lado, e as ciências psicológicas, humanas e sociais, por outro, e lançando uma nova luz sobre a relação das chamadas ciências "formais" e "empíricas" (Morris, 1938, p.56).

A semiótica como *organon* das ciências, de acordo com Morris (1938, p.56-57), ocupa um espaço único entre as ciências, considerando que as ciências se concentram em investigar e encontrar dados que possam ser utilizados como signos fidedignos, e seus resultados serão expressos por meio de signos linguísticos.

A clara dependência das ciências em relação aos signos faz com que o cientista deve se atentar a seus usos sígnicos e linguísticos, o que nos leva a primeira função da semiótica em serviço das ciências: o treinamento de cientistas em semiótica, para que este seja capaz de julgar e utilizar construções linguísticas de forma crítica.

Morris partilha a ideia, comum a filósofos e epistemólogos como Hempel ou Schefler, de que o estudo da ciência coincide com o estudo da linguagem da ciência. Mas o que é efetivamente o estudo de uma linguagem? Ele comporta, é claro, o estudo da estrutura formal, mas não fica nisso “[...] impõe-se também o estudo da relação da linguagem com os objetos designados e as pessoas que os utilizam” (Armengaud, 2006, p. 41).

Outra função semiótica de aplicação ao campo científico tem como base o estudo das linguagens de ciências especiais, com análise de suas estruturas, termos e problemas, demonstrando a instrumentalização da semiótica e sua capacidade de aplicação relevante a cada domínio.

Por fim, as implicações humanísticas da semiótica partem do pressuposto de que os signos podem ser utilizados além da aquisição de conhecimentos, e seu estudo engloba mais que o estudo das linguagens científicas.

Pensando nos fins únicos dos signos, são desenvolvidas linguagens especializadas, ainda seguindo as dimensões do processo da semiose: a linguagem matemática, que tem como finalidade destacar a relação dos termos da linguagem enquanto sua relação com seus objetos e intérpretes permanece em segundo lugar; a linguagem da ciência empírica, adaptada para descrições da natureza; linguagens como a da moralidade, das belas artes e das artes aplicadas, que são moldadas, respectivamente, para o controle comportamental, a apresentação de objetos e situações de acordo com o interesse e a manipulação de coisas que possam implicar em acontecimentos desejados.

Sobre essas linguagens, Morris (1938, p. 58) salienta que em todas elas, independentemente de sua especialização, todas as dimensões do processo da

semiose encontram-se presentes, mesmo que alguma delas seja mais enfatizada em comparação às outras.

A partir deste delineamento, Morris busca evidenciar que a semiótica "[...] fornece uma base para a compreensão das principais formas de atividade humana e suas inter-relações, uma vez que todas essas atividades e relações se refletem nos signos que mediam essas atividades".

Em síntese, Morris buscou o desenvolvimento de uma teoria geral dos signos com base nos estudos desenvolvidos por Charles Peirce e Gottlob Frege, tendo em vista uma semiótica ampla e que viesse a ser utilizada como um "instrumento" metodológico nas ciências humanas e naturais. Assim, a semiose que antes era um processo triádico, como disposto por Peirce, passa a resultar das interações decorrentes das dimensões sintática, semântica e pragmática dos signos, o que confere um papel fundamental aos contextos na ação de significação e interpretação sígnica.

2.2.3. Austin e Searle

A partir de contribuições do filósofo britânico John Langshaw Austin (1911-1960) e o filósofo e escritor norte-americano John Rogers Searle (1932), nasce a teoria dos atos de fala a partir do seguinte postulado: "A unidade mínima da comunicação humana não é nem a frase nem qualquer outra expressão. É a realização (performance) de alguns tipos de ato" (Armengaud, 2006, p. 99).

Silva (2008, p. 236, tradução nossa) aponta o seguinte sobre as investigações iniciadas por Austin:

> A problemática identificada por Austin (1996) - conciliar os atos de fala com a semântica - e posteriormente continuada por Searle (1991a, 1991b, 1994) e outros filósofos da linguagem é a base de uma das questões-chave na filosofia da linguagem [...] Essa discussão visa o entendimento do significado, seja vinculado ao significado léxico dado ou na interação entre a língua e o uso.

O princípio que guiou os estudos de Austin (1996) foi a afirmação de que "dizer é fazer", isto é, o pronunciamento de uma sentença gera uma ação no mundo, como é reforçado por Silva (2008, p. 238, tradução nossa), já que "[...] os enunciados performativos ou realizativos desempenham uma ação no mundo".

Austin ainda foi responsável por apontar a existência de enunciados realizativos - ou não descritivos-, ou seja, enunciados que não podem ser avaliados

com base em valores de verdade e falsidade, que, ao serem enunciados, executam uma ação e indicam seu verbo realizativo a partir do ato que realizam.

Assim, surgem os atos de fala. Desta forma, entende-se que para toda emissão existe uma intenção que condiciona a realização de um ato na pronunciação de uma expressão, e não apenas uma descrição (Bachioqui, 2010).

Alamillo (2007, p. 181) esclarece que Austin chega a desconsiderar a segmentação entre enunciados realizativos e enunciados constatativos, considerando que a emissão de todo e qualquer enunciado leva à realização de uma ação. Tal constatação é exemplificada da seguinte forma: no enunciado realizativo “te prometo que irei” realiza-se a ação de prometer algo a alguém, enquanto no enunciado constatativo “está chovendo” realiza-se a ação de afirmar algo a alguém.

Austin indica os seguintes atos de fala: afirmar, perguntar, ordenar, descrever, desculpar, agradecer, prometer, criticar, felicitar, acusar, sugerir, ameaçar, desafiar, autorizar e suplicar.

Claramente, a lista não está completa e poderia continuar, demonstrando a multiplicidade de atos possíveis. Segundo Armengaud (2006), a lista de atos proposta por Austin está apta a associar-se aos jogos de linguagem propostos por Wittgenstein.

Em busca de uma distinção entre os enunciados, Austin sugere a partição entre enunciados realizativos primários, os quais possuem sua base realizativa intrínseca, e os realizativos explícitos, sendo estes os enunciados realizativos diretos. Silva (2008, p. 239, tradução nossa) apresenta o seguinte exemplo, desenvolvido por Austin (1996): 1) expressão primária “estarei ali”, 2) realizativo explícito “te prometo que estarei ali”.

Esta teoria seria uma forma sistematizada de investigação dos signos em relação a seus intérpretes, e seu objetivo é “[...] saber o que fazem os intérpretes-usuários, que atos eles realizam pelo uso de certos signos [...] em certo sentido, nada é mais diretamente pragmático que esse estudo”, argumenta Armengaud (2006, p. 100).

Neste contexto, o pronunciamento de uma frase é capaz de fomentar a realização de um ou diversos atos, mas não deve confundir-se. A expressão linguística ou frase é utilizada para a concretização do ato, e não o ato em si.

Com isso em mente, Austin sistematizou os atos de fala em atos ilocucionários e atos perlocucionários. Os atos ilocucionários, representados logicamente por Searle

e Vanderveken por "*F(p)*"[3], ocorrem dentro do discurso e são caracterizados no fazer discursivo, é aquilo que se faz quando se enuncia, podendo ser suficientes a partir das delimitações das regras linguísticas ou dependentes de condições contextuais, como as condições institucionais e sociais[4].

Já os atos perlocucionários são caracterizados como efeitos decorrentes dos atos ilocucionários em seus respectivos alocutórios - receptores do discurso -, assim, atos como o de convencer, intimidar e culpar têm como efeitos perlocucionários ser convencido, intimidado ou culpado.

Armengaud (2006) ressalta que, normalmente, os atos ilocucionários já são concretizados tendo os atos perlocucionários em vista de seu propósito final, e também retoma a afirmação de que, enquanto os atos ilocucionários são, essencialmente, atos de fala, os atos perlocucionários podem resultar de outras ações, inclusive de natureza extralinguística.

Os atos ilocucionários possuem distintamente os atos subsidiários como forma de realizar referências a objetos e expressar alguma proposição e, pragmaticamente, "[...] o mesmo ato de referência, a mesma expressão de uma proposição pode aparecer em atos ilocucionários diferentes", como exemplificado em Armengaud (2006): Uma mesma declaração, como "Por favor, sirva-se de mais caldeirada!", "Você vai se servir de mais caldeirada" e "Você vai se servir de mais caldeirada?", é capaz de desempenhar diferentes atos ilocucionários, neste caso, uma solicitação, uma predição e uma pergunta, respectivamente[5].

Mesmo que, para fins de análise, os atos de fala sejam sistematicamente isolados, eles podem dar início ou integrarem sequências reguladas de atos, como uma pergunta que exigirá uma resposta. Os atos não se encontrarão desconexos em uma interação discursiva, o que, aos olhos de Armengaud (2006), aproxima a teoria à análise do discurso e de análises conversacionais.

Austin (1996) propõe um teste para a identificação dos enunciados realizativos explícitos: uso da primeira pessoa do singular do presente do indicativo em voz passiva. A partir disso, o dicionário seria consultado para a identificação de uma lista de verbos que expressam este realizativo. Este teste também seria usado para a

---

[3] *F* para força ilocucionário e *p* para conteúdo proposicional.
[4] Austin demonstra as condições institucionais a partir da seguinte ideia: para declarar que a sessão está aberta, é necessário ser o presidente da sessão.
[5] Mesmo que produzam diferentes atos ilocucionários, as sentenças exemplificadas ainda pretendem produzir os mesmo efeitos - atos perlocucionários - em seus alocutórios.

identificação destes verbos, responsáveis pela expressão da força ilocucionária de uma sentença. Com isto em mente, Silva (2008) declara que o interesse de Austin seria reconhecer os atos ilocucionários realizados na emissão de determinado verbo.

Com este “teste” como fundamento, Austin propõe a classificação dos atos de fala em cinco tipos: veridictivos, exercitivos, comissivos, expositivos e comportamentais.

Os atos veridictivos correspondem a enunciação de um veredicto, isto é, um julgamento, a partir de evidências e razão, como avaliar, caracterizar, descrever, analisar e classificar; os atos exercitivos estabelecem decisões para uma sequência de ações seguintes, como ordenar, recomendar, implorar, comandar e aconselhar; os atos comissivos determinam um comprometimento com uma ação ou sequência de ações, como prometer, jurar, se comprometer e garantir; os atos expositivos podem expressar ideias, encaminhar argumentos, esclarecimentos e uso de certas palavras e expressões, como exemplificar, afirmar, negar, responder e relacionar; por fim, os atos comportamentais são as reações às atitudes e acontecimentos, como se desculpar, felicitar, agradecer, desejar, criticar e desafiar[6].

1) Enquanto Austin contribuiu com as classificações dos atos ilocucionários, Searle define doze critérios para esses atos que, de acordo com Villanueva (1975), são as condições necessárias para que a emissão de um ato ilocucionário seja “feliz” e não “defeituosa”, sendo eles: Finalidade do ato - Busca responder o questionamento “qual é o ponto deste ato?”;
2) Orientação de ajustes entre mundo e palavras - Alguns atos buscam ajustar as palavras ao mundo (↓), enquanto outros buscam o ajuste do mundo às palavras (↑);
3) Estados psicológicos expressos - É a expressão do falante, suas atitudes e ideias através do conteúdo da proposição[7], sendo os principais estados psicológicos a crença, intenção, desejo ou necessidade e o prazer;
4) Intensidade de comprometimento - Frases com um mesmo ponto ilocucionário podem apresentar forças diferentes, como em “Sugiro que vá ao médico” e “Insisto que vá ao médico”;

[6] Armengaud (2006) expõe que essa classificação, assim como em Searle (1975), se caracteriza mais como uma classificação de verbos ilocucionários de uma língua determinada, e não atos.
[7] “Aquele que promete fazer *a*, que ameaça fazer *a*, exprime a intenção de realizar *a*” (Armengaud, 2006, p. 106).

5) Posição do falante e do ouvinte - As posições sociais dos indivíduos de uma interação também determinam a força do ato, como quando um chefe pede a um funcionário a realização de uma atividade, como um comando, e o funcionário pede ao chefe a realização dessa mesma atividade, como uma sugestão;
6) Relação entre sentença e interesse dos participantes da interação - Enquanto as lamentações são atos que satisfazem interesses dos falantes, as condolências são atos que atendem aos ouvintes;
7) Relação com o discurso e o contexto - Ao concluir, deduzir ou responder, o falante conecta seu ato a uma sequência de atos anteriores e subsequentes;
8) Marcas da força ilocucionária - Relatar dirá respeito a uma enunciação sobre o passado, enquanto prever se relacionará ao futuro;
9) Atos de fala em potencial - Existem expressões que, mesmo com o potencial de realizarem atos, em razão do silêncio e da escolha de não proferir, não chegam a se concretizar;
10) Relação com o extralinguístico - Alguns atos necessariamente se fundamentam em instituições extralinguísticas para se concretizarem;
11) Verbos ilocucionários performativos e não performáticos - Atos como prometer devem ser performados (“eu prometo que…”), na mesma medida em que atos como ameaçar não necessitam de performance (“eu te ameaço…”);
12) Estilo de realização do ato - Proclamar e confidenciar se enquadram como atos diferentes não pela divergência entre pontos ilucocionários, mas pela distinção no modo como os realizamos.

As investigações de Searle, segundo Villanueva (1975), partem do pressuposto de que o ato de falar uma língua caracteriza-se como um comportamento regulado por regras. Assim, ao oferecer caracterizações e explicações como falante de uma língua, estas devem seguir regras intrínsecas a essa língua. Essas regras se distinguem em regulativas, que regulam as formas de conduta, e constitutivas, que criam novas formas de conduta.

Silva (2008) destaca que Searle considera o ato ilocucionário como a unidade mínima da comunicação linguística. Searle também afirma que a proposição não é o ato em si, mesmo que a expressão desta proposição seja parte de determinados atos ilocucionários.

Searle (1991) indica que a realização do ato ilocucionário também depende da adequação a uma conduta, sendo esta governada por regras, isto é, "[...] conhecer o significado é conhecer as regras para seu uso, não apenas o significado ancorado no nível léxico das palavras empregadas nas proposições".

Para Searle, a questão do significado seria também uma questão de convenção, dessa forma, a emissão de uma proposição só será capaz de produzir um ato caso este esteja em concordância com as regras convencionalizadas. Em uma análise ideal dos atos ilocucionários, deve ser possível a observação de aspectos intencionais e convencionais desses atos, e ainda a relação entre estes aspectos (Silva, 2008, p. 241, tradução nossa).

Na teoria de Searle, uma das bases do significado é a distinção entre o conteúdo e a força ilocucionária da proposição. Garmendia (2009) atribui esse fundamento como herdado da semântica de Frege, e Silva (2008) realça a afirmação de Searle de que, mesmo que nem toda expressão possua um conteúdo proposicional, toda expressão possui uma força ilocucionária.

A emissão de diferentes atos de fala - como uma ordem, uma pergunta ou enunciado - com o mesmo conteúdo proposicional exige, além de sua emissão, a distinção de seus atos proposicionais - predicado e referência -, que formam o conteúdo da sentença, e da força ilocucionária, que constitui os atos de enunciar, perguntar, ordenar, entre outros (Villanueva, 1975).

Para além dos critérios dos atos de fala, Searle ainda propôs a classificação dos atos ilocucionários em cinco categorias, já que, para ele, a identificação de verbos seria apenas uma forma de identificação de atos, e não uma metodologia para a distinção de seus tipos:

- Os atos assertivos cria um comprometimento entre o falante à condição de verdade da sentença e são simbolizados por "$\vdash \downarrow B(p)$"[8].
- Atos diretivos têm como ponto ilocucionário conseguir que o ouvinte realize algo, sendo simbolizados por "! ↑ *W* (*H* realiza A)"[9].

[8] ⊢ para asserção, ↓ para o ajuste das palavras ao mundo, *B* para crença (*belief*) e *p* para conteúdo proposicional.
[9] ! para marca dos diretivos, ↑ para ajuste do mundo às palavras, *W* para estado de desejo (*want*), *H* para ouvinte (*hearer*) e *A* para ação futura.

- Atos comissivos fazem com que o falante se comprometa com uma ação e são representados por "*C* ↑ (*S* realiza *A*)"[10].
- Atos expressivos demonstram estados psicológicos sobre o estado das coisa e simbolizado como " *E* ∅ (*P*) (*S/H* + propriedade)"[11].
- Declarações forjam correspondências entre o conteúdo da proposição e a realidade, normalmente fundamentadas em uma instituição extralinguística, e são representadas através da simbologia " *D* ↕ ∅ (*p*)"[12].
- Em última instância, existem casos em que as declarações se alinham aos atos assertivos, em instituições definidas, onde, além da afirmação de um fato, precisa-se de uma autoridade que pronuncie uma decisão, e estas também possuem uma representação "*Da* ↓↕ *B*(*p*)"[13].

As colaborações de Searle ainda se estendem aos atos de fala indiretos, quando um falante, ao proferir a proposição, quer dizer algo além do que diz, além daquilo que está sendo dito. O problema, como sintetiza Armengaud (2006, p. 121), surge ao tentar compreender como o ouvinte de tal proposição pode assimilar tal intencionalidade quando aquilo que é proferido significa outra coisa.

Searle sugere que o falante, ao concretizar este tipo de ato, comunica aquilo que enuncia em uma situação baseada em informações de natureza linguística e extralinguística compartilhadas durante a interação, além da capacidade de inferência de seu ouvinte.

Analogamente, Blair (1992) encara como maior distinção entre as classificações dos atos ilocucionários de Austin e Searle o fato de que a classificação desenvolvida por Austin, na realidade, refere-se a uma classificação de verbos ilocucionários, enquanto Searle classifica atos ilocucionários.

Bachioqui (2010) reforça que os elementos pragmáticos do discurso, entre eles os atos de fala, cumprem uma função imprescindível ao ouvinte e intérprete,

---

[10] *C* para marca dos comissivos, ↑ para ajuste do mundo às palavras, *I* para intenção e *S* para falante (*speaker*).
[11] *E* para o ponto ilocucionário em comum dos expressivos, ∅ para a não existência de ajuste entre mundo e palavras, *P* para variável da diferença entre valores e estados psicológicos, *S* para falante (*speaker*) e *H* para ouvinte (*hearer*).
[12] *D* para o ponto ilocucionário das declarações, ↕ para a falta de sentido do ajuste entre mundo e palavras, ∅ para a ausência da condição de sinceridade e *p* para o conteúdo proposicional.
[13] *Da* para o ponto ilocucionário das declarações assertivas, ↓ para o ajuste da asserção, ↕ para o ajuste da declaração, *B* para a condição de sinceridade de crença (*belief*) e *p* para o conteúdo proposicional.

compondo o sentido de uma declaração a partir de elementos como a entonação, o modo verbal, a ênfase e a ordem das palavras, responsáveis pela expressão da força ilocucionária deste discurso. A intencionalidade que fundamenta o discurso compõe o reflexo da força ilocucionário do ato e acaba por relativizar o significado das orações à interpretação de seus significados proposicionais.

Resumidamente, a teoria dos atos de fala de Austin e Searle tem como pressuposto fundamental os atos realizados como unidades mínimas comunicativas, e não as palavras, frases ou expressões, que foram integrados junto à semântica por Austin, isto é, que realizam atos ou ações ao serem pronunciados no contexto comunicacional. Já Searle contribui para a ampliação da teoria inicial, propondo que o significado é resultado dos mais diversos aspectos, entre eles o pragmático, além das regras e estruturas.

Assim, conclui-se que o campo dos estudos pragmáticos mostra sua amplitude por meio dos mais diversos autores e teorias, sejam elas fundamentalmente lógicas, filosóficas, transcendentais ou linguísticas, exercendo “[...] desde já um poder integrador, antes de ter alcançado a própria unificação”, além de constatar que “[...] nela se veem ajustados o prático e o teórico, visto que a atitude para com a verdade está ligada ao movimento da reta comunicação do sentido” (Armengaud, 2006, p. 154).

2.3. Pragmática e ciência da informação

Nesta seção, é apresentada uma síntese das produções sobre as confluências entre ciência da informação e pragmática, e está organizada da seguinte forma: 1) considerações gerais sobre a ciência da informação e a representação como atividade organizadora da informação; 2) o papel da linguagem na ciência da informação; 3) introdução da pragmática à ciência da informação; 4) relações entre linguagem, pragmática e busca da informação; 5) caráter pragmático dos processos, atividades e instrumentos da ciência da informação.

No seio da sociedade da informação, a ciência da informação compromete-se com a responsabilidade social de buscar, coletar, tratar, representar, armazenar, preservar, disponibilizar e possibilitar o uso da informação para aqueles que dela necessitam, assim como disposto por Wersig e Nevelling (1975).

Como forma de atender tal responsabilidade, a ciência da informação busca novos aportes teórico-metodológicos para a solidificação da área, visando reagir aos problemas informacionais manifestados na sociedade da informação (Ribeiro, 2010).

Como forma de atender as demandas e necessidades informacionais sociais, processos de organização, como a representação da informação – sendo este o procedimento responsável pela "substituição" ou "tradução" de expressões linguísticas completas por termos e conceitos representativos de seu conteúdo – se fazem centrais e de grande relevância.

Para além da informação, Putnam (1988, p. 22, tradução nossa) alude a inexistência de um método representativo que seja capaz de manter a relação entre representações e aquilo que elas se referem de forma intrincada, já que as representações se associam de modo contingente e mutável ao seu referente, "[...] capaz de mudar à medida que a cultura muda ou conforme o mundo muda".

Nesta mesma linha de pensamento, Novellino (1998) chega a aludir sobre o significado da informação não estar predeterminado por aquele que a produz, mas vai sendo determinado durante sua comunicação, e pondera, além do conteúdo informacional, os contextos de produção e prováveis contextos de utilização.

> Segundo os linguistas, os diferentes contextos participam da construção da significação. Vão intervir o contexto externo e social, que gera as significações e o contexto expressivo individual, que gera o estilo pessoal. Os 'valores', no sentido da semântica, são as associações contextuais que precisam o sentido das palavras (Mucchielli, 1995, p. 114).

Azevedo Netto (2002) também discute e expõe um "alargamento das discussões" na ciência da informação, expandindo-se rumo às considerações sobre os meios de compartilhamento da informação, as questões de contexto de produção e suas condições de consumo.

Inerente e primordial a essas atividades encontra-se a linguagem, que cumpre uma função mediadora entre indivíduos de uma comunidade através de um código comum e arbitrário, tal qual sugere Ribeiro (2010, p. 65):

> A mediação surge, portanto, através da emergência de uma linguagem, de um sistema de representações comum a toda a sociedade, a toda a cultura, e, ao mesmo tempo, a emergência deste sistema de representação constrói um sistema social, coletivo, de pensamento, de relações, de vida – uma sociabilidade, entendida como o conjunto de condutas, de representações e de práticas pelas quais é reconhecida numa pessoa a sua pertença a uma sociedade ou

> que são comuns a todos os que pertencem a uma mesma comunidade.

Capurro (2003 e 2007) chega a indicar o papel inerente da linguagem nas funções essenciais da representação e, por conseguinte, na recuperação de símbolos científicos e culturais. Ainda, Martines (2023) destaca a existência de pesquisas que buscam a identificação de teorias da linguagem na teoria da informação.

As interações entre linguagem e CI são explicadas por Saldanha (2008, p. 180) partindo da condição da linguagem como sistema de informação primeiro, constituindo um meio de armazenamento, acesso e compartilhamento de informações, funcionando, assim, como suporte, local de guarda e preservação da informação, de modo que “a linguagem estrutura a informação, assim como permite – funciona como – a representação, assim como propicia a comunicação” (p. 228).

Além do mais, com base em Le Coadic (1996, p. 5), Martines (2023, p. 10) constrói a correspondência entre informação e sua materialidade na linguagem, já que a informação apresenta o elemento de sentido, independentemente de sua materialização em linguagem verbal ou não verbal. A inscrição dessa informação ocorre por meio de um sistema de signos, isto é, linguagem.

Além de conteúdo, os documentos de registros informacionais tem um propósito, um uso concebido por seu criador e outros em potencial, determinados de acordo com cada usuário e suas necessidades.

Seguindo esse enfoque, a pragmática surge trazendo uma nova abordagem para o campo da ciência da informação, na qual as propriedades e processos informacionais são compreendidos através de um viés social-cognitivo e comunicacional, levando em conta a esfera em que as ações de informar, se informar e ser informado ocorrem.

Assim nasce uma pragmática informacional que, segundo González de Gómez (1996), se estabelece como o conjunto de procedimentos para a concepção e uso da informação em situações efetivas intermediadas por indivíduos interligados pelo contexto social e comunicacional que se encontram.

> O homem se conecta ao mundo da informação através também de objetos sensíveis, mas à Biblioteconomia – e, por extensão, a CI – interessam certos objetos específicos, criados especialmente para conectar ao mundo da informação: os signos linguísticos articulados – ou seja, a linguagem em todas as suas manifestações (Saldanha, 2008, p. 221).

Fernandes, Pinto e Azevedo Neto (2005) já haviam estipulado o atrelamento entre as questões de representação da informação aos estudos de natureza linguística e semiótica.

A partir da chamada “virada pragmática” verificada no campo da linguística observa-se a passagem de uma visão da linguagem para além de um paradigma estruturalista de considerações acerca do signo, significante, significado e funcionalismo, e passa a focar nas questões próprias do texto em seu sentido integral, dos contextos (situacionais, temporais, espaciais, culturais e sociais) e da recepção das interações comunicativas (Kobashi; Fernandes, 2008).

Segundo Marcondes (2000, p. 40), os estudos da linguagem - tal qual seus usos - devem pressupor a pragmática como prática social, podendo assim compreender os significados a partir da relação entre os participantes da interação comunicacional, o contexto em que essa ocorre, os elementos culturais e sociais, seu objetivo e os efeitos dessa interatividade.

Putnam (1988) ainda reafirma o **c**aráter interacional do significado, aliado a função do ambiente na determinação das referências da comunicação de um determinado falante ou de uma comunidade de falantes.

Saldanha (2008, p. 147) também reafirmou o papel teórico-epistemológico da pragmática na ciência da informação concluindo que esse novo aporte leva à alternância da questão fundamental de “o que é a informação?” por “o que é a informação para?”.

A partir disso, a CI passa à exploração das “[...] formas de comunicação no cotidiano dos indivíduos em suas comunidades específicas, comungando da multiplicidade de jogos de linguagem que permitem suas trocas simbólicas” (Saldanha, 2008, p. 147).

No entendimento de Kobashi e Fernandes (2008, p. 4), “[...] a pragmática favorece as reflexões sobre os contextos em que ocorrem as operações de organização de registros documentais; favorece, igualmente, a análise e a compreensão dos processos de busca e recepção de informação”, e sem a compreensão dos elementos contextuais em sua função de delimitadores dos significados de expressões e conceitos linguísticos “[...] o usuário torna-se refém dos efeitos indesejáveis provocados pela ambiguidade das informações armazenadas”.

A premissa de que a informação nasce a partir da ação habitual de indivíduos de uma determinada comunidade discursiva torna as interações sociais, as linguagens

específicas e as condições comunicativas objetos da ciência da informação, constituindo processos informacionais que colaboram para a caracterização da informação e de seus fluxos, “[...] assim, toda análise informacional é – antes de uma ‘análise informacional’ pura - uma observação de contextos e suas movimentações” (Saldanha, 2008, p. 216; Borko, 1968).

Rendón Rojas (1996) é relembrado por Weiss e Bräscher (2014, p. 685) nas considerações acerca da informação pragmática, onde “[...] a dialética do sujeito com o mundo que o rodeia é levada em consideração”, sendo esta capaz de contribuições significativas para o campo da ciência da informação.

Para além, a informação passa a ser encarada como o resultado de elementos objetivos e subjetivos, enquanto seu acesso dependerá da elaboração coletiva do significado do conteúdo de um documento, “[...] logo, acesso, para esta tradição, não é representação, mas um coletivo de interpretações” (Saldanha, 2008, p. 177).

Fundamentando-se em Popper, Pacheco (1995) visualiza a informação como resultado de uma construção humana para um uso dentro de contextos culturais e sociais determinados, sendo assim um artefato que a criação está disposta em condições espaciais e temporais específicas - formando seu contexto ideal de interpretação - e, por poder ser utilizada em condições diferentes da qual foi criada, sujeitando-se à (re)contextualizações.

Em um mesmo movimento de raciocínio, Hjørland (1998) também observa o condicionamento do significado às condutas sociais, resultando nas formas de comunicação e interação e nos hábitos verbais e não-verbais.

O autor também estabelece que, a partir deste ponto de vista, as relações entre signos e suas referências deixará de ser relacionada aos objetos e suas existências individuais e passará a ser determinada pela cultura, as disciplinas de uma área científica e pelas comunidades discursivas, pressupondo que “[...] a sociedade é composta por diversas ‘comunidades discursivas’ que desenvolvem sua própria linguagem especial, seus tipos documentais, seus sistemas informacionais, suas instituições, e seus papeis profissionais para manter suas funções” (Hjørland, 1998, p. 22, tradução nossa).

Será a partir do trabalho de Armengaud (2006) que Kobashi e Fernandes (2008) elaboram a correspondência entre a classificação dos contextos e a organização e recuperação da informação, como apresentado no quadro 6.

Quadro 6 – Tipos de contextos associados à organização e recuperação da informação

CLASSIFICAÇÃO DO CONTEXTO	CORRESPONDÊNCIAS NA OI/RI
Circunstancial, factual, existencial, referencial	Dispositivos e unidades informacionais institucionalizados
Situacional ou paradigmático	Áreas do conhecimento
Interacional	Interface de mediação
Pressuposicional	Expectativas

Fonte: Kobashi e Fernandes (2008, p. 7).

Nesta analogia, o contexto circunstancial, factual, existencial, referencial - definido pelos interlocutores, o tempo e o local que encontram-se - se emparelhará aos dispositivos e unidades informacionais institucionalizados; o contexto situacional ou paradigmático se ligará às áreas do conhecimento em razão de sua consideração do contexto cultural; o contexto interacional diz respeito às interfaces de mediação em razão de seu objetivo pragmático; e o contexto pressuposicional caracterizará as expectativas compartilhadas pelos interlocutores (Kobashi; Fernandes, 2008).

Da mesma forma os produtos dos processos de representação da informação, como os índices e resumos, também serão dotados de uma linha pragmática, considerando seu desenvolvimento a partir do objetivo específico de recuperação da informação, assim como o conceito de “termo preferido” em tesauros, que também é selecionado a partir de elementos contextuais e de ordem pragmática (Kobashi, 2007).

Blair (2003) aponta que o processo de representação informacional está sujeito a empecilhos, como a utilização de termos incorretos na descrição, o uso de termos generalizados, ou até mesmo a seleção incorreta de informações relevantes em razão da interpretação incorreta desta representação. Todos esses são problemas linguísticos, e em certa abordagem, pragmáticos. Assim, é retomada a afirmação de Gracioso (2010, p. 140):

> Os esquemas de representação da informação pertencem à agenda de estudos da Ciência da Informação há décadas e autores, sob diferentes abordagens, já sugeriram a expansão sobre o uso de conceitos como descritores nos sistemas de representação da informação, que se aproximassem de seu contexto de uso.

O uso da linguagem no processo de organização e recuperação da informação foi articulado em duas maneiras principais: a descrição e a discriminação da informação. Em conjunto, os processos caracterizam os processos de busca pela

informação. É também nestes processos em que o uso da linguagem está mais propenso a falhas, como foi demonstrado por Blair (2003).

No processo de busca pela informação, a linguagem se apresenta como instrumento de interação com o sistema de informação, para a descoberta e interpretação dos recursos resultantes da busca, modelando ainda o relacionamento entre os agentes da interação, como demonstra Benoît (2001).

Da mesma forma, delineada como um processo que auxilia as práticas sociais, culturais, profissionais e diárias de um usuário, a busca por informações, “[...] o uso da linguagem, à luz de algum conceito, influencia a interpretação do ouvinte do discurso” (Benoît, 2001, p. 5, tradução nossa).

Quanto à descrição, as maiores dificuldades surgirão quando a descrição é realizada de forma errônea, seja por desconhecimento do assunto ou uso de descritores inadequados. Outro erro que pode ocorrer durante esta atividade, mesmo que de forma sutil, nas palavras de Blair (2003), é o uso de descritores corretos em situações em que o usuário que busca a informação não é capaz de compreender o uso do descritor e, de forma consequente, não é capaz de recuperar as informações necessárias.

Será neste último cenário que a pragmática emergirá, já que o processo de escolha de descritores não deve ser restrito à perspectiva semântica de cada termo, mas deve levar em conta aspectos relacionados aos usuários que são o alvo final da descrição e os usos que estes farão da informação. De nada vale o uso de um descritor semanticamente correto se este termo não se adequa pragmaticamente à busca do usuário.

Já na discriminação - responsável pela separação entre informações relevantes e irrelevantes à busca do usuário, a partir da descrição dos conteúdos -, a complicação surgirá a partir da descrição demasiadamente geral da informação, que impedirá que esta seja discriminada de informações irrelevantes à necessidade do usuário.

Nestas situações, o componente pragmático está presente na consideração das necessidades informacionais dos usuários - ou o uso que estes pretendem dar à informação - que guiará a discriminação destes conteúdo.

De uma mesma forma, a informação só possuirá atributos significativos em situações e práticas diárias onde está se torna útil para um receptor. Essa ideia de utilidade dependerá das delimitações de cada um dos contextos do usuário: as atividades que estão sendo realizadas, o momento em que a informação é necessária,

os conhecimentos que ele já possui, suas interações com outros interlocutores, o ambiente que se encontra, sua cultura e crenças etc.

Blair (1992) sintetiza tal contemplação ao pronunciar que, além de conteúdo, documentos possuem usos que não são dedutíveis unicamente pelo texto, dependendo também de condições contextuais.

Foram delineados por Benoît (2001) os diferentes pontos de vista de compreensão entre os desenvolvedores de sistemas informacionais e os usuários dos ditos sistemas.

Enquanto o designer de um sistema busca a adequação de unidades semânticas a um modelo fundamentalmente matemático como forma de validação do sistema, o usuário final do sistema busca pela veracidade, confiabilidade, racionalidade e clareza. No mesmo sentido, os designers de sistemas buscam na linguagem uma ferramenta puramente técnica e isolada de contextos, enquanto, na perspectiva daqueles que usam o sistema, a linguagem representará uma forma de expressão que se situa em um ambiente contextual (Benoît, 2001).

Com isso em mente, sistemas informacionais devem ser aproximados destes contextos, atividades e práticas de seu público-alvo e usuários em potencial, como Blair (2003) esclarece. Essa aproximação possibilitará a observação das particularidades desse domínio, permitindo a modelagem e adequação dos sistemas de busca e recuperação da informação a essas especificidades, sendo a pragmática capaz de fazer a ponte entre contexto e sistema a partir de descrições e representações do sistema.

Hjørland (1998) também estabelecerá a importância da compreensão dos documentos, suas particularidades, tipos, estruturas e funções para o desenvolvimento de sistemas informacionais.

Além do mais, Blair (2003) ainda recomendará, como forma de aproximação dos sistemas informacionais e os contextos em que estes são utilizados, a modelagem de sistemas com a possibilidade de interação em tempo real entre aqueles que buscam pela informação e os responsáveis pelo sistema e representações nele presentes. Esta abordagem ainda contaria com ferramentas de feedback dos usuários do sistema para a adaptação das representações informacionais.

Blair (1992, p. 202) também forja a ligação entre a teoria dos atos de fala e a recuperação da informação, que, segundo o autor, seria capaz de indicar os possíveis

usos documentais, para além do conteúdo documental presente no texto, que normalmente não se mostram claramente através do texto de um documento.

De certa perspectiva, os sistemas de classificação também contam com um viés pragmático. Após a definição do assunto a ser classificado, o profissional bibliotecário se baseará em uma perspectiva e buscará pela notação mais apropriada para a representação.

Esse processo idealmente se fundamentará nas considerações do público-alvo do documento, seu ponto de vista em relação ao assunto, os objetivos do profissional e o propósito último dessa atividade - a recuperação para o uso.

Aliado a isso, volta-se às orientações Beghtol (1986) e ao conceito de "hospitalidade cultural" voltados à inclusão de conceitos e representações associados a representações de seus contextos de uso nos sistemas de classificação, aprofundando o vínculo classificação e pragmática.

Hjørland e Nissen Pedersen (2005) explicitaram relações entre a perspectiva pragmática e os fundamentos da teoria da classificação a partir dos seguintes argumento:

a) A descrição - ou qualquer forma representativa - é a síntese do objeto descrito e das funções e contextos da disciplina responsável pela descrição;
b) A escolha - pragmática - das propriedades do objeto que darão base à classificação devem refletir os propósitos classificatórios, já que essa escolha tornará alguns de seus usos mais fáceis, enquanto limita outros;
c) Dentro de uma mesma área ou domínio, encontram-se diferentes teorias, abordagens, paradigmas e interesses que também irão delimitar os processos representativos e classificatórios, e a mudança de pontos de vista e perspectivas dentro de um domínio acarretará a reformulação e mudança de seus significados.

Martines (2023, p. 8) enfatiza a aplicação da linguagem como instrumento na análise do discurso nos processos de tratamento documental e no desenvolvimento dos sistemas de organização do conhecimento (SOCs), caracterizando-se como "[...] um ponto de conexão entre os dois vetores: a elaboração e a aplicação das linguagens documentárias".

Kobashi (2007) destaca a importância dos contextos de produção nos processos de busca e recuperação da informação, sendo fundamental também na representação das informações nos sistemas informacionais.

Dessa forma, o desenvolvimento de linguagens próprias para a representação informacional também devem contar com abordagens de cunho histórico, social, econômico, pessoal, ambiental e temporal.

> Como construtores e usuários de sistemas de informação somos, antes, estudiosos de uma linguagem capaz de transmitir significados e representações. Logo, linguagem é uma categoria de análise fundamental na CI (Saldanha, 2008, p. 227).

As linguagens documentárias nascem com o propósito de representar a informação, sendo dessa forma importantes instrumentos mediadores, propiciando a interação entre usuário e sistema informacional. Essas linguagens são dotadas de um caráter pragmático, já que as escolhas de termos preferidos são puramente pragmáticas, preocupando-se primeiramente com circunstâncias próprias do domínio e área de aplicação da linguagem. Tal fato se reflete ao observarmos as notas de escopo que acompanham cada um dos descritores de uma linguagem representativa, responsáveis pela aproximação do termo de seu contexto ideal de uso.

Dessa forma, “[...] no contexto da organização da informação, informação é, sobretudo, elemento de sentido a ser comunicado e interpretado. Essas operações só podem ser viabilizadas por meio de linguagens apropriadas” (Kobashi, 2007, p. 7).

A premissa pragmática de que uma língua poderá ser compreendida em suas formas de uso concomitante às práticas sociais de uma comunidade se estenderá também às linguagens documentárias, que serão amplamente compreendidas através do conhecimento das formas de uso de cada um de seus elementos a partir das notas de escopo, que aproximarão o profissional indexador e até mesmo o usuário do contexto de cada um dos descritores de determinada linguagem.

> Se quisermos saber o que as descrições usadas para representar um documento *significam*, devemos observar como essas descrições são usadas nas atividades e práticas que fazem uso dessa informação (Blair, 2003, p. 37, tradução nossa).

Esse tipo de afirmação também foi trabalhada por Putnam (1988), que segmentou o conhecimento do significado em (1) ser capaz de realizar sua tradução, (2) conhecer sua referência, ou (3) possuir conhecimento tácito, ou seja, prático, de um significado, isto é, conhecer as formas de uso de uma palavra em um discurso.

São evidenciadas por Saldanha (2008, p. 234-242) alguns dos reajustes que a chamada tradição pragmática promoveu na ciência da informação:

a) Tentativas do alinhamento da faceta prática das investigações informacionais e da multiplicidade de discursos que colaboram para a (re)formulação da informação;
b) Ponderação de representação em relação a um contexto específico; possibilidade de investigações acerca de discursos e comunicação cotidianas como forma de compreensão do fenômeno informacional como social e comunicativo;
c) Estudos das questões subjetivas que cercam a informação e seus usuários, como relações sociais, a cultura, e a linguagem em determinadas comunidades; expansão dos estudos da CI para além do científico e formal, incorporando discursos, conhecimentos e modelos ordinários.

O trabalho de Saldanha (2008, p. 255) ainda aponta o deslocamento do objeto da CI, no qual a informação e o homem informado não são seus objetos, assim como não são sua missão; o objeto da CI passa a ser o homem contextualizado em uma comunidade discursiva determinada, enquanto sua missão torna-se a (re)contextualização os usos da informação pelo homem em sua comunidade.

Paralelamente, Martines (2023) propõe a mudança de relação e tratamento do objeto na ciência da informação, que adquiriu preocupações relativas ao processo de significação e o discurso como forma de ponderar as necessidades, interações sociais e o cotidiano das comunidades discursivas.

Para além, a atividade científica dos processos de organização busca formas de traduzir o conhecimento expresso em um documento em conceitos “[...] com o potencial de recuperar não somente o documento, mas também todo o conhecimento já construído em torno dele, o qual o descritor assume a função social e convencional de recuperar toda a malha de informações e conhecimentos que dialogam” (Martines, 2023, p. 15).

Por fim, além dos processos de tratamento documental, surgem novos temas de investigação na organização da informação e do conhecimento, como o papel dos sistemas de informação e a representatividade de grupos culturais, a mediação da informação, a concepção de objetos informacionais e a atuação de sujeitos informacionais, tal como demonstra Martines (2023).

Em suma, a ciência da informação, em busca de ampliações teórico-metodológicas, incorpora conceitos e abordagens de natureza pragmática como forma de melhor atender as mais diversas demandas informacionais.

A linguagem, como elemento intrínseco à informação, mostra-se como um elemento fundamental ao ciclo informacional, estando presente nos mais diversos contextos, desde a produção, passando pela organização e armazenamento, até a busca e o uso da informação, destacando seu papel formador e mediador do significado e das interpretações contextuais da informação.

Dessa forma, a ciência da informação, como ciência social, engaja-se nas considerações interacionais que constituem a relação entre usuário - como sujeito de um determinado contexto social, histórico e comunicacional - e informação - registrada, organizada e recuperada por meio da linguagem.

2.4. Processamento da linguagem natural

Como ponto de convergências entre linguística e ciência da computação, o processamento de linguagem natural (PLN) concentra seus estudos para que máquinas sejam capazes de analisar a linguagem natural com o objetivo de "[...] aprender, atender, reconhecer e produzir conteúdo da linguagem humana" (Falcão; Lopes; Souza, 2021, p. 6).

Tal ferramenta é capaz de, a partir de dados da linguagem humana, extrair conhecimentos e informações de forma automatizada (Zeroual; Lakhouaja, 2018; Hirschberg; Manning, 2015).

É através de um modelo linguístico gerado que o PLN é possível. O desenvolvimento de um modelo efetivo da linguagem depende, inicialmente, da definição prévia das seguintes questões: Que tipo de processamento pretendo realizar? Que tipo de dados são necessários para representá-lo? Onde posso recuperá-los? Qual a melhor forma de organizá-los? (Casado-Mancebo, 2021).

Casado-Mancebo (2021) discute que o *corpus* do processamento da linguagem natural é organizado em listas de objetos - elementos abstratos representantes de comportamentos ou propriedades comuns -, formando conjuntos de elementos que possuem uma mesma propriedade, gerando arquivos de dados estruturados, permitindo sua gestão e acesso.

> O modelo linguístico é, para o linguista, o núcleo do PLN: por um lado, é o primeiro passo no desenvolvimento de qualquer sistema de processamento e, por outro, é onde são necessárias competências de um linguista para ser capaz de captar amostra de linguagem adequada, variada e suficiente (Casado-Mancebo, 2021, p. 749, tradução nossa).

Para isso são utilizados algoritmos, uma série de códigos e procedimentos seguidos em uma ordem lógica e essencial para a execução de tarefas. Dentre os mais diversos algoritmos utilizados, as redes neurais artificiais (RNA) se sobressaem. Seu desenvolvimento fundamenta-se no modelo biológico cerebral com “células inteligentes” configuradas para o processamento de determinado tipo de dados ordenadas em camadas, tornando possível o processamento simultâneo de diferentes tipos de dados (Silva, 2022).

A origem do PLN remonta à expansão das máquinas e computadores capazes de processar textos em linguagem natural, fato que acarretou o avolumamento de dados textuais difícil de ser tratado, pesquisado e recuperado.

Hirschberg e Manning (2015) apontam quatro fatores chaves responsáveis pelos avanços do PLN: 1) aumento da capacidade computacional; 2) grande disponibilidade de dados linguísticos; 3) desenvolvimento do método de aprendizagem de máquina; 4) compreensão da linguagem humana e suas aplicações.

Hirschberg e Manning (2015) ainda mencionam que a PLN pode possuir múltiplas finalidades, como a assistência na comunicação entre humanos, como no caso da tradução automática, facilitadora da comunicação humano-máquina, ou auxílio tanto de humanos quanto de máquinas, como a análise de dados de linguagem humana.

Bräscher (2002) indica a premissa de a utilização da linguagem natural em ferramentas de busca deve basear-se no conhecimento das expressões, suas relações e significados desta linguagem, tornando necessário que essas ferramentas sejam capacitadas para o tratamento de fenômenos linguísticos que possam vir a afetar a qualidade da recuperação da informação, como a ambiguidade.

Vallez e Pedraza-Jimenez (2007) apontam que, no início de seu desenvolvimento, os métodos da PLN foram bem-sucedidos, porém, com sua aplicação em linguagem natural sem vocabulários controlados, muitas dificuldades surgiram.

Entre os problemas que podem surgir durante o uso das técnicas do processamento de linguagem natural, os autores destacaram a variação linguística, assim como a ambiguidade linguística, ambos afetando no processo de recuperação da informação.

A variação linguística, a partir do exposto por Vallez e Pedraza-Jimenez (2007), refere-se à possibilidade de uma ideia ou mensagem ser expressa por diferentes

palavras e expressões, dando origem ao chamado silêncio documental, responsável pela omissão de documentos relevantes à necessidade documental do usuário.

A ambiguidade linguística diz respeito a palavras e frases que permitem mais que uma interpretação de seu significado, fato que gera o ruído documental, incluindo na recuperação da informação documentos que não são significativos para a busca do usuário.

A ambiguidade pode ser classificada, a partir de sua causa, como: morfológica, ao tratar-se de palavras que podem pertencer a mais de uma categoria gramatical; lexical, quando se origina pelo fato e uma palavra possuir mais de uma interpretação; sintática, quando uma construção linguística pode ser expressa a partir e mais de uma estrutura; semântica, gerada pela interpretação das relações significativas entre os termos de uma frase; pragmática, causada durante a tentativa de reconstrução dos valores enunciativos do momento de enunciação do falante (Fuchs, 1996; Beardon; Lumsden; Holmes, 1991; Ferneda, 2003; Bräscher, 2002).

Ambos os fenômenos expostos podem incidir nos mais diversos níveis linguísticos durante a recuperação da informação.

No nível morfológico temos as palavras que seus papeis na estrutura morfossintática se alteram de acordo com seu contexto de aplicação, causando ambiguidades (Vallez; Pedraza-Jimenez, 2007).

No nível sintático, a possibilidade de uma mesma mensagem ter sua estrutura sintática alterada e expressar a mesma ideia também gera ambiguidades (Vallez; Pedraza-Jimenez, 2007).

O nível semântico enfrenta a ambiguidade pois uma mesma palavra pode contar com diversos sentidos, como a polissemia (Vallez; Pedraza-Jimenez, 2007).

À nível pragmático temos alguns termos e palavras que não podem ser interpretados em sua forma literal, dificultando seu processamento automático (Vallez; Pedraza-Jimenez, 2007).

A normalização das variações linguísticas que incidem sobre o processo de recuperação da informação pode ocorrer de três formas distintas (Jacquemin; Klavans; Tzoukermann, 1997).

Na normalização morfológica temos a redução de itens lexicais através da conflação, ou seja, a combinação de representações de termos variantes em uma única forma. Esse processo pode ser realizado a partir do *stemming* (redução da

palavra a seu radical) ou da lematização (redução de verbos ao infinitivo e adjetivos ao masculino singular (Orengo; Huyck, 2001; Arampatzis, 2000; Ferneda, 2003).

Já a normalização sintática há a padronização e frases semanticamente correspondentes em uma forma representativa unificada. A normalização léxico-semântica utiliza fenômenos semânticos, como a sinonímia e hiponímia, para agrupar as semelhanças semânticas (Ferneda, 2003).

Concomitante aos níveis linguísticos da linguagem, temos os níveis de análise dos modelos computacionais e algoritmos, sendo eles: análise morfológica, responsável por determinar as palavras que integram o texto e suas “etiquetas” morfossintáticas; análise sintática, que ocorre com a junção das palavras em sintagmas e frases; análise semântica, identificando o significado das palavras em construções textuais; e análise pragmática, identificando pessoas e objetos presentes no texto, gerenciando o diálogo e definido a estrutura discursiva (Vilares, 2005).

Considerando a grande massa de dados textuais e a necessidade de manejarmos esses dados, há um alinhamento ao *big data*, cujo uma das tendências é proporcionar as competências necessárias para a extração e análise de informações apresentadas nas mais diversas linguagens, entre elas a humana (Jin *et al*., 2015).

O processamento da linguagem passa a ser considerado um campo do *big data* pela disponibilidade de dados linguísticos em formatos digitais desde o início, facilitando esse processo para os pesquisadores, e pelo fomento de melhorias de desempenho através de competições de tarefas compartilhadas, onde empresas e pesquisadores compartilham seus dados e estatísticas e competem para a criação do melhor modelo de linguagem.

O processamento da linguagem natural tem como uma de suas limitações o fato de que suas ferramentas e sistemas só possuem aplicações para línguas de “alto-recurso” (HRLs), como o inglês, chinês, alemão e francês, marginalizando línguas que são usadas por milhões de pessoas (Hirschberg; Manning, 2015).

Ferneda (2003) também aponta a restrição de uso do PLN a determinadas línguas, como o inglês, alemão e, em uma correspondência menor, o português. Tal fato, aliado ao custo elevado e suas técnicas, causa a diminuição de interesse em sua aplicação nos sistemas de recuperação de informação tradicionais.

A indexação, uma das tarefas da ciência da informação em que o PLN é aplicado, se fundamenta na precisão e revocação da recuperação, para que mais documentos relevantes – e pragmáticos - às necessidades do usuário, sejam

recuperados. Porém, a recuperação de documentos ainda passa por algumas dificuldades, como o fato da maioria dos documentos recuperados não serem relevantes para o usuário, a diversidade de formas que um conceito pode ser expresso e a redução do documento a sua descrição (Lewis e Jones, 1996).

Para um melhor rendimento e eficiência, o processamento de linguagem natural é aplicado na indexação e recuperação da informação em tarefas como a sumarização automática, criação de índices para busca, classificação de documentos por sua relevância e avaliação do sistema de recuperação.

Cabe ressaltar que o PLN compreende outras técnicas, sendo escolhidas e até mesmo combinadas para que, de acordo com o objetivo final, seus resultados sejam satisfatórios. Apontadas por Falcão, Lopes e Souza (2021), podemos observar algumas de suas aplicações na tradução automática, resumos, desambiguação de sentido, reconhecimento de entidade nomeada, extração de conteúdo, estrutura de perguntas e respostas e até mesmo na análise de sentimentos, entre outras.

Em sua atuação nos sistemas de recuperação da informação, o objetivo do PLN:

> [...] é melhorar a compreensão e representação de textos por meio da construção de modelos computacionais de linguagem, de forma que as pessoas possam desenvolver programas de computador capazes de realizar diversas tarefas envolvendo a linguagem natural (Zhou; Zhang, 2003, p. 117, tradução nossa).

Vallez e Pedraza-Jimenez (2007, p. 3) descrevem o funcionamento desses sistemas da seguinte maneira no sistema de recuperação de informação: 1) a coleção de documentos do sistema são indexados, dando origem a um índice de suas descrições; 2) análise da formulação de busca do usuário e sua representação para equipará-la ao conteúdo documental; 3) comparação da descrição do conteúdo dos documentos com a representação da busca do usuário e recuperação dos documentos correspondentes em relação a sua relevância e similaridade.

Por sua riqueza de expressões em todas as construções linguísticas, a linguagem natural é repleta de ambiguidades e significados, dificultando o funcionamento efetivo dos sistemas de recuperação de informação, se tornando um empecilho para as tarefas de compreensão de uma consulta, analisar os documentos, combinar a busca e do usuário com os documentos do sistema para definir suas semelhanças.

Assim, Zhou e Zhang (2003) expõem a suposição de que a eficiência do sistema de recuperação da informação é limitado pela distância existente entre a representação da busca do usuário e a descrição dos documentos.

Na recuperação da informação (RI), as técnicas de processamento de linguagem natural são amplamente utilizadas, principalmente, para a descrição do conteúdo dos documentos do sistema e para a representação da busca do usuário, comparando ambas as descrições para recuperar os documentos que mais se adequam àquilo que o usuário busca (Baeza-Yates, 2004; Vallez; Pedraza-Jimenez, 2007).

> O Processamento da Linguagem Natural (PLN) surge como uma possível solução aos problemas relacionados à recuperação de informação pela simples observação de que os documentos e as expressões de busca são objetos lingüísticos. O PLN é um conjunto de técnicas computacionais para análise de textos em um ou mais níveis lingüísticos, com o propósito de simular o processamento humano da língua (Ferneda, 2003, p. 82).

A incorporação da PLN nos sistemas de recuperação e busca da informação revela-se notavelmente benéfica, principalmente pela eficiência do processamento de linguagem natural na análise de palavras, cerca de um milhão de palavras por segundo, e substancial, sendo possível o processamento de grandes quantidades de dados de texto em um curto espaço de tempo (Moreira *et al*, 2020; Gamallo; García, 2012).

Cabe citar que, em sua maioria, as técnicas de PLN são aplicadas na recuperação da informação apenas como uma forma de tornar suas tarefas mais eficientes (Faloutsos; Oard, 1995).

Zhou e Zhang (2003) mencionam que as técnicas de processamento de linguagem natural podem ser empregadas em todas as fases da recuperação da informação.

Na etapa de processamento dos documentos, o PLN auxilia na representação dos documentos para indexação e busca; já no processamento de consultas, auxilia na análise da consulta dos usuários e aprimora a identificação das necessidades dos usuários; por fim, na etapa de correspondência, as respostas são elaboradas a partir de diferentes partes de uma informação, a classificação é feita a partir do nível estrutural e significativo e a correspondência conta com maior precisão e flexibilidade.

Porém, os autores também foram capazes de identificar dois problemas que afetam a relação entre PLN e os sistemas de recuperação da informação que não são

discutidos, sendo a falta de uma estrutura para as pesquisas sobre a PLN em sistemas de RI, resultando em compreensões incompletas, e o fato de que os campos de estudos possuem uma ligação fraca, com a RI sendo um dos diversos campos de aplicação da PLN, e a PLN sendo uma das várias técnicas aplicadas nos sistema de RI.

Ainda no âmbito da recuperação da informação, cabe ressaltar que seu desempenho depende diretamente da indexação, processo de atribuição de termos significantes que representam o conteúdo do documento, ou seja, “indexação é a base para a recuperação de documentos que são relevantes para a necessidade do usuário”, isto é, é a partir de uma representação bem sucedida que os documentos são pragmaticamente recuperados, cumprindo as necessidades do usuário (Lewis; Jones, 1996, p. 2, tradução nossa).

O processamento estatístico da linguagem natural consiste no modelo mais aplicado nos sistemas de recuperação de informação (Manning; Schütze, 1999), onde os documentos se encontram descritos através de termos em um índice.

Nessa abordagem, a estrutura, sua ordem, e os significados não são levados em consideração e consiste em duas etapas, sendo elas a eliminação dos elementos não significativos do documento, e a quantificação das características, ou termos, do documentos, após identificá-los (Vallez; Pedraza-Jimenez, 2007).

O TF.IDF, frequência do termo (TF) e frequência inversa do documento (IDF), configura um dos métodos mais utilizados para determinar a importância de um termo no documento, utilizando sua frequência de aparição no documento, porém dependente de sua frequência de conjunto de documentos de uma coleção (Sandoval, 1998; Casado-Mancebo, 2021).

Já o processamento linguístico da linguagem natural analisa, através de ferramentas linguísticas, os documentos a partir de seus níveis linguísticos - morfológico, sintático, semântico e pragmático - com notas adicionadas a cada nível. Cada análise é detalhada por Vallez e Pedraza-Jimenez (2007): a análise morfológica atribui, a partir de etiquetadores (*taggers*), as categorias gramaticais identificadas pelos traços morfológicos; na análise sintáticas, são identificadas as relações entre as palavras do texto e como essas formam as frases e os sintagmas; a análise semântica busca estabelecer o significado de cada frase ou sintagma que compõe o documento.

Mais recentemente, o PLN passa por mudanças significativas com o uso das redes *Transformer* na execução de suas tarefas. Essas redes, como exposto por

Rothman (2021), apresentam um desempenho superior e um treinamento mais rápido. A partir de seus *encoder* e *decoder*, responsáveis pela criação de modelos de linguagem, outros modelos de IA foram criados para geração de textos automática, como os modelos *GPT*, ou para sumarização, como o *BERT* (Silva, 2022).

Na recuperação da informação, o processamento de linguagem natural (PLN) não é aplicado como um modelo por não possuir uma estrutura de representação dos documentos e nem a função de busca, como demonstra por Ferneda (2003), porém, suas técnicas trazem um campo de convergência de métodos entre a RI e a inteligência artificial e soluções para alguns os problemas do processo de RI.

Ademais, o processamento de linguagem natural pode ter sua aplicação maximizada ao ultrapassar os níveis sintáticos e semânticos, passando a uma abordagem fundamentada no nível pragmático da linguagem.

No quesito interação homem-máquina, torna-se de extrema importância que o sistema de processamento seja capaz de reconhecer a intencionalidade comunicativa e o objetivo do falante durante a interação para que a geração do modelo linguístico seja capaz de correspondê-lo (Casado-Mancebo, 2021).

Diante da possível e crescente aplicação das técnicas de PLN nas atividades da ciência da informação, em sua grande maioria apoiadas na linguagem, nos deparamos com a oportunidade de aperfeiçoamento da representação, organização, armazenamento, busca e recuperação da informação e do conhecimento, garantindo uma melhor gestão da informação e do conhecimento.

Sumarizando, o processamento da linguagem natural nasce da interseção entre ciência da computação e linguística, de forma que máquinas e programas computacionais sejam capazes de processar e gerar conteúdos na linguagem humana. A associação do PLN a outras técnicas da inteligência artificial, como a utilização de algoritmos de aprendizagem de máquina, viabiliza a realização de atividades como a sumarização e tradução automática, a extração de frases e palavras-chave, e a análise de sentimento.

Por fim, na recuperação da informação, o PLN contribui nos processos de busca e indexação de documentos, auxiliando no controle de fenômenos característicos da linguagem natural, como a variação e a ambiguidade linguística.

3. PROCEDIMENTOS METODOLÓGICOS

Como forma de melhor definir os procedimentos metodológicos adotados, a seção está subdividida de forma a abordar, ordenadamente, o tipo de pesquisa utilizado, o universo e a amostra de estudo, as técnicas de coleta adotadas e as técnicas para a análise dos dados.

3.1. Tipo de pesquisa

O presente trabalho foi realizado nos moldes quali-quantitativos, tendo em vista uma possível triangulação dos dados teóricos sobre as bases linguísticas e pragmáticas dos *softwares* analisados, qualitativamente, e a comparação de seus recursos por meio de tabelas e gráficos, quantitativamente, buscando "[...] abranger a máxima amplitude na descrição, explicação e compreensão do fato estudado" (Marconi; Lakatos, 2022, p. 344).

A pesquisa puramente qualitativa busca a compreensão, descrição e explicação dos conceitos e fenômenos estudados por meio de interpretações de dados de natureza não numérica, como afirma Gil (2022). Já a pesquisa quantitativa obtém seus resultados principalmente por meio de dados numéricos e modelos estatísticos de interpretação, que se "[...] apresentará como um conjunto de quadros, tabelas e medidas" (Sabino, 1996, p. 204).

3.2. Universo e amostra

O universo da pesquisa constituiu-se a partir de todos os *softwares* e programas computacionais encontrados na fase de levantamento, descrito na subseção seguinte.

São registrados 38 *softwares* encontrados na fase preliminar do levantamento que, através de critérios predeterminados de seleção, foram reduzidos a uma amostra de 13 *softwares*.

Focou-se principalmente em: a) *softwares* livres, que estivessem amplamente disponíveis aos possíveis usuários (22 *softwares)*; b) programas que realizassem, necessariamente, análises de conteúdo de documentos completos, sejam eles individuais ou como parte de um *corpus* (19 *softwares)*; c) sistemas que possuíssem, no mínimo, uma funcionalidade que fizesse a identificação/extração de palavras-chaves, independentemente do método aplicado (13 *softwares)*.

A amostra de análise contou com os seguintes *softwares*: Yake!, MAUI, Annif, *Medical Text Indexer*, KEA, WEKA, Finto AI, *Voyant Tools*, Cortext, IRaMuTeQ, AntConc, *Orange Data Mining* e Yoshikoder.

Já os *softwares* que foram identificados na fase de levantamento que não serão contemplados nesta análise são: Google *Cloud* NLP API, MAXQDA, ATLAS.ti, *MonkeyLearn*, NVivo, QDA *Miner*, Transana, IBM Watson, *RapidMiner*, *TagCrowd*, Nuvem Larhud, Aylien, Amazon *Comprehend*, VOSviewer, StArt, RQDA, NLTK, GenSim, SpaCy, TextBlob, Stanford *Core* NLP, Google *Ngram Viewer*, SISA, *The Coupler* e SISTRA.

3.3. Técnicas de coleta

A coleta de dados foi realizada em dois momentos diferentes, justamente por corresponderem a objetivos diferentes do estudo.

Primeiramente, realizou-se uma revisão teórica, a partir de uma pesquisa bibliográfica, nas bases de dados Brapci, Dialnet e Scopus, para o levantamento de produções em português, inglês e espanhol no que se refere aos estudos pragmáticos e semióticos aplicados no tratamento da linguagem, buscando pelos termos “tratamento da informação”, “semiótica informacional”, "processamento da linguagem”, “pragmática”, “análise documental”, entre outros.

A pesquisa bibliográfica (Manzo, 1973, p. 32) seguiu as seguintes etapas: a) levantamento bibliográfico a partir da busca dos termos nas bases de dados indicadas; b) análise do título, resumo e palavras-chave para a seleção de materiais pertinentes; c) leitura dos artigos, capítulos de livros, livros, trabalhos em andamento, teses, entre outros.

A seguir, realizou-se a busca e seleção dos *softwares* que se fundamentam nos elementos linguístico-pragmáticos, principalmente, através da leitura dos materiais do levantamento em bases de dados e pesquisas em andamento, recuperadas, analisadas e selecionadas por meio de redes sociais acadêmicas, como *Research Gate* e outros.

Essas redes destacam-se por seu caráter científico-acadêmico, responsável por formar uma rede de comunicação entre cientistas das mais diversas áreas e permitir a busca, seleção e leitura de estudos em andamento ou já finalizados por meio da divulgação dos próprios pesquisadores, possibilitando a localização de

estudos e pesquisas que tratam de *softwares* de representação e recuperação da informação, mesmo que estes ainda não tenham sido concluídos.

Para tanto, foram realizadas pesquisas no *Research Gate* utilizando os termos “processamento da linguagem natural”, “PLN”, “sumarização automática”, “classificação automática” e “indexação automática”, em português, inglês e espanhol, em busca de aplicações de *softwares* de PLN nas atividades de organização da informação.

3.4. Técnicas de análise

Para a etapa de análise, os dados sobre os *softwares* foram dispostos em uma planilha para a sistematização das informações coletadas. A seguir, foi desenvolvida uma ficha que reúne informações essenciais sobre cada *software*, permitindo uma visualização global das informações e a comparação entre sistemas e recursos presentes em cada um. Apresenta-se um exemplo da ficha no quadro 7.

Quadro 7 - Modelo de ficha de informações para a avaliação de *softwares*

Nome: *Software* X
Desenvolvedores: Universidades, grupos e indivíduos responsáveis pela criação
Data de lançamento: XX/XX/XXXX
Programação: Linguagem de programação utilizada no sistema
Antecedentes: *Softwares* anteriores que possam ter baseado o sistema estudado
Disponibilidade: Licença gratuita ou com assinatura
Objetivo: Meta orientadora do sistema
Repositório: Para *softwares* de código aberto
Extensões: Programas ou APIs que incrementam funcionalidades ao sistema
Áreas de aplicação: Campos científicos, disciplinas e técnicas nas quais o sistema pode ser aplicado
Funcionalidades: Principais funções e ferramentas do *software*
Link: Página inicial do sistema ou página de download
Notas: Informações extras ou observações

Fonte: Elaborado pela autora (2025).

Finalmente, os dados obtidos dos *softwares* e dos materiais teóricos e suas devidas interpretações são apresentados na seção de resultados, que se encontra subdividida entre as informações dos *softwares* e as interpretações destas informações junto às considerações teóricas apresentadas na seção de desenvolvimento.

4. RESULTADOS

Para fins de apresentação, a seção de resultados está dividida em duas subseções, respectivamente: os resultados relativos aos *softwares* analisados, como suas informações de identificação, recursos e funcionalidades, e os resultados das comparações e interpretações feitas sobre as funções dos *softwares* e as bases linguísticas-pragmáticas identificadas.

4.1. Análise dos *softwares*

Nesta subseção serão apresentados os resultados de análises de funcionalidades dos *softwares* selecionados e suas respectivas abordagens, buscando principalmente as funções e métodos baseados em conceitos da pragmática-linguística. Além disso, realiza-se também a apresentação de cada *software* e a exposição dos resultados das análises em gráfico.

4.1.1. Yake!

O *software* de indexação automática Yake! é um programa do *Institute for Systems and Computer Engineering, Technology and Science* (INESC TEC), lançado em 2018, com possibilidade de utilização de versão demo diretamente na *web* (http://yake.inesctec.pt/).

Através de métodos estatísticos, o Yake! tem como objetivo a extração de palavras-chave para a indexação automática de textos de diferentes tipos, domínios e línguas. Suas aplicações vão desde as tarefas de processamento de linguagem natural e análise de conteúdo, chegando até a sumarização de textos e a produção de índices.

Diferente de outros *softwares*, o Yake! não se limita a línguas únicas, a necessidade de um *corpus* externo, instruções anteriores à utilização, extensão do texto, ou linguagens e dicionários, e funciona a partir de um método não supervisionado baseado unicamente em recursos estatísticos de extração em textos.

O *software* realiza as análises dos textos individuais, o que limita o tratamento de grandes volumes textuais, mas, ao mesmo tempo, torna sua aplicação independente da existência de um *corpus* pré-determinado, como explica Campos *et al.* (2019):

> O fato de depender de documentos individuais permite que opere independentemente da existência de um *corpus*. Em uma era de vastas, mas frequentemente não rotuladas, coleções, isso pode ser

> uma grande vantagem sobre outras abordagens em casos onde o acesso a um *corpus* de treinamentos é limitado ou restrito (p. 258, tradução nossa).

A dispensabilidade de *corpus* de treinamento evita certa limitação na utilização do *software*, mas também apresenta certos desafios, entre eles a necessidade de determinar uma lista concisa de descritores - ou termos descritivos, de acordo com os autores - baseada única e exclusivamente no documento, considerando fatores como a estrutura documental, a extensão dos textos e as nuances próprias da linguagem natural (Campos *et al.*, 2019).

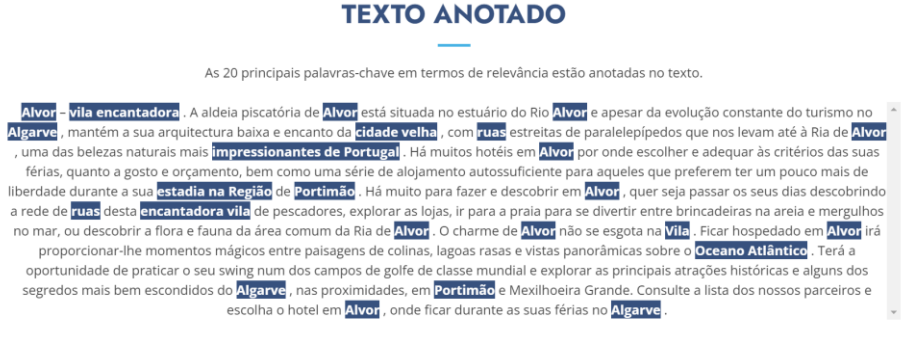
As vantagens do *software* são sistematizadas a seguir: 1) abordagem não supervisionada; 2) independência de *corpus*; 3) independência de língua e domínio; 4) extração de *stopwords*; 5) código aberto para o uso e avaliação.

O processo de extração do Yake! distingue-se em um pré-processamento do texto, junto com a identificação dos primeiros termos candidatos a descritores, extração de suas características, cálculo da pontuação dos termos, geração de n-gramas (sequências de “n” itens de um discurso), e a classificação dos termos[14].

Quanto ao uso, o programa permite que sejam adicionados textos de até 250.000 mil caracteres ou páginas URL, mas não possibilita o *upload* de textos em formatos de arquivo (PDF, DOCX, HTML, XLS etc.). Também é possível selecionar a quantidade de palavras-chave desejadas em uma escala de 1 a 100 termos, e o tamanho máximo dos n-gramas em uma escala de 1 a 10.

As palavras-chave selecionadas como termos descritores são apresentadas em destaque em suas ocorrências no texto analisado (figura 5), além da criação de uma “nuvem de palavras”, na qual o tamanho do termo se relaciona diretamente com sua relevância.

Figura 5 - Ocorrência dos termos descritores destacada pelo sistema Yake!



[14] Para maiores detalhes dos processos e algoritmos do *software*, consultar Campos *et al.* (2019).

Fonte: Yake! (2024).

Por fim, a lista dos termos selecionados são ordenados de acordo com a pontuação de relevância e ocorrência de cada termo, além de a lista ser, de certa forma, interativa, destacando as relações entre os termos à medida que estes são selecionados.

Em síntese, o Yake! proporciona uma análise estatística de textos individuais, sem o apoio de um *corpus* exterior ou instrumentos de linguagem normalizada, e extrai termos que, com base em sua relevância e frequência de ocorrência, se apresentam como candidatos a descritores para a representação do texto.

4.1.2. MAUI

O sistema MAUI (*Multi-purpose Automatic Topic Indexing*) foi desenvolvido por Alyona Medelyan em seu projeto de doutorado em 2009. O sistema faz uso de métodos estatísticos e linguísticos (lexicais) para a seleção de termos baseado em características como ocorrência, probabilidades de ser palavra-chave, e relações semânticas dos termos.

De acordo com Silva, Correa e Gil-Leiva (2020), o sistema possui os seguintes requisitos para a entrada de dados para seu funcionamento: a) adição de um tesauro em formato SKOS; b) radicalizador de palavras para o idioma dos textos que serão indexados; c) lista de palavras consideradas vazias (*stopwords*) no idioma dos textos; d) conjunto de textos para o treinamento do modelo de análise.

Com base nestes requisitos, o sistema realiza a extração dos termos candidatos a descritores, faz o cálculo das características dos termos candidatos (frequência, ocorrência, relações semânticas), treinamento dos modelos de indexação a partir do algoritmo de aprendizagem de máquina e a aplicação do modelo para a seleção de palavras-chave em outros documentos (Silva; Correa; Gil-Leiva, 2020).

Segundo Medelyan (2009), o MAUI é capaz de realizar tarefas como a atribuição de termos de um vocabulário controlado (tesauros), indexação por assunto, indexação baseada nos termos da Wikipédia, extração de terminologia, marcação automática, extração de conceitos, entre outros. O sistema ainda faz uso de algoritmos de aprendizagem de máquina para o desenvolvimento de modelos de seleção dos termos de indexação com base em um conjunto de treinamento inicial.

O sistema de Medelyan também é caracteristicamente independente do domínio de aplicação, e parcialmente independente do idioma. O MAUI já foi aplicado

com êxito em documentos das áreas da ciência da computação, agricultura, medicina, bioinformática, biologia e física.

Além disso, o sistema também é potencialmente independente do idioma dos documentos, sendo capaz de fornecer radicalizados e listas de palavras vazias (*stoplist*) em inglês, espanhol e francês, além de aplicações em português, como o realizado por Silva, Correa e Gil-Leiva (2020).

São disponibilizadas três coleções de documentos[15] utilizadas em experimentos conduzidos com o MAUI, sendo eles: conjunto FAO-30, com 30 documento sobre agricultura indexados com termos do Agrovoc; conjunto WIKI-20 com 20 relatórios da ciência da computação indexados com termos da Wikipédia; conjunto CiteULike-180 extraído das *tags* da plataforma *bookmarking CiteULike*.

As aplicações do MAUI são apoiadas por outros *softwares* constituintes do sistema, sendo eles o KEA (*Keyphrase Extraction Algorithm*), que é utilizado para a seleção de termos candidatos à palavras-chave e a filtragem dos termos utilizados, o WEKA (*Waikato Environment for Knowledge Analysis*) no treinamento dos modelos para a indexação de novos conjuntos de documentos, a biblioteca Jena para o uso de tesauros como o Agrovoc, o MeSH e os cabeçalhos de assunto da Biblioteca do Congresso, e o sistema utiliza o *Wikipedia Miner*, criando uma base de dados.

Resumidamente, o MAUI é um sistema que, além da indexação automática por extração e atribuição, realiza processos de extração conceitual, uso de vocabulários controlados, cruzamento com dados do Wikipédia e extrações de terminologia, baseando-se não somente em métodos estatísticos - como as estatísticas de frequência e ocorrência dos termos -, mas também considerando aspectos linguísticos, como a radicalização de palavras para a diminuição da variedade linguística, e as relações semânticas forjadas entre as palavras e expressões, se aproximando mais do processo de indexação intelectual.

4.1.3. Annif

O Annif é uma ferramenta multilíngue e desenvolvido pela Biblioteca Nacional da Finlândia em 2017 que, a partir de vocabulários controlados e modelos treinados por meio de algoritmos de aprendizagem de máquina e métodos de processamento

[15] Todas as coleções podem ser acessadas e instaladas através da página https://code.google.com/archive/p/MAUI-indexer/wikis/MultiplyIndexedData.wiki.

da linguagem natural, realiza os processos de indexação de assuntos e classificação de forma automatizada.

Suominen (2019) explica que a estrutura do Annif conta com quatro outros algoritmos de indexação de assunto, sendo eles o MAUI para abordagem lexical, e TF-IDF para contagem de frequência, *FastText* para a aprendizagem de máquinas, e *Vowpal Wabbit* para abordagens associadas.

Sua linha de comando (*Command Line Interface*), segundo Suominen (2019), é utilizada para a configuração, treinamento e avaliação dos modelos desenvolvidos. O treinamento destes modelos é realizado através de documentos de treinamento no formato de texto.

A linha de comando ainda pode ser utilizada na avaliação dos resultados da indexação automática com a indexação manual ou de outros *softwares*, além de atribuir assunto a textos e coleções de documentos, e calcular métricas de avaliação de precisão, revocação, pontuação F1 e ganho cumulativo descontado normalizado.

Em relação às avaliações, o Annif já foi avaliado utilizando os vocabulários Tesauro Finlandês YSA e a Ontologia Geral Finlandesa YS, além do treinamento dos algoritmos TF-IDF e *FastText* (utilizando dados do Finna.fi), MAUI (utilizando os dados de treinamento do próprio algoritmo), e PAV específico e genérico (utilizando os conjuntos de treinamento específicos de cada corpus), e dos *corpus* documentais Arto (artigos da base de dados profissional Arto), JYU *Theses* (teses e dissertações da Universidade de Jyväskylä entre 2010 e 2017), *AskLib* (perguntas e respostas do serviço *Ask a Librarian* das bibliotecas públicas da Finlândia) e Satakunnan Kansa (arquivos digitais do jornal regional de Satakunnan Kansa)[16].

Em relação a suas aplicações potenciais, Suominen (2019) lista a indexação automática sem a validação de um profissional indexador, e a indexação semiautomática com as sugestões de termos que podem ser verificados e alterados pelo indexador.

Porém, formas de uso “não convencionais”, como são designadas, também são possíveis, como o apoio na indexação de materiais impressos a partir de aplicativos para dispositivos móveis, recomendações de documentos baseados em páginas da

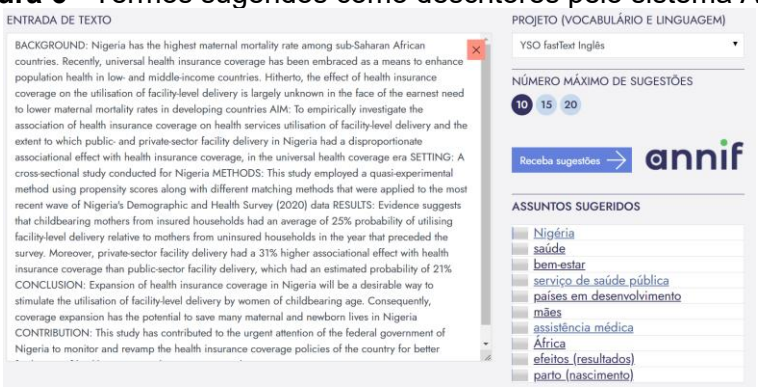
[16] Mais detalhes sobre as avaliações realizadas podem ser encontradas em Suominen (2019).

web por meio de uma extensão para o navegador *Chrome*, e a recomendação de documentos baseados nos interesses do usuário através de um *Chatbot*[17].

O sistema possui uma versão demo que pode ser acessada pela *web* (https://annif.org/) em que é necessário adicionar um texto (sem limite de caracteres predeterminado), escolher entre os vocabulários disponíveis em finlandês, sueco ou inglês, e selecionar a quantidade de sugestão de termos desejada, variando entre 10, 15 e 20 termos.

Após o processamento, as sugestões de termos são apresentadas e fazem ligações diretas com o vocabulário controlado selecionado, sendo possível ver os termos dentro de sua hierarquia, suas notas de escopo e relações (figura 6).

Figura 6 - Termos sugeridos como descritores pelo sistema Annif



Fonte: Finland (2024a).

Resumidamente, o sistema Annif utiliza algoritmos e métodos de análise lexical, estatística de frequência e ocorrência e processamento da linguagem natural em conjunto com vocabulários controlados para o treinamento de modelos de indexação e classificação automatizada, realizando também as métricas de avaliação dos modelos treinados, além de futuras aplicações em potencial que saem do domínio único da representação da informação.

4.1.4. *Medical Text Indexer*

O *Medical Text Indexer* (MTI) foi desenvolvido em 2002 pela Biblioteca Nacional de Medicina dos Estados Unidos (NLM) como produto do projeto *Indexing Initiative*[18], dos anos 1990. O sistema realiza atividades como a indexação automática

[17] As aplicações do Annif na indexação automática e semiautomática já são realidades e são exemplificadas por Suominen (2019), enquanto as aplicações “não convencionais” não passam de protótipos.
[18] Essa iniciativa explorou diversos métodos de indexação, implementando os dois melhores como protótipos, que deram origem ao *Medical Text Indexer*.

e semiautomática e classificação de materiais da área biomédica, usando como base de suas recomendações no vocabulário controlado MeSH (*Medical Subject Headings*), assistindo cotidianamente indexadores e catalogadores da NLM.

As sugestões de representação geradas pelo MTI são listas ordenadas formadas de Lista de Cabeçalhos Principais do MeSH (MH), Subcabeçalhos (SH) e *CheckTags* (CT). Os MHs são os descritores autorizados do vocabulário, enquanto os SHs são qualificadores destes descritores, e os CTs são descritores especiais que abrangem aspectos como períodos históricos, sexo, espécies e faixas etárias, inerentes a todos os materiais, como apontam Mork, Yepes e Aronson (2013). Os resultados do processamento do MTI são armazenados em uma base de dados para o uso posterior pelos profissionais indexadores.

O sistema do MTI é formado por diversos componentes com funções definidas, como o *MetaMap Indexing* (MMI), que classifica os conceitos indicando sua relevância para o texto a partir de fatores de frequência e relevância, o *PubMed Related Citations*, que calcula a similaridade entre documentos através das palavras em comum, a restrição das sugestões de termos ao MeSH, partindo de sinônimos e relações entre os conceitos, e o agrupamento e classificação apropriado dos termos.

Mork, Yepes e Aronson (2013) apontam que a performance e os resultados obtidos com o MTI são aperfeiçoados com algoritmos de aprendizagem de máquina como forma de melhorar a qualidade de revocação e precisão dos termos mais frequentes do MeSH.

A filtragem[19] dos termos selecionados pelo MTI pode ser realizada em três níveis: filtragem de alta recuperação, que gera cerca de 25 recomendações mais inclinadas à recuperação que à precisão; filtragem balanceada de recuperação/precisão, mediando entre recomendações mais genéricas (foco na recuperação) e recomendações mais específicas (foco na precisão); filtragem de alta precisão, que gera uma lista pequena de recomendações mais específicas.

As operações de pós-processamento dos textos têm como objetivo “assear” a lista de sugestões finais, removendo termos inadequados ao público final da representação e adicionando CTs e outros descritores MHs.

[19] Para mais detalhes sobre os níveis e procedimentos de filtragem, consulte Mork, Yepes e Aronson (2013).

Primeiramente, identifica-se os usuários finais daquele material para a remoção de termos inadequados das recomendações, com base em uma lista de exclusão predeterminada. A seguir, revisa-se os MHs geográficos e CTs da lista, além da adição de outros MHs e SHs potencialmente relevantes. Finalmente, ocorre mais uma adição de "lista forçada" de termos que funcionam como "gatilhos" para outros descritores no texto[20] (Mork; Yepes; Aronson, 2013, p. 7)

Por fim, o MTI anexa SHs a partir de algoritmos baseados em métodos estatísticos e linguísticos que, em conjunto, são capazes de determinar os qualificados adequados para cada descritor, baseando-se fundamentalmente nas regras especificadas pelo MeSH.

Sintetizando, o *Medical Text Indexer*, a partir de análises de frequência, ocorrência e similaridades lexicais, apoiado pelo uso do vocabulário MeSH, e aperfeiçoado a partir de técnicas de aprendizagem de máquina, gera uma lista ordenada e valorada de sugestões de descritores e qualificadores próprios para a representação de materiais bibliográficos da área médica, considerando também o público final da representação, além dos aspectos preferidos da recuperação, sejam eles a recuperação de mais materiais ou a recuperação de materiais mais precisos.

4.1.5. KEA

O sistema KEA (*Keyphrase Extraction Algorithm*) foi desenvolvido como parte do projeto de pesquisa desenvolvido na Universidade de Waikato, Nova Zelândia. A partir de métodos estatísticos e algoritmos de aprendizagem de máquina para a determinação automática de expressões-chave representativas ao documento analisado, o algoritmo pode realizar tarefas de indexação - com e sem o suporte de vocabulários controlados - e marcação automática de documentos.

Witten *et al.* (1999) afirmam que as expressões-chave (*keyphrases*) extraídas pelo KEA oferecem dados semânticos capazes de "sumarizar e caracterizar documentos" (p. 254) automaticamente. Para isso, o algoritmo identifica candidatos a expressões-chave utilizando métodos lexicais de análise, atribuindo - com base em métodos estatísticos - valores às características de cada candidato utilizando o

[20] Os autores apresentam o conceito de termos "gatilho" a partir do exemplo de um documento que possui o termo "HIV" que, necessariamente, leva à adição do termo "Síndrome da Imunodeficiência Adquirida" na lista de recomendações.

modelo treinado pelo aprendizagem de máquina para estimar as expressões potencialmente mais representativas.

A partir da avaliação do algoritmo, realizada por Witten *et al.* (1999), o uso efetivo do KEA tem como requisitos básicos um conjunto de treinamento de 20 documentos, o texto completo de cada documento, o que melhora sua performance em relação ao uso de resumos, e um *corpus* de 10 documentos para o cálculo das métricas TF-IDF. Todos os documentos processados devem estar em formato TXT.

O funcionamento do algoritmo acontece em duas etapas essenciais: treinamento e extração. O treinamento diz respeito ao desenvolvimento do modelo de identificação de expressões-frase a partir de um *corpus* documental de treinamento com as expressões e palavras-chave atribuídas pelo autor do texto. Para cada documento do *corpus*, o algoritmo identifica expressões candidatas que têm seus atributos valorados, gerando um modelo para a extração e classificação dos termos e expressões extraídos.

A linguagem padrão do sistema é o inglês, porém, pode ser alterada com o uso de *stoplists*, radicalizadores e vocabulários de outras linguagens. Caso um vocabulário controlado seja oferecido, o algoritmo tende a combinar as frases e palavras do documento com as apresentadas no tesauro[21], realizando a indexação por atribuição.

A etapa de extração refere-se à escolha das expressões de novos documentos a partir do modelo treinado anteriormente. Neste processo são identificados e escolhidos a partir da 1) “limpeza” do texto com a divisão do texto em sentenças e palavras; 2) identificação dos candidatos a partir da sequência das sentenças; 3) uniformização das palavras em letras minúsculas e radicalização.

A consideração do valor das expressões e palavras identificadas considera o cálculo de duas características básicas dos termos candidatos: a medida da frequência da expressão no documento em comparação com seu uso geral (TF-IDF), e a distância da primeira ocorrência da expressão em relação ao documento (Witten e*t al*., 1999).

Por fim, pode-se dizer que o KEA é um sistema que busca a extração de expressões e palavras-chave de documentos e coleções a partir de modelos treinados

[21] Alguns exemplos do uso do KEA apoiado em vocabulários controlados podem ser encontrados em https://community.nzdl.org/kea/examples1.html.

por algoritmos de aprendizagem de máquina, métricas estatísticas e análises lexicais, podendo ser aplicado em indexação por atribuição e por extração, além da possibilidade de uso de vocabulários controlados para apoiar a representação dos documentos.

4.1.6. WEKA

O *software* WEKA (*Waikato Environment for Knowledge Analysis*) foi desenvolvido por professores, funcionários e alunos da Universidade de Waikato, na Nova Zelândia, e lançado ao público em 1996. O projeto viabiliza interfaces que permitem a aplicação de técnicas de aprendizagem de máquina sem a necessidade de programar e tem como objetivos o desenvolvimento de algoritmos de aprendizagem de máquina, torná-los disponíveis e aplicá-los a problemas de ordem prática.

> Foi previsto que o WEKA era que ele não apenas fornecesse uma caixa de ferramentas de algoritmos de aprendizado, mas também um *framework* no qual pesquisadores pudessem implementar novos algoritmos sem que precisassem se preocupar com infraestrutura de suporte para a manipulação de dados e avaliação de esquemas (Hall *et al.*, 2009, p. 10, tradução nossa).

Por ser um ambiente que reúne uma diversidade de técnicas e algoritmos de aprendizagem de máquina e processamento de dados, o WEKA permite que pesquisadores e outros usuários utilizem e comparem múltiplos algoritmos e técnicas para o processamento de diferentes tipos de dados, o que também viabiliza o desenvolvimento de métodos complexos e sofisticados com a associação de diferentes técnicas (Witten *et al.*, 1999).

Entre os métodos presentes no *software* pode-se citar as técnicas de *clustering*, classificação, seleção de atributos, e a mineração de regras de associação, contando também com a possibilidade de exploração de dados com ferramentas de pré-processamento e visualização.

Segundo Hall *et al.* (2009), essa visualização preliminar pode ser combinada com as avaliações estatísticas de esquemas de aprendizagem de máquina e com visualizações de resultados para o desenvolvimento de modelos de mineração de dados.

O núcleo do WEKA conta com a adição dos atributos valorados por relação, definido por Hall *et al*. (2009) como um “saco no contexto de multi-instância”, que é uma técnica que, para a seleção de características, calcula a relevância dos atributos

segundo suas relações com as classes de saída, identificando os atributos de maior importância para a classificação dos dados, introduzindo análises contextuais no processamento de dados e sua classificação.

Por se tratar de um *software* disponibilizado em código aberto, surgem diversos projetos e aplicações que se baseiam no WEKA de alguma forma, entre eles destacam-se as aplicações em sistemas de processamento de linguagem natural, como o GATE, o Balie para identificação da linguagem, seus limites e elementos, o Senseval-2 para desambiguação da linguagem, e o KEA para a extração automática de palavras-chave (Hall *et al.*, 2009).

Embora seja um sistema cheio de técnicas e possibilidades de aplicações, a principal funcionalidade do sistema é a mineração de dados que, por meio das aplicações de aprendizagem de máquina, é capaz de analisar grandes quantidades de dados e identificar padrões e informações relevantes, o que também pode aprimorar a extração de palavras-chave para a indexação de documentos, a sumarização de textos e a classificação de dados textuais.

4.1.7. Finto AI

O sistema Finto AI é um serviço de descrição e indexação automática baseado em algoritmos de inteligência artificial que foi desenvolvido pela Biblioteca Nacional da Finlândia, a partir do *software* Annif, e lançado em 2020.

O Finto AI surge como ferramenta de redação e extração de palavras-chave automático que tem como objetivo dinamizar os processos de descrição de conteúdo de textos de forma automática. Atualmente, é capaz de processar textos em finlandês, inglês e sueco, e suas sugestões de assunto e descrição têm como base vocabulários controlados, classificações e até mesmo ontologias.

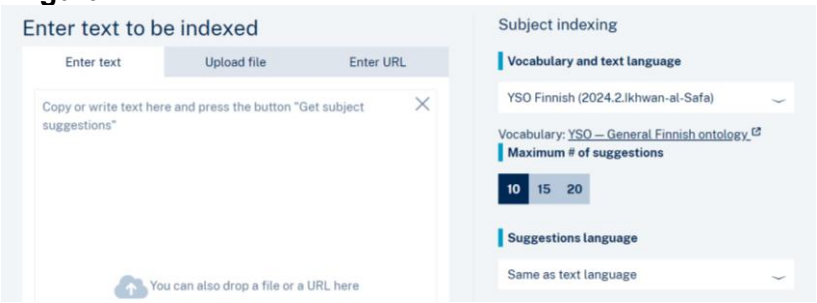
O uso do sistema é realizado a partir do próprio navegador em sua página oficial (https://ai.finto.fi/?locale=en). Os dados, isto é, o texto, pode ser escrito na caixa de texto, também pode ser recuperado por *links* ou ser carregado diretamente no formulário disponibilizado nos formatos TXT, PDF, Doc, DocX, ODT, RTF, PPTX, EPUB e HTML.

O usuário deve escolher o vocabulário controlado que será utilizado para o processamento do texto, o que sustentará diretamente os termos e conceitos sugeridos para a indexação. Por se tratar de uma ferramenta trilíngue, também é possível para o indexador escolher o idioma de saída das palavras-chaves sugeridas,

independentemente do idioma do texto, além de escolher a quantidade de termos desejados, variando de 10 a 20 sugestões.

Após a entrada dos dados, o sistema faz a extração do texto do documento para identificar seu idioma, para então enviá-lo ao Annif, que fará as sugestões de termos de indexação de assunto, que serão selecionados e até mesmo modificados pelo indexador (Suominen; Inkinen; Lehtinen, 2022).

Figura 7 - Interface de entrada de dados do sistema Finto AI



Fonte: Finland (2024b).

O *software*, por se encontrar em uma interface aberta, ser integrado em diferentes sistemas de informação, como é o caso da Universidade de Jyväskylä que utiliza o serviço de interface do Finto AI como suporte para a descrição de conteúdo de teses e no arquivo de publicações Osuva da Universidade de Vaasa.

Tendo em vista a limitação do Finto AI em relação aos idiomas e vocabulários suportados, Suominen, Inkinen e Lehtinen (2022) sugerem que, para o uso de vocabulários próprios e customizados e linguagens ainda não suportadas pelo Finto AI, é recomendado o uso do Annif para o treinamento de modelos próprios de indexação.

Sobre os vocabulários controlados, o Finto AI utiliza os dicionários da ontologia geral finlandesa YSO em finlandês, sueco e inglês, a ontologia KAUNO em finlandês para a descrição de obras de ficção, e o sistema de classificação de bibliotecas públicas YKL também em finlandês, sueco e inglês[22].

Em síntese, o Finto AI é uma ferramenta de indexação automática baseada em algoritmos de inteligência artificial que realiza o processamento e sugere termos de indexação para documentos de acordo com vocabulários controlados, esquemas de classificação e ontologias em finlandês, sueco e inglês. Sua interface *online* e sua capacidade de integrar os sistemas de informação, destacando seu aspecto útil.

[22] Todos os vocabulários, ontologias e classificações suportados pelo Finto AI podem ser encontrados no site da Finto.fi (https://finto.fi/en/).

Também é importante destacar sua limitação ao que se trata de idiomas e vocabulários suportados, porém, a utilização do Annif pode vir a ajudar com esse tipo de limitação.

4.1.8. *Voyant Tools*

O *Voyant Tools* (https://voyant-tools.org/) foi desenvolvido por Stefan Sinclair e Geoffrey Rockwell e lançado em 2003, tendo como antecedentes os *softwares HyperPo*, *Taporware* e TACT. Seu desenvolvimento teve como objetivo aprimorar a leitura por meio de análises de texto leves, como listas de frequência de palavras, gráficos de distribuição de frequência e exibições KWIC para a análise de documentos e *corpus* textuais.

Segundo Alhudithi (2021, p. 43, tradução nossa), “[...] usando algoritmos computacionais, a plataforma extrai informações linguísticas e estatísticas de textos de diferentes tamanhos, tipos e linguagens em questão de segundos”.

A entrada dos dados pode ser realizada por *links* ou pelo carregamento dos arquivos em MS *Word*, Excel, HTML, RTF, PDF ou texto simples. A análise textual pode ser realizada de textos únicos e até mesmo conjuntos de textos que serão processados como coleção.

Sobre os idiomas, o *Voyant Tools* é um sistema multilíngue que suporta textos em português, espanhol, inglês, árabe, francês, japonês, alemão, italiano, sérvio, tcheco, hebraico, bósnio e croata (Alhudithi, 2021).

Alhudithi (2021) destaca algumas das ferramentas do *softwares* que fazem análises qualitativas de contextos linguísticos, como a *WordTree*, para a visualização das frases onde determinada palavras ocorre, *Topics*, que identifica conexões entre palavras e o conteúdo do texto, *Contexts*, que apresenta todo o enunciado no qual uma palavra aparece, e *RezoViz*, que apresenta as conexões entre pessoas, lugares e organizações citadas em um texto.

Além das ferramentas de análise contextual, pode-se ressaltar algumas outras funcionalidades do *Voyant Tools*, entre elas os *Links*, que possibilita a visualização das ligações entre as palavras de grande frequência, *TermsBerry*, que apresenta palavras-chaves junto com palavras que ocorrem próximas a elas, *Veliza*, que traz

respostas às buscas dos usuários, e *Collocates*, que gera listas de conexões entre palavras co-ocorrentes[23].

Quadro 8 - Ferramentas do *Voyant Tools*

FUNCIONALIDADE	FERRAMENTA	DESCRIÇÃO
Co-ocorrência	Arco Textual	Visualização dos termos com um centro ponderado de termos e arcos que seguem os termos na ordem de ocorrência
	Links	Visualização gráfica de conexões entre palavras de alta ocorrência
	TermsBerry	Explora os termos mais frequentes e palavras que ocorrem em sua proximidade
	Colocações	Gráfico de ocorrência das palavras-chaves que ocorrem com certa proximidade
	Correlações	Variação conjunta inversa ou diretamente proporcionais da frequência de termos
Contexto linguístico	Árvore de Palavras	Permite explorar o uso das palavras nas frases e orações
	Frases	Sequência de palavras organizadas por frequência de repetição
	Tópicos	Agrupa termos semelhantes de um *corpus* para a verificação de distribuição nos documento
	Contextos	Palavras-chaves são apresentadas junto ao contexto linguístico onde ocorrem
	Mandala	Relações e conexões entre termos e documentos
	Termos	Palavras que são únicas em um documento
	Micro Pesquisa	Visualização ampla da frequência e distribuição dos termos em um *corpus*
	Paisagem dos Sonhos	Ferramenta experimental de exploração dos locais geográficos presentes nos documentos
	RezoViz	Relações e conexões entre pessoas, locais e organizações citados nos documentos
	Leitor	Forma alternativa de realizar a leitura do texto de um *corpus* sob demanda
Visualização	Documento	Visualização em tabela dos textos que formam um *corpus*

[23] Os detalhes sobre todas as ferramentas disponíveis no *software* estão na página guia do sistema (https://voyant-tools.org/docs/#!/guide/tools).

	Resumo	Visualização simplificada do *corpus*, informações sobre termos e documentos
	Termos do Documento	Visualização em tabela da frequência dos termos de um documento
	Termos do *Corpus*	Visualização em tabela da frequência dos termos de um *corpus*
	Nós (*Knots*)	Visualização dos termos de um documento em formato de linhas torcidas
	Termos Rádio	Visualização descritiva da variação de frequência das palavras de um documento
	Gráfico de Fluxo	Visualização gráfica da variação de frequência dos termos em um documento ou *corpus*
	Tendências	Visualização gráfica da distribuição de ocorrência de palavras em um *corpus* ou documento
	Cirrus	Visualização gráfica das palavras mais frequentes em uma nuvem de palavras
	Linhas de Bolhas	Visualização gráfica da frequência e distribuição dos termos em um *corpus* textual
	Bolhas	Visualização gráfica da frequência de termos no documento
	Gráfico de Dispersão	Visualização gráfica do agrupamento de palavras em um *corpus*
Resultados e análises	Veliza	Ferramenta experimental para diálogos em linguagem natural baseada no *corpus* analisado
	Catálogo	Interface de exploração de conteúdo de um *corpus* e a criação de subconjuntos de acordo com critérios de busca

Fonte: Baseado em Alhudithi (2021) e *Voyant Tools* (2024b).

Além de algoritmos, o *Voyant Tools* utiliza técnicas de processamento da linguagem natural em suas ferramentas, e suas análises contam com técnicas de tokenização para a divisão de unidades linguísticas, a normalização e radicalização de palavras, análises de concordância das palavras em relação a seus contextos, análises de co-ocorrência, análise de tendências, análise de sentimentos e uso de TF-IDF para contagens de frequência e ocorrência de palavras.

Resumidamente, o *software Voyant Tools* utiliza algoritmos e técnicas de processamento da linguagem natural para a análise e visualização de dados textuais a partir de métodos estatísticos e linguísticos. O sistema conta com mais de 20 ferramentas para a realização de análises de frequência e ocorrência, visualizações

contextuais de dados e análises gráficas em mais de 10 idiomas de documentos e conjuntos de textos.

4.1.9. Cortext

Iniciado pela Unidade INRA SenS do Instituto Nacional de Pesquisa para a Agricultura, Alimentação e Meio Ambiente (INRAE) da França, continuado pelo LISIS (*Laboratoire Interdisciplinaire Sciences Innovations Sociétés*), e apoiado pelo Instituto IFRIS (*Institut Francilien Recherche Innovation Société*) e pelo LABEX SITES (Laboratório de Excelência em Ciências, Inovações e Tecnologias para o Meio Ambiente e a Sociedade), o projeto Cortext tem como objetivo promover métodos quantitativos e qualitativos para as ciências sociais e humanas, fornecendo ferramentas, dados, métodos e algoritmos úteis.

A plataforma Cortext, lançada em 2008, nasce como uma proposta de laboratório digital para o processamento e exploração de diferentes tipos de dados textuais através de uma diversidade de ferramentas, técnicas e metodologias. Seus métodos e aplicações são desenvolvidas através de projetos de ciência da computação comunitários sobre as ciências da complexidade, inteligência artificial e processamento automático da linguagem.

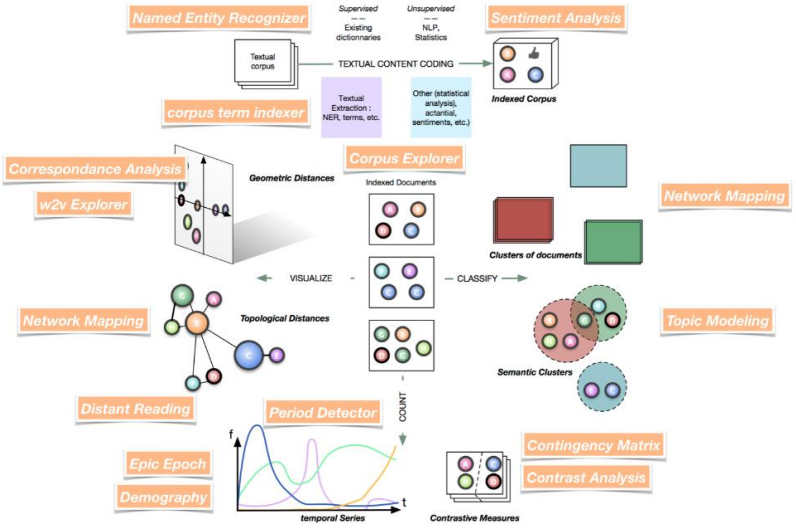
O Cortext *Manager* (https://managerv2.cortext.net/login) é um serviço disponibilizado na *web* que fornece métodos e ferramentas de análise de dados, onde um *corpus* textual pode ser carregado e processado para a análises de discurso, entidades nomeadas e localidades geográficas “[...] com métodos para mapeamento de ciência/controvérsia/questões, leitura distante, agrupamento de documentos, visualizações geoespaciais e de rede” (Cortext, 2024b, online, tradução nossa).

Esse espaço viabiliza que o usuário componha seu próprio método analítico para os dados a partir de uma série de perguntas predeterminadas, possibilitando que cientistas enriqueçam o sistema com suas próprias ferramentas.

As análises numéricas dos dados vão além das abordagens bibliométricas padrões, indicando a posição ocupada pelo indivíduo em determinada comunidade científica, além de caracterizar o coletivo a partir das produções científicas, como artigos, e as citações. As análises também podem ser distributivas, relacionais e geocodificadas, no qual analisam-se dados textuais heterogêneos para a identificação de relações entre conceitos e sujeitos para a caracterização de assuntos e até mesmo áreas científicas.

Entre as funcionalidades do Cortext, pode-se destacar o reconhecimento de entidades nomeadas, a análise de sentimentos, a indexação de termos do *corpus*, explorador de *corpus*, mapeamento de rede, modelagem de tópicos, *clustering* de documentos, agrupamentos semânticos, análise de contraste, leitura distante, codificação de conteúdo textual, análise de correspondência, entre outros.

Figura 8 - Ferramentas e técnicas do Cortext *Manager*



Fonte: Cortext (2024c).

Antes da aplicação das ferramentas, o *corpus* que será analisado passa por um processamento preliminar que permite a identificação, extração e combinação de informações para o processamento preliminar, distinto em processamento de dados, processamento textual, processamento temporal e processamento espacial.

Assim, tem-se a extração lexical, o reconhecimento das entidades nomeadas, como pessoas e organizações, a indexação de bancos de dados a partir de listas personalizadas de termos, a criação de subcorpus, filtragem de listas categóricas e a leitura distante (Cortext, 2024c).

Quadro 9 - Ferramentas do Cortext

FUNCIONALIDADE	FERRAMENTA	DESCRIÇÃO
Pré-processamento	Extração Lexical	Extração automática de termos usando técnicas de PLN
	Reconhecimento de Entidades Nomeadas	Identifica entidades como pessoas, locais e organizações
	Detector de Período	Identifica automaticamente a composição dos dados de períodos estruturais distintos

	Consulta de *Corpus*	Geração de subcorpus para consultas complexas
	List Builder	Filtrar e limpar listas categóricas
	Corpus List Indexer	Permite que o usuário controle itens que podem ser posteriormente analisados ou mapeados
	Leitura Distante	Comparar os perfis das palavras de um *corpus* dinâmico
	Geocodificação de Endereços	Adiciona nomes de cidades, regiões e coordenadas no *corpus*
	Exploração Geoespacial	Analisa a distribuição geográfica de coordenadas
Exploração de Dados	Demografia	Geração de estatísticas descritivas da estrutura de conjuntos de dados
	Word2Vec Explorer	Mapeamento de grandes volumes de palavras com o treinamento de modelos
	Análise de Contraste	Visualização de termos representados no *corpus*
Análise de Dados	Mapeamento de Redes Heterogêneas	Análise de redes que produz representações ajustáveis
	SASHIMI	Combinação de modelos estatísticos e mapas interativos hierárquicos do *corpus* e relacionamentos entre termos e documentos
	Modelagem de Tópicos	Análise da estrutura semântica do *corpus*
	Matriz de Contingência	Visualização de correlações entre diferentes tipos de dados
	Análise de Correspondência	Análise de correspondência entre variáveis

Fonte: Baseado em Cortext (2024c).

Assim, o Cortext disponibiliza técnicas e ferramentas quali e quantitativas para as ciências sociais e humanas para o processamento e análise de dados textuais variados em um laboratório digital - o Cortext *Manager* -, e possibilita que seus usuários desenvolvam seus próprios métodos para a análise dos dados, passando por um pré-processamento e conta com métodos estatísticos, lexicais, computacionais, da inteligência artificial e de processamento da linguagem natural.

4.1.10. IRaMuTeQ

Como forma de realizar análises de dados textuais e identificar padrões e tendências em *corpus* textuais, o IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) foi desenvolvido pelo pesquisador francês Pierre Ratinaud e lançado nos anos 2000 (http://www.iramuteq.org/).

O sistema francês conta com dicionários completos em inglês, português e italiano, além do francês, e dicionários experimentais em alemão, espanhol, grego e sueco, e também possui uma diversidade de ferramentas de análises, desde as mais simples, como a lexicografia, até algumas mais complexas, como as classificações hierárquicas descendentes e análises de correspondências (Camargo; Justo, 2013).

Os textos a serem analisados devem passar por um pré-processamento, onde estarão em um único arquivo *LibreOffice* ou *OpenOffice*, que terá seus elementos segmentados a partir de linhas de comando e será normalizada e uniformizada seguindo os padrões do *software*, mantendo os textos sem recursos de destaque, como negrito e itálico, siglas em minúsculo, palavras compostas hifenizadas, numeral em algarismos, entre outros.

Segundo Camargo e Justo (2013), uma das diferenças do IRaMuTeQ para outros *softwares* de análise textual é o formato dos dados que podem ser analisados, que pode ser um conjunto de textos em um mesmo arquivo - um *corpus* - ou tabelas, nas quais as linhas contém indivíduos e as colunas contém as palavras, preferencialmente gerados pelos *softwares OpenOffice* ou *LibreOffice*.

O IRaMuTeQ conta com análises textuais estatísticas, de especificidades, classificação, de similitude e nuvens de palavras. As análises estatísticas, ou análises lexicográficas, tem funcionalidades como a identificação de unidades textuais, quantificação e frequência de palavras e lematização.

Já as análises de especificidade e análise fatorial e de correspondência (AFC) realiza análises com variáveis de diferentes modalidades e faz a comparação entre produções textuais. A classificação é realizada a partir do método de Reinert, que classifica segmentos textuais de acordo com seus vocabulários e dividido seguindo a frequência das palavras lematizadas.

A análise de similitude é realizada com as relações dos objetos de um conjunto, identificando correspondências entre as palavras. Por fim, a nuvem de palavras

apresenta uma visualização gráfica das palavras agrupadas de acordo com sua frequência (Camargo; Justo, 2021).

O sistema IRaMuTeQ é um *software* aberto de análises estatísticas e textuais utilizada principalmente na ciências humanas e sociais que possibilita diversos tipos de análises de textos, como as análises de correspondência, e a visualização de dados textuais, como a nuvem de palavras e se destaca de outros *softwares* especialmente por processar e analisar dados dispostos em tabelas, além de documentos individuais e *corpus* textuais.

4.1.11. AntConc

Desenvolvido pelo professor inglês Dr. Laurence Anthony e lançado em 2002, o *software* AntConc (https://www.laurenceanthony.net/software/antconc/) realiza análises de grandes quantidades de textos e dados textuais, auxiliando operações como a análise de discurso em pesquisas qualitativas em ciências humanas e sociais.

O AntConc realiza, fundamentalmente, análises de *corpus*, que gera comparações entre diferentes textos em grande escala, também chamado de leitura distante. A análise de *corpus* permite a identificação de padrões gramaticais em diferentes textos de uma mesma coleção ou frases recorrentes em um *corpus*, por exemplo. Esse método analítico é útil principalmente para “[...] testar intuições sobre textos e/ou triangulação de resultados de outros métodos digitais” (Froehlich, 2015, tradução nossa).

As análises são realizadas unicamente em documentos em TXT, sendo necessário que textos em DOC, DOCX e PDF sejam convertidos. Além do mais, recomenda-se que os textos do *corpus* sejam lematizados antes das análises, o que, junto com as listas de lemas disponibilizadas no site do *software*, viabiliza a geração de listas de palavras e lemas do *corpus* analisado.

Sobre suas funcionalidades, temos a ferramenta concordância, que é uma visualização de determinadas palavras-chaves em seu contexto de uso (KWIC), a visualização de arquivos, que apresenta o arquivo completo para contextualização dos resultados, o gráfico de concordância, para uma visualização alternativa dos resultados do KWIC, os *clusters*, que apresenta palavras que co-ocorrem com certa frequência, os *collocates*, que apresenta palavras que, estatisticamente, ocorrem juntas, a lista de palavras, com todas as palavras presentes no *corpus* analisado, e a

lista de palavras-chaves, com uma comparação entre as palavras presentes em dois textos diferentes.

A ferramenta de concordância, segundo Anthony (2005), facilita a aquisição de vocabulário, compreensão gramatical e identificação de estilos de escrita, o que a caracteriza como um recurso que pode ser aplicado no estudo de línguas estrangeiras, permitindo a visualização do uso de uma mesma palavra ou expressão em diversos textos e contextos.

Já a plotagem de termos de busca utilizada na concordância demonstra a utilização dos termos de busca em um determinado *corpus*, possibilitando que usuários visualizem em que parte do discurso um termo aparece. Embora seja parecida com a concordância, a plotagem de termos de busca oferece uma visualização, que apresenta a distribuição dos termos ao decorrer dos arquivos.

Os *clusters* fazem o agrupamento de palavras a partir de um termo de busca e as classificam em ordem alfabética ou por frequência. A busca pode ser realizada através de palavras, expressões, frases, n-gramas e *substrings* de busca.

Tanto a lista de palavras quanto a lista de palavras-chaves são utilizadas para destacar possíveis áreas de interesse, identificar palavras e lemas relacionados, classificar palavras alfabeticamente ou por sua frequência no *corpus*. Porém, Anthony (2005) ressalta que, mesmo sendo ferramentas úteis, as listas de palavras não são adequadas para compreender a importância de determinadas palavras no *corpus*.

Froehlich (2015) ressalta a importância da construção do *corpus*, e como os diferentes *corpus* podem gerar diferentes resultados, principalmente nas considerações sobre a amostragem e a representatividade dos textos que constituem os conjuntos analisados. Assim, a baixa representatividade do *corpus* destaca-se como uma das possíveis limitações do *software*.

Ademais, Anthony (2005) enfatiza algumas limitações do AntConc, primeiramente, por analisar diretamente o texto bruto, isto é, não construir índices para as análises posteriores, o *software* torna-se limitado a *corpus* relativamente pequenos e especializados. Além disso, o *software* também não oferece estatísticas em detalhes sobre o *corpus* e as buscas realizadas, mas permite que os resultados sejam copiados e transferidos para planilhas. Por fim, Anthony considera que a maior limitação do AntConc seja no processamento de dados codificados, como os formatos HTML/XML.

Resumidamente, o *software* AntConc realiza o processamento e análises de grandes quantidades de dados textuais, principalmente no que diz respeito às análises

de *corpus*, a partir de ferramentas como concordâncias, nuvem de palavras, plotagem de termos de busca, entre outras. Os resultados das análises dependem diretamente do *corpus* construído, que deve ser o mais representativo possível, dentro das limitações do sistema, que trabalha com dados brutos de *corpus* menores e mais especializados. Os resultados do processamento e análises auxiliam em pesquisas qualitativas, como a análise do discurso, principalmente nas ciências humanas e sociais.

4.1.12. *Orange Data Mining*

O sistema *Orange Data Mining* (https://orangedatamining.com/) foi desenvolvido pela Universidade de Ljubljana e preliminarmente lançado com uma versão estabilizada em 2023, oferecendo ferramentas de mineração e visualização de dados a partir de algoritmos de aprendizagem de máquina e análises com programação visual para a exploração de dados qualitativos.

Demšar e Zupan (2013) explicam que a parceria com pesquisadores de outras áreas, como a biomedicina, impactou na decisão de criar um sistema de exploração de dados que possibilitasse a criação de fluxos de análise próprios sem precisar saber programar em *Python*.

O *Orange Data Mining* é capaz de processar qualquer arquivo de dados que sejam delimitados por vírgula ou tabulação, e arquivos do Excel e Google Sheets, assim como documentos em um formato de dados próprio do sistema, que normalmente são tabelas com amostras de dados nas linhas, atributos dos dados em colunas, sejam eles numéricos, categóricos, de data e hora e texto, e tem suas funções, sendo recursos de entrada, meta atributos e classes.

O grande diferencial do *Orange*, que atrai grandes quantidades de usuários, é sua interface gráfica, na qual os usuários são capazes de adicionar *widgets*, cada um dotado de uma ferramenta/método/técnica de processamento, conectando-os por canais de comunicação e compondo seu fluxo de análise. Por isso, Demšar e Zupan (2013) afirmam, destacou-se o desenvolvimento dos *widgets* para a visualização e interação com os dados.

Segundo Demšar *et al.* (2013), o *Orange* é como uma “caixa de ferramentas” organizada hierarquicamente por elementos de mineração de dados. Essa hierarquia é composta principalmente pelo gerenciamento e pré-processamento de dados (entrada e saída de dados, filtragem, amostragem, manipulações de características e

seleção), classificação (algoritmos supervisionados de aprendizagem e ligação com bibliotecas externas), regressão (linear, lasso, árvores e floresta de regressão e *splines* de regressão multivariada), associação (regras de associação e mineração de conjuntos), conjuntos (*wrappers* para *bagging*, *boosting*, árvores de florestas e *stacking*), agrupamento (*k-means* e agrupamento hierárquico), avaliação (validação cruzada, qualidade dos métodos de previsão e estimativa de confiabilidade), e projeções (análise de componentes, mapas auto-organizáveis e escalonamento multidimensional).

Para a análise dos dados, é possível realizar ações de domínios; para "transformar" conta com o amostrador de dados, tabela dinâmica e pré-processamento; para visualizar gráficos de caixa, dispersão, barras e linhas, diagrama de Venn e de peneira, e projeções; também pode-se citar os modelos de árvore, floresta aleatória, regressão linear e logística e redes neurais; para avaliar tem-se previsões, curvas de desempenho e gráficos de permutação e calibração; entre os não supervisionados estão as correlações, aprendizagem múltipla, análises de correspondência e digitalização de banco de dados; em espectroscopia os espectros de pré-processamento, série espectral e hiper espectros; para mineração de texto a criação de *corpus*, ligações com *PubMed*, *NY Times* e Wikipédia, saco de palavras, análises de sentimento, concordância e extração de palavras chaves; na análise de sobrevivência a regressão de Cox, coortes e características de sobrevivência de classificação; análise de imagens; gerador, explorador, análise e clusterização de redes, entre outros recursos[24].

O *software* pode ser enriquecido por meio de extensões de *widgets* para novos tipos de análises ou para a análise de dados especializados, entre eles métodos de bioinformática, exploração de redes e conjuntos para a mineração de dados.

Assim, conclui-se que o *Orange Data Mining oferece*, por meio de uma interface gráfica interativa, uma gama de ferramentas e técnicas, além de possibilitar a criação de novos fluxos de análise, sem a necessidade de programar, através dos *widgets*, que podem ser aprimorados por meio de expansões para a análise de dados especializados.

[24] Para mais funcionalidades, ferramentas e recursos, indica-se a consulta do catálogo de *widgets* do *Orange Data Mining* (https://orangedatamining.com/widget-catalog/).

4.1.13. Yoshikoder

Desenvolvido por Will Lowe como parte do Projeto Identidade do Centro de Relações Internacionais de Harvard e lançado em 1989, o Yoshikoder (https://www.yoshikoder.org/) é um *software* de análises de conteúdo multilíngue e multiplataforma que processa a linguagem natural e possibilita, além de análises textuais, a criação e utilização de dicionários em diversos idiomas.

De modo geral, o sistema permite a criação de dicionários de análise de conteúdo em qualquer língua, oferece resumos de documentos e tabelas de frequência de palavras, permite a criação e a visualização de palavras-chaves dos documentos, além de possibilitar a aplicação dos dicionários junto à ferramenta de concordância, colaborando para a compreensão de palavras e seus diferentes contextos de uso no documento.

Os desenvolvedores do sistema mantém dicionários[25] especializados e multilíngues disponíveis já traduzidos para o formato aceito pelo Yoshikoder, como o dicionário *Linguistic Inquiry* and *Word Count,* e os dicionários de imagens regressivas em diversos idiomas. Além disso, o *software* também possui um conversor[26] de arquivos HTML, *Word*, PDF, *Powerpoint*, Excel para o formato texto simples, legível pelo sistema.

Lowe (2006) afirma que o Yoshikoder realiza análises de conteúdo clássicas, isto é, realiza ações como a análise de frequência de palavras, análise de concordância e a criação dos dicionários do conteúdo analisado, antes realizadas manualmente com a contagem e classificação de palavras, mas que foram facilmente automatizadas.

A análise de frequência de palavras gera uma lista da quantidade de ocorrências das palavras proporcionalmente ao texto, sendo útil para uma análise preliminar e exploratória dos documentos. Já os dicionários são construídos como um conjunto de categorias que presumivelmente seguem algum tipo de padrão, que podem ser modificados diretamente pelos usuários e também podem ser aplicados nas análises de palavras-chaves em contexto (Lowe, 2006).

Além disso, Lowe (2006) ainda discute a questão do processamento de idiomas realizado pelo sistema. Como já apontado, o Yoshikoder realiza análises multilíngues,

---

[25] Para conhecer todos os dicionários disponíveis, recomenda-se que visite o site do projeto Yoshikoder (https://www.yoshikoder.org/).
[26] Disponível em https://www.yoshikoder.org/ykconverter/.

porém, o autor ressalta a importância de especificar a localidade da língua do documento, como no português falado no Brasil, para garantir a segmentação apropriada das unidades linguísticas durante o processamento do texto, além de contar com o desenvolvimento de "*plugins* de tokenização" para esse processamento, como é o caso do *plugin* tokenização para chinês.

Concluindo, o Yoshikoder realiza análises de conteúdo clássicas a partir do processamento multilíngue da linguagem natural, viabilizando a criação e modificação de dicionários categóricos próprios para as análises de conteúdo, além de possuir dicionários especializados disponíveis para o uso, produzindo listas de frequência e dicionários a partir de dados textuais qualitativos.

4.2. Bases pragmático-linguísticas dos *softwares*

Na presente subseção, apresentam-se visual e descritivamente os resultados consequentes da análise dos *softwares* estudados através de gráficos, porcentagens e descrições de dados, baseando-se no referencial teórico e nas descrições e funções dos *softwares* anteriormente apresentados.

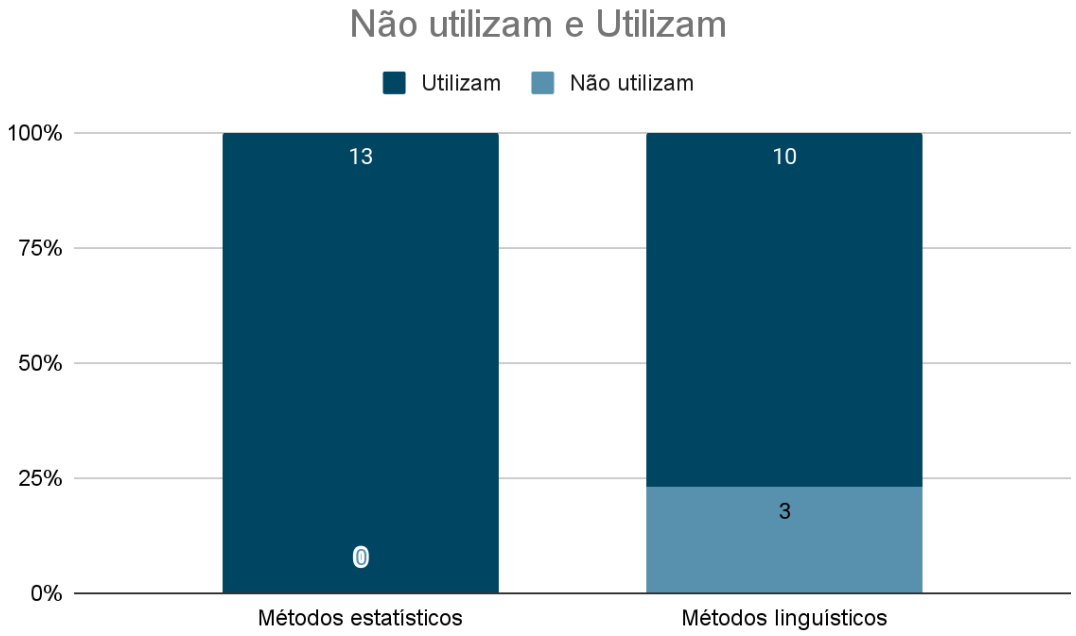
Por conseguinte, apresentam-se as discussões como produto das seções de referencial teórico, as apresentações e descrições dos *softwares* e dos resultados expostos quantitativamente e visualmente por meio de gráficos. Por fim, encontra-se a seção de considerações gerais.

A análise individual de cada *software*, anteriormente apresentados, possibilitou uma visão mais detalhada de cada um dos seus sistemas, permitindo a identificação de suas principais funções, aplicações reais e virtuais, métodos e análises realizadas, o que objetivou a identificação, ainda que inicial, das bases teórico-metodológicas que os constituem.

Um dos pontos de maior interesse para esta análise é o método utilizado para as análises textuais, que podem ser, embora não exclusivamente, métodos estatísticos e métodos linguísticos, lexicais em sua maioria.

Entre os 13 *softwares* analisados, todos utilizam abordagens estatísticas em suas análises, seja de forma exclusiva ou em associação com outros métodos, e 10 *softwares* (76,9%) utilizam abordagens linguísticas, exclusivamente análises léxico-morfológicas, associadas aos métodos estatísticos (figura 9).

Gráfico 9 - Métodos de análise textual aplicados pelos *softwares*



Fonte: Elaborado pela autora (2025).

Segundo Gasperin (2001), os modelos estatísticos para o processamento da linguagem surgem como alternativa para o tratamento de grandes volumes de dados textuais no formato digital, utilizando métodos quantitativos para o processamento automatizado de textos, determinando probabilidades de ocorrência, sequências e categorias de palavras.

A abordagem linguística, comumente aplicada em associação com métodos estatísticos, realiza análises textuais a partir dos níveis linguísticos - morfológico, sintático, semântico e pragmático -, sendo mais utilizadas as análises léxico-morfológicas, que decompõem textos em elementos lexicais que são decompostos em morfemas para a examinação de sua estrutura e características (Vallez; Pedraza-Jimenez, 2007; Vieira; Lima, 2001).

A análise individual dos *softwares* também apontou que todos os *softwares* são capazes de realizar análises de textos em diferentes idiomas, como o caso do MTI que é capaz de adaptar sua aplicação a todos os vocabulários compatíveis com o MeSH, o FintoAI que realiza análises em finlandês, inglês e sueco, e o *Orange Data Mining* capaz de processar 50 idiomas.

Tal característica mostra-se relevante para determinar as possibilidades de aplicação dos sistemas no tratamento de dados textuais em diferentes formatos,

assuntos, áreas do conhecimento, e idiomas, demonstrando sua flexibilidade de utilização.

Outro fator de importância que foi identificado é a utilização de linguagens documentárias para o processamento de textos, principalmente em operações como a extração de palavras-chaves para a indexação de documentos, e a classificação de textos por assunto.

Nesse sentido, os vocabulários controlados são instrumentos que auxiliam as atividades de organização e recuperação da informação, controlando e padronizando termos e descritores utilizados na representação de assunto, especialmente na indexação, de documentos (Davanzo; Moreira, 2014).

Assim, sete (53,8%) dos *softwares* examinados fazem o processamento de textos e documentos tendo como base vocabulários controlados, como é o caso do MeSH, utilizado pelo *Medical Text indexer*, e o Annif, que utiliza desde dicionários até tesauros e ontologias em suas análises.

Outro fator que impacta diretamente na qualidade das análises textuais é a utilização de algoritmos de radicalização (*stemmer*) e lematização na padronização das entidades linguísticas dos textos.

Os radicalizadores, segundo Xavier, Silva e Gomes (2013), possibilita a identificação de palavras relacionadas, de forma morfológica ou semântica, a partir de remoção de sufixos, como é o caso das palavras “FLORISTA” e “FLORAL” que, embora sejam palavras com conotações distintas, se relacionam mutuamente através do radical “FLOR”.

Já a lematização é responsável pela padronização da representação de palavras por meio de suas formas canônicas, transformando os verbos para seu modo infinitivo - transformação de “tenho”, “tiveram” e “teremos” para a forma “ter” -, e os substantivos e adjetivos para suas formas no masculino singular, como apresentado em Lucca e Nunes (2002), tendo em vista a simplificação do processo de agrupamento lexical.

Ambos os processos de radicalização e lematização têm em vista a normalização das diferentes aspectos gramaticais para suas formas invariáveis, evitando variações linguísticas de termos e conceitos com mesmo significado.

Assim, foi possível observar que oito dos *softwares* (61,5%) não utiliza nenhum dos dois métodos de normalização da linguagem, o que pode levar a resultados afetados por fenômenos como a variação e ambiguidade linguística.

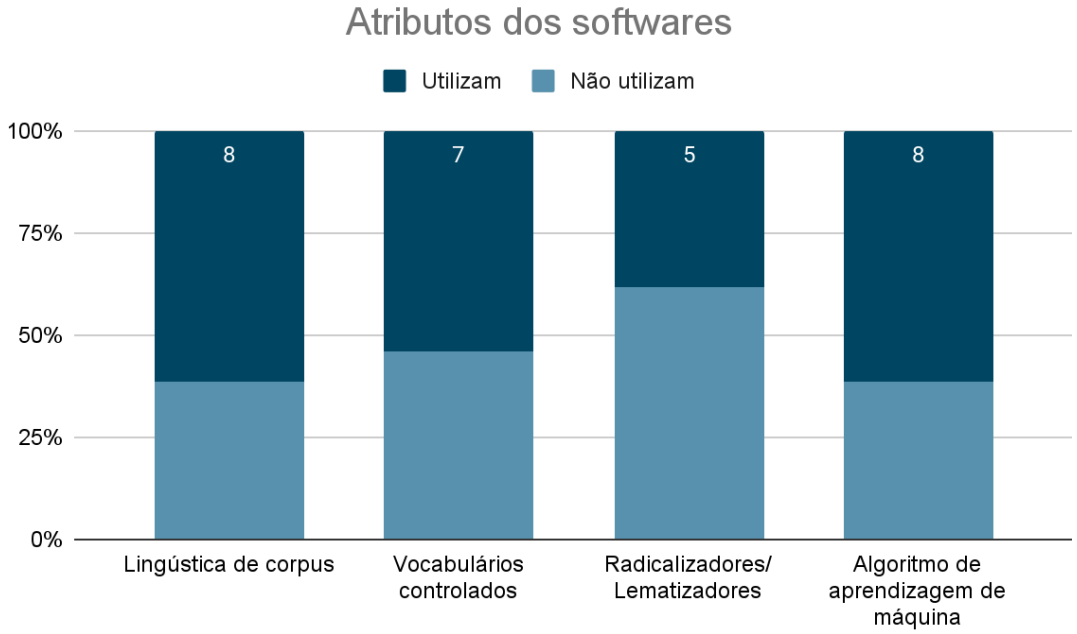
Outro tipo de algoritmo que pode aprimorar o desempenho do processamento textual é a aprendizagem de máquina (*machine learning*), subcampo da inteligência artificial que procura capacitar os computadores para lidar e aprender com os dados sem programação prévia e de forma mais eficiente, embora aprender, nesse sentido, esteja associado a uma nova maneira de "[...] encontrar regularidades estatísticas e outros padrões nos dados" (Ayodele, 2010, p. 19; Manesh, 2020).

A aplicação de algoritmos de aprendizagem de máquina nos sistemas de processamento de dados possibilita que o *software* seja treinado para a análise de conteúdos textuais a partir de um conjunto inicial de dados, também conhecido como *corpus* de treinamento. Entre os 13 *softwares* estudados, oito (61,5%) são constituídos por algoritmos de aprendizagem de máquina, o que pode indicar sistemas mais "modeláveis" aos tipos de dados e textos analisados.

Por fim, considerando que o treinamento de algoritmos de aprendizagem de máquina e a necessidade de análises linguísticas mais contextualizadas, entende-se a importância da linguística de *corpus* que, por meio de um conjunto representativo de textos criteriosamente selecionados, é capaz de proporcionar uma compreensão variada de uma língua, suas palavras e expressões, proporcionando uma "[...] metodologia empírica de trabalho, baseada no emprego de dados reais de amostras do uso da língua" (Llamazares, 2008, p. 329; Sardinha, 2000).

Ainda com as possibilidades de aprimoramento das análises textuais com o uso da linguística de *corpus*, seja para o treinamento de algoritmos ou para a contextualização linguística dos dados, oito dos *softwares* avaliados (61,5%) fazem seu uso, seja como forma de aplicar a aprendizagem de máquinas ou para o processamento dos dados dos documentos.

Gráfico 10 - Presença de funções/bases nos *softwares*



Fonte: Elaborado pela autora (2025).

Para melhor visualização dos atributos e resultados das análises de cada um dos *softwares*, suas funções e recursos, desenvolveu-se o quadro 10, para a sumarização desta subseção.

Quadro 10 - Síntese dos resultados

	Análise estatística	**Análise linguística**	**Análise multilíngue**	**Linguística de *corpus***	**Vocabulário controlado**	**Normalização linguística**	**Aprendizagem de máquina**
Yake!	X		X				
MAUI	X	X	X	X	X	X	X
Annif	X	X	X	X	X		X
Medical Text Indexer	X	X	X		X		X
KEA	X	X	X	X	X	X	X
WEKA	X		X				X
Finto AI	X		X		X		X

Voyant Tools	X	X	X	X		X	
Cortext	X	X	X	X			X
IRaMuTeQ	X	X	X	X	X		
AntConc	X	X	X	X		X	
Orange Data Mining	X	X	X				X
Yoshikoder	X	X	X	X	X	X	

Fonte: Elaborado pela autora (2025).

De maneira geral, a partir da apresentação dos recursos de cada um dos sistemas e suas características, pode-se evidenciar a proeminência dos *softwares* MAUI, KEA e Annif em relação aos atributos analisados.

No âmbito do tratamento e representação da informação, propriedades como a possibilidade de uso de vocabulários controlados, análises em *corpus* e a aplicação de métodos linguísticos apresentam-se como possibilidades para o processamento mais eficiente de materiais em linguagem natural, considerando questões como a variação e ambiguidade linguística.

Como discutido anteriormente, os fenômenos de variação e ambiguidade linguística são responsáveis pela ocorrência dos chamados silêncio e ruído documental, respectivamente. Com isso em vista, recursos como o uso de vocabulários controlados e métodos de radicalização e lematização linguística se tornam possibilidades para a atenuação destas manifestações características da linguagem natural.

Assim, o uso de métodos de radicalização e lematização de variedades linguísticas possibilita que a análise dos elementos linguísticos de um texto ou de um conjunto de informações não se restrinja à conjugação apenas de palavras e unidades idênticas, mas de toda uma categoria de palavras semântica, e até mesmo pragmaticamente, reduzindo o silêncio documental resultante das variações linguísticas.

Nesta categoria, podemos citar os *softwares* MAUI, KEA, *Voyant Tools*, AntConc e Yoshikoder, que utilizam a radicalização e/ou lematização das formas

linguísticas dos dados linguísticos-textuais em uma etapa de pré-processamento textual.

Com relação à ambiguidade linguística e as possibilidades de ruído documental, tem-se o uso dos vocabulários controlados. Buscando limitar os significados das palavras dentro de um domínio, as linguagens documentárias caracterizam-se como instrumentos de ordem pragmática no tratamento da informação, auxiliando na delimitação do significado de palavras e expressões em uma área ou contexto especificado.

Desta forma, a possibilidade de significados semânticos é limitada de acordo com o contexto da própria informação ou da área de conhecimento para que se processa, permitindo a padronização de termos de indexação e palavras-chaves atribuídas no processamento da informação.

Aqui podemos dar destaque aos *softwares* MAUI, Annif, *Medical Text Indexer*, KEA, FintoAI, IRaMuTeQ e Yoshikoder, que permitem - e em alguns casos necessitam - de vocabulários controlados, como tesauros e ontologias, para a análise textual dos dados de entrada no sistema.

Durante a análise dos sistemas, tornou-se claro a utilização da linguística de *corpus* como método nas análises textuais das informações. Com esse fato em vista, entende-se a linguística de *corpus* como parte da linguística aplicada que tem como objeto a língua por meio de seus usos em grande escala, baseada também no tratamento estatístico de dados e em processos informatizados (Sardinha, 2004).

Assim, compreende-se a utilização da linguística de *corpus* pelos *softwares*, sendo eles MAUI, Annif, KEA, *Voyant Tools*, Cortext, IRaMuTeQ e AntConc, estabelece uma relação direta com a pragmática, isto é, por meio de um contexto determinado de usos de uma língua - um *corpus* - o sistema é capaz de identificar, selecionar e extrair partes significativas de documentos com maior fundamentação e assertividade (Mokva; Comassetto; Fontana, 2004).

Porém, esse tipo de processamento ainda é essencialmente diferente da ideia de um processamento pragmático por duas razões: 1) um *corpus*, para contribuir para a compreensão de uma língua, depende da seleção de um conjunto de textos representativos à área e língua em questão, que pode ser uma tarefa altamente subjetiva e dependente de um controle de qualidade rigoroso; 2) o contexto produzido pelo *corpus*, diferente da ideia de contexto da pragmática, que compreende os contextos social, econômico, cultural e de domínio do discurso, o conjunto de textos

que forma o *corpus*, se bem selecionado, cria um panorama linguístico ao texto ou palavra analisado, ainda limitado em comparação à pragmática (Bird; Klein; Loper, 2009).

Dessa forma, embora ainda insuficiente para a realização de processamentos linguísticos essencialmente pragmáticos, a aplicação da linguística de *corpus* surge como um tipo de aproximação entre os programas de análise textual, normalmente limitados aos métodos de processamento estatístico, e a esfera linguística, ainda que não fundamentalmente pragmática, para a maior dinamização e compreensão da linguagem e seu processo de significação.

Também é importante destacar o modelo de análise estatísticas de dados textuais, seus aspectos positivos, e suas limitações para o processamento da linguagem. A partir da década de 1990, máquinas e computadores passaram a ter cada vez mais capacidades de processar dados, concebendo o chamado paradigma estatístico que, a partir de conjuntos de textos, foram programados sistemas capazes de representar línguas a partir de modelos probabilísticos calculados a partir de frequências e ocorrências, possibilitando a classificação, tradução e geração de textos (Caseli e Nunes, 2024).

Porém, assim como qualquer uma das abordagens, possui suas limitações, o que abre espaço para os chamados paradigmas híbridos, que fazem a junção de dois métodos distintos, normalmente com o estatístico, buscando aprimorar suas aplicações.

E além das limitações, Caseli e Nunes (2024) afirmam, e aqui busca-se reforçar, que os modelos de processamento da linguagem, em sua generalidade, fazem o processamento de padrões de caracteres e não de unidades de natureza linguística, isto é, os modelos não aprendem uma língua para analisá-la, mas a reproduzem, baseados em regras probabilísticas explícitas e implícitas, e criam generalizações que são possíveis a partir do *corpus* de treinamento.

Por fim, salienta-se a insuficiência dos modelos de processamento de dados textuais ao lidarem com textos e documentos em linguagem natural, como afirmam Caseli e Nunes:

> Como vimos, nos diversos anos de pesquisa e desenvolvimento em PLN, as estratégias e abordagens vão e vêm, mas a linguagem natural é muito mais complexa de se aprender e compreender do que uma simples contagem de frequências e coocorrências. Assim, [...] o conhecimento linguístico e de uso da língua ainda não foi

> completamente representado/capturado por nenhum dos métodos atuais (2024, p. 15).

Nesse ponto, retomam-se as questões que, inicialmente, direcionaram os objetivos da pesquisa: Quais os atributos da pragmática estão presentes nos *softwares* de representação e recuperação da informação? Quais *softwares* oferecem um processamento pragmático da linguagem, atendendo as demandas dos processos de representação da informação?

Tendo em vista as análises empreendidas nos 13 *softwares* da amostra, conclui-se que a abordagem mais próxima, embora que ainda não sendo parte, da pragmática foi a aplicação da linguística de *corpus* utilizada como forma de delinear um contexto, embora apenas linguístico, para os significados dos termos e expressões analisadas do discurso.

Além disso, também se destaca os sistemas que fazem uso de vocabulários controlados no processamento da linguagem de textos e documentos, aplicação essa que delimita o contexto de análise dos dados, o que colabora para a determinações dos significados contextuais de unidades e expressões linguísticas, o que se aproxima de um campo mais pragmático.

Por fim, por meio dos resultados apresentados e das discussões realizadas, finalmente é possível afirmar que, entre os *softwares* que foram analisados, nenhum deles demonstrou ser capaz de realizar processamentos essencialmente pragmáticos de dados textuais, porém, entre todos os recursos e aspectos abordados no estudo, os sistemas MAUI, KEA e Annif apresentam-se como opções viáveis para a aplicação nos processos e atividades da organização da informação, como a indexação, sumarização, classificação e extração de palavras-chave de documentos a partir de abordagens supervisionadas por profissionais, seja na atuação profissional de pessoas bibliotecárias, para o ensino em escolas de biblioteconomia, e até mesmo em cursos e treinamentos para formação continuada dos profissionais que já se encontram no mercado de trabalho.

5. CONSIDERAÇÕES FINAIS

Inicialmente, entende-se que o objetivo geral de identificar os *softwares* e sistemas que se concentram no processamento técnico da linguagem natural e suas bases pragmáticas, que guiou o desenvolvimento da pesquisa foi atingido através do cumprimento dos objetivos específicos, apresentado a seguir:

a) Revisão da literatura produzida sobre a pragmática linguística

Através do levantamento bibliográfico foi possível desenvolver o referencial teórico relativos à organização da informação, sua distinção da organização do conhecimento e definir o campo do tratamento temático da informação, também buscou-se definições do conceito de pragmática, tanto no campo linguístico quanto no semiótico, e apresentaram-se alguns dos principais teóricos e abordagens pragmáticas, a semântica de Gottlob Frege, a tripartição semiótica de Charles Morris, e a teoria dos atos de fala de Austin e Searle.

Também foi possível desenvolver uma sistematização inicial das contribuições pragmáticas ao campo da ciência da informação já indicadas na literatura. Por fim, apresentou-se a conceituação e caracterização do processamento da linguagem natural e suas aplicações.

b) Identificação das aplicações de *softwares* de processamento da linguagem para o tratamento da informação

Para o levantamento dos *softwares* e sistemas para análise, contou-se com o uso da rede social acadêmica *Research Gate*, além do levantamento por trabalhos e revisões preexistentes na literatura da ciência da informação e ciência da computação, totalizando 38 *softwares* que, com a aplicação de critérios de solução (apresentados na seção de metodologia), constituiu uma amostra de 13 sistemas para a análise.

c) Identificar quais dessas aplicações se fundamentam nos estudos realizados no campo pragmático

Para este objetivo, foi necessário a realização de análises individuais de cada um dos *softwares*, apontando suas principais funções, características, recursos e possibilidades de aplicação, processo que teve como produto uma ficha avaliativa de cada um dos sistemas com as informações essenciais de cada um (nome, desenvolvedores, lançamento, linguagem de programação, *softwares* antecedentes, disponibilidade, objetivo, repositórios, extensões, áreas de aplicação, funcionalidades, link e notas).

Assim desenvolveu-se a seção de resultados, com subseções específicas para a descrição de cada um dos *softwares*, o que foi sintetizado em um quadro com os principais recursos analisados em cada um dos sistemas, sendo eles os tipos de análises realizadas (estatística e/ou linguística), idiomas suportados, uso de vocabulários controlados, pré-processamento textual (radicalização e lematização), algoritmos, entre outros.

d) Relacionar os fundamentos pragmáticos dos *softwares* para apoiar com os estudos da semiótica informacional

Por fim, foi possível afirmar a presença - neste caso, ausência - dos conceitos e ideias dos estudos pragmáticos nos *softwares* analisados, mas também permitiu a identificação de outras abordagens de linguagem, como a linguística de *corpus*.

As etapas atingidas por meio dos objetivos específicos levam às respostas dos problemas indicados inicialmente: **1)** Quais os atributos da pragmática estão presentes nos *softwares* de representação e recuperação da informação?

Tendo em vista os *softwares* estudados e os resultados das análises realizadas em cada um deles, pode-se afirmar que nenhum dos 13 sistemas se fundamenta nos conceitos e estudos empreendidos no campo pragmático, resultado esse que enfatiza a subvalorização das contribuições da semióticas e linguísticas na concepção de sistemas de análise textual e de processamento linguísticos, que ainda demonstram preferência para os métodos estatísticos para o processamento de dados textuais.

Embora possa se afirmar a aplicação de abordagens linguísticas em uma parte dos *softwares*, também está claro sua limitação no que diz respeito a amplitude em relação ao campo linguístico, restringindo-se às análises léxico-morfológicas, pouco se estendendo à dimensão da significação nos domínios semântico e pragmático.

A presença de princípios da linguística de *corpus*, embora ainda não chegue ao campo pragmático, é uma possibilidade de considerações da significação para além do significado semântico, alargando as potencialidades linguísticas-semióticas no campo do processamento da linguagem.

2) Quais *softwares* oferecem um processamento pragmático da linguagem, atendendo as demandas dos processos de representação da informação?

Como discutido na seção de resultados, embora não tenham sido reconhecidos *softwares* com funcionalidades pragmáticas, de fato, pode-se destacar os sistemas MAUI, KEA e Annif como os mais promissores em relação ao processamento e tratamento de informações.

Assim, determina-se a importância da descoberta e aplicação de novas metodologias e instrumentos na organização da informação, seja para o ensino de alunos de biblioteconomia, para o uso de profissionais durante suas atividades de tratamento técnico de documentos, e até mesmo para a formação continuada de profissionais da informação, como alternativa para uma abordagem semi-automatizada das atividades de representação e organização, dinamizando parte do ciclo informacional.

Embora o campo pragmático não esteja conceitualmente sistematizado e possa ser visto como uma miscelânea de estudos, teorias e conceitos, cada uma de suas facetas tem algo a oferecer à ciência da informação, tendo como elo mais forte desta relação os estudos de organização da informação, seja o princípio de contexto e a ideia de referente na lógica de Frege, a fragmentação do signo semiótico de Morris, e os atos ilocucionários de Austin e Searle.

Por fim, compreende-se que a pragmática, mesmo com as contribuições já reconhecidas na literatura da ciência da informação e com as perspectivas e possibilidades em vista, ainda não foi explorada o suficiente nos estudos informacionais. No processamento da linguagem natural, ainda não foi possível reconhecer abordagens e recursos que sejam associados aos fundamentos pragmáticos, mesmo que a relação entre as áreas da informação, tecnologia e linguagem tenha muito a oferecer conceitual, teórico e metodologicamente.

REFERÊNCIAS

ABBAGNANO, N. **Dicionário de filosofia**. 2 ed. São Paulo: Editora Mestre Jou, 1962.

ALAMILLO, A. R. Los verbos realizativos, la situación comunicativa y los actos declarativos de Searle. **Pragmalingüística**, n. 15-16, p. 179-199, 2007.

ALBUQUERQUE, A. C. Em foco a classificação: abordagens conceituais na Arquivologia, Biblioteconomia e Museologia. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da informação, Florianópolis, v. 20, n. 43, p. 20-46, 2015. Disponível em: https://brapci.inf.br/#/v/37094. Acesso em: 18 nov. 2025.

ALHUDITHI, E. Review of Voyant Tools: See through your text. **Language Learning & Technology**, v. 25, n. 3, 2021.

ANTHONY, L. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. *In*: IEEE INTERNATIONAL PROFESSIONAL COMMUNICATION CONFERENCE, 2005, Limerick. **Proceedings of the International Professional Communication Conference** [...]. Los Alamitos: IEEE Computer Society, 2005. p. 729-737,

ANTHONY, L. **AntConc**. Laurence Anthony's Site, 2024. Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 16 jul. 2024.

ARAÚJO, C. A. Á. Fundamentos teóricos da classificação. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 11, n. 22, p. 117-140, dez. 2006.

ARMENGAUD, F. **A pragmática**. São Paulo : Parábola Editorial, 2006.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: Métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6028**: Informação e documentação — Resumo, resenha e recensão — Apresentação. Rio de Janeiro, 2021.

AUSTIN, J. **¿Cómo hacer cosas con palabras?**. Madrid: Paidós, 1996.

AYODELE, T. O. Types of machine learning algorithms. *In*: Zhang, Y. (ed.). **New advances in machine learning**. Reino Unido: InTech, 2010. p.19-48.

AZEVEDO NETTO, C. X. de A. SIGNO, SINAL, INFORMAÇÃO: as relações de construção e transferência de significados. **Informação & Sociedade**: Estudos, v. 12, n. 2, 2002. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/143. Acesso em: 9 jan. 2024.

BACHIOQUI, E. L. La interpretación y los actos de habla. **Mutatis Mutandis**, v. 3, n. 2, p. 333-348, 2010.

BAEZA-YATES, R. Challenges in the Interaction of Information Retrieval and Natural Language Processing. *In*: **International Conference on Computacional Linguistics and Intelligent Text Processing** (CICLing) 5., 2004, Seoul.

BARANOW, U. G. Perspectivas na contribuição da lingüística e de áreas afins à Ciência da Informação. **Ciência da Informação**, v. 12, n. 1, 1983. DOI: 10.18225/ci.inf..v12i1.191 Acesso em: 05 mai. 2023.

BARBOSA, A. **Teoria e prática dos sistemas de classificação** bibliográfica. Rio de Janeiro: Instituto Brasileiro de Bibliografia e Documentação, 1969.

BATLEY, S. **Classification in theory and practice**. Oxford: Chandos, 2005.

BEARDON, C.; LUMSDEN, D.; HOLMES, G. **Natural Language and Computational Linguistics**. England: Ellis Horwood Ltd., 1991.

BEGHTOL, C. Semantic validity: concepts of warrant in bibliographic classification systems. **Library Resources & Technical Services**, v. 30, n. 2, p. 109-25, 1986.

BENOÎT, G. Critical theory as a foundation for pragmatic information systems design. **Information**, v. 6, n. 2, 2001. Disponível em: http://informationr.net/ir/6-2/paper98.html. Acesso em: 14 dez. 2023.

BENTHEM, J. V. **Handbook of the Philosophy of Information**. Amsterdã: Elsevier, 2008.

BIRD, S; KLEIN, E; LOPER, E. **Natural language processing with Python**: analyzing text with the natural language toolkit. Sebastopol: O'Reilly Media, 2009.

BLAIR, D. C. Information retrieval and the philosophy of language. **Annual Review of Information Science and Technology**, p. 3-50, 2003.

BLAIR, D. C. Information retrieval and the philosophy of language. **Computer Journal**, v. 35, p. 200-207, 1992.

BORKO, H. Information Science: what is it?. **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

BRÄSCHER, M. A ambiguidade na recuperação da informação. **Datagramazero**, v. 3, n. 1, 2002.

BRÄSCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento?. *In*: ENANCIB, IX. 2008, São Paulo. **Diversidade cultural e políticas de informação**. São Paulo: USP, 2008.

BRÄSCHER, M.; GUIMARÃES, J. A. C. Tratamento temático da informação (TTI): influência dos paradigmas físico, cognitivo e social em artigos de revisão de literatura no período de 1966-1995. **LIINC EM REVISTA**, v. 14, p. 241-258, 2018.

BRUGGER, W. **Dicionário de filosofia**. São Paulo: Editora Herder, 1962.

BURKE, P. **Uma história social do conhecimento**: de Gutenberg a Diderot. Rio de Janeiro: Jorge Zahar, 2003.

CAFÉ, L. M. A.; SALES, R. de. Organização da informação: conceitos básicos e breve fundamentação teórica. *In*: ROBREDO, J.; BRÄSCHER, M. (org.). **Passeios pelo bosque da informação**: estudos sobre representação e organização da informação e do conhecimento. Brasília: IBICT, 2010. p. 115-129.

CAMARA JR., J. M. **Dicionário de linguística e gramática**: referente à Língua Portuguesa. Petrópolis: Editora Vozes, 2009.

CAMARGO, B. V.; JUSTO, A. M. **Tutorial para o uso do software IRaMuTeQ**. Florianópolis: Laboratório de Psicologia Social da Comunicação e Cognição (UFSC), 2021. Disponível em: http://www.iramuteq.org/documentation/fichiers/Tutorial%20IRaMuTeQ%20em%20portugues_22.11.2021.pdf. Acesso em: 15 jul. 2024.

CAMARGO, B. V.; JUSTO, A. M. IRAMUTEQ: Um software gratuito para análise de dados textuais. **Temas em Psicologia**, v. 21, n. 2, 2013.

CAMPOS, R. *et al*. YAKE! Keyword Extraction from Single Documents using Multiple Local Features. **Information Sciences Journal**, v. 509, p. 257-289, 2019.

CAPURRO, R. Epistemologia e Ciência da informação. *In*: V ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5., Belo Horizonte, 2003. **Anais**... Belo Horizonte: Escola de Ciência da informação da UFMG, 2003.

CAPURRO, R. Epistemología y ciencia de la información. **Enlace**, v. 4, n. 1, p. 11-29, 2007.

CASADO-MANCEBO, M. Una aproximación a la lingüística computacional. **Sincronía**, v. 26, n. 81, p. 746-76, 2021. http://dx.doi.org/10.32870/sincronia.axxvi.n81.34a22.

CASELI, H. de M.; NUNES, M. das G. V. **Processamento da Linguagem Natural**: Conceitos, Técnicas e Aplicações em Português. 2. ed. São Carlos: Brasileiras em LN, 2024.

CHAUMIER, J. Indexação, conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. ½, p. 63-79, 1988.

COFFA, J. A. **The semantic tradition from Kant to Carnap**: to the Vienna Station. Indiana: Cambridge University Press, 1991.

CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Inteligência artificial e ferramentas da web semântica aplicadas à recuperação da informação: um modelo conceitual com foco na linguagem natural. **Informação & Informação**, v. 27, n. 1, p. 625-651, 2022. DOI: 10.5433/1981-8920.2022v27n1p625. Acesso em: 17 abr. 2023.

CORTEXT. **About us**. A research infrastructure for the social sciences and humanities. França: LISIS, 2024a. Disponível em: https://www.cortext.net/about-us/. Acesso em: 15 jul. 2024.

CORTEXT. **Cortext**. A research infrastructure for the social sciences and humanities. França: LISIS, 2024b. Disponível em: https://www.cortext.net/. Acesso em: 15 jul. 2024.

CORTEXT. **Cortext Manager Documentation**. Learn about CorText methods and share your experience. França: LISIS, 2024c. Disponível em: https://docs.cortext.net/. Acesso em: 15 jul. 2024.

CORTEXT. **Cortext Manager**. França: LISIS, 2024d. Disponível em: https://www.cortext.net/activities-projects/cortext-manager/. Acesso em: 15 jul. 2024.

CRYSTAL, D. **Dicionário de linguística e fonética**. Rio de Janeiro: Jorge Zahar Editor, 1985.

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, 2008.

DAHLBERG, I. Teoria da classificação, ontem e hoje. *In*: **CONFERENCIA BRASILEIRA DE CLASSIFICAÇÃO BIBLIOGRÁFICA**, Rio de Janeiro: IBICT, 1979.

DAVANZO, L.; MOREIRA, W. O vocabulário controlado como ferramenta do processo de organização e recuperação da informação. *In:* **XV ENANCIB**, 2014, Belo Horizonte-MG.

DEMŠAR, J.; CURK, T.; ERJAVEC, A.; GORUP, C.; HOČEVAR, T.; MULUTINIVIČ, M.; MOŽINA, M.; PONAJNAR, M.; TOPLAK, M.; STARIČ, A.; ŠTAJDOHAR, M.; UMEK, L.; ŽAGAR, L.; ŽBONTAR, J.; ŽITNIK, M.; ZUPAN, B. Orange: Data Mining Toolbox in Python. **Journal of Machine Learning Research**, v. 14, 2013.

DEMŠAR, J.; ZUPAN, B. Orange: Data Mining Fruitful and Fun - A Historical Perspective. **Informatica**, v. 37, 2013.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. 2. ed. rev. Brasília: Briquet de Lemos, 2013.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116p.

DIGITAL LIBRARIES AND MACHINE LEARNING LABS. **KEA** - Keyphrase extraction algorithm. [S.l.]. Digital Libraries and Machine Learning Labs, 2024. Disponível em: https://community.nzdl.org/kea/index.html. Acesso em: 6 jun. 2024.

FALCÃO, L. C. de J.; LOPES, B.; SOUZA, R. R. Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. **Em Questão**, Porto Alegre, v. 28, n. 1, p. 13–34, 2021. DOI: 10.19132/1808-5245281.13-34. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/111323. Acesso em: 13 jul. 2023.

FERNANDES, J. C.; BENTES PINTO, V.; AZEVEDO NETTO, C. X. de. Representação indexal na web: estudo do sintagma "História da Paraíba" nos sites Alta Vista e Google. **Informação & Sociedade**: Estudos, v. 15, n. 2, p. 181-212. 2005.

FERNEDA, E. Recuperação de Informação: **Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 2003. Tese de Doutorado. Universidade de São Paulo.

FERRÁNDEZ, A.; ROJAS, Y.; PERAL, J. The successful application of natural language processing for information retrieval. **Journal of Computer Science and Technology**, v. 7, n. 1, p. 79-85, 2007.

FINATTO, M. J. B.; LOPES, L.; CIULLA, A. Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida. **DOMÍNIOS DE LINGU@GEM**, v. 9, n. 5, 2015.

FINLAND, National Library Of. **Annif**: tool for automated subject indexing. Tool for automated subject indexing. 2024a. Disponível em: https://annif.org/. Acesso em: 26 mar. 2024a.

FINLAND, National Library Of. **Finto AI**: a service for automated subject indexing based on artificial intelligence. A service for automated subject indexing based on artificial intelligence. 2024b. Disponível em: https://ai.finto.fi/?locale=en. Acesso em: 26 mar. 2024.

FINTO. **YSO** - Yleinen suomalainen ontologia. Finlândia: Finto.fi, 2024. Disponível em: https://finto.fi/yso/fi/. Acesso em: 4 jun. 2024.

FREGE, G. Begriffsschrift: a formula language of pure thought modelled on that of arithmetic. *In*: BEANEY, M. **The Frege Reader**. Oxford: Blackwell Publishers, 1997, p. 47-78.

FROEHLICH, H. Corpus Analysis with Antconc. **Programming Historian**, v. 4, 2015. Disponível em: https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc. Acesso em: 16 jul. 2024.

FUCHS, C. **Les ambiguïtés du français**. Paris: Ophrys, 1996. 184 p. Coleção L'Essentiel Français.

FUJITA, M. S. L. **A análise documentária no tratamento da informação**: as operações e os aspectos conceituais interdisciplinares. Marília: Departamento de Ciência da Informação, FFC/UNESP, 2003.

FUJITA, M. S. L. **A leitura documentária na formação inicial do indexador**: a abordagem sociocognitiva na investigação de estratégias de ensino. 2004. Descrição detalhada (Projeto Integrado de Pesquisa) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista; CNPq, Marília.

FUJITA, M. S. L.; NARDI, M. I. A.; SANTOS, S. A leitura em análise documentária. **Transinformação**, v. 10, n. 3, p. 13 – 31, set./dez. 1998.

GAMALLO, P.; GARCÍA, M. **Técnicas de Procesamiento del Lenguaje Natural en la recuperación de información**. 2012.

GARDIN, Jean-Claude. Les analyses de discours. **Neuchâtel**: Delachaux et Niestlé, 1974, p. 81-82.

GARMENDIA, J. Revisiting Searle. **Teorema**: Revista internacional de filosofía, v. 28, n. 3, p. 191-195, 2009.

GASPERIN, C. V. **Fundamentos do Processamento Estatístico da Linguagem Natural**. Porto Alegre, PPGCC-PUCRS: Trabalho Individual, 2000.

GIASSON, J. **A compreensão na leitura**. Lisboa: Asa, 1993.

GIL, A. C. **Como elaborar projetos de pesquisa**. 7. ed. São Paulo: Atlas, 2022.

GIL, F. Classificações. *In*: _____. **Enciclopédia Einaudi**. Portugal: Imprensa Nacional/Casa da Moeda. 2001. 432p. v.41.

GONZÁLEZ DE GÓMEZ, M. N. Comentários ao artigo “Hacia un nuevo paradigma en Bibliotecologia”. **Transinformação**, v. 8, n. 3, 1996.

GONZÁLEZ DE GÓMEZ, M. N. AS AÇÕES DE INFORMAÇÃO E SEUS CONTEXTOS: aportes da pragmática ao campo investigativo da Ciência da Informação. **Informação & Sociedade**: Estudos, v. 30, n. 4, 2020.

GOOGLE CODE ARCHIVE. **KEA** - Keyphrase Extraction Algorithm. Google, 2024. Disponível em: https://code.google.com/archive/p/kea-algorithm/. Acesso em: 06 jun. 2024.

GOOGLE CODE ARCHIVE. **Maui** - Multi-purpose automatic topic indexing. Google, 2024. Disponível em: https://code.google.com/archive/p/maui-indexer/. Acesso em: 30 mai. 2024.

GRACIOSO, L. S. Parâmetros teóricos para elaboração de instrumentos pragmáticos de representação e organização da informação na Web: considerações preliminares sobre uma possível proposta metodológica. **InCID**: R. Ci. Inf. e Doc., Ribeirão Preto, v. 1, n.1, p. 138-158, 2010.

GREIMAS, A. J.; COURTÉS, J. **Dicionário de semiótica**. São Paulo: Editora Contexto, 2008.

GUIMARÃES, J. A. C. Abordagens teóricas em tratamento temático da informação: catalogação de assunto, indexação e análise documental. *In*: GARCÍA MARCO, F. J. (Org.). **Avances y perspectivas en sistemas de información y de documentación**. Zaragoza: Prensas Universitárias de Zaragoza, 2009. p. 105-117.

GUIMARÃES, J. A. C. A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knwoledge Organization (ISKO). **Revista Ibero-Americana** de Ciência da Informação, v. 1, p. 77-99, 2008.

GUIMARÃES, J. A. C. *et al*. A dimensão conceitual da organização do conhecimento no universo científico da ISKO: uma análise de domínio a partir dos congressos da ISKO-Brasil, ISKO-Espanha, ISKO-América do Norte e ISKO-França. **Scire**, Zaragoza, v. 21, n. 2, p. 13-26, 2015.

GUIMARÃES, J. A. C.; FERREIRA, G. M.; FREITAS, M. F. M. Correntes teóricas do tratamento temático da informação: uma análise de domínio da presença da catalogação de assunto e da indexação nos congressos de ISKO-España. *In*: CONGRESO ISKO CAPÍTULO ESPAÑOL, 10., 2011. **Actas** [...]. Ferrol, 2012.

GUIMARÃES, J. A. C.; SALES, R. de. Análise documental: concepções do universo acadêmico brasileiro em Ciência da Informação. **DataGramaZero**, Rio de Janeiro, v. 11, n. 1, fev. 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 11, n. 1, 2009.

HARTMANN, R. R. K.; STORK, F. C. **Dictionary of language and linguistics**. Londres: Applied Science Publishers LTD, 1972.

HERNÁNDEZ SAMPIERI, R.; FERNÁNDEZ COLLADO, C.; BAPTISTA LUCIO, M. del P. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261-266, 2015.

HJØRLAND, B. Fundamental of knowledge organizations. **Knowledge Organization**, v. 30, n. 2, p. 87-111, 2003.

HJØRLAND, B. Information retrieval, text composition, and semantics. **Knowledge Organization**, v. 25, n. 1-2, 1998.

HJØRLAND, B.; NISSEN PEDERSEN, K. A substantive theory of classification for information retrieval. **Journal of Documentation**, v. 61, n. 5, p. 582-597, 2005.

HJØRLAND, B. Semantic and Knowledge organization. **ARIST**, v. 41, p. 367-405, 2007.

INESCTEC. **Tecnologia Yake!**. Extrator de palavras-chave automático e multilíngue plug-and-play. Porto: NQ Digital Agency, 2024. Disponível em: https://www.inesctec.pt/pt/tecnologias/yake#. Acesso em: 29 mai. 2024.

JIN, X. *et al.* Significance and Challenges of Big Data Research. **Big Data Research**, v. 2, n. 2, p. 59-64, 2015. Elsevier BV. http://dx.doi.org/10.1016/j.bdr.2015.01.006.

KAMP, H.; STOKHOF, M. Information in Natural Language. *In*: ADRIAANS, P.; BENTHEM, J. V. **Handbook of the Philosophy of Information**. Amsterdã: Elsevier, 2008.

KÉDROV, B. M. **Clasificación de las ciencias**: Engels y sus predecesores. Moscou: Editorial Progreso, 1974.

KOBASHI, N. Y. **Análise documentária**: metodologias para indexação e resumo. [S.l: s. n.], 1995.

KOBASHI, N. Y. **Elaboração de informações documentárias**: em busca de uma metodologia. 1994. 195 f. Tese (Doutorado) – Curso de Ciência da Informação, Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 1994.

KOBASHI, N. Y.; FERNANDES, J. C. Pragmática Lingüística e organização da informação. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais**... São Paulo: Ancib, 2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/171978. Acesso em: 13 abr. 2023.

KOBASHI, N. Y. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. **DataGramaZero**, v. 8, n. 6, 2007. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6666. Acesso em: 13 abr. 2023.

KOBASHI, N. Y. Organização da informação e do conhecimento: interfaces informacionais, comunicacionais e linguísticas. *In*: SIMÕES, M. da G.; BORGES, M. M. (coord.). **Tendências atuais e perspectivas futuras em organização do conhecimento: atas do III Congresso ISKO Espanha e Portugal – XIII Congresso ISKO Espanha**. Universidade de Coimbra: Centro de Estudos Interdisciplinares do Século XX, 2017. p. 85-95.

KOBASHI, N. Y. Resumos documentários: uma proposta metodológica. **Revista de Biblioteconomia de Brasília**, v. 21, n. 2, p. 201-210, jul./dez. 1997.

KOCHANI, A. P.; BOCCATO, V. R. C.; RUBI, M. P. Política de indexação para clippings: otimização do tratamento temático em coordenadorias de comunicação social brasileiras. **Revista Interamericana de Bibliotecología**, Medellín, v. 35, n. 3, p. 257-273, 2012.

LANCASTER. F. W. **Indexação e resumos: teoria e prática**. Tradução de: Antonio Agenor Briquet de Lemos. 2 ed. ver. atual. Brasília: Briquet de Lemos, Livros, 2004a.

LANCASTER, F. W. **Indexação**: teoria e prática. 2. ed. rev. atual. Brasília, DF: Briquet de Lemos/Livros, 2004b.

LEITÃO, H.; SIMÕES, M. G. O resumo científico como recurso de acesso equitativo à informação. *In*: Congresso ISKO Espanha, 13, 2017, Coimbra. **Anais**... Coimbra: Universidade de Coimbra, 2017, p.825-839.

LEWIS, D. D.; JONES, K. S. Natural language processing for information retrieval. **Communications of the ACM**, v. 39, n. 1, p. 92-101, 1996.

LLAMAZARES, M. V. LINGÜÍSTICA CON CORPUS. **E.H. Filología**, n. 30, 2008.

LOWE, W. **Conjugate prior**. Softwares. Will Lowe, 2024. Disponível em: https://conjugateprior.org/. Acesso em: 17 jul. 2024.

LOWE, W. Yoshikoder: An Open Source Multilingual Content Analysis Tool for Social Scientists. *In*: Annual meeting of the American Political Science Association, 2006, Philadelphia. **Proceedings of the American Political Science Association** [...]. Philadelphia: AMERICAN POLITICAL SCIENCE ASSOCIATION, 2006.

LUCCA, J. L. de; NUNES, M. G. V. **Lematização versus stemming**. Relatórios Técnicos do ICMC 14, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos-SP, 2002.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v. 2, n. 2, p. 159-165, 1958.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309-317, 1957.

MAIMONE, G. D.; KOBASHI, N. Y.; MOTA, D. Indexação: teoria e métodos. *In*: SILVA, J. F. M. da; PALETTA, F. C. (org.). **Tópicos para o ensino de biblioteconomia**: volume I. São Paulo: ECA-USP, 2016.

MANESH, B. Machine Learning Algorithms - A Review. **International Journal of Science and Research** (IJSR), v. 9, n. 1, 2020.

MANNING, C.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: MIT Press, 1999. 620 p.

MANZO, A. J. **Manual para la preparación de monografías**: una guía para presentar informes y tesis. 2. ed. Buenos Aires: Humanitas, 1973.

MARCONDES, D. Desfazendo mitos sobre a pragmática. **ALCEU**, Rio de Janeiro, v. 1, n. 1, p. 38-46, 2000.

MARCONI, M. de A.; LAKATOS, E. M. **Metodologia Científica**. 8. ed. São Paulo: Atlas, 2022.

MARI, H. Dos fundamentos da significação à produção do sentido. **Perspectivas em Ciência da Informação**, v. 1, n. 1, p. 93-109, 1996.

MARTINES, A. R. Ciência da Informação e seus paradigmas: a presença da linguagem na organização do conhecimento. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 19, p. 1-29, 2023.

MARTÍNEZ VÁSQUEZ, L. A. El lenguaje como sentido y significación en la filosofía de Gottlob Frege y Xavier Zubiri. **Revista Praxis**, v. 1, n. 82, 2020.

MEDEIROS, R. Resumo de artigo científico: sua estrutura composicional. **Revista de Pesquisa em Arte**, v. 5, n .1, 2018. p. 67-73.

MEDEIROS, R. **Resumo de dissertações e teses**: a estrutura composicional à luz da análise textual dos discursos. 2015. Tese (Doutorado em Estudos da Linguagem). Centro de Ciências Humanas, Letras e Artes, Universidade Federal do Rio Grande do Norte, Natal, 2015.

MEDELYAN, O. **Human-competitive automatic topic indexing**. PhD Thesis. University of Waikato, New Zealand, 2009.

MILSTEAD, J. L. Indexing for subject cataloguers. **Cataloging e Classification Quarterly**, New York, v. 3, n. 4, p. 37-44, 1983.

MOKVA, A. M. D. Z.; COMASSETTO, L. R.; FONTANA, N. I. Linguística de corpus: como utilizá-lo para compreensão em gramática. **REVISTA VOZ DAS LETRAS**, Concórdia: Universidade do Contestado, n. 1, 2004.

MORA, J. F. **Dicionário de filosofia**. São Paulo: Martins Fontes, 1993.

MOREIRA, D. *et al*. Análisis del Estado Actual de Procesamiento de Lenguaje Natural. **Revista Ibérica de Sistemas e Tecnologias de Informação**, n. 42, p. 126-136, 2021.

MOREIRA, W.; MORAES, I. S. O assunto "classificação" na literatura brasileira de Ciência da Informação: uma análise nos anais do ENANCIB (2003-2014). **Informação & Informação**, Londrina (PR), v.24, n.1, p. 226-246, 2019.

MORK, J. G.; YEPES, A. J. J.; ARONSON, A. R. **The NLM Medical Text Indexer System for Indexing Biomedical Literature**. 2013 Disponível em: https://lhncbc.nlm.nih.gov/ii/information/Papers/MTI_System_Description_Expanded_2013_Accessible.pdf. Acesso em: 05 jun. 2024.

MORRIS, C. W. Foundations of the Theory of Signs. **International encyclopedia of unified science**, p. 1-59, 1938.

MOTA, D. A. R.; KOBASHI, N. Y. Web semântica e web pragmática: discussão crítica sobre versionamento na web e limites conceituais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n. 2, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119679. Acesso em: 12 abr. 2023.

MUCCHIELLI, A. **Les sciences de l'information et de la communication**. Paris: Hachette, 1995.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação Automatizada de Artigos de Periódicos Científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. **Informação & Sociedade**: Estudos, João Pessoa, v. 19, n. 2, p. 99-118, 2009.

NATIONAL LIBRARY OF FINLAND. **Annif**. Finlândia: National Library of Finland, 2024. Disponível em: https://annif.org/. Acesso em: 4 jun. 2024.

NATIONAL LIBRARY OF FINLAND. **Finto AI**. a service for automated subject indexing based on artificial intelligence. [S.l.]. Kansalliskirjasto, 2024. Disponível em: https://ai.finto.fi/?locale=en. Acesso em: 11 jul. 2024.

NATIONAL LIBRARY OF FINLAND. **Finto.fi**. Centralized service for interoperable thesauri, ontologies and classification schemes. [S.l.]. Kansalliskirjasto, 2024. Disponível em: https://finto.fi/en/. Acesso em: 11 jul. 2024.

NATIONAL LIBRARY OF FINLAND. **FINTO - SANASTO- JA ONTOLOGIAPALVELUT**. Finto AI, automaattisen sisällönkuvailun palvelu. [S.l.]. Kansalliskirjasto, 2016. Disponível em: https://www.kiwi.fi/pages/viewpage.action?pageId=165380606. Acesso em: 11 jul. 2024.

NATIONAL LIBRARY OF MEDICINE. **NLM Medical Text Indexer (MTI)**. [S.l.]. National Library of Medicine, 2024. Disponível em: https://lhncbc.nlm.nih.gov/ii/tools/MTI.html. Acesso em: 5 jun. 2024.

NEVEU, F. **Dicionário de ciências da linguagem**. Petrópolis: Editora Vozes, 2008.

NORTE, M. B. **Glossário de termos técnicos em ciência da informação**: inglês/português. São Paulo: Cultura Acadêmica, 2010.

NÖTH, W. **A semiótica no século XX**. 3. ed. São Paulo: ANNABLUME, 1996.

NOVELLINO, M. S. F. A linguagem como meio de representação ou de comunicação da informação. **Perspectivas em Ciência da Informação**, v. 3, n. 2, 1998. Disponível em: http://hdl.handle.net/20.500.11959/brapci/38505. Acesso em: 19 abr. 2023.

OLIVEIRA, L. P. de. **Bases conceituais e formativas do tratamento temático da informação no Brasil**. 2021. 490 f. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Filosofia e Ciências, Programa de Pós-Graduação em Ciência da Informação, 2021.

OLIVEIRA, L. P. de; GRÁCIO, M. C. C.; MARTÍNEZ-ÁVILA, D. Instrumentos, processos e produtos do Tratamento Temático da Informação: um estudo nos anais da ISKO-Brasil (2012-2019). **Scire**, Zaragoza, v. 26, n. 1, p. 47-56, en./jun. 2020.

OLIVEIRA, L. P. de; MARTÍNEZ-ÁVILA, D. Tratamento temático da informação: análise sobre abordagens e enfoques em artigos científicos da área de Ciência da Informação. **Informação em Pauta**, Fortaleza, v. 4, n. 2, p.83-100, jul./dez. 2019. DOI: 10.32810/2525-3468.ip.v4i2.2019.42654.83-100.

ORANGE DATA MINING. **Loading your Data**. University of Ljubljana, 2015. Disponível em: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/loading-your-data/index.html. Acesso em: 16 jul. 2024.

ORANGE DATA MINING. **Orange Data Mining**. University of Ljubljana, 2024a. Disponível em: https://orangedatamining.com/. Acesso em: 16 jul. 2024.

ORANGE DATA MINING. **Widget Catalog**. University of Ljubljana, 2024b. Disponível em: https://orangedatamining.com/widget-catalog/. Acesso em: 16 jul. 2024.

PACHECO, L. S. Informação enquanto artefato. INFORMARE - Cadernos do Programa de **Pós-graduação em Ciência da Informação**. Rio de Janeiro: IBICT/CNPq, v. 1, n. 1, p. 20-24, 1995.

PEREGRIN, J. Pragmatization of Semantics. *In*: TURNER, K. **The semantics/Pragmatics Interface from Different Points of View**. Amsterdam: Elsevier, 1999.

PINHEIRO. L. V. R. Medidas de consistência indexação: interconsistência. **Ciência da Informação**, v. 7, n. 2, p. 109-114, 1978.

PINTO MOLINA, M.; GARCÍA MARCO, F. J.; AGUSTIN LACRUZ, M. del C. **Indización y resumen de documentos digitales y multimedia**: técnicas y procedimientos. Gijón: Trea, 2002.

PITT, E. A. A influência da tradição semântica analítica na filosofia de Frege. **Griot**: Revista de Filosofia, v. 18, n. 2, p. 325-337, 2018.

PUTNAM, H. **Representation and reality**. Cambridge: MIT Press, 1988.

RATINAUD, P. **Iramuteq** - Documentação. LERASS, 2024a. Disponível em: http://www.iramuteq.org/documentation/html. Acesso em: 15 jul. 2024.

RATINAUD, P. **Iramuteq**. Interface R para análises multidimensionais de textos e questionários. LERASS, 2024b. Disponível em: http://www.iramuteq.org/. Acesso em: 15 jul. 2024.

REDIGOLO, F. M. **O processo de análise de assunto na catalogação de documentos**: a perspectiva sociocognitiva do catalogador em contexto de Biblioteca Universitária. 2010. 176 f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

REDIGOLO, F. M. **O processo de análise de assunto na catalogação de livros em bibliotecas universitárias**: aplicação do protocolo verbal. 2014. 262 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.

REITZ, J. M. **Dictionary for Library and Information Science**. Westport: Greenwood Publishing Group, 2004.

RENDÓN ROJAS, M. A. Hacia um nuevo paradigma em bibliotecologia. **Transinformação**, Campinas, v. 8, n. 3, p. 17-31, set./dez. 1996.

REQUENA, E. Frege: sentido y referencia. **Teorema**: Revista Internacional de Filosofía, v. 4, n. 4, 1974, p. 579-586.

RIBEIRO, F. Da mediação passiva à mediação pós-custodial: o papel da ciência da informação na sociedade em rede. **Informação & Sociedade**: Estudos, v. 20, n. 1, p. 63-70, 2010.

ROTT, H. Words in Contexts: Frege Elucidations. **Linguistics and Philosophy**, v. 23, n. 6, 2020.

RUSS, J. **Dicionário de filosofia**. São Paulo: Editora Scipione, 1993.

SABINO, C. A. **El proceso de investigación**. Buenos Aires: Lumen-Humanitas, 1996.

SALDANHA, G. S. **Viagem aos becos e travessas da tradição pragmática da Ciência da Informação**: uma leitura em diálogo com Wittgenstein. 2008. 302f. Dissertação (Mestrado) – Universidade Federal de Minas Gerais, 2008.

SANDOVAL, A. M. Lingüística computacional: introducción a los modelos simbólicos, estadísticos y biológicos. **Síntesis**, 1998.

SAN SEGUNDO MANUEL, R. **Sistemas de organización del conocimiento**: la organización del conocimiento en las bibliotecas españolas. Madrid: Imprenta Nacional del Boletín Oficial del Estado, 1996.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, 1996. Disponível em: http://hdl.handle.net/20.500.11959/brapci/37415. Acesso em: 26 abr. 2023.

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. **DELTA**: Documentação de Estudos em Lingüística Teórica e Aplicada, v. 16, n. 2, p. 323–367, 2000.

SARDINHA, T. B. **Lingüística de corpus**. São Paulo: Manole, 2004.

SEARLE, J. ¿Qué es un acto de habla?. *In*: VÁLDES, L. (edit.). **La búsqueda del significado**. Madrid: Ed. Tecnos, 1991.

SILVA, M. A. A. Austin y searle: la relación entre verbos y actos ilocucionarios. **Literatura y Lingüística**, n. 19, p. 235-250, 2008.

SILVA, S. R. de B.; CORREA, R. F.; GIL-LEIVA, I. Avaliação direta e conjunta de Sistemas de Indexação Automática por Atribuição. **Informação & Sociedade**: Estudos, v. 30, n. 4, p. 1-27, 2020.

SILVA, T. F. L. O Pão Que O Viado Amassou: contribuições da semiótica para o processamento de língua natural. **Estudos Semióticos**, v. 18, n. 3, p. 70-92, 2022.

SINCLAIR, S.; ROCKWELL, G. **Voyant Tools**. See through your text. Canadá: Stéfan Sinclair & Geoffrey Rockwell, 2024. Disponível em: https://voyant-tools.org/. Acesso em: 11 jul. 2024.

SUOMINEN, O. Annif: DIY Automated Subject Indexing Using Multiple -Algorithms. **Liber Quarterly**, v. 29, 2019.

SUOMINEN, O; INKINEN, J; LEHTINEN, M. Annif and Finto AI: Developing and Implementing Automated Subject Indexing. **JLIS.it**, v. 13, n. 1, 2022.

THELLEFSEN, T. L; THELLEFSEN, M. M. Pragmatic semiotics and knowledge organization. **Knowledge Organization**, v. 31, n. 3, p. 177-187, 2004.

TRASK, R. L. **Dicionário de Linguagem e Lingüística**. São Paulo: Contexto, 2004.

UNIVERSITY OF WAIKATO. **The Weka Workbench**. Waikato: University of Waikato, 2024. Disponível em: https://ml.cms.waikato.ac.nz/index.html. Acesso em: 8 jul. 2024.

VALLEZ, M.; JIMENEZ-PEDRAZA, R. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. **Hipertext.net**, 2007.

VARELA, A. V.; BARBOSA, M. L. A. Aplicação de teorias cognitivas no tratamento da informação. **Revista Brasileira de Biblioteconomia e Documentação**, v. 3, n. 2, 2007.

VIEIRA, R. de M. **Introdução à teoria geral da biblioteconomia**. Rio de Janeiro: Interciência, 2014.

VIEIRA, R.; LIMA, V. L. S. Lingüística computacional: princípios e aplicações. *In*: **IX Escola de Informática da SBC-Sul**. NEDEL, L. (ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001.

VILLANUEVA, L. M. V. J. R. Searle: Una teoría general de los actos de habla. **Teorema**: Revista internacional de filosofía, v. 5, n. 3/4, p. 501-513, 1975.

VILARES, J. **Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español**. 2005. 297 f. Tese (Doutorado) - Curso de Ciência da Computação, Departamento de Computação, Universidade da Coruña, Coruña, 2005.

VOYANT TOOLS. **About**. Canadá: Stéfan Sinclair & Geoffrey Rockwell, 2024a. Disponível em: https://voyant-tools.org/docs/#!/guide/about. Acesso em: 11 jul. 2024.

VOYANT TOOLS. **List of Tools**. Canadá: Stéfan Sinclair & Geoffrey Rockwell, 2024b. Disponível em: https://voyant-tools.org/docs/#!/guide/tools. Acesso em: 11 jul. 2024.

WERSIG, G., NEVELING, U. The phenomena of interest to information science. **Information Scientist**, v. 9, p. 127-140, 1975.

WISCHIN, K. Sentido, significado y el principio de contexto en Frege. **Philosophical Research Bulletin**, v. 1, n. 2, 2012, p. 94-104.

WITTEN, I. H.; FRANK, E.; TRIGG, L.; HALL, M.; HOLMES, G.; CUNNINGHAM, S. J. **Weka**: practical machine learning tools and techniques with Java implementations. 1999. Disponível em: https://ml.cms.waikato.ac.nz/index.html.

WITTEN, I. H.; PAYNTER, G. W.; FRANK, E.; GUTWIN, C.; NEVILL-MANNING, C. G. KEA: Practical automatic keyphrase extraction. *In*: ACM conference on Digital libraries, 4., 1999, Berkeley. **Proceedings** [...] New York: Association for Computing Machinery, 1999. p. 254-255.

XAVIER, B. M.; SILVA, A. D. da; GOMES, G. R. R. Análise comparativa de algoritmos de redução de radicais e sua importância para a mineração de texto. **Revista Eletrônica Pesquisa Operacional para o Desenvolvimento**, v. 5, n. 1, 2013.

YAKE!, 2024. Disponível em: http://yake.inesctec.pt/demo.html. Acesso em: 26 mar. 2024.

YOSHIKODER. **YKConverter**. Weatherhead Center for International Affairs, 2024. Disponível em: https://www.yoshikoder.org/ykconverter/. Acesso em: 17 jul. 2024a.

YOSHIKODER. **Yoshikoder**. Weatherhead Center for International Affairs, 2024. Disponível em: https://www.yoshikoder.org/. Acesso em: 17 jul. 2024b.

ZERBINATTI, L. **Extração de Conhecimento de Laudos de Radiologia Torácica Utilizando Técnicas de Processamento Estatístico de Linguagem Natural**. 2010. 160 f. Tese (Doutorado) - Curso de Engenharia, Departamento de Engenharia de Telecomunicações e Controle, Escola Politécnica da Universidade de São Paulo, São Paulo, 2010.

ZEROUAL, I.; LAKHOUAJA, A. Data science in light of natural language processing: an overview. **Procedia Computer Science**, [S.L.], v. 127, p. 82-91, 2018. Elsevier BV. http://dx.doi.org/10.1016/j.procs.2018.01.101.

ZHOU, L.; ZHANG, D. NLPIR: A theoretical framework for applying natural language processing to information retrieval. **Journal of the American Society for Information Science and Technology**, v. 54, n. 2, p. 115-123, 2003.

APÊNDICE A – CONTRIBUIÇÕES DE CRITÉRIOS DE QUALIDADE DE *SOFTWARES* DE PROCESSAMENTO DA LINGUAGEM NATURAL À INDEXAÇÃO[27]

A análise realizada dos *softwares* Yake!, MAUI, Annif, *Medical Text Indexer*, KEA, WEKA, Finto AI, *Voyant Tools*, Cortext, IRaMuTeQ, AntConc, *Orange Data Mining* e Yoshikoder identificou a insuficiência dos sistemas no que diz respeito à incorporação de fundamentos pragmáticos.

Com base nesta limitação, a pesquisa aqui apresentada sugere critérios de qualidade de *softwares* como forma de garantir a qualidade da automação das atividade de tratamento temático da informação e preservar as relações de sentido e relevância entre documentos e suas representações – produtos documentários – realizadas de maneira automatizada.

Nas últimas décadas, foi possível observar o crescimento exponencial do interesse pelos estudos sobre a aplicação e aperfeiçoamento de produtos tecnológicos nos serviços e operações próprias do ciclo informacional, além do esforço conjunto para o desenvolvimento de novos instrumentos tecnológicos para a automação dos serviços informacionais, como vem sendo realizado no processo de indexação.

O objetivo da atividade indexadora é a condensação do conteúdo temático documental através de sua tradução em termos de indexação representativos por meio do uso de linguagens documentárias, realizada em duas etapas fundamentais: análise conceitual ou de assunto, na qual é realizada a leitura e compreensão do conteúdo documental, e a tradução de conceitos em descritores ou termos da linguagem documentária (Lancaster, 2004[28]; Cintra, 2002[29]).

Por ser entendida como fase fundamental ao processo de indexação, a análise de assunto deve ser realizada com excelência, tendo em vista a qualidade dos produtos documentários resultantes, o que depende diretamente de fatores subjetivos, como a experiência e conhecimento do profissional indexador, a adequação do ambiente de trabalho e a qualidade dos instrumentos e estratégias auxiliadoras.

---

[27] Pesquisa financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 128763/2024-8.

[28] LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

[29] CINTRA, A. M. M. **Para entender as linguagens documentárias**. 2. ed. rev. e atual. São Paulo: Editora Polis, 2002.

Justamente por sua dependência da atividade intelectual humana, a indexação caracteriza-se por sua subjetividade e sujeição aos erros humanos, afetando diretamente a consistência e qualidade de seus produtos, refletindo principalmente na fase de análise de assunto, como destacam Lancaster (2004) e Pinheiro (1978)[30].

Tendo em vista a natureza subjetiva e altamente intelectual da indexação, sua vulnerabilidade às falhas humanas do profissional, o tempo e recursos necessários para sua execução e o crescimento cada vez mais acelerado na produção documental, é possível destacar e justificar o processo de automação da atividade indexadora, como argumenta Suominen (2019)[31].

Em busca de mitigar os efeitos do fator humano na indexação, iniciam-se tentativas de automação a partir dos trabalhos de Luhn (1957[32] e 1958[33]), que propõe que as análises temáticas sejam realizadas com base no vocabulário do documento, sugerindo uma solução para os problemas de subjetividade na indexação ao eliminar o fator humano da equação.

Segundo Gil Leiva (1999)[34], as propostas de automação iniciaram-se com a aplicação de métodos estatísticos e de frequência , passando pela utilização de análises linguísticas no processo, e chegando então no uso de métodos mistos ou híbridos, que mesclam ambos os métodos estatísticos e linguísticos (Gil Leiva, 2008[35]; Guimarães, 2000[36]; Vieira, 1988[37]).

Lapa e Corrêa (2014)[38] definem a indexação automatizada como conjunto de operações linguísticas, computacionais e matemáticas que realizam o processamento de conteúdos textuais para a seleção, extração e atribuição de termos representativos

---

[30] PINHEIRO. L. V. R. Medidas de consistência indexação: interconsistência. **Ciência da Informação**, v.7, n.2, p. 109-114, 1978.
[31] SUOMINEN, O. Annif: DIY Automated Subject Indexing Using Multiple Algorithms. **Liber Quarterly**, v. 29, 2019.
[32] LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. **IBM Journal of Research and Development**, v.1, n.4, p. 309-317, 1957.
[33] LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v.2, n.2, p. 159-165, 1958.
[34] GIL LEIVA, I. **La automatización de la indización de documentos**. Gijón: Trea, 1999.
[35] GIL LEIVA, I. **Manual de indización**: teoría y práctica. Gijón: Trea, 2008.
[36] GUIMARÃES, J. A. C. **Indexação em um contexto de novas tecnologias**. 2000. 10 p. Texto didático.
[37] VIEIRA, S. B. Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação. **Revista de Biblioteconomia**, Brasília, v.16, n.1, p. 83-94, 1988.
[38] LAPA, R.; CORRÊA, R. F. Indexação Automática no Âmbito da Ciência da Informação no Brasil. **Informação & Tecnologia** (ITEC). Marília/João Pessoa, v. 2, n. 1, p. 1-18, 2014.

ao documento, possibilitando “[...] identificar e selecionar os termos que representam o conteúdo dos documentos, sem a intervenção direta do indexador”, de acordo com Robredo (2005, p. 170[39]).

Nesse sentido, Bandim e Corrêa (2019)[40] caracterizam três tipos distintos de indexação realizada por computadores, a indexação automática por extração, na qual os termos são extraídos diretamente do vocabulário e conteúdo do documento, a indexação automática por atribuição, que os termos representativos de um documento são atribuídos através dos termos de uma linguagem documentária, e a indexação semiautomática, que tem seus resultados validados pelo profissional indexador.

Embora a indexação automatizada tenha surgido como possibilidade de transpor as limitações da atividade manual, também é preciso lidar com suas próprias limitações, como a insuficiência dos sistemas de indexação e a sua inequivalência ao processo manual pela impossibilidade de replicação dos processos mentais realizados durante a indexação humana, destacado por Moreiro González (2004)[41].

Porém, Lancaster (2004) salienta que, por mais que os sistemas de indexação automatizada ainda não consigam alcançar o mesmo nível de qualidade da indexação realizada por indexadores profissionais, eles ainda podem ser empregados como forma de reduzir o tempo e os recursos necessários para a realização manual da indexação, reduzindo os gastos e a carga de trabalho e realizando uma indexação preliminar que poderá ser validada pelo indexador profissional.

Dessa forma, compreendendo as limitações também presentes nos sistemas de indexação automatizada, principalmente no que diz respeito à compreensão e interpretação de significados implícitos aos dados textuais, as técnicas de processamento da linguagem natural são uma oportunidade de aprimoramento no processamento documental, particularmente na etapa de análise de assunto.

Nesse sentido, os estudos da linguagem surgem como marcos teóricos na ciência da informação capazes de apoiar a compreensão de aspectos como a

---

[39] ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. e ampl. Brasília DF: Edição de autor, 2005.

[40] BANDIM, M. A. S.; CORRÊA, R. F. A. Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação. **Transinformação**, v. 31, e180004, 2019.

[41] MOREIRO GONZÁLEZ, J. A. **El contenido de los documentos textuales**: su análisis y representación mediante el lenguaje natural. Gijón: Trea, 2004.

produção, representação e uso dos signos na comunicação, possibilitando o conhecer como condição indissociável tanto do informar quanto do representar.

Tendo em vista a presença substancial da linguagem nos estudos de organização e tratamento temático da informação, e considerando o processo de automação das atividades de representação, nasce mais uma possível relação interdisciplinar: a automação de atividades de organização e tratamento temático da informação através de técnicas e sistemas de processamento da linguagem natural.

Nos sistemas de informação as tarefas processamento da linguagem natural são técnicas de grande valor, como salientam Vallez e Pedraza-Jimenez (2007)[42], e podem contribuir diretamente nas atividades de tratamento e indexação de documentos e dados textuais, no processamento das expressões de busca formuladas em linguagem natural e na recuperação de informações e documentos a partir de representações documentárias.

Considerando que a "[...] indexação é a base para a recuperação de documentos que são relevantes para a necessidade do usuário", como ressaltam Lewis e Jones (1996, p. 2[43], tradução nossa), entende-se que o desempenho e eficiência dos sistemas de informação estão diretamente relacionados à qualidade da indexação dos documentos que compõem esses sistema.

Assim, o processamento da linguagem natural acaba ganhando ainda mais significância no processo de organização da informação por se tornar uma possibilidade de aprimoramento das atividades de tratamento temático da informação.

Assim, Moens (2000)[44] destaca as etapas que são realizadas no processo de indexação utilizando estas aplicações, sendo elas: 1) análise morfológicas de unidades lexicais; 2) desconsideração de unidades funcionais; 3) normalização; 4) análise sintática de sintagmas e expressões; 5) análise semântica para tradução; 6) atribuição de valor aos descritores.

De certa maneira, independente das abordagens ou métodos escolhidas para a indexação realizada por processamento da linguagem natural, a atividade de

---

[42] VALLEZ, M.; JIMENEZ-PEDRAZA, R. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. **Hipertext.net**, 2007.
[43] LEWIS, D. D.; JONES, K. S. Natural language processing for information retrieval. **Communications of the ACM**, v. 39, n. 1, p. 92-101, 1996.
[44] MOENS, M. F. **Automatic indexing and abstracting of document texts**. Norwell: Kluwer Academic Publishers, 2000.

avaliação destas aplicações acabam por dificultar a elaboração de técnicas de avaliação geral destes sistemas e de seu desempenho na indexação.

Assim, a presente pesquisa encontra-se contextualizada entre as ponderações sobre os laços interdisciplinares entre a informação, linguagem, representação e automatização dos processos representativos da informação, evidenciando a necessidade de critérios de qualidade de sistemas de processamento da linguagem natural para sua aplicação na atividade de indexação.

Ainda se considera que o estabelecimento de parâmetros para a avaliação e escolha de *softwares* para a automação da indexação podem contribuir para a preservação das relações de sentido entre o texto e suas representações documentárias (Araújo; Santos, 2021[45]).

O delineamento aqui traçado traz à luz a seguinte questão: Tendo como base a literatura, quais os critérios de avaliação de sistemas de processamento da linguagem natural que podem ser aplicados e contribuírem para a avaliação e seleção de sistemas para a indexação automática?

Em busca de responder o questionamento formulado, a presente pesquisa teve como objetivo propor critérios de avaliação de sistemas de processamento da linguagem natural para a seleção de sistemas para a automação da indexação, de acordo com a literatura.

Para isso, foi realizada pesquisa bibliográfica para que, com base na bibliografia já existente sobre determinado assunto, seja possível definir e resolver problemas já definidos, além de ser possível descobrir e reconhecer novas questões e aspectos ainda não cristalizados.

Esta etapa teve como fonte de informação as bases de dados Scopus, que conta com referenciais com resumos, BRAPCI (Base de Dados em Ciência da Informação), com produções brasileiras sobre ciência da informação, e Dialnet, com referenciais com resumos e textos completos.

Assim, buscou-se pelas seguintes palavras-chave em português, inglês e espanhol: “indexação”, “indexação automática”, “indexação automatizada”, “automatização da indexação”, “avaliação da indexação”, “qualidade da indexação”, “critérios de indexação”, “processamento da linguagem natural”, “PLN”, “avaliação do

[45] ARAÚJO, V. J.; SANTOS, C. A. C. M. dos. Elaboração do resumo documentário: qualidade, parâmetros e normas. **InCID**: Revista de Ciência da Informação e Documentação, v.12, n.1, p. 60-81, 2021.

processamento da linguagem natural", "qualidade do processamento da linguagem natural" e "critérios de processamento da linguagem natural".

Os textos resultantes do levantamento bibliográfico foram avaliados e selecionados a partir do título, palavras-chave e resumo, considerando principalmente: 1) estudos disponíveis em acesso aberto; 2) estudos que citam necessariamente algum critério de qualidade para indexação e/ou processamento da linguagem natural.

Como resultado destes procedimentos, foi possível compreender que a complexidade da atividade indexadora impede a definição de diretrizes verdadeiramente úteis ao processo de tratamento documental. Mas, mesmo com a impossibilidade da definição de normas e regras para a qualidade e objetividade do processo de indexação, surgem princípios gerais e orientadores para a realização da indexação, como os delineados por Mendes e Simões (2001)[46].

O primeiro princípio de Mendes e Simões (2001) entende que a qualidade da indexação depende, indissociavelmente, da análise de assunto e sua fidelidade ao conteúdo e pertinência informacional em relação ao utilizador final, garantindo que as buscas informacionais no sistema não tenham como resultados documentos não pertinentes - ruído documental -, e que documentos pertinentes não escapem dos resultados - silêncio documental.

O segundo princípio define que a qualidade da indexação também depende da pertinência dos conceitos selecionados para a realidade do utilizador do sistema e da informação, isto é, o usuário.

Outro princípio é o de simplicidade formal, o qual estipula que, sempre que possível, os termos de indexação devem ser o mais simples possível, ainda que termos compostos sejam úteis e necessários para a representação conceitual, o que relaciona este ao princípio de pertinência dos conceitos.

No quarto princípio, de coerência e uniformidade, determina a coerência na aplicação dos princípios para a indexação e a uniformidade dos critérios de escolha para a solução de casos particulares - uniformidade intrínseca -, e a uniformidade nas escolhas de termos em casos de sinonímia - uniformidade extrínseca.

---

[46] MENDES, M. T. P.; SIMÕES, M. da G. Indexação por assuntos: princípios gerais e normas. **Páginas a&b**, n.8, p. 7-74, 2001.

Seguindo a linha do princípio de coerência e uniformidade, o princípio de controle da sinonímia estabelece que, para a representação de um mesmo conceito, deve ser selecionado um único termo de indexação e, através de equivalências permitir o acesso através de sinônimos não descritores.

O princípio da analogia diz respeito também ao princípio de coerência e uniformidade, determinando que, no caso de dúvidas quanto à aplicação dos princípios, deve-se buscar uma solução análoga em casos parecidos, aplicando o princípio de coerência e uniformidade.

Por fim, o princípio de controle da ambiguidade aborda os cuidados necessários para evitar a ambiguidade em sistemas de informação, principalmente em sistemas pós-coordenados, no qual é necessário atenção com termos de indexação demasiadamente genéricos, polissemia e termos homógrafos.

Também se destaca o trabalho de Moreiro González (2002)[47], que se preocupa com a padronização da qualidade dos produtos informacionais em sistemas de informação.

Para isso, define critérios que compreendem as diferentes fases de construção de produtos informacionais como indicadores de qualidade para avaliação, sendo eles a competências, atenção e interesse do profissional, as características das fontes utilizadas, os métodos, procedimentos, tempo gasto e custos da análise, o produto obtido e sua adequação aos objetivos documentais, o esforço necessário ao usuário, e as formas de apresentação.

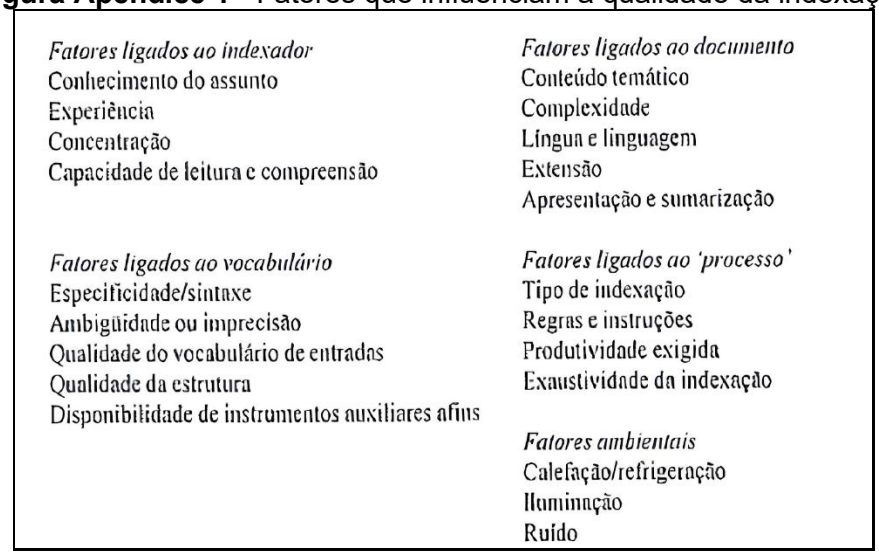
Moreiro González (2002, p. 54, tradução nossa) considera como indicadores de qualidade critérios que medem "[...] a coerência, a pertinência ou precisão, a exaustividade ou resposta, a consistência, a densidade informativa, a profundidade, a extensão ou tamanho, assim como os indicadores temporais, de custos, de esforço do usuário e de erros".

Por fim, Lancaster (2004), ao trabalhar com a teoria e prática da indexação, lista fatores que influenciam a qualidade da indexação, partindo dos fatores relacionados ao profissional indexador, passando pelos fatores relacionados ao vocabulário, fatores relacionados ao documento, fatores relacionados ao processo da indexação, alcançando até os fatores relacionados ao ambiente.

---

[47] MOREIRO GONZÁLEZ, J. A. Criterios e indicadores para evaluar la calidad del análisis documental de contenido. **Ciência da Informação**, v.31, n.1, 2002.

Figura Apêndice 1 - Fatores que influenciam a qualidade da indexação



Fatores ligados ao indexador
Conhecimento do assunto
Experiência
Concentração
Capacidade de leitura e compreensão

Fatores ligados ao vocabulário
Especificidade/sintaxe
Ambigüidade ou imprecisão
Qualidade do vocabulário de entradas
Qualidade da estrutura
Disponibilidade de instrumentos auxiliares afins

Fatores ligados ao documento
Conteúdo temático
Complexidade
Língua e linguagem
Extensão
Apresentação e sumarização

Fatores ligados ao 'processo'
Tipo de indexação
Regras e instruções
Produtividade exigida
Exaustividade da indexação

Fatores ambientais
Calefação/refrigeração
Iluminação
Ruído

Fonte: Lancaster (2004, p. 89).

Entre os fatores relacionados ao indexador, o autor destaca a importância de que o profissional possua conhecimentos acerca do conteúdo temático, assim como conhecimentos sobre os interesses e necessidades dos usuários finais do produto informacional, experiência como indexador profissional, capacidade de concentrar-se no processo, e competências de leitura, interpretação e compreensão (Lancaster, 2004, p. 88-90).

Os fatores relacionados ao vocabulário, segundo Lancaster (2004, p. 90), compreendem a clareza da linguagem do autor, questões gramaticais e sintáticas, ocorrências de ambiguidade, variação ou imprecisão linguística, a forma de apresentação do texto e sua estrutura, além da qualidade e disponibilidade de instrumentos auxiliares, como os vocabulários controlados.

Já os fatores relacionados ao documento incluem o conteúdo temático e sua complexidade, a correspondência entre conteúdo temático e interesse ou conhecimentos do indexador, a extensão e exaustividade do texto, questões ligadas à língua e linguagem do documento, além de questões de apresentação e sumarização do conteúdo, como menciona o autor (2004, p. 90).

Nos fatores relacionados ao processo de indexação listam-se o tipo de indexação definida pela instituição ou sistema, seja ela mais profunda ou exaustiva, as regras e normas definidas pela política de indexação da instituição, e a produtividade exigida ao profissional indexador (Lancaster, 2004, p. 91).

Por fim, os fatores relacionados ao ambiente dizem respeito ao contexto físico em que o profissional se encontra no momento da indexação, como é o caso das

condições de iluminação, climatização, espaço de trabalho, sons e ruídos, como indicado por Lancaster (2004, p. 91).

Assim, apesar das adversidades que dificultam a delimitação de regras e normas de qualidade da indexação, é possível identificar autores que buscam reconhecer aspectos que influenciam a atividade indexadora e, com base nisso, propor diretrizes para garantir a qualidade deste processo, como os princípios gerais de Mendes e Simões (2001), os critérios e indicadores de Moreiro González (2002), e os fatores de influência de Lancaster (2004).

Histórica e internacionalmente, a avaliação de métodos e modelos de processamento da linguagem natural torna-se objeto de estudo na década de 1980 no Programa de Fala e Linguagem Natural financiado pela ARPA (*Advanced Research Projects Agency)* e DARPA (*Defense Advanced Research Projects Agency*), e em 1993 com o trabalho do grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*), baseado no padrão ISO/IEC 9126, como forma de aprimorar os métodos de avaliação e estabelecer padrões para os produtos de engenharia de linguagem, como explicam Balkan, Arnold e Meijer (1994[48]).

Destacam-se as seguintes motivações que justificam a relevância da avaliação de sistemas de processamento de linguagem natural: 1) podem estimar comportamentos do sistema e possíveis desvios; 2) é de grande relevância para o design dos sistemas; 3) possibilita o diagnóstico de problemas e falhas do sistema durante sua aplicação; 4) amplia os conhecimento sobre projetos de PLN e outras aplicações computacionais da linguagem (Guida; Mauri, 1986[49]).

As avaliações das aplicações do processamento da linguagem natural se diferem de acordo com o contexto e objetivo da avaliação: a avaliação de diagnóstico busca identificar deficiências na aplicação, enquanto que a avaliação de progresso faz comparações entre os estágios de desenvolvimento de um sistema, e a avaliação de adequação tem como objetivo de determinar em que medida o sistema atende os requisitos de uma aplicação específica de acordo com as necessidades dos usuários (Balkan; Arnold; Meijer, 1994).

---

[48] BALKAN, L.; ARNOLD, D.; MEIJER, S. Test suites for natural language processing. *In*: **TRANSLATING AND THE COMPUTER**, 16, 1993, Londres. *Proceedings* [...]. Londres: Aslib, 1994. p. 51-58.

[49] GUIDA, G.; MAURI, G. Evaluation of natural language processing systems: issues and approaches. **Proceedings of the IEEE**, v.74, n.7, 1986.

Independentemente dos métodos ou técnicas escolhidos, uma boa avaliação de sistemas de PLN deve concentrar-se nas seguintes propriedades, segundo Wang *et al.* (2019, p. 4)[50]: dados variados, representativos, objetivos e confiáveis para os testes; abrangência no testes de propriedades para maior eficácia; boa correlação entre as tarefas específicas e o pré-processamento da linguagem natural; eficiência no desempenho e resolução de problemas computacionais específicos; resultados com variação e significância estatística para comparação.

O padrão de avaliação estabelecido na ISO/IEC 9126:1991 define seis características de qualidade definidas para "[...] capturar os principais componentes que constituem a qualidade de um software", de acordo com King e Maegaard (1998, p. 226[51], tradução nossa), sendo elas: funcionalidade, confiabilidade, usabilidade, eficiência, manutenibilidade e portabilidade.

A definição destas características tem como finalidade servir como uma estrutura básica para avaliações que podem se desenvolver especificamente, de acordo com a avaliação desejada e a aplicação do sistema, a partir de sub-características, decompondo e detalhando cada uma das características de modo a alcançar uma ampla gama de sistemas e aplicações a serem avaliados.

Com a necessidade de atualizações, a ISO/IEC 9126[52] foi substituída pelo modelo de qualidade ISO/IEC 25010, que evoluiu para a atualização ISO/IEC 25010:2023[53], oferecendo

> [...] uma estrutura detalhada para avaliar a qualidade de softwares e sistemas [...] define um modelo de qualidade de produto que categoriza características essenciais e suas sub características, fundamentais para medir a capacidade de um sistema atender a requisitos funcionais,

[50] WANG, B.; WANG, A.; CHEN, F.; WANG, Y.; KUO, C. C. J. Evaluating Word Embedding Models: Methods and Experimental Results. **APSIPA Trans. Signal Inf. Process**., v.8, 2019.
[51] KING, M.; MAEGAARD, B. Issues in natural language systems evaluation. *In*: **First International Conference on Language Resources & Evaluation** (LREC), 1, 1998, Granada. Proceedings [...]. Espanha: European Language Resources Association, 1998. p. 225-230.
[52] ISO/IEC 9126. **ISO/IEC 9126:1991**: Software Engineering - Product Quality Intl Standard Quality Model. 1991.
[53] ISO/IEC 25010. **ISO/IEC 25010:2023**: Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — System and Software Quality Models. 2023.

> de desempenho e específicos do usuário (Ariningsih; Muhammad, 2024, p. 88[54], tradução nossa).

Este novo padrão apresenta nove características de qualidade de sistemas:

- Adequação funcional: capacidade do sistema de realizar funções que atendam requisitos específicos do usuário pré-definidas, garantindo a disponibilidade dos recursos necessários, resultados precisos e eficiência;
- Eficiência de desempenho: capacidade do sistema de desempenhar suas funções proporcionalmente ao consumo de recursos, como o tempo de resposta, uso dos recursos computacionais e capacidade de execução;
- Compatibilidade: capacidade do sistema de se “comunicar” e trocar informações com outros sistemas, sua operação em ambientes compartilhados e interoperabilidade;
- Capacidade de interação: experiência e usabilidade do usuário na interface do sistema, a facilidade de entendimento, prevenção de erros, operabilidade geral e satisfação do usuário;
- Confiabilidade: capacidade do sistema de realizar suas funções sem falhas, sua tolerância a falhas, restauração após falhas e tempo de resposta;
- Segurança: capacidade do sistema de proteger recursos e dados contra ameaças, como a privacidade dos dados, integridade das informações, comprovação de ações, verificação de identidade e registro de atividades;
- Manutenibilidade: facilidade do sistema de ser modificado, como a modularidade, reaproveitamento de recursos, modificabilidade e testabilidade;
- Flexibilidade: capacidade do sistema de adaptação a novos requisitos e ambientes, ajustes de novas condições, expansão e implementação;
- Segurança operacional: capacidade do sistema de prevenir riscos contra propriedades, ambientes e pessoas, como alertas de perigo, identificação de riscos e minimização de acidentes.

De modo conclusivo, Gatt e Krahmer (2018, p. 134[55]), ao estudarem a avaliação em sistemas de processamento e geração de linguagem natural, recomendam o uso

---

[54] ARININGSIH, P.; MUHAMMAD, A. H. Quality Evaluation of Ticketing Management System Using ISO/IEC25010:2023 Standards and AHP Method. **Intechno Journal**: Information Technology Journal, v.6, n.2, 2024.

[55] GATT, A.; KRAHMER, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research** 61, 2018.

de múltiplos métodos e técnicas de avaliação, favorecendo a diversidade da avaliação e a correlação entre resultados de diferentes avaliações tendo em vista a complexidade da avaliação destes sistemas dentro das múltiplas possibilidades de aplicação.

Considerando a subjetividade do processo de indexação, a definição de critérios de qualidade e de avaliação torna-se uma tarefa de difícil execução, partindo de uma diversidade de pontos de vista e focando em diferentes aspectos do processo da análise de assunto.

Mendes e Simões (2001) partem da consideração das dificuldades intrínsecas ao processo de indexação e concebem sete princípios fundamentais tendo em vista a busca pela normalização e uniformidade na análise de assunto e representação temática na indexação.

Já Moreiro González (2002) entende que os indicadores de qualidade da indexação devem considerar e, de certa maneira, medir a coerência, precisão, consistência, profundidade, exaustividade e eficiência como proposta para garantir a recuperação da informação frutífera e adequada.

Por fim, Lancaster (2004) não define critérios de qualidade para a indexação, mas lista fatores de influência na qualidade do processo, considerando sua complexidade e subjetividade além do texto, levando em conta elementos relacionados ao contexto físico, profissional e linguístico da análise de assuntos.

De todo modo, independente dos aspectos focados em cada trabalho ou a nomenclatura escolhida para os critérios - princípios, indicadores ou fatores -, todos os autores buscam delimitar elementos concretos como forma de tornar objetivo, mesmo que em partes, um processo tão subjetivo quanto a indexação.

Uma importante convergência entre os critérios a abordagens de Mendes e Simões (2001), Moreiro González (2002) e Lancaster (2004), é a **adequação do processo de indexação** às necessidades dos usuários finais do sistema de informação, expressa pelo princípio de finalidade de conteúdo e pertinência informacional em Mendes e Simões (2001), o indicador de adequação do produto obtido aos objetivos documentais em Moreiro González (2002), e o conhecimento das necessidades dos usuários pelo indexador em Lancaster (2004).

Outra similaridade é o enfoque dado à **uniformização da indexação**, destaca por Mendes e Simões (2001) a partir do princípio de consistência e uniformidade, por Moreiro González (2002) ao incluir os indicadores de coerência e consistência, e por

Lancaster (2004) que aborda a consistência do trabalho do profissional entre os fatores de influência na atividade indexadora.

Ainda na linha de consistência da indexação, todos os critérios trabalhados destacam a **atuação profissional do indexador** durante o processo, principalmente no que diz respeito à qualidade da análise de assunto.

Embora Mendes e Simões (2001) não citem diretamente a atuação do indexador, tocam nesta temática intrinsecamente em outros princípios. Já Moreiro González (2002) destaca entre seus indicadores as competências, atenção e interesse do profissional durante o processo, enquanto Lancaster (2004) possui uma categoria dedicada aos fatores que influenciam a indexação que se relacionam ao profissional indexador, englobando sua experiência e habilidades.

O aspecto da **linguagem na indexação** também é um ponto em comum dentre os trabalhos, sendo retratado nos princípios de controle de sinonímia e controle de ambiguidade propostos por Mendes e Simões (2001), nos indicadores de forma de apresentação, de coerência e de precisão trabalhados por Moreiro González (2002), e nos fatores relacionados ao vocabulário, e até mesmo alguns relacionados ao documento, definidos por Lancaster (2004).

Por fim, as convergências se estendem até aos **usuários finais da indexação** e seus produtos, sendo explorado por Mendes e Simões (2001), embora que de maneira inerente, no princípio de simplicidade formal, e mais explicitamente no princípio de pertinência informacional, sem esquecer do indicador de esforço necessário ao usuário de Moreiro González (2002), e, embora não seja um elemento claro no trabalho de Lancaster (2004), se manifesta nos fatores relacionados ao indexador, como o conhecimento do usuário e suas necessidades.

Quadro Apêndice 1 - Síntese das convergências entre critérios, indicadores e fatores dos autores

CONVERGÊNCIAS TEÓRICAS	AUTORES	CRITÉRIOS, INDICADORES E FATORES
Adequação da indexação	Mendes e Simões (2001)	Princípio de finalidade de conteúdo e pertinência informacional
	Moreiro González (2002)	Indicador de adequação do produto obtido aos objetivos documentais
	Lancaster (2004)	Fatores relacionados ao indexador - conhecimento das necessidades dos usuários

Uniformização da indexação	Mendes e Simões (2001)	Princípio de consistência e uniformidade
	Moreiro González (2002)	Indicadores de coerência e consistência
	Lancaster (2004)	Fatores relacionados ao indexador - consistência do trabalho do profissional
Atuação profissional na indexação	Mendes e Simões (2001)	Intrínseco aos princípios
	Moreiro González (2002)	Indicadores de competências, atenção e interesse do profissional
	Lancaster (2004)	Fatores relacionados ao indexador
Linguagem na indexação	Mendes e Simões (2001)	Princípios de controle de sinonímia e controle de ambiguidade
	Moreiro González (2002)	Indicadores de forma de apresentação, de coerência e de precisão
	Lancaster (2004)	Fatores relacionados ao vocabulário e ao documento
Usuários finais da indexação	Mendes e Simões (2001)	Princípios de simplicidade formal e pertinência informacional
	Moreiro González (2002)	Indicador de esforço necessário ao usuário
	Lancaster (2004)	Fatores relacionados ao indexador - conhecimento das necessidades dos usuários

Fonte: Elaborado pela autora (2025).

Pensando no **foco dos critérios**, os indicadores de Moreiro González (2002) e os fatores de Lancaster (2004) partem de um enfoque mais prático de mensuração da eficiência e dos resultados, enquanto os princípios propostos por Mendes e Simões (2001) partem de uma perspectiva representacional e conceitual da qualidade do processo de indexação, destacando também as características linguísticas desse processo.

Pensando em **mensuração de qualidade**, o trabalho de Moreiro González (2002) se destaca na definição de critérios próprios para a mensuração quanti-qualitativa da indexação, enquanto Mendes e Simões (2001) e Lancaster (2004) se preocupam mais com a qualidade intelectual da indexação, de modo mais abstrato em Mendes e Simões, e mais extrínsecos em Lancaster.

O **conceito de qualidade** dos autores também divergem, sendo mais abstrato e restrito à qualidade intelectual e linguística em Mendes e Simões (2001), mais ampla, integrando fatores externos ao processo em Lancaster (2004), e até mesmo um meio termo entre a ideia de qualidade intelectual focada no processo e a qualidade prática do produto, como sugere Moreiro González (2002).

Finalmente, as **considerações sobre o usuário**, embora seja um ponto convergente entre os critérios, varia de acordo com a relevância a ele atribuído, de maneira mais intrínseca em Lancaster (2004) e Mendes e Simões (2001), que abordam as questões relacionadas aos usuários de maneira indireta, e de maneira mais explícita e direta nos indicadores de esforço necessário ao usuário e de adequação do produto final aos objetivos documentais de Moreiro González (2002).

Quadro Apêndice 2 - Síntese das divergências entre critérios, indicadores e fatores dos autores

DIVERGÊNCIAS TEÓRICAS	AUTORES	DETALHAMENTO
Foco dos critérios	Mendes e Simões (2001)	Enfoque conceitual e representacional do processo
	Moreiro González (2002)	Enfoque prático na mensuração da eficiência e dos resultados
	Lancaster (2004)	Enfoque prático na identificação e categorização de elementos externos
Mensuração de qualidade	Mendes e Simões (2001)	Mensuração da qualidade intelectual de modo mais abstrato
	Moreiro González (2002)	Mensuração quanti-qualitativa do processo e produtos
	Lancaster (2004)	Mensuração da qualidade intelectual a partir de elementos externos
Conceito de qualidade	Mendes e Simões (2001)	Mais abstrata e restrita à qualidade intelectual e linguística
	Moreiro González (2002)	Meio termo entre a ideia de qualidade intelectual focada no processo e a qualidade prática do produto
	Lancaster (2004)	Mais ampla integrando fatores externos ao processo intelectual
Consideração sobre o usuário	Mendes e Simões (2001)	Mais intrínseco e indireto como o princípio de pertinência informacional

	Moreiro González (2002)	Mais explícito e direto nos indicadores de esforço necessário ao usuário e de adequação do produto final aos objetivos documentais
	Lancaster (2004)	Mais intrínseco e indireto como o conhecimento do indexador sobre o usuário

Fonte: Elaborado pela autora (2025).

No tocante às relações entre os critérios de qualidade de indexação dos autores supracitados - Mendes e Simões (2001), Moreiro González (2002) e Lancaster (2004) - e os critérios de qualidade de *softwares* definidos na ISO/IEC 25010:2023, também podem ser identificadas convergências fundamentais de qualidade.

Principiando pelas similaridades, ambos os grupos de critérios têm como fundamento a **relevância da adequação do processo aos objetivos finais** como parte essencial da avaliação de qualidade, independente do objeto de análise.

Na ISO/IEC 25010:2023 enfatiza-se o parâmetro de adequação funcional na mensuração da capacidade do sistema de executar funções que atendam as necessidades de seus usuários, e nos critérios de indexação, tanto Mendes e Simões (2001), Moreiro González (2002) e Lancaster (2004) abordam esta questão a partir do princípio de fidelidade ao conteúdo e pertinência informacional, os indicadores de produto obtido e sua adequação aos objetivos documentais, e o conhecimento dos usuários pelo indexador, respectivamente.

Também é possível constatar os conceitos de **confiabilidade, coerência e/ou consistência**, na ISO/IEC 25010:2023 a partir do parâmetro de confiabilidade para determinar a capacidade do sistema de desempenhar suas funções de maneira consistente e sem falhas, em Mendes e Simões (2001) com o princípio de coerência e uniformidade garantindo a consistência da indexação, em Moreiro González (2002) com os indicadores de coerência, de consistência e de uniformidade da representação de assunto, e mais implicitamente em Lancaster (2004) nos fatores ligados ao processo e na atuação do profissional indexador.

A **interação ou usabilidade pelo usuário** também é uma questão abordada de maneira semelhante nos critérios analisados partindo do parâmetro de capacidade de interação da ISO/IEC 25010:2023, tendo em vista a experiência do usuário na utilização de sistemas, seguindo com o indicador de esforço necessário ao usuário e interação determinado por Moreiro González (2002).

Neste quesito, tanto Mendes e Simões (2001) quanto Lancaster (2004) não definem princípios ou fatores próprios a questão de usabilidade e interação, porém este ponto não passa despercebido aos outros elementos abordados pelos autores, sendo mencionado o princípio de interesse do utilizador, ainda que não seja diretamente relacionado à usabilidade.

Finalizando as convergências, cita-se a **adaptabilidade**, trabalhada no parâmetro de manutenibilidade da ISO/IEC 25010:2023 ao abordar a capacidade de adaptação eficiente de *softwares* a novos requisitos e necessidades, e os fatores relacionados ao processo, definidos por Lancaster (2004), que atribui relevância aos diferentes tipos e níveis de indexação para as diferentes necessidades de representação.

Mendes e Simões (2001) e Moreiro González (2002) não fazem referência direta à adaptabilidade ou manutenibilidade, mas seus princípios e indicadores são de natureza geral, com flexibilidade suficiente para serem adaptados e aplicados na indexação de diferentes tipos de documentos e assuntos.

Quadro Apêndice 3 - Síntese das convergências entre critérios de qualidade de indexação e critérios de qualidade de *softwares*

CONVERGÊNCIAS TEÓRICAS	TIPO	AUTORES	PARÂMETROS, CRITÉRIOS, INDICADORES E FATORES
Relevância da adequação do processo aos objetivos finais	Qualidade de *softwares*	ISO/IEC 25010:2023	Parâmetro de adequação funcional
	Qualidade da indexação	Mendes e Simões (2001)	Princípio de fidelidade ao conteúdo e pertinência informacional
		Moreiro González (2002)	Indicadores de produto obtido e sua adequação aos objetivos documentais e de pertinência ou precisão
		Lancaster (2004)	Fatores relacionados ao indexador - conhecimento das necessidades dos usuários
Confiabilidade, coerência e/ou consistência	Qualidade de *softwares*	ISO/IEC 25010:2023	Parâmetro de confiabilidade
	Qualidade da indexação	Mendes e Simões (2001)	Princípio de coerência e uniformidade

		Moreiro González (2002)	Indicadores de coerência, de consistência e de uniformidade
		Lancaster (2004)	Implícito aos fatores relacionados ao processo e ao indexador
Interação ou usabilidade pelo usuário	Qualidade de *softwares*	ISO/IEC 25010:2023	Parâmetro de capacidade de interação
	Qualidade da indexação	Mendes e Simões (2001)	Indicadores de esforço necessário ao usuário e interação
		Moreiro González (2002)	Implícito ao princípio de interesse do utilizador
		Lancaster (2004)	Implícito aos fatores relacionados ao indexador - conhecimento das necessidades dos usuários
Adaptabilidade	Qualidade de *softwares*	ISO/IEC 25010:2023	Parâmetro de manutenibilidade
	Qualidade da indexação	Mendes e Simões (2001)	Princípios de natureza geral e flexíveis
		Moreiro González (2002)	Indicadores de natureza geral e flexíveis
		Lancaster (2004)	Fatores relacionados ao processo - diferentes tipos e níveis de indexação

Fonte: Elaborado pela autora (2025).

Dando início a discussão das divergências entre os critérios, salienta-se sobretudo a **natureza dos objetos de avaliação**, com Mendes e Simões (2001), Moreiro González (2002) e Lancaster (2004) analisando o processo de representação de assunto da indexação e a qualidade de seus produtos, enquanto os critérios definidos na ISO/IEC 25010:2023 estabelecem um padrão de qualidade de produtos com foco em sistemas e aplicações computacionais.

Seguindo esta perspectiva, a **ênfase das análises** também muda, concentrando-se no fator humano da indexação, que configura o ponto mais subjetivo

do processo, sendo inerente aos princípios de Mendes e Simões (2001), e mais explícita no indicador de competências, atenção e interesse do profissional de Moreiro González (2002) e nos fatores relacionados ao indexador de Lancaster (2004), enquanto que a ISO/IEC 25010:2023 tem como foco nas características e funcionalidades dos *softwares* como produtos a serem adquiridos.

Outro fator é a **objetividade** dos critérios na avaliação, com métricas mais concisas e quantitativas na avaliação de sistemas pela ISO/IEC 25010:2023, embora também considere questões subjetivas, ao passo que os critérios de indexação analisam aspectos de difícil mensuração e dependente de elementos extrínsecos aos processo em si, como as competências de profissionais, questões relativas à linguagem e a dependência de regras e normas institucionais.

O quesito **linguagem e controle terminológico** também configura a diferença entre os conjuntos de critérios, sendo mais proeminentes os aspectos relacionados à linguagem na qualidade da indexação, como nos princípios de controle da sinonímia, analogia e controle da ambiguidade de Mendes e Simões (2001), os indicadores de forma de apresentação, profundidade e precisão de Moreiro González (2002) e os fatores relacionados ao vocabulário de Lancaster (2004), enquanto que este fator não é abordado diretamente na ISO/IEC 25010:2023.

Quadro Apêndice 4 - Síntese das convergências entre critérios de qualidade de indexação e critérios de qualidade de *softwares*

DIVERGÊNCIAS TEÓRICAS	TIPO	AUTORES	DETALHAMENTO
Natureza dos objetos de avaliação	Qualidade de *softwares*	ISO/IEC 25010:2023	Estabelece um padrão de qualidade de produtos com foco em sistemas
	Qualidade da indexação	Mendes e Simões (2001)	Analisam o processo de representação de assunto da indexação e a qualidade de seus produtos
		Moreiro González (2002)	
		Lancaster (2004)	
Ênfase das análises	Qualidade de *softwares*	ISO/IEC 25010:2023	Concentra-se nas características e funcionalidades dos softwares como produtos

	Qualidade da indexação	Mendes e Simões (2001)	Concentra-se no fator humano de maneira mais implícita
		Moreiro González (2002)	Concentra-se no fator humano - indicador de competências, atenção e interesse do profissional
		Lancaster (2004)	Concentra-se no fator humano - fatores relacionados ao indexador
Objetividade	Qualidade de *softwares*	ISO/IEC 25010:2023	Métricas mais concisas e quantitativas
	Qualidade da indexação	Mendes e Simões (2001)	Aspectos de difícil mensuração e dependente de elementos subjetivos e extrínsecos aos processo
		Moreiro González (2002)	
		Lancaster (2004)	
Linguagem e controle terminológico	Qualidade de *softwares*	ISO/IEC 25010:2023	Não é abordado diretamente
	Qualidade da indexação	Mendes e Simões (2001)	Mais presente nos critérios de indexação - princípios de controle da sinonímia, analogia e controle da ambiguidade
		Moreiro González (2002)	Mais presente nos critérios de indexação - indicadores de forma de apresentação, profundidade e precisão
		Lancaster (2004)	Mais presente nos critérios de indexação - fatores relacionados ao vocabulário

Fonte: Elaborado pela autora (2025).

A análise de convergências e divergências teóricas entre os critérios de qualidade definidos pela ISO/IEC 25010:2023 para a avaliação de *softwares* nos permite observar as possibilidades de sua aplicação para a avaliação de qualidade no processo de indexação, principalmente por certas proximidades teóricas observadas previamente.

Em um primeiro momento, entre os critérios destacados pela norma ISO/IEC 25010:2023, a avaliação da **adequação funcional** mostra-se a mais relevante para a avaliação de *softwares* de processamento da linguagem natural para a indexação,

identificando a capacidade do sistema de executar de maneira eficiente funções determinantes para a representação temática, como a seleção de informações representativas ao assunto, como a aplicação de linguagens documentárias.

Outro ponto que sobressai é a análise de **eficiência de desempenho** de *softwares*, sendo a eficiência um dos pontos principais a favor da automatização da indexação. Seguindo este princípio, observa-se o tempo necessário e a capacidade do *software* para o processamento de assunto e os recursos necessários para a determinação de eficiência do sistema na indexação.

A **capacidade de interação** entre sistema e usuários também é um fator de grande importância para a determinação da qualidade de um *software*, principalmente para processos que demandam a supervisão humana e validação dos resultados, como a indexação.

Questões como a capacidade de aprendizado e a operabilidade do sistema influencia diretamente na utilização do sistema pelo profissional, possibilitando o uso intuitivo e diminuindo o esforço necessário ao indexador para isso, além da importância da prevenção contra possíveis erros humanos durante o processo de análise de assunto.

Tendo em vista o caráter complexo e subjetivo do processo de análise de assunto, o critério da ISO/IEC 25010:2023 de **confiabilidade** assegura a robustez e consistência na execução do *software*, garantindo a integridade, fidelidade e qualidade da representação temática do produto final, avaliando a tolerância do sistema às possíveis falhas operacionais e sua capacidade de restaurar os dados caso essas falhas ocorram.

Finalizando, os critérios de **flexibilidade** e **manutenibilidade** dos *softwares* avaliados na ISO/IEC 25010:2023 levam em consideração as possibilidades de adaptação do sistema para novos requisitos e necessidades do usuário.

A indexação, por estar estritamente vinculado às linguagens documentárias e vocabulários controlados, que passam por constantes modificações de modo a se manterem relevantes, necessita que os sistemas de automatização possam acompanhar as atualizações do processo, sempre em busca de aperfeiçoar seu processo e produtos.

Quadro Apêndice 5 - Síntese das aplicações dos critérios da ISO/IEC 25010:2023 na avaliação de *softwares* para a automatização indexação

CRITÉRIOS	APLICAÇÃO NA INDEXAÇÃO

Adequação funcional	Determinar a capacidade do sistema de executar de maneira eficiente funções determinantes para a indexação
Eficiência de desempenho	Determinar a eficiência da automatização da indexação a partir da capacidade, tempo e recursos necessários
Capacidade de interação	Determinar a operacionalidade pelo indexador na supervisão do processo de indexação e validação dos resultados
Confiabilidade	Assegurar a robustez e consistência na execução e garantir a integridade, fidelidade e qualidade do produto final
Flexibilidade e Manutenibilidade	Assegurar atualizações seguindo as modificações dos instrumentos de indexação e aperfeiçoar o processo e produto

Fonte: Elaborado pela autora (2025).

Em suma, em meio a uma gama de convergências, e mesmo com as divergências, é possível conceber uma possibilidade de vínculos e aplicações metodológicas da norma ISO/IEC 25010:2023 na avaliação e seleção de sistemas para a automatização da indexação.

Embora o processo de avaliação da indexação passe por desafios no delineamento de princípios e regras de qualidade, é possível destacar os trabalhos de Mendes e Simões (2001), Moreiro González (2002) e Lancaster (2004) que definiram princípios gerais, critérios e indicadores, e fatores para o direcionamento da indexação em meio a toda sua subjetividade.

Paralelamente, a avaliação de sistemas de processamento da linguagem natural emerge com a finalidade de demonstrações de desempenho e funcionalidades para o destaque entre outros produtos da mesma natureza, dando origem aos parâmetros de avaliação do padrão ISO/IEC 25010:2023.

Pensando na utilização dos critérios da ISO/IEC 25010 na avaliação e seleção de sistemas para a automação da indexação, destacam-se as convergências e divergências entre os critérios de qualidade da indexação e os critérios de qualidade dos sistemas.

Nas convergências tem-se a relevância da adequação do processo aos objetivos finais, a confiabilidade, coerência e/ou consistência, a interação ou usabilidade pelo usuário, e a adaptabilidade dos critérios. Já nas divergências tem-se a natureza dos objetos de avaliação, a ênfase das análises, a objetividade e a linguagem e controle terminológico dos critérios.

Por fim, fundamentando-se nas convergências e divergências entre estes critérios, salienta-se a possibilidade de aplicação dos critérios de **adequação funcional**, **eficiência de desempenho**, **capacidade de interação**, **confiabilidade**, e **flexibilidade e manutenibilidade**, definidos na ISO/IEC 25010, no processo de avaliação e seleção de sistemas para a automação da indexação.

É possível concluir que, além da natureza e objetivos distintos entre os critérios para qualidade da indexação e os critérios para qualidade de *softwares* de processamento da linguagem natural, é possível identificar convergências fundamentais que permitem sua confluência no que se refere à avaliação e seleção de sistemas para a automação da indexação via processamento da linguagem natural, fornecendo novas bases teóricas-metodológicas capazes de contribuir com a eficiência, desempenho e qualidade no desenvolvimento de produtos documentários.