UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS SOCIAIS
CURSO DE BIBLIOTECONOMIA

JOÃO MATHEUS NASCIMENTO RODRIGUES

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS COMO FERRAMENTA DE REPRESENTAÇÃO E ORGANIZAÇÃO DA INFORMAÇÃO

São Luís
2020

JOÃO MATHEUS NASCIMENTO RODRIGUES

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS COMO FERRAMENTA DE REPRESENTAÇÃO E ORGANIZAÇÃO DA INFORMAÇÃO

Monografia apresentada ao Curso de Biblioteconomia da Universidade Federal do Maranhão, como requisito para obtenção do grau de Bacharel em Biblioteconomia.

Orientadora: Profa. Dra. Valdirene Pereira da Conceição

São Luís

2020

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo (a) autor (a).

Núcleo Integrado de Bibliotecas/UFMA

Rodrigues, João Matheus Nascimento.
Sumarização automática de textos como ferramenta de representação e organização da informação / João Matheus Nascimento Rodrigues. - 2020.
108 f.

Orientador(a): Valdirene Pereira da Conceição.
Monografia (Graduação) - Curso de Biblioteconomia, Universidade Federal do Maranhão, São Luís, 2020.

1. Organização do conhecimento - PLN. 2. Processamento automático de línguas naturais. 3. Sumarização automática de textos. 4. Sumarização automática - Indexação. I. Conceição, Valdirene Pereira da. II. Título.

JOÃO MATHEUS NASCIMENTO RODRIGUES

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS COMO FERRAMENTA DE REPRESENTAÇÃO E ORGANIZAÇÃO DA INFORMAÇÃO

Monografia apresentada ao Curso de Biblioteconomia da Universidade Federal do Maranhão, como requisito para obtenção do grau de Bacharel em Biblioteconomia.

Aprovado em 30/11/2020

BANCA EXAMINADORA

_____

Profa. Dra. Valdirene Pereira da Conceição (Orientadora)

Doutora em Linguística e Língua Portuguesa

Universidade Federal do Maranhão

_____

Profa. Dra. Cenidalva Miranda de Sousa Teixeira

Doutora em Engenharia Elétrica

Universidade Federal do Maranhão

_____

Profa. Me. Juliana Rabelo do Carmo

Mestre em Ciência da Informação

Universidade Federal de Santa Catarina

Aos meus Avós. Elos de toda uma geração em um laço forte de amor e carinho. Em especial à Maria Madalena Silva (*in memorian*).

AGRADECIMENTOS

Ao Deus eterno e imortal, invisível, mas real, porque d'Ele, por meio d'Ele e para Ele são todas as coisas.

Aos meus pais, Antonio Carlos e Ana Cláudia, meu irmão Alysson Klynsmann e a minha família, por todo apoio, alicerce e incentivo.

À minha orientadora, Valdirene Pereira da Conceição, que durante o último ano muito me ensinou, por toda atenção, paciência, conhecimentos compartilhados e pela amizade e instrução.

Às professoras que integram a banca examinadora, Cenidalva Miranda de Sousa Teixeira e Juliana Rabelo do Carmo, pelo aceite no convite, críticas construtivas e contribuições oportunas que permitiram o aprimoramento do estudo.

Aos professores do Departamento de Biblioteconomia da UFMA, pelos conhecimentos adquiridos nas disciplinas cursadas.

À Karolina Costa Cavalcante, pelo suporte constante, amizade, companheirismo e paciência.

Ao Projeto de Extensão Formação Discente Pesquisador, em nome da Profa. Claudia Pecegueiro, pelos aprendizados e conhecidos adquiridos na condição de monitor voluntário.

À amizade que se fortaleceu durante a participação no Projeto de Extensão Formação Discente Pesquisador, em especial à Iraceles Cardoso Luzo e Taynara de Sousa Mendes, parceria firme e boas risadas. Obrigado, T.P.D.

Aos amigos que fiz durante o curso de Biblioteconomia, em especial à Danyelle Lobo, Leandro Lima, Isabele Rodrigues, Jefferson Cruz, Jacira Soares, Francymar Abreu e todos os outros, cujo apoio foi fundamental para a conclusão desta graduação.

À Universidade Federal do Maranhão, pelas oportunidades, acolhimento e por ter proporcionado uma excelente formação acadêmica.

Às instituições por qual estive como bolsista e estagiário, e que muito contribuíram para minha formação pessoal, profissional e acadêmica, em especial ao Departamento de Pós Graduação da AGEUFMA, em nome de Luciana Soares Santos, ao Centro Universitário UNDB, em nome de Adriana Cabral e ao Tribunal Regional do Trabalho da 16ª Região, em nome de Raimunda Nonata Teixeira e Mary Rose Viana Machado.

E a todos que, ao longo desses anos de curso, contribuíram direta ou indiretamente, para minha formação profissional, acadêmica e pessoal. Obrigado!

A tecnologia avança e é preciso acompanhá-la,
ou usufruir de seus benefícios.
Mey e Silveira, 2009, p. 77.

RESUMO

Estudo sobre a Organização da Informação e do Conhecimento. Trata da aplicação de ferramentas de Sumarização Automática de Textos no auxílio e desenvolvimento de atividades de representação temática da informação, em especial a Indexação. Apresenta os fundamentos teóricos-metodológicos referentes às áreas de Organização da Informação e do Conhecimento, Processamento Automático de Línguas Naturais e Sumarização Automática de Textos. Expõe as bases teóricas da organização e representação da informação e do conhecimento, com o intuito de compreender a evolução e aplicação dos instrumentos e ferramentas que auxiliam na representação temática do conhecimento notadamente os oriundos de outras áreas do saber humano, levando em consideração o caráter interdisciplinar do estudo. Entende que a Organização da Informação e do Conhecimento, passa por mudanças significativas em busca de adaptar-se frente as novas demandas e evoluções advindas do desenvolvimento das sociedades. Objetiva com o estudo conhecer o processo de sumarização automática de textos e as ferramentas disponíveis para a sua realização, assim como, perceber as contribuições da sumarização automática de textos para a representação temática, organização e recuperação da informação. Adota como procedimentos metodológicos de investigação a pesquisa bibliográfica e documental, com abordagem de natureza aplicada, qualitativa e exploratória; o corpus de análise é constituído por cinco artigos científicos sobre Covid-19. Utiliza como ferramenta de sumarização automática de texto o Turbine Text e o Intellex Summarize NE, e como instrumento de avaliação dos resumos resultantes do processo, um Modelo de Avaliação Subjetiva Intrínseca sugerida pela DUC (*Document Understanding Conference*). Apresenta como resultados indícios de viabilidade da sumarização automática como ferramenta de apoio no processamento de grandes volumes de informações textuais digitais. Revela também a necessidade de estudos e pesquisas na área de Biblioteconomia e Ciência da Informação sobre a abordagem teórica e aplicada de ferramentas baseadas em PLN; além do desenvolvimento de propostas metodológicas mais sólidas e a concepção de softwares específicos para a realização de tarefas de SA, que potencializem a organização da informação, em especial no ambiente web.

Palavras-chave: Sumarização automática de textos. Sumarização automática - Indexação. Processamento automático de línguas naturais. Organização do conhecimento - PLN.

ABSTRACT

Study on Information and Knowledge Organization. It deals with the application of Automatic Text Summarization tools in the assistance and development of activities of thematic representation of information, especially Indexation. It presents the theoretical-methodological fundamentals concerning the areas of Information and Knowledge Organization, Automatic Natural Language Processing and Automatic Text Summarization. It exposes the theoretical foundations of the organization and representation of information and knowledge, in order to understand the evolution and application of instruments and tools that help in the thematic representation of knowledge, especially those from other areas of human knowledge, taking into account the interdisciplinary nature of the study. It understands that the Organization of Information and Knowledge, goes through significant changes in search of adapting itself to the new demands and evolutions coming from the development of societies. It aims with the study to know the process of automatic summarization of texts and the tools available for its realization, as well as to perceive the contributions of the automatic summarization of texts for the thematic representation, organization and recovery of information. It adopts as methodological procedures of investigation the bibliographical and documental research, with applied, qualitative and exploratory approach; the corpus of analysis consists of five scientific articles on Covid-19. It uses as an automatic text summarization tool the Turbine Text and the Intellex Sumar, and as an instrument of evaluation of the summaries resulting from the process, an Intrinsic Subjective Evaluation Model suggested by DUC (Document Understanding Conference). The results show signs of the feasibility of automatic summarization as a support tool in the processing of large volumes of digital textual information. It also reveals the need for studies and research in the area of Librarianship and Information Science on the theoretical and applied approach of tools based on PLN; in addition to the development of more solid methodological proposals and the design of specific software for the performance of SA tasks, which enhance the organization of information, especially in the web environment.

Keywords: Automatic text summarization. Automatic summary - Indexing. Automatic processing of natural languages. Knowledge organization - PLN.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AACR2	Anglo American Catalogue Rules 2 ed.
CC	Ciência da Computação
CI	Ciência da Informação
DC	Dublin Core
IA	Inteligência Artificial
LC	Library of Congress
MARC	Machine Readable Catalogue
NILC	Núcleo Interinstitucional de Linguística Computacional
OC	Organização do Conhecimento
OI	Organização da Informação
OIC	Organização da Informação e do Conhecimento
PLN	Processamento Automático de Línguas Naturais
RDA	Resource: Description and Access
SA	Sumarização Automática de Textos
SSA	Sistema de Sumarização Automática
SGB	Sistema de Gerenciamento de Bibliotecas
SPLN	Sistemas de Processamento Automático de Línguas Naturais
SRI	Sistema de Recuperação da Informação
TA	Tradução Automática
TDI	Tratamento Descritivo da Informação
TIC	Tecnologias de Informação e Comunicação
TTI	Tratamento Temático da Informação

SUMÁRIO

1 INTRODUÇÃO

O mundo está profundamente marcado por mudanças de ordem informacional e tecnológica. As Tecnologias de Informação e Comunicação – TICs mudaram e estão mudando de forma significativa o modo de fazer e viver do homem. A cada dia surgem novas tecnologias, ferramentas e maneiras de realizar processos e atividades, frutos da vasta produção informacional e da expansão em larga escala do mundo digital.

Mudanças estas que se fazem presentes em todas as esferas da sociedade, pois todos estamos inseridos de forma direta ou indireta nesse grande processo de avanço tecnológico e de globalização que o mundo atualmente tem passado, embora ainda seja comum, que parcelas da sociedade vivam à margem de todo esse processo.

Assim, a tamanha produção informacional aliada à expansão em larga escala do mundo digital, provocam uma produção desenfreada de informações nos dias atuais, o que gera cada vez mais a necessidade de instrumentos, ferramentas e métodos que garantam o tratamento dessas informações para seu posterior acesso e uso.

Instrumentos e ferramentas que, desde sua concepção, até os dias atuais, foram evoluindo para dar conta do número cada vez mais crescente de informações, assim como, da necessidade cada vez mais ágil por parte dos usuários, de acesso a essas informações disponíveis nos acervos de instituições.

A título de exemplo, temos os códigos de catalogação e classificação, que tiveram e desempenham um papel muito importante no tratamento da informação. E são utilizados em processos de tratamento e organização, das informações disponíveis em acervos físicos de bibliotecas, arquivos e museus, para possibilitar a organização, recuperação e acesso da informação.

Processos que, com a introdução do computador em meados dos anos 1950, sofreram significativas mudanças e evoluções, além de serem potencializados, com o advento das TICs e o emprego de softwares e hardwares, no sentido de dar suporte no desenvolvimento e realização de atividades de representação, organização e recuperação da informação, não apenas em ambientes tradicionais ou físicos, mas também agora em ambientes digitais. Assim, tornou-se um grande desafio conseguir acompanhar esse novo momento e desenvolver instrumentos que pudessem tratar essas informações em formato tradicional e ao mesmo tempo, em formato digital.

Brascher e Café (2008, p. 5), definem organização da informação como "[...] um processo que envolve a descrição física e de conteúdo dos objetos informacionais." com o

intuito de “[...] arranjá-los sistematicamente em coleções, nesse caso, temos a organização da informação em bibliotecas, museus, arquivos, tanto tradicionais quando eletrônicos.” (BRASCHER; CAFÉ, 2008, p. 6).

Assim, é possível conceber a organização da informação, e a necessidade de novos métodos e técnicas de processamento de dados/informação que auxiliem e acompanhem os avanços tecnológicos para permitir o acesso e uso às informações disponíveis nos diversos tipos de suportes.

Nesse sentido, tendo como pressuposto a interdisciplinaridade presente na CI, estudos que envolvem aplicações de técnicas e métodos de outros domínios do conhecimento tem sido cada vez mais comum. Citamos o Processamento Automático de Línguas Naturais – PLN, vertente da Inteligência Artificial, que se constitui a partir de contribuições teórico-metodológicas das Ciências da Computação, da Ciência da Informação, Linguística e outras áreas do conhecimento, e que estuda a possibilidade de computadores simular à compreensão e o entendimento humano no que se refere ao tratamento da língua natural.

Os estudos acerca do PLN tiveram seu início no campo da Ciência da Computação e Linguística Computacional, e logo depois se estenderam para outros domínios como a Ciência da Informação, resultando assim, em “[...] um campo de estudos bastante heterogêneo e fragmentado, acumulando uma vasta literatura e agregando pesquisadores das mais variadas especialidades, com formação acadêmica, embasamento teórico e interesses também bastante diversos.” (DIAS DA SILVA et al. 2007, p. 5).

Em Ciência da Informação, tem sido estudado na perspectiva teórica e prática, em especial no campo da “[...] Indexação e Recuperação da Informação, pela possibilidade de softwares baseados nesse modelo propiciarem a extração de termos com maior precisão semântica para recuperação da informação em sistemas de buscas automatizados [...]” (CARMO; CONCEIÇÃO, 2018, p. 317), potencializando as tarefas automatizadas com o emprego do PLN.

Entretanto, além dos sistemas baseados em perspectivas de extração e mineração de dados, existem também sistemas baseados em processos que permitem a condensação de textos, com a finalidade de possibilitar a compreensão do conteúdo mais importante existente no texto (PARDO, 2008).

Existe uma diversidade de usos e métodos que estão sendo desenvolvidos nos estudos em PLN, para utilização dessa ferramenta em vários campos do conhecimento e com suas aplicações em diversas tarefas e processos comuns ao homem.

Assim, no que se refere à perspectiva deste trabalho, podemos destacar que a Sumarização Automática de Textos (SA) "trata da produção automática de sumários a partir de um ou mais textos fontes" tendo em vista que em SA, "[...] sumário pode tanto se referir a índice quanto a resumo propriamente dito." (PARDO, 2008, p. 2), assim, ponderamos que a concepção adotada nesta pesquisa é a de sumários se referirem especialmente a resumos, levando em consideração a utilização do resumo como instrumento de representação, divulgação e recuperação da informação, tendo em vista a sua característica de sintetizar conteúdos e/ou informações. Assim, sempre que utilizarmos os termos sumarização e/ou sumários, estaremos nos referindo à concepção de resumo.

Nesse sentido, uma atividade inerente ao trabalho dos bibliotecários é a indexação de documentos, que é parte integrante do Tratamento Temático da Informação. Guinchat e Menou (1994, p. 175), destacam que a indexação "[...] é uma das formas de descrição de conteúdo. É a operação pela qual escolhe-se os termos mais apropriados para descrever o conteúdo dos documentos [...]", descrição que poderá possibilitar a recuperação da informação disponíveis nos documentos a partir dos termos de indexação.

Desse modo, a atividade de indexação, em geral, é composta por duas etapas: análise conceitual e tradução. A primeira implica em decidir qual o assunto de determinado documento, enquanto a segunda, envolve a conversão da análise conceitual de um documento num determinado conjunto de termos de indexação. (LANCASTER, 2004).

Por conseguinte, a elaboração de resumos dos documentos que estão sendo indexados, é uma tarefa importante na determinação do conteúdo temático destes, tendo como pressuposto que o "principal objetivo do resumo é indicar de que trata o documento ou sintetizar seu conteúdo." (LANCASTER, 2004, p. 6).

Acerca desta aplicação dos resumos no processo de indexação, Lancaster (2004, p. 6) afirma que "A indexação de assuntos e a redação de resumos são atividades intimamente relacionadas, pois ambas implicam a preparação de uma representação do conteúdo temático dos documentos.".

O avanço das pesquisas em tecnologias de informação e comunicação, assim como descobertas e desenvolvimento de softwares que auxiliam o processamento de dados, além dos estudos em inteligência artificial e processamento automático de línguas naturais, fazem surgir novas formas de realizar atividades manuais com o auxílio de sistemas operacionais.

Faz-se necessário portanto, a existência e aplicação de métodos e recursos que possibilitem a organização das informações disponíveis, para o seu acesso e uso, por parte daqueles que lhe interessam, os usuários. Considerando esses argumentos surge o desejo de

compreender o funcionamento de sistemas baseados em processamento automático de línguas naturais, e a contribuição destes para a Organização da Informação e do Conhecimento em processo de representação temática da informação, notadamente na indexação de assuntos. Nesse contexto, surge como problema de pesquisa, os seguintes questionamentos:

a) Como ocorre o processo de sumarização automática de textos e quais as ferramentas disponíveis para a sua realização?

b) Quais as contribuições da sumarização automática de textos para a representação temática da informação, bem como para a organização e recuperação da informação?

Assim, a motivação para realização desta pesquisa, se deu a partir de inquietações e interesses de ordem pessoal, profissional e acadêmica do autor, acerca de ferramentas baseadas em PLN, e a partir de pesquisas e leituras realizadas sobre a temática de Sumarização Automática de Textos e a viabilidade do emprego dessa ferramenta no processo de indexação de assuntos, a partir da produção de resumos com aplicações de SA.

Logo, estabelecemos como objetivo geral, conhecer o processo de sumarização automática de textos e as ferramentas disponíveis para a sua realização, assim como, perceber as contribuições da sumarização automática de textos para a representação temática, organização e recuperação da informação.

E por objetivos específicos:

a) identificar as principais aplicações de SA e demonstrar aplicações;

b) apresentar os usos do PLN na Ciência da Informação e na Organização do Conhecimento;

c) compreender os fundamentos teóricos e metodológicos da Sumarização Automática e do PLN;

d) analisar a qualidade dos sumários gerados automaticamente, com o intuito de compreender o grau de contribuição para identificação do conteúdo temático da informação.

Nesse sentido, destacam-se os autores que sustentaram a realização da pesquisa, dando suporte teórico para as reflexões e discussões suscitadas. No âmbito da Biblioteconomia e Ciência da Informação, nas argumentações voltadas para a Organização da Informação e do Conhecimento, autores como Mey (2003); Mey e Silveira (2009); Brascher e Café (2008); Dias e Naves (2007); Guichat e Menou (1994); Araújo (2017); Fugita, Rubi e Boccato (2009); Lancaster (2004); Café e Sales (2010), Alvarenga (2003) e outros. E no âmbito do PLN e da SA, a pesquisa sustenta-se a partir dos estudos de Dias da Silva (2006); Dias da Silva *et al*

(2007); Sousa (2015); Nunes (2008); Rosa (2011); Vieira e Lopes (2010); Pardo (2008); Souza *et al* (2017); Ribeiro (2016); Simonassi (2015); Cardoso (2014); Rino e Pardo (2008) e outros.

Afim de permitir uma melhor visualização da pesquisa em torno da questão problema e dos objetivos apresentados, o presente trabalho encontra-se estruturado em sete seções.

A primeira seção, apresenta as demarcações iniciais que implicam a realização deste estudo, assim como, anuncia a questão problema de pesquisa, objetivos, geral e específicos, justificativa e as possíveis contribuições que o estudo oferecerá para área.

Os limites, desafios e tendências da Organização da Informação, são tratados na segunda seção deste estudo, e objetiva fundamentar a pesquisa na literatura da área. É a partir do referencial teórico realizado neste momento, que as discussões levantadas nas seções seguintes se sustentam. Assim, o referencial teórico contempla os conhecimentos relacionados a organização da informação e do conhecimento, discorrendo sobre aspectos históricos e conceituais, técnicas e estratégias empregadas na organização da informação, bem como, questões de interdisciplinaridade na Ciência da Informação.

O Processamento Automático de Línguas Naturais na Organização do Conhecimento, é tratado na terceira seção, e possui por finalidade, apresentar os pressupostos teóricos e metodológicos do PLN, assim como, caracterizá-lo enquanto domínio do conhecimento, e expor os usos e aplicações no domínio da Ciência da Informação, resgatando inclusive, estudos e pesquisas na CI sobre PLN.

O percurso metodológico adotado para a realização da pesquisa é apresentado na quarta seção, etapa em que se descrevem os instrumentos utilizados, bem como esquematiza em forma de fluxograma os procedimentos.

A Sumarização Automática de Textos na Representação da Informação, é apresentada na quinta seção, evidenciando os sistemas de sumarização automática, as bases, tipos de sumários e aplicações.

Na sexta seção, é apresentada o processo da sumarização automática de textos, e seu uso na representação temática da informação, assim como, análise e discussão dos resultados obtidos.

Por fim, a sétima seção do estudo, trata-se da conclusão da pesquisa. Nesta seção, são expostas as questões e limitações encontradas no desenvolvimento do estudo, e as indagações que surgiram no decorrer da pesquisa.

2 ORGANIZAÇÃO DA INFORMAÇÃO: limites, desafios e tendências

A informação nos dias atuais é considerada como grande insumo e fator estratégico determinante para o desenvolvimento e crescimento das sociedades. Aliada ao conhecimento e à inovação tecnológica, são considerados o motor do crescimento econômico. (MARTINS, 2014).

O início do século XXI é marcado profundamente por mudanças advindas das tecnologias digitais de comunicação, que propulsionaram o crescente número de publicações, caracterizada por intensa produção informacional. Nesse sentido, a concepção do que seja a informação passa por mudanças, e gera dificuldades no que tange ao tratamento adequado dessas informações para a premissa básica de uso.

Entretanto, a informação sempre esteve presente nas sociedades, e o início de sua democratização, "[...] atribui-se ao surgimento da escrita que possibilitou a preservação do conhecimento permitindo dessa forma a expansão cultural e científica [...]", e deve-se à invenção da Prensa de Gutemberg em meados do século XV, e o surgimento da Internet e consequentemente, o advento das Tecnologias de Informação e Comunicação, advindos do século XXI, têm transformado de forma significativa as possibilidades de acesso às fontes de informação e conhecimento. (MARTINS, 2014).

Trata-se de períodos que romperam paradigmas e que transformaram significativamente a maneira como a sociedade vê a informação. O primeiro período é marcado com a mudança quanto aos suportes primitivos que permitiam o registro da informação, desde as tábuas de argila ao papiro e pergaminho, até a invenção dos tipos móveis, que permitia a produção semi-industrial dos primeiros registros em forma de livro impresso em papel.

O segundo momento tem como marco o advento da internet e, consequentemente, as tecnologias de informação e comunicação em meios digitais, possibilitando novos suportes de registro para a veiculação de informações, como os livros digitais, as páginas na web, as bases e bancos de dados em formato eletrônico e outros suportes que estão surgindo e outros que surgirão.

Levando em consideração as relações estabelecidas entre informação, conhecimento e comunicação, que sempre estiveram presentes nas sociedades, essas primeiras aproximações são necessárias para que o entendimento sobre o impacto da informação fique claro. Necessárias também para o entendimento acerca de que a Organização da Informação (OI) teve de adaptar-se e preparar-se para as proposições de novas soluções que fossem

satisfatórias às mudanças de suportes pelos quais a escrita incidiu e têm incidido. (AGANETTE; TEIXEIRA; AGANETTE, 2017).

Logo, é fundamental entender o que é informação, e o que ela representa. Definir informação, diante de sua complexidade e infinitas possibilidades conceituais e etimológicas, assim como diversos outros aspectos que servem de suporte e insumo para as pesquisas desenvolvidas em Biblioteconomia e Ciência da Informação, tal qual em outras áreas do conhecimento, torna-se um processo que aflora inquietações e descobertas.

No contexto da Biblioteconomia e Ciência da Informação, destacamos alguns expoentes nessa temática, autores como Meadows (1991), Capurro e Hjorland (2007), Robredo (2003), Davenport (1998), Araújo (2001), Brascher e Café (2008), entre outros.

A abordagem mais clássica do que seja informação, pauta-se na Teoria Matemática da Informação, que é estabelecida a partir de uma analogia com a transmissão de sinais elétricos por meio de canais mecânicos de comunicação, onde o conceito de informação no contexto da Ciência da Informação é descrito por Claude Shannon e Waren Weaver, em meados da década de 1940, em que entendem como se dá o processo de comunicação com a transmissão de uma mensagem entre uma fonte (emissor) e um destino (receptor) utilizando de um canal para o fim desejado. (MEADOWS, 1991).

Já em uma perspectiva etimológica, informação para Araújo (2001, p. 1), vem de "[...] origem latina, do verbo '*informare*', que significa dar forma, colocar em forma, criar, representar, construir uma ideia ou uma noção.", Enquanto que ainda nessa perspectiva, Capurro e Hjorland (2007, p. 155), afirmam que a "[...] palavra informação tem raízes latinas (*informatio*).", e que deve ser considerada em dois contextos básicos em que o termo é usado, "[...] o ato de moldar a mente e o ato de comunicar conhecimento.".

Assim, informação é "dar forma", ação humana de "informar" e registrar, conceito que percorre todo esse processo, e tem sua origem na produção de registros informacionais e se prolonga nas atividades humanas sobre esses registros. E é, assim, objeto de estudo presente em várias áreas do conhecimento.

Brascher e Café (2008, p. 4), destacam ainda que para entender informação, é necessário "[...] englobar aspectos no nível semântico (cognitivo) e pragmático (real), incluindo assim as propriedades relativas tanto ao conteúdo e significado como sua função social", permitindo-nos o entendimento de que no primeiro aspecto, refere-se ao conteúdo propriamente dito, ao possibilitar a transformação das estruturas do conhecimento dos indivíduos. Enquanto que no segundo aspecto, o pragmático, relaciona-se ao problema ou questão que a informação deve satisfazer.

Culminando na função social da informação, no sentido de possibilitar a apropriação desta pelos indivíduos, o que propiciará uma mudança nas estruturas do conhecimento, e consequentemente na sociedade em que estão inseridos. Com o intuito de assegurar o exercício de uma cidadania ativa, consciente e participativa, com o uso da informação para o fim que o indivíduo desejar. (MARTINS, 2014).

Assim, "[...] a informação é vislumbrada como uma possibilidade de transformar estruturas do conhecimento, e, portanto, o conhecimento pode ser visto como algo provisório e em permanente revisão". (BRASCHER; CAFÉ, 2008, p. 4), entendida como um processo cognitivo, subjetivo, relativo e que necessita da interpretação do receptor.

Em se tratando das dificuldades conceituais acerca do termo informação, o conceito de conhecimento por vezes é atribuído à informação. Lima e Alvares (2012, p. 23), destacam que: "Conhecimento e informação são termos de difícil conceituação devido à amplitude semântica e às diversas perspectivas de análise, domínio e concepções de cada área [...]", como já problematizado na literatura, a exemplo de Rodrigues (2015), Davenport (1998) e outros.

Rodrigues (2015) destaca que, em muitos casos, o conceito de informação e conhecimento, ora aparecem de forma conjunta, ora se complementam e ora se opõem, além de que estarão "[...] invariavelmente, associados à ideia de comunicação, com a tríade emissor/receptor/mensagem e também com a ideia de cognição [...]", como destacam Robredo (2003) e Capurro e Hojorland (2007).

Enquanto que para Brascher e Café (2008, p. 3), para o entendimento dos termos informação e conhecimento é necessário: "a) relacionar seus conceitos às funções que damos a eles nos contextos em que se inserem; b) diferenciá-los de conceitos próximos a eles incluídos no sistema referencial", apontando assim, para a necessidade de inicialmente entender o contexto em que o termo é empregado, e consequentemente distingui-los dos conceitos que estão empregados por semelhanças.

Nesse sentido, entre informação e conhecimento, Davenport (1998) elenca ainda o conceito de dado, reconhecendo que na prática não é tão simples diferenciá-los, principalmente pelas proximidades dos contextos em que são empregados, assim, assinala, nesse caso, para discussão de Brascher e Café (2008), que discorrem sobre a necessidade de entender o contexto em que ambos são empregados.

O Quadro 1, a seguir, apresenta a distinção apresentada por Davenport (1988), entre os conceitos em discussão.

Quadro 1 – Dados, informação e conhecimento

DADOS	INFORMAÇÃO	CONHECIMENTO
Simples observações sobre o estado do mundo	Dados citados de relevância e propósito	Informação valiosa da mente humana
- Facilmente estruturado	- Requer unidade de análise	- Inclui reflexão, síntese, contexto
- Facilmente obtido por máquinas	- Exige consenso em relação ao significado	- De difícil estruturação
- Frequentemente quantificado	- Exige necessariamente a mediação humana	- De difícil captura em máquinas
- Facilmente transferível	-	- Frequentemente tácito
-	-	- De difícil transferência

Fonte: adaptado de Davenport (1998).

Dados, portanto, seriam observações sobre o estado do mundo, em sua condição bruta, pois informação, a partir da visão de Drucker, são dados dotados de relevância e propósito, logo, munidos de significados, conhecimento, passa a ser a informação em um contexto, imbuída de significado e interpretação. (DAVENPORT, 1998).

Partindo dessas exposições, que versam sobre informação e conhecimento no âmbito da Biblioteconomia e Ciência da Informação, fica evidente a distinção conceitual. Logo, é possível compreender que quando tecemos alusões sobre Organização da Informação e Organização do Conhecimento (OC), estamos falando de áreas de estudo com atuações distintas, conforme Rodrigues (2015, p. 13), ao destacar que "[...] para podermos falar de Organização da Informação e Organização do Conhecimento é preciso antes, retomar a diferenciação entre informação e conhecimento adotados na Ciência da Informação.".

Subsidiando os pensamentos expostos anteriormente, tendo em vista o que é informação, Brasher e Café (2008, p. 5, grifo das autoras), destacam que

> A organização da informação é, portanto, um processo que envolve a descrição física e de conteúdos dos objetos informacionais. O produto desse processo descritivo é a **representação da informação**, entendida como um conjunto de elementos descritivos que representam os atributos de um objeto informacional específico.

Entendendo a descrição física, conhecida popularmente como catalogação, ou ainda como representação descritiva, que consiste em um conjunto de informações que simbolizam e representam um registro do conhecimento. (MEY, 2003). "E pode utilizar-se de linguagens específicas, normas e formatos que padronizam esse tipo de descrição [...]", a exemplo do *Anglo-American Cataloguing Rules 2.ed.* (AACR2). (CAFÉ; SALES, 2010).

Desse modo, a descrição de conteúdo, que também é reconhecida como descrição temática da informação, possui como principal objeto a representação temática dos objetos

informacionais, e é realizada pelos processos de classificação, indexação e resumo. Ambas utilizadas na OI para possibilitar a recuperação da informação. (CAFÉ; SALES, 2010).

Como sustentam Café e Sales (2010, p. 118), "A organização da informação é um processo de arranjo de acervos tradicionais ou eletrônicos, realizado por meio da descrição física e de conteúdo (assunto) de seus objetos informacionais", processo que possui como objetivos a compreensão melhor do acervo e a recuperação dos recursos, objetos e conteúdos informacionais pelos usuários de determinada unidade de informação. (CAFÉ; SALES, 2010).

Em relação à Organização do Conhecimento, Brascher e Café (2008, p. 6) dizem que "[...] por sua vez, visa à construção de modelos de mundo que se constituem em abstrações da realidade". Já a Representação do Conhecimento, enquanto produto desta "[...] se constitui uma estrutura conceitual que representa modelos de mundo [...]", dos quais nos permitem descrever e fornecer explicações sobre fenômenos que observamos. (BRASCHER; CAFÉ, 2008, p. 6).

Lima e Alvares (2012, p. 27), ainda destacam que

> No sentido mais genérico do termo, organização do conhecimento é o modo como ele é disposto em assuntos em toda parte onde se deseja a sua sistematização ordenada para atingir determinado propósito. Pelo seu caráter interdisciplinar, a organização do conhecimento é estudada também em outras áreas, como antropologia, computação, filosofia, linguística, psicologia, sociologia, entre outras.

O esquema apresentado na Figura 1 demonstra de forma sucinta os objetos da Organização da Informação e Representação da Informação, assim como da Organização do Conhecimento e Representação do Conhecimento, com base nas exposições de Brascher e Café (2008).

Figura 1 – Processos e produtos da O.I e O.C

O.I e R.I > objetos = registros de informação (objetos físicos)

O.C e R.C > objetos = conceito > cognição > conhecimento

Fonte: adaptado de Brascher e Café (2008).

Logo, OI e OC diferem quanto aos seus objetos que, respectivamente são apontados como sendo o mundo dos objetos físicos e o mundo da cognição, cujo produto é o conhecimento. Assim sendo, seguiremos dando ênfase para os processos e produtos referentes à Organização da Informação, que se constitui como componente do objeto deste estudo.

O período histórico que caracteriza a Organização da Informação remonta ao processo histórico das bibliotecas, instituições que em sua gênese possuíam o objetivo de salvaguardar

o conhecimento produzido pelo homem, para permitir o acesso e uso, materializados em registros do conhecimento, que perpassaram historicamente desde a tábua de argila, ao pergaminho, papiro, códices, livros impressos, digitais e até as informações disponíveis em meios digitais como conhecemos nos dias atuais.

Logo, Araújo (2010), Café e Sales (2010), Mey (2003) e Mey e Silveira (2009), destacam os períodos históricos que retratam os percursos e avanços que decorrem aos processos e produtos da Organização da Informação, no que tange à descrição física e de conteúdos dos objetos informacionais.

Em se tratando aos aspectos físicos, desde a Antiguidade até a Renascença, incluindo a Idade Média, temos as primeiras tentativas de representar o acervo das bibliotecas em forma de catálogos, que na época muito se pareciam com o que conhecemos nos dias atuais, como as listas de inventários, iniciativas primitivas que continham informações básicas, porém necessárias para a recuperação dos registros bibliográficos. (MEY, 2003).

Quanto aos aspectos temáticos, temos as primeiras iniciativas que se referem ao tratamento temático da informação, tendo como pioneiras as primeiras classificações bibliográficas, em que se destaca a Classificação Decimal de Dewey, que surgiu no início do ano de 1876, e outros instrumentos como as linguagens documentárias, listas de cabeçalhos de assunto, vocabulários controlados, tesauros, e outros. (GUICHAT; MENOU, 1994).

Assim, no intuito de sustentar as exposições anteriores, no que diz respeito ao período histórico da organização da informação, o Quadro 2, abaixo, apresenta de uma forma mais detalhada essas informações:

Quadro 2 – Períodos Históricos da Organização da Informação

PERÍODOS HISTÓRICOS	CONTRIBUIÇÕES PARA A OI	INSTRUMENTOS DESENVOLVIDOS
Idade Antiga	Primeiras iniciativas de construção de catálogos.	**260 - 240 a.C** - *Pinakoi* – Catálogo de Calímaco; **669 - 626 a.C** - Tabletes de argila. (Classificados em Ciências da Terra, do Céu e outras informações bibliográficas)
Idade Média	Tentativas incipientes de organização de catálogos.	**822 - 842** - Biblioteca de Richenau (Alemanha); (compilação de catálogos). **831** - Mosteiro de Saint Requier (França); **1389** - Catálogo de Saint Martin (Inglaterra).
Idade Moderna	Primeiras iniciativas para a construção de um código de catalogação – inexistência de padronização e consolidação de princípios.	**1410 - 1412** - Catálogo de Amplonius Rating de Berka; **1498** - Catálogo de Aldo Manúcio; **14** - Bibliografia de Johann Tritheim; (Organizada em ordem cronológica, com índice alfabético de autor); **1545 - 1548** - Publicação da *Bibliotheca Universalis* e a *Pandectarum*, de Konrad Gesner (considerada a primeira classificação bibliográfica);

		1627 -1643 - Advis pour dresser une bibliothèque de Gabriel Naudé; **1780 - 1782** - Publicação da Encyclopédie por Diderot e d´Alembert.
Idade Contemporânea	Consolidação e desenvolvimento de padrões e princípios, assim como de instituições que tiveram um papel importante para a solidificação da área da organização da informação e do conhecimento.	**1810** - Criação do Sistema Brunet; **1839** - 91 regras de Panizzi; **1850** - Código de Munique (Alemanha); **1850** - Código de Catalogação de Charles C. Jewett; **1876** - Classificação Decimal de Dewey; **1891** - Expansive Classification (Classificação de Cutter); **1895** - Criação do Instituto Internacional de Bibliografia (IIB); **1899** - Início da criação da Classificação da Biblioteca do Congresso (EUA); **1905** - Classificação Decimal Universal; **1906** - Subject Classification (Classificação de Brown); **1908** - 1ª edição das Regras de catalogação: entradas de autores e títulos; **1912** - Bibliographic Classification (Classificação de Bliss); **1920** - Código da Vaticana **1927** - Criação da Federação Internacional de associações Bibliotecárias (IFLA); **1931** - Código da Vaticana; **1933** - Classificação de dois pontos (Classificação de Ranganathan); **1954** - Criação do Instituto Brasileiro de Bibliografia e Documentação (IBBD); **1961** - Conferência de Paris; **1960** - MARC; **1967** - Publicação do Código de Catalogação Anglo-Americano (AACR); **1969** - Reunião Internacional de Especialistas em Catalogação (RIEC); **1971** - International Standard Bibliographic Description - ISBD; **1978** - Publicação da AACR2; **1995** - Protocolo Z39.50; **1996** - Dublin Core; **1998** - Requisitos Funcionais para Registros Bibliográficos (FRBR); **1994** - MARC 21; **2010** - Desenvolvimento do Resource: Description and Acess (RDA); **2011** - Início dos estudos do Bibliographic Framework (BIBFRAME).

Fonte: adaptado de Alves (2010) e Sales (2019).

Percebemos que a busca por avanços nas ferramentas, processos e instrumentos empregados na Organização da Informação são evidentes. Leva-se consequentemente em consideração o papel e o valor que a informação possui para o desenvolvimento das sociedades, e a necessidade de soluções que acompanhem a evolução dos recursos

informacionais, e possibilite o emprego adequado para o fim que se deseja: a recuperação da informação.

Logo, nos dias atuais, muitas são as iniciativas que buscam dar continuidade no desenvolvimento desses processos, produtos e ferramentas/instrumentos voltados para a Organização da Informação, tendo em vista a expansão do espaço digital, o surgimento de novos suportes, recursos, objetos e conteúdos informacionais, além do crescente número de publicações. Avanços esses que estão intrinsecamente ligados às tecnologias de informação e comunicação, e que modificam a natureza da Organização e Representação da Informação. (KURAMOTO, 2006; ANNA, 2015).

Nessa perspectiva, o volume de informações disponibilizadas, tanto no formato tradicional quanto no digital, inviabiliza o seu tratamento por especialistas e, nesse sentido, surgem os aspectos ligados aos limites, desafios e tendências que fazem parte dos processos e produtos relacionados à organização da informação, que contemplam os aspectos de forma e de conteúdo. (KURAMOTO, 2006).

Limites, desafios e tendências que permeiam questões como o avanço desenfreado das tecnologias de informação e comunicação, assim como o número cada vez mais crescente de publicações em diversos suportes, recursos e objetos informacionais, a exemplo dos livros impressos e digitais, materiais audiovisuais, cartográficos, museológicos, materiais disponíveis em bases e bancos de dados, em repositórios temáticos e institucionais, páginas na *Web*, e afins que possuem o objetivo de veicular a informação.

Panorama de limites esses, que fazem suscitar desafios na O.I, a exemplo de questões relacionadas à concepção e desenvolvimento de ferramentas, processos e instrumentos atuais que alcancem tanto os recursos informacionais disponíveis em bibliotecas tradicionais, híbridas e digitais, além de recursos disponíveis em ambiente web, e que ao passo disso, acompanhem os novos suportes que estão surgindo, dando também ênfase ao caráter teórico e conceitual dessas ferramentas de organização da informação. (AGANETTE; TEIXEIRA; AGANETTE, 2017).

No que se refere aos desafios, estes refletem diretamente na formação e ensino de bibliotecários, como profissionais da informação. Podemos destacar, ainda, questões relacionadas a dificuldades no emprego de tecnologias da informação e comunicação em atividades de Organização e Representação da Informação; investimentos insuficientes que, em alguns casos, configuram-se em obstáculos referentes à produção lenta de estudos teóricos, que culminam e dificultam o crescimento da OI, enquanto campo científico e que

traz sérias consequências também para o desenvolvimento de atividades relacionadas à Recuperação da Informação.

Tendências e novas abordagens surgem em resposta aos limites e desafios que se apresentam em torno da Organização e Representação da Informação e, consequentemente os processos, produtos, ferramentas/instrumentos e abordagens existentes no âmbito da descrição física e temática da informação, principalmente nos dias atuais, em virtude das TIC disponíveis, em que discussões acerca da gestão de informações e conteúdos digitais se tornam frequentes.

Quanto ao tratamento descritivo, podemos citar como tendências os padrões de Metadados, a exemplo do formato MARC, padrão Dublin Core (DC), e outros formatos e padrões como o Mods e Mets. Metadados que se referem comumente à informação descritiva sobre recursos da Web, em que o formato MARC tem seu registro bibliográfico legível por máquina, enquanto o DC é um elemento básico de descrição de recursos em ambiente Web. (MEY; SILVEIRA, 2009; ALVES, 2010).

Os protocolos e padrões de intercâmbio, a exemplo do ISO 2709 e Z39.50 também se inserem nessa perspectiva, pois são cada vez mais importantes, tendo em vista menção feita anteriormente, quando se afirma a necessidade de ferramentas que possibilitem adequação a diferentes tipos de itens informacionais e que auxiliam no intercâmbio de dados e na interoperabilidade entre sistemas, levando em consideração o acelerado avanço das TIC. (AGANETTE; TEIXEIRA; AGANETTE, 2017; ALVES, 2010).

Atualmente, abordagens acerca da representação descritiva têm apresentado discussões a respeito da necessidade de agregar semântica à descrição bibliográfica, com o intuito de permitir um tratamento mais efetivo dos conteúdos digitais. Surgindo assim, pesquisas sobre as chamadas Linguagens de Representação, que é uma "[...] linguagem estruturada e padronizada, pois as mesmas permitem aos computadores uma melhor compreensão e interpretação semântica dos termos utilizados nesse processo." (MARTINS, 2018).

Nesse sentido, linguagens como a Standard *Generalized Markup Language* (SGML), *Hypertext Markup Language* (HTML), *eXtensible Markup Language* (XML), *Resource Description Framework* (RDF) e *Web Ontology Language* (OWL) têm tido destaque no Campo da Ciência da Informação e são consideradas as principais linguagens usadas nos processos de representação e criação de modelos de dados.

Ainda na perspectiva da descrição física, temos os formatos bibliográficos de representação, aos quais como MARC21, e o DC, expostos anteriormente, os quais também se configuram dentro desta categoria, e também o *Bibliographic Framework* ou BIBFRAME,

que “[...] fornece uma base para a descrição bibliográfica contemporânea na Web ao ser fundamentado na proposta de dados interligados”. (MARTINS, 2018, p. 72).

Importa destacar também o *Resource: Description and Acess* (RDA), o novo código de catalogação que visa substituir as AACR2, e que embora mantenha uma forte relação com o seu antecessor, a RDA difere muito, devido ser baseada numa estrutura teórica, ter sido projetada para o ambiente digital e seu escopo ser mais abrangente que o das AACR2. (AGANETTE; TEIXEIRA; AGANETTE, 2017).

E por fim, acentuamos como tendência relacionada à descrição física, os requisitos funcionais, a exemplo dos *Functional Requeriments for Bibliographic Records* (FRBR), que são modelos conceituais do tipo entidade-relacionamento, que se tornaram base conceitual utilizada para o aprimoramento de normas, regras e formatos relacionados ao tratamento descritivo da informação, direcionando o foco aos usuários. (MEY; SILVEIRA, 2009; AGANETTE; TEIXEIRA; AGANETTE, 2017).

Referindo-se à descrição temática da informação, as tendências versam desde as soluções automatizadas de indexação, que possuem expoentes desde a década de 90, abordadas de forma rudimentar por Guinchat e Menou (1994), e que com o avanço das TIC e da informática sofreram mudanças e avanços significativos.

Ressaltamos também nessa perspectiva o emprego de ferramentas de Processamento de Linguagem Natural (PLN), como a sumarização automática e a mineração e extração de termos. O PLN no âmbito da C.I tem sido estudado principalmente pela “[...] possibilidade de softwares baseados nesse modelo propiciarem a extração de termos com maior precisão semântica para a recuperação da informação em sistemas de busca automatizados.” (CARMO; CONCEIÇÃO, 2018, p. 317).

Araújo (2017, p. 16) aponta que em face da produção acelerada de informações e das necessidades dos usuários, surge a *Folksonomia,* apresenta uma forma alternativa de indexação social, pois é “[...] articulada a uma dinâmica descentralizada das ações de representação da informação.”, com o intuito de permitir a representação colaborativa da informação, realizada pelos próprios usuários em ambiente *Web*, com a finalidade de propiciar melhorias na recuperação da informação.

Dias e Naves (2007), apontam para o uso das ontologias, como instrumentos para a organização dos recursos eletrônicos em ambiente digital com base no conteúdo. À medida que o ambiente *Web* vai evoluindo, esses instrumentos vão se adaptando e estão em processo de desenvolvimento contínuo.

Neste sentido, é notável que os avanços impostos pelas tecnologias de informação e comunicação mudam de maneira circunstancial o entendimento do que seja a informação e dos suportes, recursos e objetos em que ela está disponível. Mudanças que culminam diretamente sobre as ferramentas, processos e instrumentos da Organização da Informação, e que tecem novas abordagens e tendências nas atividades de descrição, tratamento, depósito, disseminação e recuperação da informação.

Logo, a busca por soluções que atendem essas necessidades são cada vez mais necessárias. Assim, a subseção a seguir dará continuidade a respeito das técnicas e estratégias atuais utilizadas na descrição física e temática da informação.

2.1 Técnicas e estratégias utilizadas na organização da informação

A Organização da Informação faz uso de processos, produtos e instrumentos/ferramentas para a plena realização das atividades que são direcionadas para os fins a que se destina. Assim, técnicas e estratégias são empregadas no âmbito da OI para que, juntamente com as ferramentas, processos e instrumentos possibilitem as questões em torno do tratamento e representação da informação com vistas à recuperação.

Santos (2017, p. 23), é enfática ao esclarecer que "O tratamento da informação é uma grande área a qual engloba diversas atividades que têm por objetivo tratar e organizar a informação, a fim de disponibilizá-la de forma recuperável aos usuários de um sistema de recuperação da informação". Certificando a necessidade e a importância deste para possibilitar a disseminação, acesso e uso da informação.

Portanto, de maneira sintetizada, Dias e Naves (2007, p. 9), definem o Tratamento da Informação, como

> [...] expressão que engloba todas as disciplinas, técnicas, métodos e processos relativos a: a) descrição física e temática dos documentos numa biblioteca ou sistema de recuperação de informação; b) desenvolvimento de instrumentos (códigos, linguagens, normas, padrões) a serem utilizados nessas descrições; e c) concepção/implantação de estruturas físicas ou bases de dados destinadas ao armazenamento dos documentos e de seus simulacros (fichas, registros eletrônicos, etc.).

O Tratamento da Informação é subsidiado por atividades relacionadas ao Tratamento Descritivo da Informação (TDI) e o Tratamento Temático da Informação (TTI), conhecidos ainda por representação descritiva e representação temática. Compreendendo as atividades de catalogação propriamente dita, e indexação, classificação, bem como outras atividades delas derivadas, respectivamente.

Assim, o TDI cobre os aspectos mais objetivos capazes de serem identificados extrinsecamente, como, autor, o título, a editora, e elementos similares. (DIAS E NAVES, 2007; FUGITA; RUBI; BOCCATO, 2009). Nesse contexto, a catalogação ou a representação descritiva, é definida por Mey e Silveira (2009, p. 7) como:

> "O estudo, preparação e organização de mensagens, com base em registros do conhecimento, reais ou ciberespaciais, existentes ou passíveis de inclusão em um ou vários acervos, de forma a permitir a interseção entre as mensagens contidas nestes registros do conhecimento e as mensagens internas dos usuários."

Nessa concepção, as autoras concebem a catalogação como um conjunto de informações que simbolizam um registro do conhecimento, que podem alimentar bases de dados ou um Sistema de Recuperação da Informação (SRI), podendo ser físico ou em linha. Logo, as mensagens com base em registros, sinalizadas pelas autoras, seriam a representação destes objetos informacionais, enquanto que as mensagens internas dos usuários seriam as necessidades informacionais, empregadas em forma de termos de busca no SRI.

Para Mey e Silveira (2009), a catalogação possui três funções. A primeira é de permitir ao usuário localizar um item específico, ou seja, encontrar no SRI a informação ou item que esteja buscando em meio ao manancial de informações disponíveis; a segunda função é de permitir a um item encontrar seu usuário, indo ao encontro com a 2ª Lei de Ranganathan, e, por fim, a terceira função, é permitir que outra biblioteca localize um item específico, subsidiando os princípios da catalogação cooperativa, com o intuito de diminuir custos e tempo de processamento por meio da adoção de registros e intercâmbio de dados.

No entanto, para assegurar o cumprimento dessas funções, a catalogação deve possuir características, tais como: integridade, clareza, precisão, lógica e consistência. (MEY; SILVEIRA, 2009).

A integridade "significa fidelidade, honestidade na representação, transmitindo informações passíveis de verificação"; clareza "significa que a mensagem deve ser compreensível aos usuários"; precisão "significa que cada uma das informações só pode representar um único conceito, sem dubiedades ou dúvidas"; lógica "significa que as informações devem ser organizadas de modo lógico" e consistência "significa que a mesma solução deve ser sempre usada para informações semelhantes". (MEY; SILVEIRA, 2009, p. 10).

Referidas características possuem unicamente o objetivo de facilitar o usuário em suas buscas, oferecendo-lhe um sistema com uma linguagem clara, compreensível, e acessível no sentido de facilitar os esforços realizados para o acesso às informações.

Tendo em vista a complexidade da atividade de catalogação, e da sua abrangência em nível internacional, é necessário o uso de instrumentos que padronizem a atividade e que sejam aceitos em nível internacional, para possibilitar o intercâmbio de informações entre distintos SRI, independente da forma que se configuram, quer sejam Bibliotecas, Arquivos, Museus ou quaisquer outros tipos de Unidades de Informação.

Desse modo, a história da catalogação permeia o contexto dos catálogos que, em sua gênese, tinham apenas o intuito de servir de inventário às coleções existentes nos acervos das bibliotecas na Antiguidade, e seguiu assim, com mudanças incipientes, até o século XVIII, com a mudança dos objetivos dos catálogos, em que passa a ser desenvolvido para servir como instrumento de busca, traçando assim, novos rumos para a catalogação. (MEY, 1995).

Logo, no contexto histórico da catalogação, nota-se vários marcos que subsidiam os avanços e as técnicas empregadas nesta área, assim como o emprego de instrumentos que norteiam as atividades realizadas, a fim de possibilitar a representação descritiva da informação, e de ações que há tempos buscavam a padronização como um fator importante que deveria ser empregada na catalogação. (MEY, 1995; MEY; SILVEIRA, 2009).

Nesse contexto, podemos destacar a impressão e a venda de fichas catalográficas pela *Library of Congress* (LC), em 1901, que fora a primeira iniciativa de padronização de catálogos; o Código da Vaticana em 1920; o surgimento da UNESCO em 1946, que em seguida criou um programa de Controle Bibliográfico Universal, e elegeu como norma básica para a descrição bibliográfica a ISBD, e com o formato de intercâmbio, o UNIMARC.

Em 1960, com a evolução dos recursos informacionais, a LC apresenta o projeto MARC (*Machine Redable Cataloging*), um formato padrão para entrada de informações bibliográficas em computador. Em seguida, em 1961, houve a Conferência de Paris, primeiro evento no sentido de normalização internacional, em que se determinou por acordos e discussões, vários pontos básicos da catalogação. (MEY; SILVEIRA, 2009)

Em 1967, foi publicada em uma ação conjunta entre a ALA, C*anadian Library Association* e *Library Association* (Inglaterra), a primeira edição das *Anglo-American Cataloging Rules* (AACR), que em 1969, passa a ser editada no Brasil com a tradução para o português da versão americana com o título de Código anglo-americano de catalogação, seguido de uma segunda edição em 1978, conhecida como AACR2, código inclusive ainda utilizado nos dias atuais em âmbito internacional, embora o seu sucessor já esteja em desenvolvimento. (MEY; SILVEIRA, 2009).

Considerando o breve contexto histórico acerca da catalogação apresentado, percebemos a constante preocupação na evolução das discussões que permeiam as técnicas e

estratégias empregadas no âmbito do tratamento descritivo da informação. Assim, nos dias atuais, a busca por ferramentas, processos e instrumentos que deem continuidade às atividades oriundas do processo de catalogação são constantes, principalmente se levarmos em consideração os novos suportes e o avanço da produção das informações disponíveis em ambiente Web, como já destacado anteriormente.

Embora se configurem atividades correlacionadas, o Tratamento Temático da Informação, ao contrário, tem uma forte carga subjetiva, pois visa caracterizar os objetos informacionais sob o ponto de vista do seu conteúdo (DIAS E NAVES, 2007), reforçando a exposição de GUIMARÃES (2009, p. 105), quando afirma "[...] que a distinção entre tais abordagens reside na busca do o quê (materialização) e do sobre o quê (teor) que [...]".

Em ambientes como Bibliotecas, Arquivos e outras tipologias de Unidade de Informação, o T.T.I, relaciona-se ao assunto tratado no documento, compreendendo a análise documentária como área teórica e metodológica que engloba atividades de classificação, elaboração de resumos, indexação, levando em consideração as diferentes finalidades de recuperação da informação. (FUGITA, RUBI, BOCCATO, 2009).

Para Guimarães (2008), esta área de estudos pode ser historicamente caracterizada, em três momentos, definidos de forma metafórica como a arte, a técnica e a busca por metodologias.

A arte, caracterizada como momento inicial em que a determinação do conteúdo do documento e sua consequente nomeação, assim, inicialmente é vista como um talento especial, uma habilidade artística e um processo estritamente intuitivo. Guimarães (2008, p. 79), assegura ainda que

> Isso se confirma pela trajetória trilhada desde os envelopes de argila que descreviam o conteúdo de papiros e pergaminhos na Mesopotâmia, passando pela classificação de Calímaco em Alexandria, pelos "índices marginais" dos monges copistas medievais e, já na Idade Moderna, chegando às concordâncias bíblicas de Alexander Cruden ou mesmo à concepção alemã de Schlagwort para a representação de assunto pelos livreiros.

A técnica, tida como um segundo momento, caracteriza-se a partir do século XIX, em que a produção documental em larga escala passa a exigir um tratamento sustentado pelas técnicas, refletindo as ideias advindas da Revolução Industrial. Destacando-se nesse contexto a estrutura de notações decimais do sistema de Dewey, a concepção do sistema de classificação da *Library of Congress*, e os sistemas de indexação *Uniterm* e KWIC, constantemente considerando a necessidade de estabelecimento de regras claras para o desenvolvimento do fazer do T.T.I, notadamente a princípio em bibliotecas. (GUIMARÃES, 2008).

E, por fim, a busca por metodologias, que se configura por momentos como a consolidação acadêmica da Biblioteconomia e, posteriormente, da Ciência da Informação, as experiências de tratamento automatizado da informação, a preocupação a partir dos anos 50 dos Estados Unidos e principalmente da Europa "[...] com o desenvolvimento de bases científicas para *o fazer* do TTI, no intuito de ir além de técnicas prescritivas para buscar a construção de metodologias defensáveis para o desenvolvimento dos procedimentos da área". (GUIMARÃES, 2008, p. 80, grifo do autor).

Momentos esses, referidos há pouco, que caracterizam esta área de estudos, possibilitando-nos a compreensão da evolução e constante busca por técnicas e métodos que sustentassem a realização das atividades relacionadas ao T.T.I, em cada momento da história.

Assim, é considerável frisar que, nesse âmbito, o Tratamento Temático da Informação, apresenta-se na literatura especializada sob três eixos teóricos, sendo a catalogação de assunto (subject cataloguing), de matriz norte-americana, a indexação (indexing) de matriz inglesa e a análise documental (analyse documentaire), de matriz francesa. (GUIMARÃES, 2009; SANTOS, 2017). E a respeito do assunto, Café e Sales (2010, p. 120), esclarecem que

> Embora essas três vertentes se diferenciem quanto à ênfase dada aos seus fazeres – catalogação de assunto focada no desenvolvimento de produtos (como catálogos), indexação focada no desenvolvimento de instrumentos (como tesauros) e análise documental focada no desenvolvimento de referenciais teórico metodológicos para os procedimentos envolvidos no TTI (GUIMARÃES, 2008 e 2009) – elas constroem juntas um arcabouço conceitual que, conjuntamente às idéias classificacionistas, formam a base teórica do tratamento temático da informação.

Tendo em vista que "A análise documentária é definida como um conjunto de procedimentos efetuados com o fim de expressar o conteúdo dos documentos, sob formas destinadas a facilitar a recuperação da informação." (DIAS, NAVES, 2007, p. 11). Importa esclarecer que este trabalho segue as diretrizes da Análise Documentária como área teórica e metodológica que concentra os processos desenvolvidos durante a execução de atividades do Tratamento Temático da Informação, tendo como ênfase a indexação, e os instrumentos auxiliares utilizados, como a elaboração de resumos, as classificações, thesaurus e outros.

Como visto, o termo indexação (indexing) pertence à corrente teórica inglesa e, de acordo com os "Princípios de Indexação" do World Scientific Information Programme (UNISIST, 1981, p. 84), é "a ação de descrever e identificar um documento de acordo com seu assunto". Guinchat e Menou (1994, p. 175) constatam que "A indexação é uma das formas de descrição de conteúdo.", operação pela qual são escolhidos ou atribuídos termos que descrevem de maneira objetiva o assunto de um documento.

Os princípios da indexação remontam às tarefas realizadas pelos antigos escribas da Mesopotâmia, que na época, começaram a ter salas para a cópia das tábuas de argila, a elaboração de etiquetas e a conservação das placas. “Os textos eram armazenados em prateleiras de madeira, colocados em nichos nas paredes ou eram dispostos em caixas de madeira. Para saber o que continham, colocavam uma pequena etiqueta anexada na lateral, onde escreviam o conteúdo dos documentos”. (LEIVA, 2012, p. 65).

Nos estudos que versam sobre indexação é comum encontrarmos distintos termos para o mesmo conceito. “Não indo muito longe, na mesma definição de indexação é surpreendente a variedade de verbos empregados para descrever essa ação: reter, extrair, captar, resumir, descrever, caracterizar, escolher, analisar, identificar, traduzir, indexar, indicar, interpretar, enumerar, etc”. (LEIVA, 2012, p. 68).

Robredo (2003, p. 165), afirma que “a indexação consiste em indicar o conteúdo temático de uma unidade de informação, mediante a atribuição de um ou mais termos (ou códigos) ao documento, de forma a caracterizá-lo de forma unívoca”, desta forma Chaumier (1988, p. 63) ao ponderar que “a indexação é a parte mais importante da análise documentária”, permite o entendimento de que as atividades desenvolvidas no âmbito da indexação são de extrema importância e determinantes para o pleno funcionamento de um sistema de recuperação da informação.

Outra concepção é a da NBR 12676 da Associação Brasileira de Normas Técnicas (1992, p. 2), documento oficial brasileiro que aborda sobre a indexação, e a define como o “Ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação”. Assim, “[...] os termos atribuídos pelo indexador servem como pontos de acesso mediante os quais um item é localizado e recuperado, durante uma busca por assunto num índice publicado ou numa base de dados eletrônica”. (LANCASTER, 2004, p. 6).

Nota-se que os autores empregam termos e abordagens distintas para conceituar a atividade de indexação. Entretanto, todos apresentam um consenso no entendimento de que tal atividade consiste na determinação de assuntos de documentos, abrangendo todos os tipos de objetos informacionais. “Ficando evidente, assim, que a indexação é complexa e carrega consigo um caráter essencial no que se refere ao suprimento das necessidades informacionais dos usuários de um sistema de recuperação da informação”. (SANTOS, 2017, p. 26).

Logo, análoga às várias concepções que existem na literatura para definir a indexação, de igual modo existem vertentes distintas que apresentam as etapas desse processo, e dos

instrumentos e como estes podem ser utilizados no intuito de auxiliar a realização das etapas que a constituem.

Assim, a ABNT NBR 12.676 (ASSOCIAÇÃO..., 1992, p. 2), entende que o processo de indexação é composto por três momentos que acontecem consecutivamente durante a atividade, sendo "a) exame do documento e estabelecimento do assunto de seu conteúdo; b) identificação dos conceitos presentes no assunto e; c) tradução desses conceitos nos termos de uma linguagem de indexação".

Dias e Naves (2007, p. 15) concebem que a indexação compreende apenas duas etapas, e que as mesmas são distintas, a saber, "a extração de conceitos que possam representar o assunto de um documento e a tradução destes para termos de instrumentos de indexação, que são as chamadas linguagens de indexação ou linguagens documentárias".

Enquanto Lancaster (2004) corrobora com Dias e Naves (2007), ao compreender que a indexação é realizada em apenas duas etapas e, ao passo disso, entende de maneira diferente da NBR 12.676. Para ele, é composta de análise conceitual e tradução. De acordo com o autor (2004, p. 9) a análise conceitual consiste em "[...] decidir do que trata um documento – isto é, qual o seu assunto". Já a etapa de tradução "[...] envolve a conversão da análise conceitual de um documento num determinado conjunto de termos de indexação" (LANCASTER, 2004, p. 18). Como na norma, o autor especifica que as etapas podem ocorrer simultaneamente.

Levando em consideração as várias abordagens encontradas na literatura, o Quadro 3, a seguir, apresenta as definições acerca da indexação, e as etapas que compõem o processo na perspectiva de cada autor.

Quadro 3 – Concepções e etapas da indexação

AUTOR	DEFINIÇÃO	ETAPAS
UNISIST (1981)	Ação de descrever e identificar um documento de acordo com seu assunto.	1º Determinação do assunto 2º Tradução dos conceitos nos termos da linguagem de indexação
NBR 12.676 (1992)	Ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação.	1º Exame do documento e estabelecimento do assunto de seu conteúdo; 2º Identificação dos conceitos presentes no assunto; 3º Tradução desses conceitos nos termos de uma linguagem de indexação.
CHAUMIER (1988)	Descrição e caracterização dos conceitos contidos em um documento.	1º Reconhecimento e extração de conceitos; 2º Tradução desses conceitos em linguagem natural.
LANCASTER (2004)	Preparação de uma representação dos conteúdos temáticos dos documentos.	1º Análise conceitual; 2º Tradução.
DIAS; NAVES (2007)	Terminologia mais usada para	1º Extração de conceitos;

	designar o trabalho de organização da informação quando realizado nos chamados serviços de indexação e resumo.	2º Tradução.

Fonte: adaptado de Fujita, Rubi e Boccato (2009, p. 25).

Etapas que, embora se divirjam em alguns aspectos, possuem uma conformidade ao caracterizarem a indexação como atividade que sustenta a realização da análise do documento, no intuito de entender a tematicidade manifestada no item e, em seguida, utilizar de instrumentos para a tradução para linguagens documentárias, expressas através de termos que serão inseridos nos sistemas de recuperação da informação.

Citamos, então, Análise: momento de leitura e segmentação do texto para identificação e seleção de conceitos; Síntese: construção do texto documentário com os conceitos selecionados; e Representação: por meio de linguagens de indexação. (RUBI, 2009).

A indexação pode ainda acontecer por extração, quando as palavras e expressões são selecionadas ocorrem no documento ou objeto informacional para representar seu conteúdo, ou por atribuição, quando termos são atribuídos, mas não se apresentam de modo direto no documento, utilizando assim de instrumentos para fornecer esses termos. E há ainda elementos que caracterizam o processo e o resultado da indexação, aos quais destacamos a exaustividade, consistência, especificidade e correção. (LEIVA, 2012)

O acelerado crescimento da produção de informações e os estudos que possuem como objeto a indexação tiveram significativos avanços, inicialmente suscitando pesquisas que questionavam o emprego de aplicações computacionais em atividades que até então eram exclusivamente atribuídas aos humanos. E, consequentemente, com os avanços advindos das tecnologias digitais de informação e comunicação, e das informações disponíveis em ambiente Web, estudos com essas perspectivas são cada vez mais frequentes.

Assim, já em meados da década de 90, surgiam estudos que sustentavam a realização da indexação de modo manual; semiautomatizada e automatizada. A manual é realizada exclusivamente por um humano, em que de acordo com a perspectiva teórico-metodológica empregada pela unidade de informação, extrai ou atribui termos ao documento ou objeto informacional. Na semiautomatizada o indexador utiliza de programas computacionais para auxiliar a análise do documento e a extração de termos, assim, os termos são atribuídos seguindo a seu critério ou a partir das políticas estabelecidas pela unidade. E, por fim, a automatizada, em que o emprego de software computacional é utilizado para a indexação, sem quaisquer interferências do indexador. (GUINCHAT; MENOU, 1994).

Tendo em vista a complexidade do processo de indexação, independente da abordagem em que ela é realizada, instrumentos auxiliares são utilizados no sentido de subsidiar a compreensão dos objetos que estão sendo indexados e permitir uma representação fidedigna, para que o processo de indexação possa cumprir com seu objetivo final, que é o de alimentar o SRI de forma coerente e possibilitar a recuperação da informação.

Nesse sentido, a atividade de indexação integra um conjunto de instrumentos que auxiliam o processo de análise documentária e representação, a saber, as classificações, os thesauri, a elaboração de resumos e outros instrumentos.

O Uso dos esquemas de Classificações Bibliográficas a exemplo da Classificação Decimal de Dewey e a Classificação Decimal Universal e a da *Library of Congress*, utilizadas e conhecidas em âmbito mundial, auxiliam na indexação, no sentido de possibilitar a tradução dos termos que descreverão os documentos em representações simbólicas, que gerarão uma linguagem notacional e possibilitarão a recuperação das informações disponíveis nos documentos em ambientes tradicionais como bibliotecas.

A UNESCO (1973, p. 13-17) define o tesauro como sendo "vocabulário controlado e dinâmico de termos relacionados semântica e genericamente, que cobre de forma extensiva um campo específico do conhecimento". Assim, são instrumentos utilizados na tradução de termos em linguagens naturais para termos equivalentes em linguagens documentárias, pois estes são a linguagem do sistema, e possibilitarão assegurar o grau de polissemia e sinonímia, que consequentemente irão influenciar diretamente nos níveis de precisão e revocação dos sistemas de recuperação da informação. (ROBREDO, 2003).

A respeito da indexação e elaboração de resumos, Lancaster (2004, p. 6) afirma que estas são atividades intimamente relacionadas, "[...] pois ambos implicam a preparação de uma representação do conteúdo temático dos documentos", exercendo o resumo nessa perspectiva, uma síntese da totalidade temática do documento ou objeto informacional, quando bem elaborado.

Guinchat e Menou (1994, p. 178), afirmam ainda que "Se o resumo for bem feito, ele pode fornecer o essencial da indexação, o que representa uma economia de tempo", tendo em vista a realidade de grandes centros informacionais, de tratar consideráveis quantidades de documentos, o resumo é uma ferramenta importante para análise documentária.

Assim, levando em consideração esse panorama de instrumentos utilizados na representação da informação, nos dias atuais são recorrentes os estudos teóricos e metodológicos que buscam dar sustentabilidade a novos modos de fazer a indexação, empregando cada vez mais as tecnologias de informação e comunicação em ambiente digital,

e nos dias atuais, usando de ferramentas oriundas da ciência da computação, linguística computacional, inteligência artificial e outras áreas do conhecimento.

Desse modo, dá-se ênfase aos processos oriundos da inteligência artificial no sentido mais estrito, voltado para o Processamento Automático de Línguas Naturais (PLN), a Sumarização Automática de Textos (SA), que consiste na geração de textos de forma automática, tendo em vista o uso dessas ferramentas no processo de análise temática da informação.

Conforme exposto inicialmente na seção 1, esta pesquisa visa compreender a aplicação dos sumários automáticos, oriundos de aplicações de PLN na indexação, no intuito de auxiliar no processo de análise documentária, devido a impossibilidade de nos dias atuais, com a acelerada produção informacional, principalmente em ambiente web, da análise de documentos por conta dos bibliotecários e considerando ainda outras atividades tão importantes como as que são desenvolvidas em ambientes informacionais, tais como as bibliotecas, museus, arquivos, unidades de informação, bases e bancos de dados.

2.2 Interdisciplinaridade na OIC: aproximações necessárias da Inteligência Artificial, Linguística Computacional e Ciência da Informação

A acentuada produção informacional, o surgimento de diversos suportes em que se encontram, e a necessidade de consumo cada vez mais instantânea por parte daqueles que dela necessitam, fazem com que os estudos acadêmicos enfatizem cada vez mais na análise dessas realidades para propor soluções aos diversos problemas que recaem nos processos de organização, representação e recuperação da informação. Neves (2019, p. 10), relata que

> Além das habilidades genéricas de gerenciar informações e serviços, o papel do bibliotecário tem ficado a cada dia mais complexo. Embora esse profissional continue a desenvolver suas atividades em torno de encontrar informações específicas para seus usuários, essa ação pode levar menos tempo com a implementação de tecnologias cada vez mais rápidas e interativas.

Dito isso, percebemos que os profissionais da informação, e em especial no que diz respeito ao escopo desta pesquisa, os bibliotecários, possuem grandes desafios, uma vez que há uma busca intermitente por técnicas, processos, e estratégias que sejam satisfatórios para os processos que englobam a organização e representação da informação, subsidiando de forma paralela os recursos informacionais disponíveis em suportes tradicionais, em ambiente web e ainda os que estão surgindo com o advento do mundo digital.

Nesse sentido, Alvarenga (2003, p. 34), relata que

> O advento do mundo digital ocasionou novas mudanças no trabalho de autores e profissionais da informação, fazendo com que estes se envolvessem com novas possibilidades tecnológicas, diretamente incidentes nos processos de produção, armazenagem, representação e recuperação de documentos e informações, alterando seus processos de trabalho e produtos finais.

Novas possibilidades que não apenas os fazem ter um contato mais direto com as tecnologias, mas que ampliam a visão e o entendimento acerca da interdisciplinaridade no campo da Ciência da Informação, e consequentemente na Biblioteconomia e nos processos relativos à Organização e Recuperação da Informação e do Conhecimento, no sentido de empregar e unir forças com outros domínios do conhecimento que possam agregar qualidade e agilidade aos processos citados. E levando em consideração o escopo interdisciplinar em que a Ciência da Informação surge, essa aproximação com diversas áreas do conhecimento, faz ainda mais sentido. (ROBREDO, 2003).

Para Borko (1968, p.3) a Ciência da Informação é

> [...] uma disciplina que investiga as propriedades e o comportamento da informação, as forças que governam seu fluxo, e os meios de processá-la para otimizar sua acessibilidade e uso. A CI está ligada ao corpo de conhecimentos relativos à origem, coleta, organização, armazenagem, recuperação, interpretação, transmissão, transformação e uso de informação. Ela tem tanto um componente de ciência pura, através de pesquisa dos fundamentos, sem atentar para sua aplicação, quanto um componente de ciência aplicada, ao desenvolver produtos e serviço.

Novellino (1996. p. 37), corrobora destacando que

> A Ciência da Informação é uma disciplina voltada para o estudo de fenômenos subjacentes à produção, circulação e uso da informação. O estudo desses fenômenos tem como finalidade possibilitar a criação de instrumentos e o estabelecimento de metodologias que viabilizem a transferência de informações.

Empregando para o fim que se destina, como as possibilidades das diversas áreas do conhecimento humano existentes. Saracevic (1996), destaca quatro disciplinas, que possuem uma contribuição relevante no quesito da interdisciplinaridade na Ciência da Informação que, embora introduzidas por diferentes experiências profissionais, trouxeram significativas contribuições, a saber a Biblioteconomia; Ciência da Computação; Ciência cognitiva, incluindo inteligência artificial; e a Comunicação.

Lima (2003, p. 79, grifo nosso), aponta que

> A interdisciplinaridade inerente à CI tornou-se visível pela própria variedade de profissionais que atuam na área, apresentando como ponto comum uma dependência cada vez maior da tecnologia por profissionais diversos, como engenheiros, cientistas da informação, linguistas, cientistas da computação, filósofos e outros, cujo interesse é **compreender e comunicar a informação**. A tecnologia da informação veio auxiliar esses profissionais trazendo uma nova **potencialidade ao trabalho de processamento e agilidade na busca da informação**. Isso concretizou-se com o surgimento

> de computadores com grande capacidade de armazenamento e de grande rapidez na recuperação da informação.

Entretanto, no que diz respeito aos processos de Organização, Representação e Recuperação da Informação, nos últimos anos, tem tido destaque a Linguística Computacional, a Inteligência Artificial, e a Ciência da Computação, que contribuem cada uma dentro de seu escopo de atuação, além de favorecerem o desenvolvimento do conceito interdisciplinar. Complementando, assim, esforços com ricas possibilidades aos Sistemas de Recuperação da Informação, no sentido de melhorias nos processos realizados com o intuito de compreender e comunicar a informação.

Nesse sentido, a Linguística e a Semiótica são contributos "na criação de linguagens de recuperação da informação em sistemas de indexação e resumo automático de textos, tradução em máquina, unificação nacional e internacional de terminologia especializada, normalização (padronização) de registro de resultados de atividades criativas". (FID apud PINHEIRO, 1999, p. 165).

Almeida (2011), aponta que a organização da informação e do conhecimento possui seus fundamentos baseados na Linguística. Tendo em vista que a Linguística

> [...] fornece os principais conceitos para a análise documental, desde as noções de signo, linguagem, representação, até pontos mais específicos que tratam da estruturação de linguagens de indexação ou linguagens documentais, sistemas de classificação e suas relações internas. (ALMEIDA, 2011, p. 89).

A Ciência da Informação e a Linguística possuem ainda relações no que diz respeito à representação da informação, atrelada aos processos de redução semântica, pluralidade de significados e produção de sentido. E, com o emprego da informática, a Linguística Computacional contribui com a representação da informação, visto que o auxílio de ferramentas computacionais viabiliza esses processos pela máquina de forma automática e semiautomática. (DODEBEI, 2002; PINHEIRO, 1999).

Referidas relações, da Informática e Linguística Computacional, exercem a decodificação da linguagem com o propósito de acesso à informação, possibilitando o tratamento do objeto informacional, no sentido de compreender a obra, a expressão e manifestação desta em um item informacional, ficando clara a compreensão para o profissional da informação, que é o responsável em comunicar a informação alimentando os SRI, para promover a disseminação da informação.

No que concerne as aproximações entre a Inteligência Artificial e a Ciência da Informação, estudos indicam possibilidades de inserção de tecnologias da IA aos processos desenvolvidos, que também se referem à recuperação da informação. "A ideia geral é que a

partir de uma questão formulada pelo usuário, o sistema seja capaz de lhe apresentar os resultados que sejam compatíveis com a questão apresentada." (MARTINS, 2010, p. 10).

Assim, a utilização de mecanismos de IA, no intuito de contribuir para a RI, versa sobre aspectos da programação evolutiva; agentes inteligentes; agentes de busca; interfaces inteligentes; classificações automáticas de conteúdos com o emprego de ontologias e outros. (MARTINS, 2010, p. 10).

E em se tratando das bases que estabelecem as relações entre Ciência da Informação e Ciência da Computação (CC), estas são formadas pelo emprego de aplicações computacionais na recuperação da informação, desdobrando-se em outras associações como, produtos, serviços e redes. (SARACEVIC, 1996). "Entre ambas há uma relação de complementaridade, uma vez que a Ciência da Computação trata de processos algorítmicos que transformam a informação, e Ciência da Informação trata da 'natureza da informação e sua comunicação para pessoas". (PINHEIRO, 1999, p. 172).

Ainda sobre a CC, Lima (2003) aponta as possibilidades de interseção entre a CI e a CC que se concentram nos processos de categorização, indexação, recuperação da informação e interação homem-máquina, possibilitando o entendimento de que a CI e a CC são áreas complementares que conduzem a aplicações diversas (SARACEVIC, 1996).

Sendo assim, tendo como pressuposto a interdisciplinaridade presente na Ciência da Informação, a aplicação de técnicas e métodos de outros domínios do conhecimento tem sido cada vez mais comum, e além dos já destacados nesta subseção, citamos o PLN, microcampo da Linguística Computacional e subcampo da Inteligência Artificial, que estuda a possibilidade de computadores simularem a compreensão e o entendimento humano no que se refere ao processamento automático da língua natural.

Os estudos iniciais em torno da temática têm como ponto de partida os campos da Ciência da Computação e Linguística Computacional, e, logo depois, se estenderam para outros domínios como a Ciência da Informação, resultando assim, em que o PLN configura-se como um campo de estudo heterogêneo e fragmentado, ao agregar pesquisadores com interesses e perspectivas diversas quanto ao PLN. (DIAS DA SILVA et al. 2007, p. 5).

Em Ciência da Informação tem sido estudada na perspectiva teórica, em especial no campo da "[...] Indexação e Recuperação da Informação, pela possibilidade de softwares baseados nesse modelo propiciarem a extração de termos com maior precisão semântica para recuperação da informação em sistemas de buscas automatizados [...]" (CARMO; CONCEIÇÃO, 2018, p. 317), potencializando as tarefas automatizadas com o emprego do PLN.

Desta forma, entendemos que a informação é insumo para os processos comunicacionais que se fazem presentes entre as relações humanas. Percebe-se, inclusive, que existem muitos esforços no sentido de viabilizar o acesso a essas informações disponíveis nos diversos suportes informacionais existentes, e, para que isso aconteça, diferentes áreas do conhecimento estão unindo esforços, visando otimizar e tornar a Organização, Representação e a Recuperação da Informação, em processos céleres e assertivos.

2.3 Tecnologias da informação e comunicação aplicadas à OIC

Os princípios que norteiam a Organização da Informação e do Conhecimento, como já mencionados anteriormente, remontam aos catálogos que, em sua gênese possuíam características semelhantes às listas de inventários, e, posteriormente, os envelopes de argila que descreviam o conteúdo de papiros e pergaminhos, iniciativas presentes em instituições como Bibliotecas. Esses produtos esboçavam de maneira bastante primitiva possibilidades que viabilizavam o conhecimento acerca do que havia disponível em determinadas coleções, e, consequentemente, permitiam o acesso aos documentos e informações então disponíveis.

Iniciativas que nortearam o desenvolvimento da OI, de seus processos, produtos e instrumentos da OI, e o processo evolutivo que as sociedades passaram ao longo do tempo, aumentaram a percepção sobre a necessidade de desenvolvimento constante de técnicas consistentes e atuais para suprir as necessidades de tratamento informacional, quanto à representação e recuperação da informação. Anna (2015, p. 315), aponta que:

> As bibliotecas no decorrer dos tempos vêm sendo impactadas pelas novas tecnologias da informação e comunicação (TICs), cujo poder de tratamento informacional, de armazenamento e de distribuição se consomem de forma cada vez mais eficientes, a custos mais baixos e maior agilidade nos processos realizados.

Os avanços impostos pela tecnologia da informação e da comunicação proporcionaram profundas mudanças nos entendimentos sobre as técnicas e procedimentos da organização da informação e do conhecimento, assim como dos recursos tecnológicos utilizados para sustentarem a realização destes.

Logo, as TIC influenciam de maneira significativa os processos oriundos da Organização da Informação e do Conhecimento, no que diz respeito às representações temáticas e descritivas, o que acarreta no emprego de padrões, formatos e ferramentas, com vistas a dinamizar e subsidiar os processos de tratamento e organização da informação. (ALVARENGA, 2003; SOUSA; HILLESHEIM, 2014).

A princípio, o ponto fulcral que trouxe as conexões das TIC ao que diz respeito à Organização da Informação e do Conhecimento, aos ambientes das bibliotecas, dá-se com o emprego de computadores para a automação de bibliotecas em meados dos anos de 1960. Inicialmente a finalidade destes foi de apenas aumentar a capacidade de armazenamento, com a possibilidade de recuperação da informação.

Para Alvarenga (2006, p. 77), à medida que "as tecnologias da informação foram sendo criadas, disponibilizadas e aperfeiçoadas, os sistemas de representação e recuperação de informações documentais assistiram a uma extrapolação dos limites dos tradicionais catálogos referenciais em fichas, alcançando as bases de dados em linha".

Assim, consequentemente com a produção acelerada de informações, a inserção das TIC e uso de ferramentas computacionais, softwares e hardwares, tornaram-se cada vez mais comuns nos processos tradicionais de armazenamento, tratamento e recuperação da informação. Evidencia-se, nesse sentido, a consolidação da automação em bibliotecas a partir das contribuições significativas desta para o desenvolvimento de atividades do processamento técnico.

Assim, destacam-se os Sistemas de Gerenciamento de Bibliotecas (SGB), ferramentas utilizadas com o intuito de automatizar processos rotineiros de uma unidade de informação/biblioteca, passando pela circulação, catalogação e indexação. Estes tiveram um importante papel no que se refere aos processos comuns a OIC, tendo em vista o produto decorrente destes, os catálogos em formato online, conhecidos como OPACs. Ainda neste sentido, Rowley (2002, p. 5), diz que "a introdução de sistemas informatizados nas bibliotecas resultou em padronização, aumento da eficiência, interligação por redes e melhores serviços".

Evidentemente, atualmente, as atividades relacionadas à Organização da Informação e do Conhecimento, são cada vez mais mediadas por meio do uso de computadores e da internet. Assim, podemos citar várias iniciativas empregadas que possibilitam a realização da OIC, através de processos, produtos e instrumentos, subsidiados por padrões, formatos e ferramentas.

Logo, em se tratando da descrição de recursos em ambientes digitais, temos os metadados, que é o termo da era da internet para a informação descritiva sobre recursos da Web, sendo assim um conjunto de atributos ou elementos para descrever um item, dentre os quais, de uso universal em bibliotecas, destacamos o Dublin Core (DC) e o Formato MARC. (MEY; SILVEIRA, 2009).

O DC é uma ferramenta de descrição para registro de objetos eletrônicos em rede, desenvolvido pela OCLC e NCSA. Possui como características a simplicidade na descrição de recursos, o entendimento semântico universal, escopo internacional e a extensibilidade.

O MARC é um formato de registro bibliográfico legível por máquina, datado da década de 1960 e desenvolvido pela Library Congress e British Library, possui um padrão desenvolvido para a entrada e manuseio de informações em computador, com vistas ao intercâmbio de informações entre sistemas. (MEY; SILVEIRA, 2009).

No entanto, o crescente avanço da produção de informações, a conectividade e instantaneidade em que os sistemas possibilitavam, percebeu-se necessidades no que se referiam à conectividade entre sistemas, e, nesse processo, "as TICs auxiliam o desenvolvimento de formatos e padrões e proporcionam um intercâmbio ainda maior de informação, ressaltando como principal característica a interatividade." (SOUSA; HILLESHEIM, 2014, p. 83). Assim, podemos destacar, a norma ISO 2709, Z39.50, OAI-PMH.

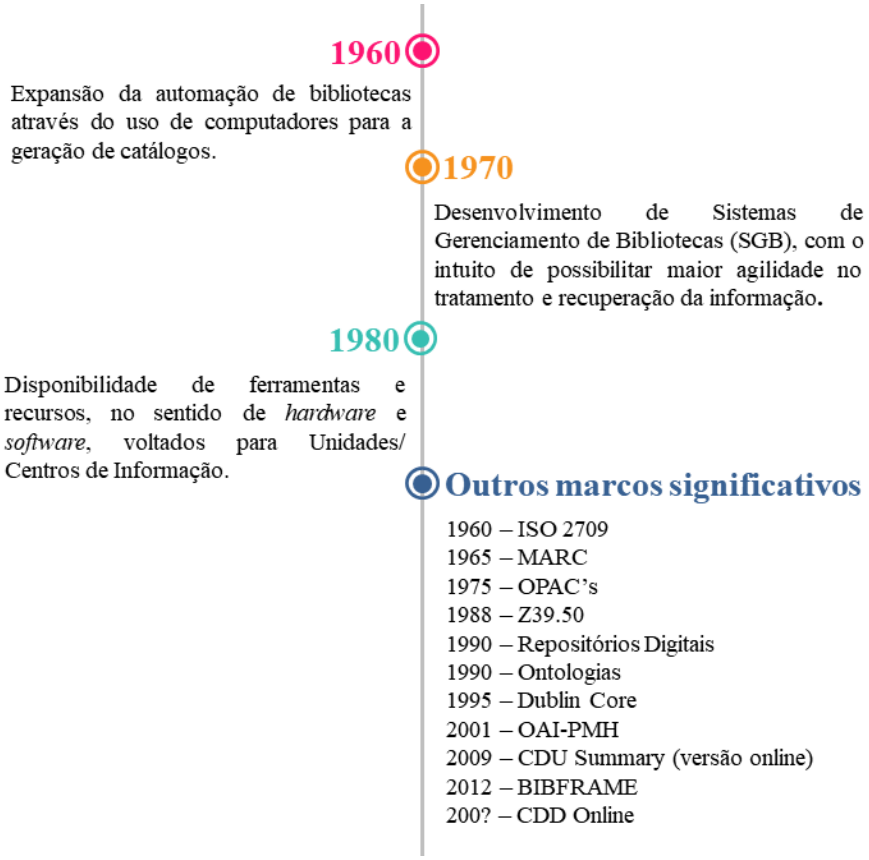
A Norma ISO 2709 é um formato padrão de comunicação para registros bibliográficos, utilizado para intercâmbio de registros em meio magnético de um sistema para outro, tornando possível através de seus padrões estabelecidos, a transferência de um item bibliográfico de um sistema ou banco de dados para outro, sem perda de informações, fazendo com que os dados sejam independentes de software e hardware, tornando os registros bibliográficos portáveis entre sistemas.

O Z39.50 é um protocolo que permite a interoperacionalização de diferentes sistemas de computação com diferentes sistemas operacionais, equipamentos, formas de pesquisa, sistemas de gerenciamento de bases de dados. Enquanto que o Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), é um protocolo desenvolvido pela Open Archives Initiative que define um mecanismo para coleta de registros de metadados em repositórios.

Nesse contexto de TICs, que viabilizam a Organização da Informação e do Conhecimento, existem ferramentas que são importantes e que estão permitindo compreender cada vez mais a necessidade de seu uso nos dias atuais, como por exemplo, os Repositórios Digitais, os Bancos e Bases de dados, os Serviços de Descobertas, os Periódicos Eletrônicos e outros, que viabilizam a organização e a disseminação de informações em ambientes digitais.

A Figura 2, a seguir, apresenta de forma sucinta o emprego de Tecnologias de Informação e Comunicação no contexto da Organização da Informação, destacando as iniciativas que trouxeram marcos significativos para a área.

Figura 2 – TICs na Organização da Informação



Fonte: o autor (2020).

É evidente o número de contribuições significativas que as Tecnologias da Informação e Comunicação trouxeram para a OIC, tendo em vista que a partir do momento em que estas foram inseridas em atividades até então realizadas de forma exclusivamente manuais, são utilizadas até os dias de hoje, passando por mudanças e desenvolvendo-se, para sustentar as realidades de cada era. Sobre o assunto, Sousa e Hillesheim (2014, p. 82), afirmam que

> Essas inovações são essenciais para acompanhar o aumento no volume de informações, principalmente em meio eletrônico, e nesse contexto, percebe-se a evolução das tecnologias de informação e comunicação - TICs interferindo diretamente na disseminação da informação e na gestão do conhecimento.

Assim, pondera-se a necessidade de aperfeiçoamento e evolução constante desses instrumentos aplicados à representação, organização e recuperação da informação. Nesse sentido, a partir dos dados representados na ilustração anterior, é possível compreender a importância da adaptação e transição dos instrumentos tradicionais de representação, para os contemporâneos, que desenvolveram-se principalmente em ambientes digitais, tais como os Sistemas de Organização do Conhecimento (SOC), a exemplo das taxonomias, ontologias, tesauros, mapas conceituais, folksonomias e outros.

Nesse sentido, há uma clara percepção, de que, para além da disseminação da informação e gestão do conhecimento, estes são de extrema importância para a Organização da Informação e do Conhecimento, tendo em vista o mundo globalizado em que vivemos. Assim, como afirmado anteriormente, é cada vez maior o número de esforços empregados para que isso seja possível, agregando possibilidades de outras áreas, como o emprego de técnicas de PLN, item que será discutido na próxima seção.

3 PROCESSAMENTO AUTOMÁTICO DE LÍNGUAS NATURAIS NA ORGANIZAÇÃO DO CONHECIMENTO: demarcações iniciais

Introduzidos na cultura ocidental no início dos anos 40, os computadores têm se mostrado grandes aliados no desenvolvimento de atividades que, até outrora, eram realizadas exclusivamente pelo homem. Desde então nota-se, grandes avanços no desenvolvimento de tecnologias digitais que possam substituir o homem em tarefas arriscadas, mecânicas, sistemáticas, repetitivas e enfadonhas, além de estarem mais presentes em nossa sociedade, sobretudo com o avanço da internet. (DIAS DA SILVA, 2006).

Transpondo o contexto para qual foram desenvolvidas, – e a necessidade cada vez maior de uma comunicação mais efetiva entre homem e máquina, superando as linguagens de programação e interfaces gráficas, desenvolvidas para facilitar sobremaneira a comunicação entre usuário e computador, – surgem novas possibilidades que sustentam essa nova realidade em que ambos estão inseridos atualmente.

Entretanto, estabelecer meios que possibilitem a comunicação entre homem e máquina, se configura um desafio de difícil realização, tendo em vista as diferentes linguagens utilizadas entre esses sujeitos. “Os computadores normalmente estão aptos a compreenderem instruções escritas em linguagens computacionais – tais como Java, C++, Python, Perl –, mas possuem muita dificuldade para entender comandos simples escritos em uma linguagem humana.” (EVERS, 2013, p. 86). Em contrapartida, temos a complexidade da língua natural utilizada pelos humanos, imbuída de facetas e formalismos linguísticos. Desafios que tomam proporções maiores, em se tratando de criar mecanismos que possibilitem a compreensão e entendimento por parte da máquina, das línguas naturais.

A esse respeito, Sousa (2015, p. 15), afirma que

> Pela forma como é constituído, o computador é incapaz de compreender comandos ambíguos, como se caracteriza a comunicação humana, sendo trabalho do PLN organizar a língua em modelos exatos que possam ser inteligíveis por máquinas.

Nasce assim, nesse contexto de interação homem e máquina, o Processamento Automático de Línguas Naturais, (doravante PLN). É compreendido como “[...] à área de pesquisa que se dedica a investigar, propor e desenvolver formalismos, modelos, técnicas, métodos e sistemas computacionais que têm a língua natural como objeto primário.”. (NUNES, 2008, p. 3).

O PLN é um domínio de pesquisa amplo e controverso, de natureza inter/multidisciplinar, tendo em vista a pluralidade de objetivos e interesses dos pesquisadores

que desenvolvem estudos nesse domínio. Embora seja uma vertente da Inteligência Artificial, que ajuda computadores a entender, interpretar e manipular a linguagem humana, tanto escrita quanto falada, compreende técnicas e estudos conceituais de diversas áreas do conhecimento, e constitui-se a partir de contribuições teórico-metodológicas das Ciências da Computação, da Ciência da Informação, Estatística, Filosofia, Linguística, Linguística Computacional, Lógica, Matemática, Psicologia e outras. (LADEIRA; ALVARENGA, 2012; SOUSA, 2015).

Dias da Silva (2006, p. 104), sinaliza ainda que uma das características do PLN é o fato de agregar uma heterogeneidade de objetivos, que vão desde:

> [...] a meta de investigar meios de empregar o computador como uma simples ferramenta auxiliar para investigar material lingüístico (por exemplo, a criação de programas de computador para calcular estatísticas de ocorrências de palavras em textos ou para identificar e indexar palavras e segmentos de texto) até a meta de criar uma inteligência artificial, nos moldes do supercomputador HAL-9000 do clássico de Stanley Kubrick – 2001: Uma Odisséia no Espaço.

Sustentando assim, a perspectiva inter/multidisciplinar dos estudos em PLN, mas levando em consideração as possibilidades que os Sistemas de Processamento Automático de Línguas Naturais (SPLN) podem oferecer, é comum encontrar na literatura diversas formas de defini-lo e designá-lo. Entretanto, todas as definições integram a noção de armazenamento em computador e manipulação de dados linguísticos. “Em outras palavras, o PLN pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia a dia”. (ROSA, 2011, p. 137).

E se tratando das designações, expressões como Processamento da Linguagem Natural; Processamento de Línguas Naturais; Linguística Computacional; e Processamento Automático das Línguas Naturais. Adotamos este último, no âmbito desta pesquisa, levando em consideração a clareza presente no termo, no sentido de evitar ambiguidades terminológicas, em se tratando do processamento automático pela máquina das línguas naturais (utilizadas pelos humanos), e não apenas o processamento da língua pelos humanos, ou ainda, ao processamento de linguagens, tendo em vista que o termo “linguagem” se estende a outras perspectivas.

Em face desses fatos, apresentar uma narrativa que tenha como cerne a evolução histórica do PLN, torna-se uma tarefa difícil, quando consideramos as diversas abordagens apresentadas pela literatura, no sentido de métodos e estratégias utilizadas. Entretanto, tem-se como marco inicial dos estudos em PLN, datada no início da década de 40, a Tradução Automática (TA), “[...] considerada pela maioria dos autores o marco inicial do uso do

computador para a investigação das línguas naturais, [o que] permite também apresentar uma síntese da evolução dos estudos nesse campo." (DIAS DA SILVA *et al*, 2007, p. 5).

Os estudos iniciais em TA, deram margem para outros estudos em PLN. Inicialmente, objetivava-se a tradução automática de textos científicos, utilizando programas tradutores baseados em criptografia, em que ocorria a tradução literal do texto de origem para o texto de destino, ou seja, tinha-se a tradução de palavra por palavra, realizada de maneira mecânica e sistemática, sem quaisquer aplicações de fundamentação linguística e analise sintática, o que ocasionava erros grosseiros de tradução, além da necessidade de revisões por tradutores humanos.

Questões essas que, por ora, desaqueceram as pesquisas sobre a área de PLN, tendo em vista as experiências mal sucedidas. Dias da Silva (2006, p. 7), pondera que

> Depois de muitas experiências negativas e concepções equivocadas em relação ao tratamento computacional das línguas naturais, a partir de meados da década de 70, os trabalhos de tradução automática foram retomados com uma atitude mais acadêmica e realista.

Então, após o reconhecimento por parte dos pesquisadores em PLN de que a técnica da criptografia era inadequada ao tratamento computacional das línguas naturais, – e que era necessário criar sistemas fundamentados em conhecimentos linguísticos, que compreendessem a polissemia das unidades linguísticas e a estrutura das línguas, linguagens e itens lexicais – que as pesquisas avançaram, abrindo margem pra consolidação do PLN. (VIEIRA; LOPES, 2010).

Nesse sentido, a década de 50 é marcada ainda por explorações no campo das traduções automáticas, em que se configuram como avanços na área a sistematização computacional das classes de palavras descritas nos manuais de gramática tradicional e a identificação computacional de constituintes oracionais. (DIAS DA SILVA *et al*, 2007).

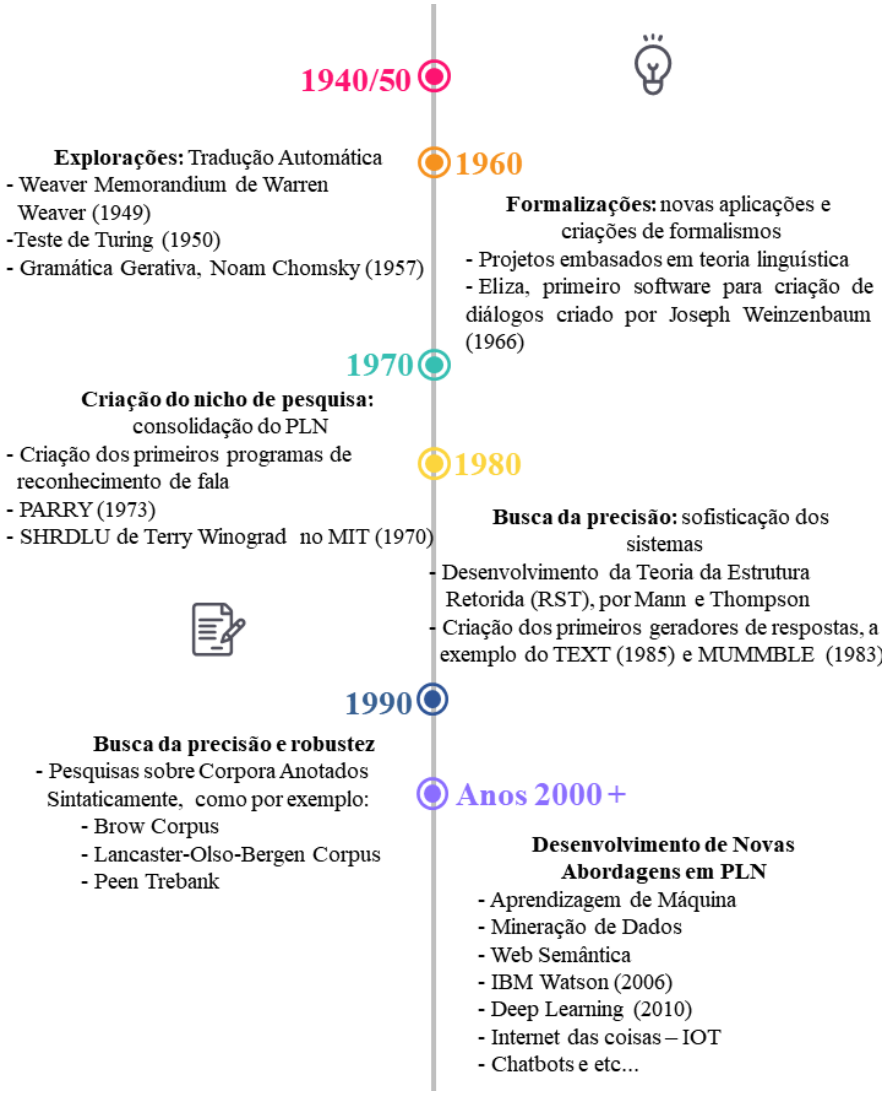
Os anos 60 são marcados pelas formalizações, novas aplicações e criação de formalismos. É nesse período que surgem os primeiros tratamentos computacionais das gramáticas livres de contexto, criação dos primeiros analisadores sintáticos (*parsers*) e as primeiras formalizações do significado em termos de redes semânticas. (DIAS DA SILVA *et al*, 2007).

Em 70, tem se a criação do nicho de pesquisa e a consolidação do PLN, caracterizados pela implementação de parcelas das primeiras gramáticas e analisadores sintáticos, baseados na gramática gerativo-transformacional e pela busca de formalizações de fatores pragmáticos e discursivos. (DIAS DA SILVA *et al*, 2007).

A década de 80 destaca-se pela busca da precisão e sofisticação dos sistemas, que resulta no desenvolvimento de teorias linguísticas motivadas pelos estudos do PLN como, por exemplo, a gramática sintagmática generalizada e a gramática léxico-funcional. (DIAS DA SILVA *et al*, 2007).

Em seguida, os anos 90, são caracterizados pela busca da precisão e robustez em que são desenvolvidos sistemas baseados em representações do conhecimento no tratamento estatístico de massa de texto. Nesse período, tem-se o desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos linguísticos e extralinguísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos linguísticos. (DIAS DA SILVA *et al*, 2007). A Figura 3, a seguir, apresenta de forma sucinta os principais eventos históricos que amparam os avanços em PLN.

Figura 3 – Evolução dos estudos em PLN



Fonte: adaptado de Dias da Silva *et al*., (2007); Vieira e Lopes, (2010); Rosa (2011); Pardo (2017).

Nesse sentido, com o desenvolvimento dos estudos em PLN ao longo do tempo, surgem novos recursos e ferramentas linguísticas computacionais, agregando com maior frequência os conhecimentos linguísticos, em especial o léxico, a gramática e a semântica, potencializando as possibilidades de interação entre homem e máquina, quanto ao uso de línguas naturais e artificiais. Estudos que foram impulsionados sobremaneira após o surgimento da internet e as várias possibilidades de aplicações em Línguas Naturais. (NUNES, 2008).

Após a consolidação do PLN, e a aplicação de fundamentação e formalismos linguísticos nos SPLN, percebeu-se a possibilidade de "[...] construir uma série de ferramentas que facilitam a interação do computador com o usuário, otimizam atividades de edição e tradução de texto ou permitem aos estudiosos investigar com maior precisão os fenômenos linguísticos". (SOUSA, 2015, p. 15).

Em síntese, as investigações iniciais sobre PLN tiveram como bojo, a II Guerra Mundial. Dias da Silva *et al.* (2007, p. 6), destacam ainda que

> As primeiras investigações institucionalizadas sobre o PLN começaram a ser desenvolvidas no início da década de 50, depois da distribuição de 200 cópias de uma carta, conhecida como Weaver Memorandum, escrita por Warren Weaver, então vice-presidente da Fundação Rockfeller e exímio conhecedor dos trabalhos sobre criptografia computacional.

Carta esta em que Weaver convida instituições e empresas, para desenvolvimento de projetos sobre um novo campo de pesquisa, que ficou conhecido como tradução automática, o que culminou, após dois anos, o interesse nas pesquisas nessa nova área por instituições importantes do mundo todo, a exemplo do Instituto de Tecnologia de Massachusetts (MIT), a Universidade da Califórnia, a Universidade de Harvard e a Universidade de Georgetown, instituições que desenvolveram estudos pioneiros em nível de mundo sobre o PLN.

E no Brasil, grande parte das pesquisas em PLN, estão associadas a grupos de pesquisas vinculados a Programas de Pós Graduação, em nível de Mestrado e Doutorado, embora também estejam, em uma menor escala, presentes em nível de graduação. E mesmo que a área de PLN se configure como um campo de pesquisa interdisciplinar, grande parcela desses estudos são desenvolvidos na área da Ciência da Computação.

Com o intuito de compreender/caracterizar as pesquisas e atividades em PLN desenvolvidas no Brasil, o Quadro 4, abaixo, apresenta um quadro síntese, destacando as iniciativas em PLN no Brasil, através dos grupos de pesquisa, e atividades, que se caracterizam como eventos, simpósios, encontros, comissões e afins da área de PLN no Brasil.

Quadro 4 – Eventos e Grupos de Pesquisa em PLN em nível de Brasil

TIPO	ANO	NOME	ÁREA
Evento	1993	Conferência Internacional de Processamento Computacional de Língua Portuguesa (PROPOR).	Linguística e Ciência da Computação
Núcleo	1993	Núcleo Interinstitucional de Linguística Computacional (NILC-ICMC-USP).	Ciência da Computação
Evento	2003	Encontro de Linguística de Corpus (ELC).	Letras/Linguística
Evento	2003	Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL).	Ciência da Computação
Grupo de Pesquisa	2004	Linguística Computacional, Língua e Literatura.	Linguística
Evento	2006	Seminário de Pesquisa em Ontologias do Brasil (ONTOBRAS).	Ciência da Computação
Grupo de Pesquisa	2006	Laboratório de Linguística e Inteligência Computacional (LALIC).	Ciência da Computação
Comissão	2007	Comissão Especial de Processamento de Linguagem Natural (CE-PLN).	Interdisciplinar
Evento	2007	Escola Brasileira de Linguística Computacional (EBRALC).	Letras/Linguística
Grupo de Pesquisa	2009	Computação e Linguagem Natural (CompLin).	Linguística
Grupo de Pesquisa	2010	Grupo de Pesquisa de Processamento de Linguagem Natural (PLN-PUCRS).	Ciência da Computação
Grupo de Pesquisa	2010	Estudos em Linguística de Corpus.	Linguística
Grupo de Pesquisa	2013	Núcleo de Pesquisas em Acessibilidade, Usabilidade, Linguística Computacional (Alcance).	Ciência da Computação
Grupo de Pesquisa	2013	Arquitetura da Informação, Linguística Computacional e Multimodalidade, Mídias e Interatividade (R.E.G.I.I.M.E.N.T.O.)	Ciência da Informação

Fonte: o autor (2020).

Isto posto, percebe-se uma expressiva presença dos pesquisadores brasileiros em PLN, oriundos das áreas da Linguística, Ciência da Computação e Ciência da Informação, evidenciando as características interdisciplinares da área também no Brasil, tanto quanto aos grupos de pesquisas quanto aos eventos existentes na área.

Por conseguinte, compreende-se, a partir do Quadro 4, que o ano de 1993 foi o ponto de partida para os eventos e grupos de pesquisa em PLN no Brasil, estendendo-se aos anos seguintes, e que em 2013 percebe-se uma pausa no que diz respeito a novos eventos e grupos de pesquisa na área.

Assim, dentre os grupos de pesquisa que possuem maior notoriedade, está o Núcleo Interinstitucional de Linguística Computacional (NILC), criado em 1993, vinculado ao Instituto de Ciências Matemáticas e da Computação da Universidade de São Paulo (ICMC/USP), que inclui, ainda, cientistas da computação, linguistas e pesquisadores de diversas universidades e centros de pesquisa, como a Universidade Federal de São Carlos (UFSCar), Universidade Estadual Paulista (UNESP) e Universidade Estadual de Maringá (UEM), entre outras. Destaca-se também o Grupo de Pesquisa de Processamento de Linguagem Natural da Pontifícia Universidade Católica do Rio Grande do Sul (PLN-PUCRS), criado em 2010.

Além dos grupos de pesquisas, destaca-se também como iniciativa no Brasil no que se refere aos estudos em PLN, a Comissão Especial de Processamento de Linguagem Natural (CE-PLN), aprovada durante o XXVII Congresso da Sociedade Brasileira de Computação em 2007, cujo objetivo é promover e representar a área de PLN no Brasil, com apoio e realização de eventos científicos, propondo e organizando meios de publicação e divulgação para a área, além de gerenciar listas e fóruns de discussão, dentre outras medidas.

Desse modo, dando continuidade nas discussões suscitadas nesta seção, será explorado na sessão seguinte, acerca dos aspectos teóricos e metodológicos adotados no processo de tratamento computacional da língua.

3.1 Fundamentos teóricos e metodológicos do PLN

Em razão da consolidação do PLN, enquanto área de pesquisa, após os avanços obtidos em consequência das aplicações de fundamentação e formalismos linguísticos, é possível observar uma significativa mudança no que diz respeito ao emprego de recursos, ferramentas e aplicações que objetivam o processamento computacional das línguas naturais.

Nesse sentido, para o desenvolvimento de sistemas computacionais capazes de tratar/processar as línguas naturais, é indispensável, antes, que a língua/linguagem tornem-se compreensíveis pela máquina, levando em consideração as suas facetas. Assim, é necessário incorporar a máquina, conhecimento linguístico para o tratamento das línguas naturais, como

estrutura linguística; níveis de processamento e outras informações linguísticas. (DIAS DA SILVA *et al*, 2007).

Consequentemente, são integrados a esses sistemas os níveis de conhecimento linguístico, em que a língua é processada por nível linguístico, sustentando a interpretação da língua natural pela máquina. Na abordagem linguística, os SPLN, aplicam essa fundamentação linguística, segmentando o texto de entrada em partes menores e essenciais, buscando entender as relações e funções de cada elemento lexical na sentença, e explora como esses pedaços funcionam juntos para criar significados. Esse modelo diverge dos sistemas baseados em abordagem estatística, que se caracterizam por empregar dados de frequência de ocorrências de elementos lexicais, ou seja, a frequência das palavras em textos.

Para tal, os sistemas são alimentados pelos níveis de conhecimento linguísticos necessários para a operacionalização e tratamento computacional das línguas naturais, – "[...] como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos" (GONZALEZ; LIMA, 2003, p. 9) – e, para isso, são associadas informações de natureza fonético-fonologica; morfológico; sintático; semântico; pragmático; e discursivo, ilustrados para melhor visualização e compreensão, na Figura 4. (BARROS; ROBIN, 1977; VIEIRA; LIMA, 2001; NUNES, 2008; DIAS DA SILVA *et al*, 2007; ROSA 2011).

Figura 4 – Níveis de conhecimento linguísticos em PLN

Fonte: o autor (2020).

Assim, o nível de natureza fonético-fonológico, está condicionado aos estudos dos sons. A fonologia estuda a representação dos sons mentalmente e a capacidade de atribuir distinções de significado às palavras. Já a fonética estuda os sons produzidos por humanos, assim como eles se manifestam, ou seja, a maneira como são executados e recebidos pela audição humana. Como IVO (2019, p. 2) explica, que "Os sons, quando indicados foneticamente, são representados segundo o alfabeto fonético entre colchetes; portanto a transcrição fonética das palavras *casa* e *vasa* é ['kaza] e ['vaza]. O símbolo ' indica que a sílaba seguinte possui acento". Embora possuam relações quanto aos estudos desenvolvidos, se divergem quanto aos objetos estudados, o fonema e os fones. (VIEIRA; LIMA, 2001). Logo, "Inseridas no PLN, ambas, Fonologia e Fonética, caracterizam um mesmo nível de processamento linguístico — o dos sons — e contribuem para o reconhecimento automático e a síntese da fala humana." (SOUSA, 2015, p. 22).

A morfologia caracteriza-se pelos estudos dos morfemas, "[...] as menores unidades de significado da língua, e os morfes, que são as formas com que esses significados se realizam" (IVO, 2019, p. 7). Em outras palavras, é o estudo a respeito da estrutura, formação e classificação das palavras, em que se relacionam as categorias gramaticais, como os adjetivos, substantivos, verbos, advérbios, etc. No Processamento Automático das Línguas Naturais, o nível morfológico, objetiva "[...] contribuir com a análise da estrutura das palavras, restando a níveis posteriores determinar o seu real sentido." (SOUSA, 2015, p. 22), ou seja, acontece "quando as unidades mínimas dotadas de significado são isoladas para a compreensão do processo de formação e flexão das palavras." (DIAS DA SILVA et al, 2007, p. 17).

Já o nível sintático, corresponde aos estudos desenvolvidos pela Sintaxe, que objetiva o estudo das frases, como elas são construídas e capazes de expressar o que se deseja dizer. Relaciona-se à disposição das palavras nas frases, das frases no discurso e a combinação entre uma frase e outra para que haja um sentido. No âmbito do PLN, a Sintaxe "determina o papel de cada uma das palavras de uma sentença e, assim, permite ao sistema convertê-la em estruturas mais facilmente manipuláveis" (COPPIN, 2004, p. 573 apud SOUSA, 2015, p. 25).

A semântica é a área da linguística que corresponde aos significados, que se estendem às unidades lexicais presentes nas palavras ou frases. Em SPLN, seu objetivo, "[...] é dar sentido à estrutura frasal já reconstituída em formatos inteligíveis por máquinas, para tanto, aproveitando aspectos de níveis anteriores e adicionando-lhes informações de cunho conceitual". (SOUSA, 2015, p. 27).

O nível pragmático-discursivo está também condicionado aos estudos dos significados, entretanto segue uma perspectiva divergente dos níveis sintático e semântico, pois é o ramo da linguística que se preocupa em analisar o uso concreto da linguagem pelos falantes da língua, em seus variados contextos. A Pragmática sobrepõe a significação dada às palavras pela semântica e pela sintaxe, observando o contexto extralinguístico em que estão inscritas. No Processamento Automático das Línguas Naturais, esse nível "[...] pode ser entendido, portanto, como o estágio de análise do material linguístico sob uma visão mais global: o contexto situacional e social confere acuidade à interpretação semântica e a avaliação do texto na íntegra permite deduzir aspectos não observáveis em sentenças isoladas". (SOUSA, 2015, p. 30).

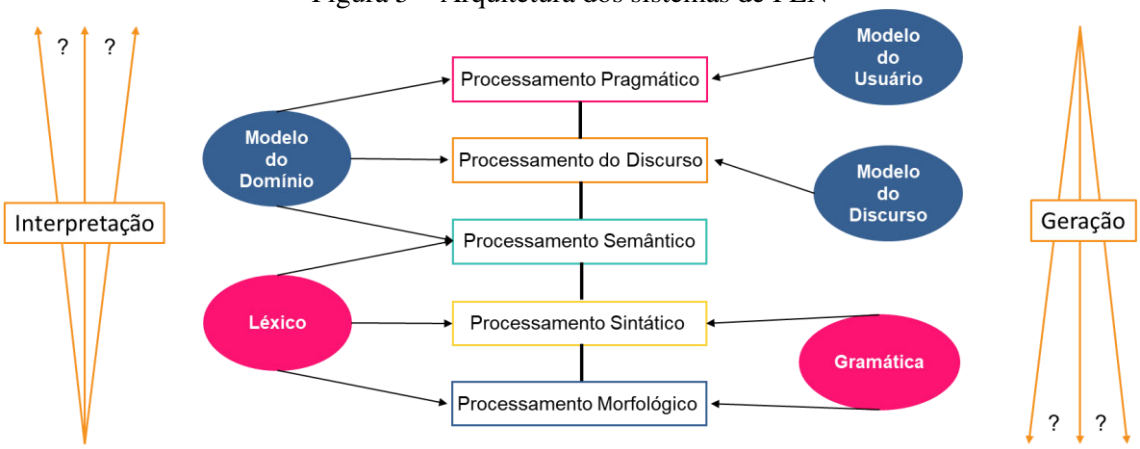
Ainda sobre os níveis de conhecimento linguístico, Vieira e Lopes (2010, p. 185), afirmam que:

> De um ponto de vista linguístico, o foco das pesquisas em PLN pode estar em um de cinco níveis de análise: (a) fonético ou fonológico, (b) morfológico, (c) sintático, (d) semântico ou (e) pragmático. Todos esses níveis possuem suas características próprias e suas dificuldades associadas, mas cada aplicação de PLN pode ter a preocupação mais voltada para um subconjunto desses níveis. Por exemplo, aplicações sobre textos científicos usualmente não têm preocupação com uma análise fonológica (a), por outro lado, aplicações que façam uma interface com reconhecimento de voz focam esse nível de análise.

Nesse sentido, com o objetivo de reduzir a complexidade de tratamento linguístico computacional por parte de determinadas aplicações, muitos sistemas de PLN dedicam-se apenas a alguns níveis de conhecimento linguístico, como é o caso das aplicações que possuem interesse em análises de textos escritos que, em geral, não fazem uso do nível fonético-fonológico.

Assim, a arquitetura dos sistemas em PLN é caracterizada conforme as atividades que executam com a língua, e subdividem-se em dois tipos, interpretação e geração. Os sistemas de interpretação possuem como objetivo principal, analisar e compreender a estrutura linguística e o significado a que está atrelado, de maneira respectiva. Enquanto os sistemas de geração destinam-se a construir sentenças ou textos maiores de forma automática. São exemplos desses sistemas analisadores de linguagens sintáticos e semânticos (*parsers*), e sumarizadores, respectivamente. (Dias da Silva *et al.,* 2007). A Figura 5, a seguir, ilustra a arquitetura de sistemas de PLN de interpretação e geração.

Figura 5 – Arquitetura dos sistemas de PLN



Fonte: o autor, adaptado de Barros e Robin, 1997.

Barros e Robin (1997) comentam sobre a característica, de modo geral, modular dos sistemas de PLN, onde são executados em módulos distintos, diferentes níveis de processamento, em que, os módulos se comunicam durante o processo do texto em análise. Logo, de acordo com a tarefa que objetiva o sistema, se é interpretar ou gerar texto, apenas o fluxo de informação muda. (BARROS; ROBIN, 1997).

Sousa (2015) destaca ainda sobre os recursos linguísticos presentes nas arquiteturas de interpretação e geração, necessários para o processamento das línguas naturais, que se caracterizam como Bases de Conhecimento (BCs) e módulos de inferência, que desdobram-se no léxico, ou seja, o conjunto das palavras que compõem uma língua e que são usadas respeitando as regras da gramática; a gramática, que a grosso modo, pode ser definida como um conjunto de regras que regem a utilização de uma língua natural; o modelo de domínio, que é a área do conhecimento em que o sistema se destina, e o modelo de usuário, referindo-se as questões que o caracteriza-o, ilustrados também na arquitetura apresentada na figura anterior.

Embora as possibilidades de contexto de uso dos Sistemas baseados em PLN tenham aumentado em decorrência dos avanços e após a sua consolidação enquanto campos de pesquisa, ainda existem desafios que limitam as possibilidades de avanços ainda maiores nesse campo. Corroborando com essa afirmativa, Oliveira Neto, Tonin e Prietch (2010, p. 4), destacam que “Ao falarmos de processamento de linguagem natural, surgem vários problemas, como a grande variação morfológica e sintática das unidades lexicais ou a ambiguidade intrínseca da língua portuguesa”.

Desafios estes que em grande parte dos casos, estão relacionados à ambiguidade intrínseca a língua natural, e que se agrava em alguns casos quando relacionada a um idioma

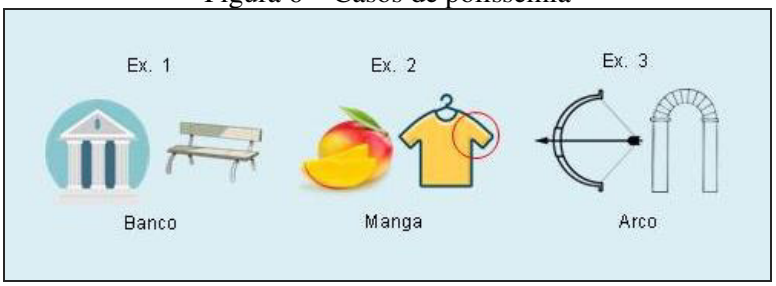
específico. Fernerda (2003, p. 86-87), fundamenta-se em Beardon, Lumsden e Holmes (1991), para apresentar as causas mais prováveis da ambiguidade, que podem ser lexical e estrutural.

A primeira está relacionada ao fato de que uma única palavra possa possuir múltiplos significados, desdobrando-se em problemas relacionados à hononímia e polissemia. Fernerda (2003, p. 87), acerca da hononímia explica que

> A hononímia ocorre entre itens lexicais com significados diferentes que possuem o mesmo som e a mesma grafia (homônimos perfeitos: como substantivo "alvo" e o adjetivo "alvo"), ou apenas o mesmo som (homônimos homófonos: como "acento" e "assento"), ou apenas a mesma grafia (homônimos homógrafos: como o verbo "seco" e o adjetivo "seco"). (Sacconi, 1999).

Já em se tratando da polissemia, ocorre quando uma mesma palavra pode adquirir diferentes significados, levando em consideração o contexto em que a mesma é empregada. Por exemplo, a palavra “banco”, pode referir-se a Instituições Bancárias ou a um objeto. A Figura 6, a seguir, ilustra essa fala no sentido de possibilitar uma melhor visualização.

Figura 6 – Casos de polissemia



Fonte: o autor (2020).

Assim, como ilustrado na figura 6, no que diz respeito à polissemia, o Ex. 1, apresenta o caso referente à palavra “banco”, que pode referir-se tanto a uma Instituição Bancária, quanto a um objeto; o Ex. 2 corresponde à palavra “manga”, que pode ser associada à manga enquanto fruta ou a manga de uma camiseta; e, por fim, o Ex. 3, ilustra a palavra “arco”, que pode se referir a um arco como objeto, mas precisamente uma arma, ou um arco enquanto um elemento arquitetônico.

No sentido da ambiguidade estrutural, esta acontece quando há a possibilidade de existir mais de uma estrutura sintática para uma única sentença, podendo ser, ainda, local, caso possa ser resolvida dispensando o contexto onde ocorre, e a global, se for necessária a análise do contexto para a resolução do problema de ambiguidade. Fernerda (2003, p. 87), explica que

> Por exemplo, na frase "ele olhou o computador com esperança" existe uma ambiguidade estrutural local. Neste caso o sentido expresso pela frase "computador com esperança" pode, em princípio, ser descartada. Em "ele olhou o colega com esperança" há ambiguidade estrutural global, sendo possível construir duas associações diferentes: "olhou com esperança" e "colega com esperança".

Outro desafio apresentado por Rosa (2011), é o que está relacionado à integração entre sistemas de conhecimento para o PLN. De acordo com o autor, a interpretação da língua natural requer cooperatividade de várias aplicações de conhecimentos linguísticos, que vão desde conhecimentos específicos sobre o uso da língua, até conhecimentos sobre o mundo real em que é utilizada, o contexto, situações em cotidianas em que a língua é empregada. Assim, Rosa (2011, p. 160, grifo do autor), destaca como problemas:

> **Ambiguidade**. A ambiguidade é o maior problema no Processamento de Línguas Naturais. Existem basicamente duas abordagens para tratamento de sentenças ambíguas: o *backtracking*, como usado nas redes de transição aumentadas (ATNs) de Woods (1970), e o *delay*, usado no analisador *espere e veja*.
>
> **Interpretação única**. Outro fenômeno interessante na interpretação da língua é que as pessoas podem considerar apenas uma interpretação de uma sentença ambígua de cada vez, mas podem facilmente saltar entre interpretações.
>
> **Erros de compreensão**. Erros na compreensão, como as sentenças enganosas (*garden path*), têm sido explicados por princípios estruturais puros. Entretanto, existem explicações mais completas e naturais como os efeitos colaterais de processos fortemente interativos. Os efeitos enganosos podem ocorrer em todos os níveis do processamento da língua.
>
> [...]
>
> **Texto não gramatical**. As pessoas são capazes de interpretar linguagem não gramatical se esta ocorrer naturalmente (devido a gramática pobre, estrangeirismos, inferência de barulho externo, interrupções, autocorreções etc.).

Em contrapartida a esses e outros desafios que se fazem presentes no Processamento Automático de Línguas Naturais, existem diversas vantagens no que se diz respeito ao uso e aplicação de sistemas baseados nesse modelo, principalmente nos dias atuais, em decorrência do massivo número de informações disponíveis e a geração em massa de dados não estruturados, principalmente em ambientes digitais/web, fazendo cada vez mais necessário o uso de PLN em conjunto com outras ferramentas, com o intuito de solucionar as necessidades dos usuários da rede de computadores. Assim, a próxima seção, apresenta os principais usos e aplicações de sistemas baseados em PLN.

3.2 Usos e aplicações do PLN

Tem-se ampliado bastante o desenvolvimento de diversas tarefas e atividades que possuem, como pressupostos, o auxílio de sistemas baseados em PLN, com objetivos e fins distintos. Nesse sentido, o impacto dessas ferramentas para o desenvolvimento da ciência,

tecnologia e até mesmo para uma vida menos mecanizada, apresenta ganhos positivos em vários âmbitos das sociedades.

Os SPLN, em conjunto com outros domínios da IA e demais áreas do conhecimento, estão ganhando espaços significativos nas pesquisas cientificas, no mercado de trabalho, e principalmente no cotidiano das pessoas, que são beneficiadas com a implementação de tecnologias que facilitam tarefas até então estritamente desenvolvidas por humanos. As quais se destacam, o *Machine Learning*, *Deep Learning*, *Big Data*, *Data Sciense*, *Machine Learning*, *Internet of Things* e outros.

Rosa (2011) expõe que geralmente, o PLN pode ser divido em seis grandes áreas, a saber, interfaces em línguas naturais para bases de dados; tradução de máquina de uma língua natural para outra; programas de indexação inteligentes para sumarização de grandes quantidades de textos; geração de texto para produção automática de documentos padrões; sistemas de fala para permitir a interação de voz com computadores; e por fim, ferramentas para desenvolver sistemas de PLN para aplicações específicas. Áreas essas que se desdobram em uma diversidade de usos e aplicações.

Nessa perspectiva, existe uma gama de possibilidades no que diz respeito aos usos e aplicações do Processamento Automático de Línguas Naturais, partindo desde sistemas de reconhecimento de fala, que objetivam a interação diretamente com o usuário, como é o caso dos Assistentes Virtuais (como por exemplo, Google Assistent – Google Inc.; Cortana – Microsoft; Alexa – Amazon; Siri – Apple, BIA – Bradesco, e outros), até projetos mais específicos de *Analytics* voltados para variados contextos de aplicação. A Figura 7, a seguir, ilustra duas possibilidades de uso em PLN, com as suas respectivas aplicações.

Figura 7 – Usos e aplicações baseadas em PLN

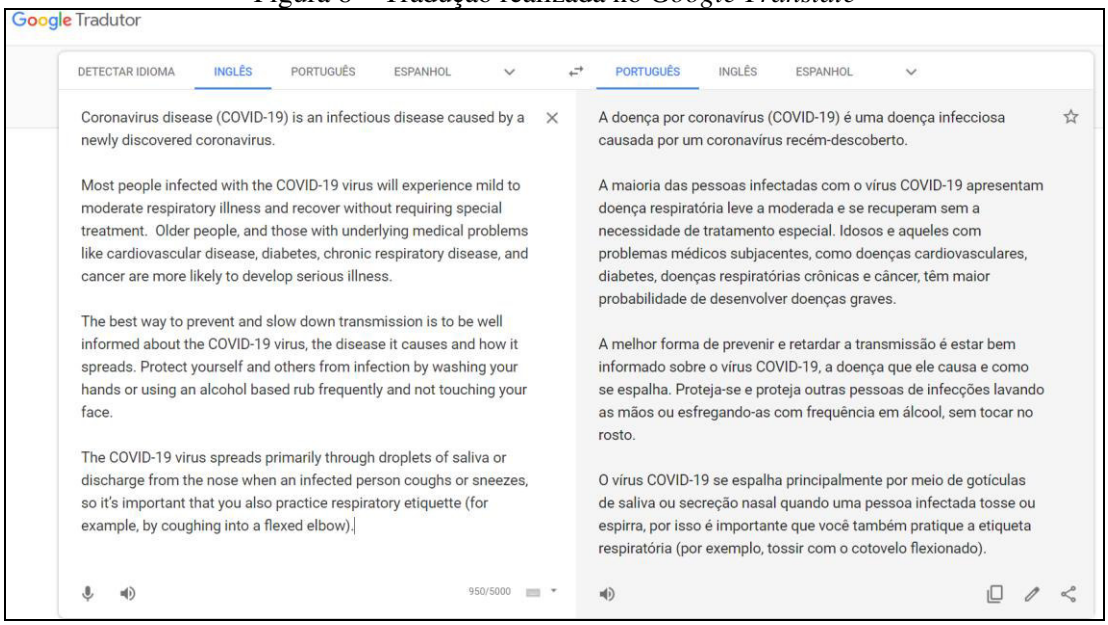
Fonte: o autor (2020).

A Figura 8, ilustra o funcionamento do sistema de Tradução Automática do Google, o *Google Translate* na sua versão *web*. A construção da interface do tradutor é simples e didática, com vistas à usabilidade.

Assim, basta apenas transferir o texto que deseja ser traduzido – (nessa versão, o limite de caracteres é de 5000) – e selecionar a opção do idioma que deseja que ele se encontre após

o fim do processo, ou fazer *upload* do documento, se assim for o caso. O *Google Translate*, é ainda um tipo de SPLN que integra à sua arquitetura outros tipos de aplicações, como por exemplo, o conversor de texto em fala e fala em texto.

Figura 8 – Tradução realizada no *Google Translate*



Fonte: *Google Translate* (2020).

A ilustração da figura 8, apresenta a tradução de um texto em inglês, retirado da página da Organização Mundial da Saúde (do inglês *World Health Organization)*[1], em que são apresentadas diretrizes e uma visão geral acerca do COVID-19.

Posto isto, os usos e aplicações apresentados na literatura, – até mesmo as terminologias utilizadas para empregá-los e as categorias em que são apresentados –, divergem bastante, tendo em vista a característica inter/multidisciplinar presente nos estudos em PLN, as inúmeras motivações que atraem pesquisadores de diversas áreas do conhecimento para os estudos em PLN, levando também em consideração as possibilidades das ferramentas baseadas nesse modelo.

Vieira e Lima (2001) apresentam como usos e aplicações de PLN, reconhecedores e sintetizadores de fala, que objetivam a transcrição da fala em texto aplicados, por exemplo, em serviços automatizados de informação por telefone; corretores ortográficos e gramaticais; tradutores automáticos; geradores de textos e resumos; interfaces em língua natural, aplicados à manipulação de base de dados; recuperação da informação; extração da informação; avaliação de sistemas de processamento de linguagem natural e processamento de linguagem natural e sistemas multiagentes (articulação em múltiplas fontes de conhecimento).

[1] https://www.who.int/health-topics/coronavirus#tab=tab_1

Ainda nessa perspectiva, Nunes (2008), exemplifica aplicações que envolvem ou podem envolver o PLN, como Sistemas de TA; Sistemas de Sumarização Automática (SA); Sistemas de Categorização de Textos; Sistemas de Recuperação da Informação (SRI); Sistemas de Extração de Informação (EI); Sistemas de Diálogos e Sistemas de Auxílio à Escrita.

Por outro lado, Oliveira Neto, Tonin e Prietch (2010), apresentam esses sistemas a partir de duas grandes categorias: aqueles que possuem a função de processar textos escritos, como é o caso dos editores de texto, sistemas de busca de páginas na *Web*, e os sistemas de ajuda à tradução; e aqueles que possuem a função de tratar a linguagem natural, que se desdobram em sistemas baseados em textos, em que destaca os sistemas de recuperação da informação; tradutores; e resumidores. E, por fim, as aplicações baseadas em diálogos.

Sousa (2015) corrobora com os autores citados acima, ao apresentar a tradução automática; sumarização automática; os sistemas de recuperação da informação; sistemas de extração da informação e sistemas de diálogos como aplicações usuais de PLN. Entretanto, colabora ao acrescentar, ainda, os sistemas de perguntas e respostas, e os analisadores de sentimentos como aplicações de PLN.

Grael (2019), ao classificar os usos e aplicações de PLN por categorias de tarefas realizadas por essas aplicações, apresenta esses sistemas por outra perspectiva. Para ele, são divididos em três grandes categorias, a partir das tarefas realizadas, que podem ser de naturezas Sintáticas, Semânticas e de Discurso e Fala. Assim, na categoria de Sintáticas, estão os Sistemas de Segmentação; Partes do Discurso *(Part-of-Speech)*; e Análise de Dependência. Na categoria Semântica, estão os sistemas de Tradução Automática; Reconhecimento de Entidades Nomeadas; Representação Semântica; Análise de Tópicos; e Análise de Sentimentos. E, por fim, na categoria de Discurso e Fala, estão os sistemas de Sumarização Automática e de Diálogos, ou como também são conhecidos, *Chat boots*.

Existem ainda algumas aplicações que estão ganhando espaço e notoriedade em SPLN, como por exemplo, os sistemas de reconhecimento de textos em imagens; categorização de conteúdos; descoberta e modelagem de tópicos e sistemas de conversão de fala-texto e texto-fala.

Em suma, há uma percepção de que os sistemas baseados em Processamento Automático de Línguas Naturais possibilitam as máquinas, computadores e outros periféricos digitais, leiam textos, ouçam e interpretem falas, identifiquem sentimentos e determinem inferências com base em contextos reais ou não reais, realizando essas tarefas em grande parte, de forma consistente e imparcial, questões que até mesmo para os humanos, em alguns

casos, são caracterizadas difíceis. O Quadro 5, a seguir, apresenta uma síntese das ferramentas baseadas em modelos de aplicações de Processamento Automático de Línguas Naturais.

Quadro 5 – Usos e aplicações de PLN

USOS E APLICAÇÕES	FUNÇÃO/OBJETIVOS	FERRAMENTAS BASEADAS NESSE MODELO
Analisadores de sentimentos	Consiste em extrair informações de textos em linguagem natural, com o objetivo de classificá-lo como positivo, negativo ou neutro. Utilizado por empresas, para análise de sentimentos de opiniões em redes sociais.	Algoritmo Naive Bayes Analisador de Sentimento LK Hugme LinguaKit
Corretores ortográficos e gramaticais	Objetiva verificar a correta construção ortográfica e gramatical das frases; as regras de concordância da língua, através de consulta em vocabulários e dicionários incorporados ao sistema.	Flip 9 Ginger Software Grammarly Hemingway App LanguageTool Microsoft Word (editor de texto que integra esse modelo de PLN) Reverso SpellBoy
Geradores de textos e resumos Sumarizadores automáticos (SA)	Objetiva a produção automática de resumos a partir de um ou mais textos.	Esummarizer LinguaKit Resoomer Smmry Text Compactor Turbine Text
Reconhecedores e sintetizadores de fala (Assistentes Virtuais) Sistemas de diálogos (Chatter Bots)	Objetiva a interação direta entre usuário-máquina/aplicação, através de comandos de texto escrito ou fala/voz.	Alexa – Amazon Cortana – Microsoft Eliza Google Assistent – Google Google Now – Google LiveChat Siri – Apple
Reconhecimento de textos em imagens (Ferramentas baseadas em OCR)	Objetiva realizar o reconhecimento de textos em imagens automaticamente.	Amazon Rekognition Camscaner Google Cloud OCR OCR Instantly OCR Tesseract
Sistemas de auxílio à escrita	Ambientes que objetivam o auxílio à produção de texto, em que os usuários podem encontrar recursos para construção de textos estruturados, de um gênero e/ou domínio específicos.	CALeSE SciEn-Produção SCiPo SCiPo-Farmácia
Sistemas de conversão texto-fala e fala-texto	Convertem textos em fala ou fala em textos.	Soar PiliApp AlReader Oratlas Bolabolka Natural Reader
Sistemas de extração de informação (EI)	Buscam a resposta em um ou mais documentos a partir de uma pergunta de entrada.	LinguaKit

Tradutores automáticos (TA)	Possuem a finalidade de traduzir textos ou falas de um idioma para outro, automaticamente pela máquina.	Babylon DeepL Google Translate; Linguee Microsoft Translator Reverso

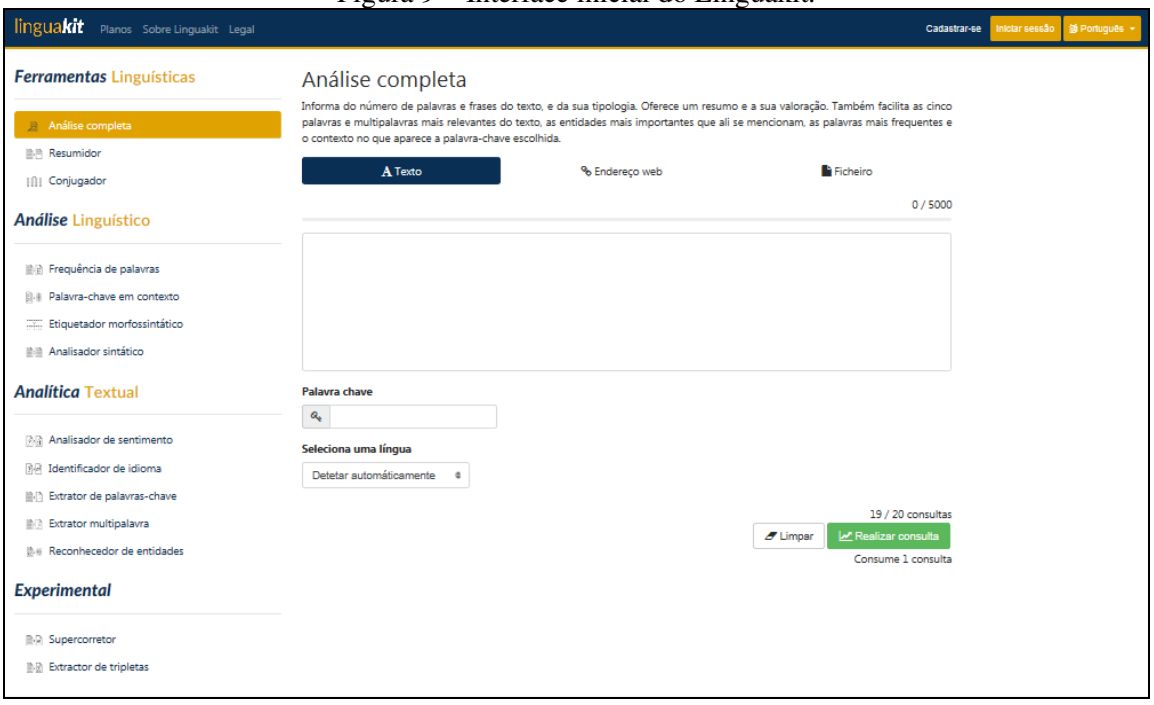
Fonte: o autor (2020).

Amparado nas informações apresentadas no quadro 5, é possível perceber que essas aplicações baseadas em PLN, possuem características bem diversas, pois são ferramentas que se desdobram em sistemas, softwares, aplicativos e aplicações propriamente ditas.

É comum, nos dias atuais, encontrar ferramentas que possuem várias aplicações de PLN integradas, como é o caso de editores de texto, por exemplo, o *Word* do Pacote *Office* da *Microsoft*, que possui integrado à arquitetura de seu sistema, a correção ortográfica e gramatical, ou ainda, o *Google Translate*, ilustrado anteriormente, que é uma aplicação de tradução automática, que incorpora também, sistema de conversão texto-fala e fala-texto, e correção ortográfica e gramatical.

Ainda nessa perspectiva, destaca-se também a ferramenta para *Web* e aplicação para dispositivos móveis, como o LínguaKit. O LinguaKit é suíte multilíngue que compõe ferramentas de análise sintática, análise de sentimentos, geração automática de resumos, extração de informações e outras aplicações baseadas em PLN, em uma única ferramenta, como apresentado na Figura 9:

Figura 9 – Interface inicial do Linguakit.

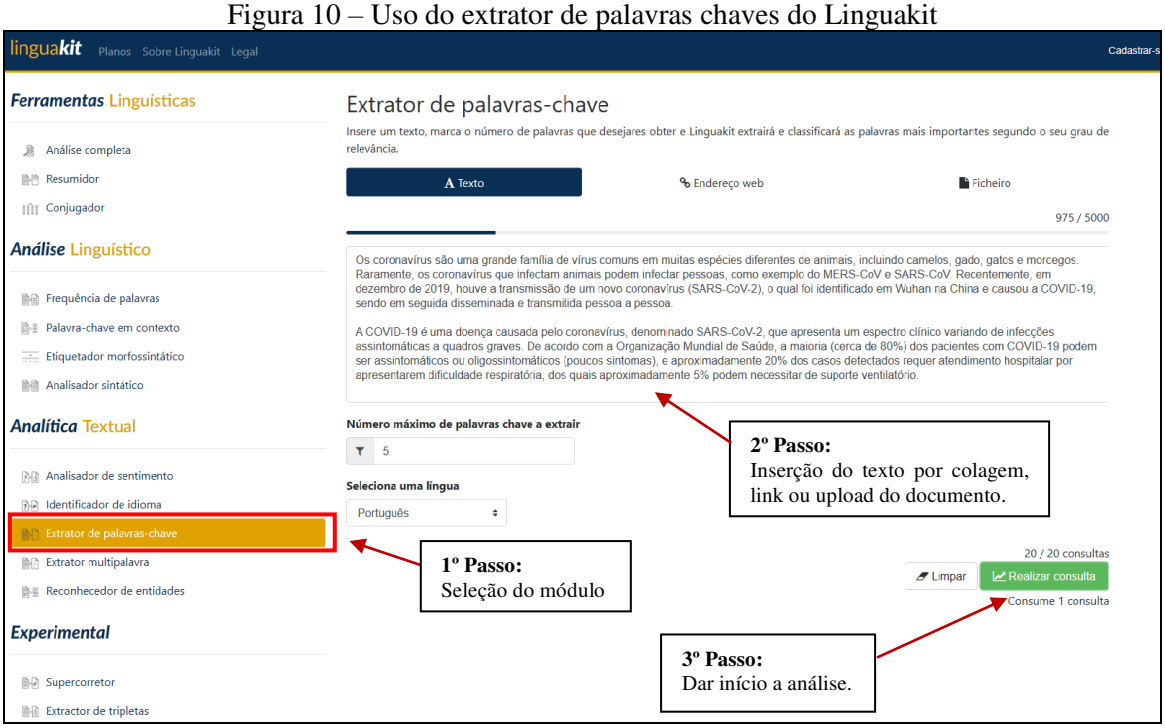


Fonte: Linguakit (2020).

Como é possível visualizar na Figura 9, o Linguakit integra à sua arquitetura várias aplicações de PLN, apresentados em módulos linguísticos, organizados por categorias, inicialmente as Ferramentas Linguísticas, que incluem Análise Completa, Resumidor e Conjugador, destinados a atender aspectos mais genéricos da linguagem. O segundo módulo, Análise Linguística, inclui frequência de palavras, palavra-chave em contexto, etiquetador morfossintático e analisador sintático. O terceiro, analítica textual, que inclui analisador de sentimento, identificador de idioma, extrator de palavras-chave, extrator multipalavra e reconhecedor de entidades; e, por fim, o módulo experimental, que apresenta as novas ferramentas do projeto, inclui o supercorretor e o extractor de tripletas.

A figura 10, e a figura 11, ilustram o processo de extração de palavras chaves do Linguakit. Inicialmente seleciona-se o módulo de palavras-chave extrator, e em seguida, é inserido o texto no campo destinado – que pode ser inserido também através de *link* ou *upload* do documento que contém o texto a ser analisado –, após isso, define-se o número máximo de palavras-chave para serem extraídas, o idioma em que o texto se encontra, e, por fim, inicia-se a análise como ilustrado nas figuras seguintes.

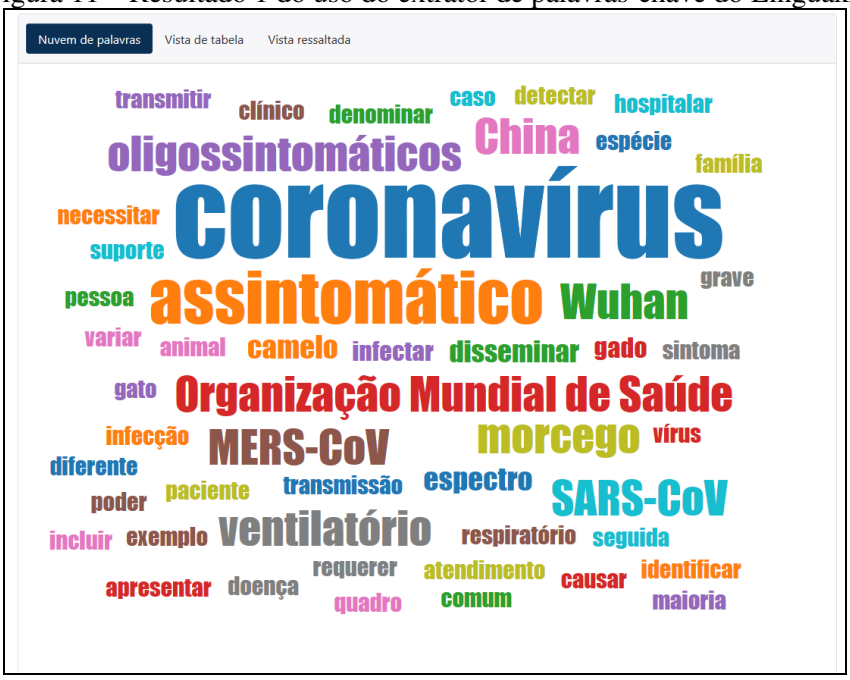
Figura 10 – Uso do extrator de palavras chaves do Linguakit



Fonte: Linguakit (2020).

Após seguir os passos acima destacados, ao concluir a análise, o Linguakit apresenta o resultado da análise do texto a partir do módulo selecionado ao qual o mesmo foi submetido, como ilustrado na figura 11, a seguir.

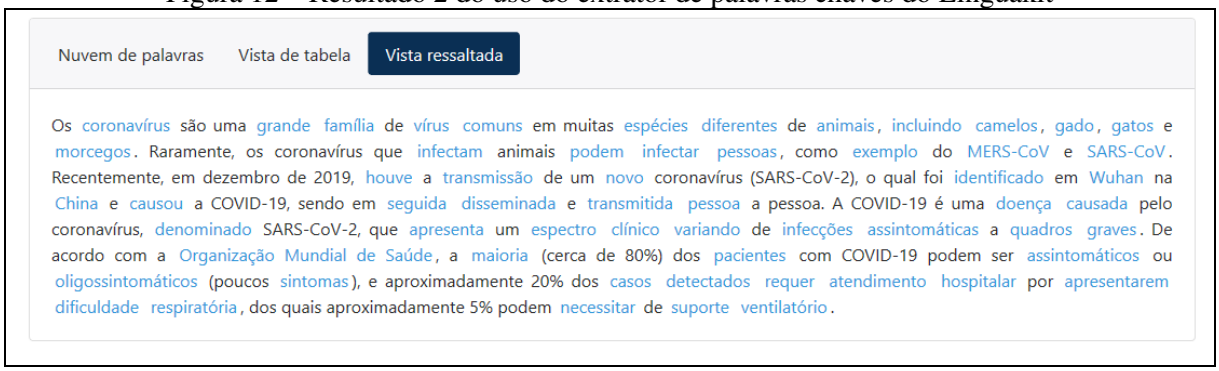
Figura 11 – Resultado 1 do uso do extrator de palavras-chave do Linguakit



Fonte: Línguakit (2020).

Logo, o resultado da análise, no módulo de extração das palavras-chave, é apresentado em três formatos. O primeiro é em nuvem de *tag*, como ilustrado na figura 11, em que o destaque dado às palavras-chave é proporcional à sua representatividade no texto fonte. O segundo formato, é a visualização em tabela, em que é apresentado um ranking com as palavras-chave mais representativas, e por fim, o terceiro formato, a visão destacada, ilustrado na figura 12, em que as palavras-chave mais representativas são destacadas em azul no próprio texto fonte.

Figura 12 – Resultado 2 do uso do extrator de palavras chaves do Linguakit



Fonte: Línguakit (2020).

Assim, tendo em vista as possibilidades das aplicações apresentadas a partir do conhecimento dos diversos usos e aplicações possíveis para os sistemas baseados em Processamento Automático de Línguas Naturais, e, em especial, as possibilidades apresentadas na ferramenta Línguakit, destacam-se também, aqueles que provocam questionamentos quanto a possibilidades e aplicações em Sistemas de Recuperação da

Informação. Nesse sentido, a subseção a seguir abordará o contexto de usos e aplicações desses sistemas na área de Ciência da Informação, voltados para processos de Organização da Informação e do Conhecimento.

3.3 PLN na Ciência da Informação

Em Ciência da Informação, o PLN tem sido estudado em uma concepção teórico-metodológica, entretanto, em virtude das possibilidades apresentadas nos avanços dos estudos nessa área, têm ganhado espaço em uma perspectiva prática, em especial no campo da "[...] Indexação e Recuperação da Informação, pela possibilidade de softwares baseados nesse modelo propiciarem a extração de termos com maior precisão semântica para recuperação da informação em sistemas de buscas automatizados [...]" (CARMO; CONCEIÇÃO, 2018, p. 317), potencializando as tarefas automatizadas com o emprego do PLN.

Nesse sentido, percebe-se um considerável destaque dessas técnicas de Processamento Automático de Línguas Naturais no contexto da Organização, Representação e Recuperação da Informação, tendo o emprego dessas ferramentas, voltado para Sistemas de Recuperação da Informação; Análise de Informações/Dados, e outros. (LADEIRA; ALVARENGA, 2012).

Desse modo, Barbosa e Kobashi (2017), salientam para as possibilidades de uso da visualização de dados no cenário da recuperação da informação, a partir das funções de extroversão e descoberta.

Enquanto Rampão e Tsunoda (2019) apontam para a mineração de dados aplicada a tarefas de classificação e associação e descoberta de padrões através da análise de grandes conjuntos de dados, no contexto específico de bases de dados jurídicas.

Vieira e Lima (2001) apontam para aplicações de PLN voltadas para SRI. De acordo com os autores, a RI é uma área que se envolve com a recuperação de documentos relevantes, dado um determinado tema, e não necessariamente com a recuperação de informação específica ou com a obtenção de respostas a dadas perguntas.

Assim, em relação ao uso do PLN em face da Recuperação da Informação, Fernerda (2003), o aponta como possível solução aos problemas referentes a RI, – levando em consideração a natureza do PLN –, pela simples observação de que, tanto os documentos como as expressões de busca, utilizados para recuperá-los em sistemas automatizados, são objetos linguísticos. Nhacuongue (2020, *online*) corrobora nessa perspectiva, ao propor "[...] estratégias de recuperação da informação baseadas no PLN, para extrair relações semânticas

da WordNet.Pt, e utilizá-las na representação de documentos e de expressões de busca dos usuários".

Ainda nesse âmbito, Câmara Júnior (2013, p. 74), fala que "A área de RI pode ser vista, sob alguns aspectos, como uma aplicação de sucesso de PLN. O crescimento rápido, desordenado e acachapante da Internet só foi possível por causa de motores de busca livres, disponíveis e efetivos, a maioria desenvolvida com técnicas de PLN", sucesso este relacionado, a possibilidade de esses sistemas de RI, baseados em modelos de PLN, realizarem a manipulação dos termos de entrada no sistema, traduzindo-os para a linguagem utilizada pelo sistema, e recuperar os documentos que sejam satisfatórios à pesquisa realizada pelo usuário.

Nunes (2008) destaca ainda que com o crescimento da Web, as aplicações em SRI ganharam bastante destaque, tendo o sistema de busca do Google, um dos mais conhecidos representantes. O grande destaque, para ela, está na possibilidade de hoje, o usuário ser livre para formalizar suas consultas em língua natural, em contraste, às linguagens rígidas de busca utilizadas anteriormente (a exemplo do SQL). Possibilidade esta, que também se faz presente em SRIs utilizados em Unidades de Informações, a exemplo de Bibliotecas, Museus e Arquivos.

Assim, em módulos de busca em Sistemas de Recuperação da Informação, as consultas realizadas de maneira erradas ou inexatas são processadas através de aplicações de PLN, que transformam os termos de busca por termos próximos ou semelhantes do correto, possibilitando uma recuperação mais precisa. (LIMA; NUNES; VIEIRA, 2007).

Transcendendo esse cenário, em torno da recuperação de informações e documentos, existem aplicações direcionadas a processos de organização e representação da informação. É o caso dos sistemas de extração e mineração de informação/dados ou mineração de textos, que podem ser aplicados no intuito de auxiliar a execução de atividades de indexação; construção de vocabulários controlados; sumarização de informação, destinado à interpretação de conteúdo dos documentos; e análise de conteúdo. (LADEIRA, 2010; LADEIRA; ALVARENGA, 2012). Tal como, o uso de ferramentas de PLN para a construção dos *Simple Knowledge Organization System* (SKOS).

Pardo (2008) corrobora com essa perspectiva, ao destacar que, além dos sistemas baseados em perspectivas de extração e mineração de dados, existem também sistemas baseados em processos que permitem a condensação de textos, com a finalidade de possibilitar a compreensão do conteúdo mais importante existente no texto, a exemplo das ferramentas de Sumarização Automática.

Em contrapartida, à luz das possibilidades apresentadas pelos fundamentos teóricos e metodológicos expostos na subseção anterior, é possível vislumbrar o uso dessas aplicações em outros ramos da Ciência da Informação, em que, podemos citar a possibilidade de uso de Chat Boots e Assistentes Virtuais em Sistemas de Referência Virtual; Aplicações de PLN em motores de busca de Bases de Dados e Repositórios Institucionais; e o uso de api's baseados em modelos de PLN voltados para tradução automática, integrados também em bases de dados.

Nesse sentido, a fim de caracterizar as pesquisas em PLN na área de Ciência da Informação, o Quadro 6, apresenta dados em formato de referência bibliográfica, de teses e dissertações desenvolvidas na área de Ciência da Informação, que possuem como foco de pesquisa vertentes do Processamento Automático de Linguagem Natural. O levantamento dos dados disponibilizados no quadro a seguir, foram possíveis através de buscas realizadas na Biblioteca Digital Brasileira de Teses e Dissertações – BDTD.

Quadro 6 – Pesquisa em PLN na CI

REFERÊNCIA
CÂMARA JÚNIOR, Auto Tavares da. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília, Brasília, 2007.
BRUZINGA, Graciane Silva. **Indexação automática de documentos textuais**: proposta de critérios essenciais. 2009. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2010.
NASCIMENTO, Gustavo Diniz. **Dos sintagmas nominais aos descritores documentais**: estudo de caso na indexação de teses e dissertações da área de Direito. 2015. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, Recife, 2015.
CELERINO, Victor Galvão. **Proposta de normalização dos sintagmas nominais em termos para indexação automática**. 2018. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, Recife, 2018.
SOUZA, Renato Rocha. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2005.
MAIA, Luiz Claudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2008.
LADEIRA, Ana Paula. **Processamento de linguagem natural**: caracterização da produção científica dos pesquisadores brasileiros. 2010. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2010.
MARTINS, Agnaldo Lopes. **O uso do sintagma nominal na recuperação de documentos**: proposta de um mecanismo automático para classificação temática de textos digitais. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

MESQUITA, Luiz Antonio Lopes. **Sintagmas nominais na indexação automática**: uma análise estrutural da distribuição de termos relevantes em teses de doutorado da UFMG. 2012. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2012.
CAMARA JUNIOR, Auto Tavares da. **Processamento de linguagem natural para indexação automática semântico-ontológica**. 2013. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2013.
SILVA, Tiago José da. **Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa**. 2014. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, Recife, 2014.

Fonte: o autor (2020).

São evidentes as múltiplas possibilidades do emprego de sistemas baseados em modelos de PLN na Ciência da Informação, levando em consideração as várias frentes que cada uma das áreas se dispõe a investigar e a natureza inter/multidisciplinar em que estão. Entretanto, após o resultado e sistematização das buscas, percebemos ainda uma predominância das pesquisas quanto às temáticas, pois grande parte delas estão voltadas para estudos entorno de processos cujo objetivo é contribuir para a indexação.

Notamos ainda que, de acordo com as informações apresentadas no quadro 6, grande parte dos estudos estão atrelados aos Programas de Pós Graduação em Ciência da Informação – a nível de mestrado e doutorado –, da Universidade Federal de Minas Gerais, em seguida da Universidade Federal de Pernambuco e Universidade de Brasília. Percebemos a região que mais tem produzido estudos nessa temática é a região Sudeste, seguida da Nordeste e Centro-Oeste.

Nesse contexto, Ladeira (2010) realiza um estudo em que caracteriza a produção científica nacional da área em PLN ao longo dos últimos 40 anos a partir da análise de assunto de artigos de revisão revelados no ARIST. Em que reforça a modesta participação da Ciência da Informação nos estudos da área, e aponta para a necessidade de tornarem-se conhecidos os estudos já realizados, com o intuito de, permitir a aplicação dessas ferramentas estudadas "[...] nos processos clássicos de catalogação e posterior recuperação nos centros informacionais, assim como na concretização de modelos abstratos de representação de informação inerentes ao campo". (LADEIRA, 2010, p. 249).

Desse modo, entendemos a importância e necessidade da presença dos estudos sobre PLN na CI. Logo, na seção a seguir, será apresentado o percurso metodológico deste estudo.

4 DESCRIÇÃO METODOLÓGICA

A pesquisa cientifica é uma atividade que tem por objetivo aproximar a teoria da realidade, a partir do relacionamento estabelecido entre aquela, e os dados desta. (MORESI, 2003). Nesse sentido, a metodologia da pesquisa compreende o conjunto de procedimentos que sustentam a realização da atividade de uma pesquisa científica, os quais serão descritos a seguir.

4.1 Caracterização da pesquisa

Prodanov e Freitas (2013, p. 48), destacam que "A pesquisa científica é uma atividade humana, cujo objetivo é conhecer e explicar os fenômenos, fornecendo respostas às questões significativas para a compreensão da natureza.". E para esta tarefa, o pesquisador necessita utilizar de conhecimentos anteriores acumulados, além de métodos, técnicas e instrumentos que garantam e proporcionem a cientificidade necessária para obter as respostas inerentes às suas indagações enquanto pesquisador.

Desse modo, este estudo, é caracterizado como uma pesquisa de natureza aplicada, exploratória, em que utiliza como procedimentos técnicos e meios de investigação a pesquisa bibliográfica e documental, optando pela abordagem qualitativa.

Entende-se por pesquisa aplicada quando esta "objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos" (PRODANOV; FREITAS, 2013, p. 51), tendo em vista a viabilidade da aplicação prática dos resultados obtidos para contribuir em determinada realidade. Em que, no presente contexto, objetiva-se gerar conhecimentos no âmbito da Sumarização Automática de Textos aplicadas a processos de organização e representação da informação.

A pesquisa exploratória por sua vez, proporciona maior familiaridade com o problema e assunto em que está sendo desenvolvida a pesquisa, pela sua natureza de "sondagem". Além de que "Seu planejamento é, portanto, bastante flexível, de modo que possibilite a consideração dos mais variados aspectos relativos ao fato estudado.", (GIL, 2010, p. 41), o que possibilita o desenvolvimento dos estudos sob diversos ângulos e aspectos.

A pesquisa bibliográfica "[...] abrange toda bibliografia já tornada pública em relação ao tema de estudo, desde publicações avulsas, boletins, jornais, revistas, livros, pesquisas, monografias, teses, material cartográfico etc. [...]" (MARCONI; LAKATOS, 2003, p. 181), e a "A pesquisa documental assemelha-se muito à pesquisa bibliográfica. A diferença essencial

entre ambas está na natureza das fontes." (GIL, 2010, p. 45). Assim, no que diz respeito ao desenvolvimento e sustentação teórica desta pesquisa, utilizou-se artigos científicos, teses, dissertações, monografias, livros e afins, e quanto à pesquisa documental, faz uso de relatórios técnicos científicos, slides de apresentações em eventos científicos e outros materiais desta natureza.

Quanto à abordagem qualitativa, Prodanov e Freitas (2013, p. 70), afirmam que esta "[...] considera que há uma relação dinâmica entre o mundo real e o sujeito, isto é, um vínculo indissociável entre o mundo objetivo e a subjetividade do sujeito que não pode ser traduzido em números.". Os autores destacam ainda que a compreensão dos fenômenos, assim como a atribuição de significados são elementos básicos no processo da pesquisa qualitativa. Em outras palavras, esse tipo de pesquisa busca entender um problema específico em profundidade, e, para isso, trabalha com descrições, comparações e interpretações.

Esta pesquisa se configura também como interdisciplinar, por oportunizar a integração de duas ou mais áreas do conhecimento humano. Assim,

> A interdisciplinaridade corresponde a uma nova consciência da realidade, a um novo modo de pensar, que resulta num ato de troca, de reciprocidade e integração entre áreas diferentes de conhecimento, visando tanto à produção de novos conhecimentos, como a resolução de problemas, de modo global e abrangente. (FAVARÃO; ARAÚJO, 2004, p. 107).

A presença da interdisciplinaridade na pesquisa cientifica, oportuniza uma relação de integração entre duas ou mais áreas do conhecimento, e, na presente pesquisa, caracterizam-se pelas áreas da Biblioteconomia e Ciência da Informação, Linguística, Linguística Computacional e Ciência da Computação, que desenvolvem estudos nas áreas da Organização e Representação da Informação e do Conhecimento, Processamento Automático de Línguas Naturais e a Sumarização Automática de Textos, respectivamente.

Logo, a escolha dos *softwares* utilizados nessa pesquisa se deu a partir de vários critérios técnicos. Inicialmente, quanto a abordagem no quesito formação, optamos por sumarizadores extrativos, que são aqueles que extraem sentenças do texto-fonte para a construção do sumário. Em seguida, quanto à categoria, houve uma escolha por um *software* fundamentado em iniciativas de *software* livre e aplicação *web*; e outro baseado em uma iniciática comercial, em sua versão para *desktop.* Assim, utilizamos o Turbine Text e o Intellexer Summarizer Network Edition.

Dando continuidade aos métodos, a organização do corpus foi realizada a partir das diretrizes de Almeida (2006) que expõe a necessidade de haver um conjunto de requisitos de forma a garantir a validade e confiabilidade do corpus. Assim, teve-se como ponto de partida, a concepção de que "Para organizar um corpus, parte-se, inicialmente, da seleção dos textos

pertinentes e relevantes para a pesquisa, bem como dos gêneros aos quais eles pertencem." (ALMEIDA, 2006, p. 88).

A autora destaca também, a importância da variação de gêneros textuais na composição do corpus, tendo em vista as possibilidades de representatividade na comunicação de determinados domínios. Assim, o corpus adotado para realização deste estudo, é constituído por cinco artigos científicos sobre a temática do Covid-19, após a realização de busca e seleção na base do Portal de Periódicos da Capes no mês de junho, com a utilização de termos como: COVID-19; Corona Vírus; Pandemia; Novo Corona Vírus, tendo como critério de seleção, a ordem de relevância em que foram recuperados. Em seguida, o material foi organizado e estruturado em arquivos .txt, que após esses passos, foram submetidos aos dois sumarizadores Turbine Text e Intellexer Summarizer Network Edition, onde, definimos o percentual de texto a ser mantida no resumo em 10%. O resultado do processamento do corpus submetido aos sumarizadores, está disponível no Apêndice A – Artigos Sumarizados.

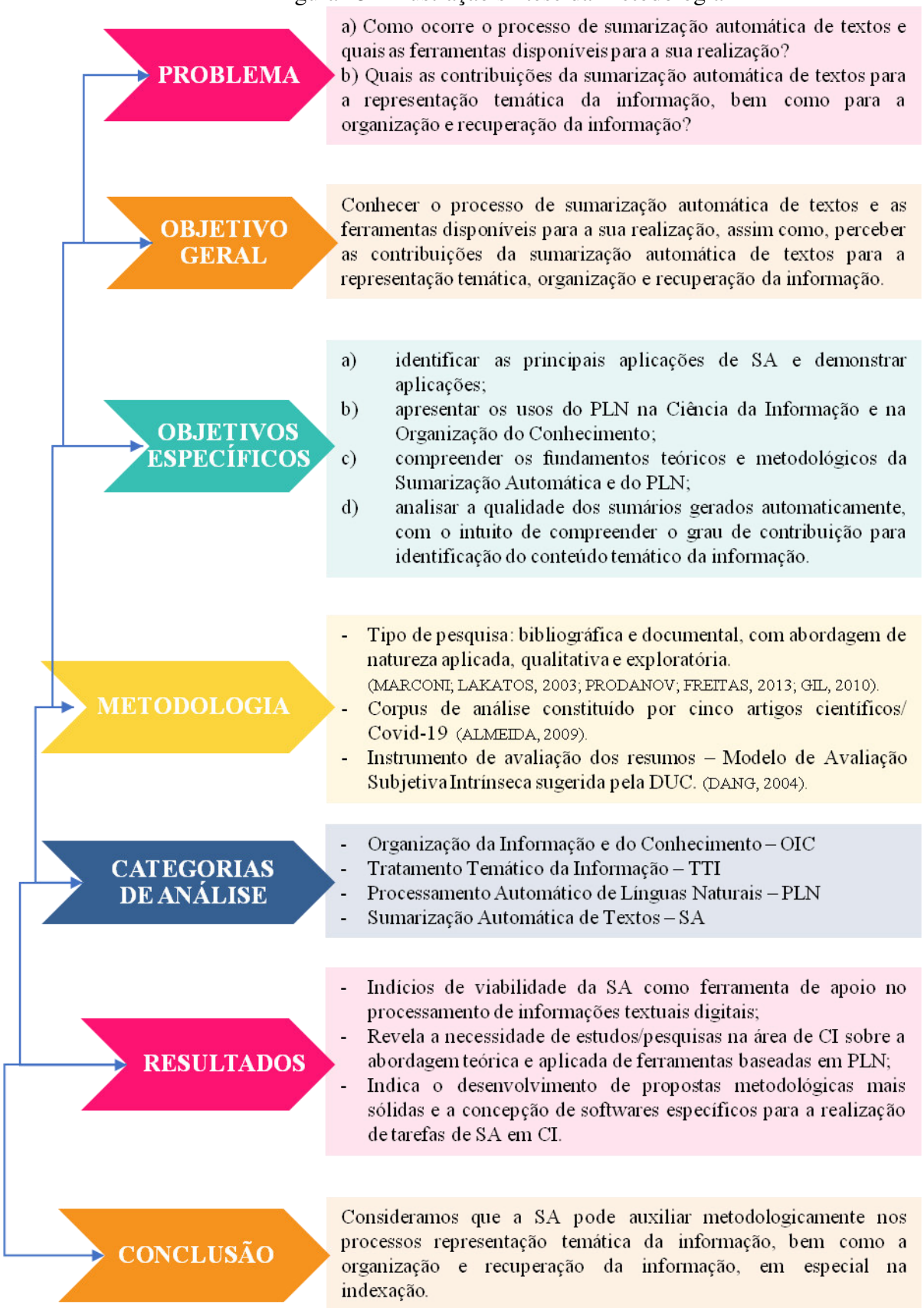
Quanto ao instrumento de avaliação destes resumos, optamos por um Modelo de Avaliação Subjetiva Intrínseca, sugerido pela DUC[2], este tipo de metodologia busca "[...] julgar os sumários de acordo com critérios qualitativos, como legibilidade, qualidade de conteúdo ou utilidade. (LEITE, 2010, p. 13-14). Assim, [...] verifica-se o desempenho do sistema por meio da análise dos seus sumários". (CARDOSO, 2014, p. 25).

A avaliação dos resumos foi pautada em cinco categorias linguísticas, a saber: gramaticalidade, redundância, clareza referencial, foco, estrutura e coerência, tendo como parâmetro uma escala de 1 (muito ruim); 2 (ruim); 3 (aceitável); 4 (bom) e 5 (muito bom). Essas categorias buscam avaliar quesitos de conteúdo e de textualidade, caracterizando-se como medidas de qualidade. (DANG, 2005; LEITE, 2010; CARDOSO, 2014).

Assim, a síntese da metodologia utilizada para a realização deste trabalho, pode ser visualizada na Figura 13.

[2] A *Document Understanding Conference* (DUC) surgiu em 2001, em resposta ao progresso dos estudos em SA e da necessidade de realização de avaliações de sumários automáticos. Em 2008, a DUC foi substituida pela *Text Analysis Conference* (TAC), que é uma conferência internacional, composta por uma série de oficinas de avaliação que objetivam promover estudos na área de PLN e de áreas similares vinculadas à ela.

Figura 13 – Ilustração síntese da Metodologia



Fonte: o autor (2020).

5 SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS NA REPRESENTAÇÃO DA INFORMAÇÃO: fundamentos

É inegável a necessidade de instrumentos que viabilizem o acesso à crescente dinâmica de informações, e que, além disso, possibilite uma pré-visualização, no sentido de dar um panorama sobre a estrutura e conteúdo dessas informações, com o intuito de auxiliar o acesso e permitir um uso acertado. (CABRAL, 2015).

Grande parte dessas informações estão disponíveis em mídias digitais, em ambiente *web* e na própria *internet*, através de bases de dados, repositórios institucionais e outras fontes de informação/sistemas de recuperação da informação, acessados de forma *online*. E, ainda hoje, destacam-se aquelas em formato textual, em que temos como exemplo os textos acadêmicos; as notícias de jornais; livros literários; blogs e outros. Nesse sentido, um tipo de ferramenta que pode contribuir para o acesso a essas informações disponíveis em formato textual, é a Sumarização Automática de Texto (SA).

A SA é uma subárea de pesquisa em PLN, e ocupa-se da produção automática de sumários, a partir de um ou mais textos-fonte, e propõe-se a simular a produção humana de resumos, levando em consideração inclusive a riqueza apresentada no desenvolvimento de sumários por humanos. O intuito é possibilitar o acesso à informação de maneira clara e concisa, propiciando economia de tempo e recursos. (PARDO, 2008).

Assim, Sumários Automáticos são produtos de Sumarizadores Automáticos, que são softwares, baseados em PLN, que possuem a intenção de condensar grandes volumes de informações, em informações mais sucintas e mais acessíveis, segundo determinados requesitos.

Nesse contexto, Rino e Pardo (2003), destacam que um sumarizador automático pode ser definido como, “Um sistema computacional cujo objetivo é produzir uma representação condensada do conteúdo mais importante de sua entrada, para consumo por usuários”. E, para isso, deve ser capaz de: identificar a representação conceitual de um texto; o que é importante, assegurando a consistência e coerência do sumário.

Submetendo assim, textos de entrada, por processos de análise, transformação e síntese, culminando em um texto de saída mais sucinto, condensado e que apresente as principais informações extraídas da estrutura de um ou mais documentos, como exemplificado na Figura 14.

Figura 14 – Etapas do processo de sumarização

Fonte: Cardoso (2014, p. 25).

Em face disso, tem despertado interesse em algumas áreas de estudo, especialmente daquelas que possuem como insumo a informação. É o caso das áreas da Biblioteconomia, Ciência da Informação, Recuperação da Informação, Ciência da Computação e outros, em face das possibilidades e das aproximações com atividades desenvolvidas nessas áreas do conhecimento. Em Biblioteconomia e Ciência da Informação, por exemplo, vêm ganhando espaço principalmente pela possibilidade de serem empregados com o intuito de auxiliar em processos inerentes a representação da informação e do conhecimento.

Sabendo disso, Souza *et al* (2017), ressaltam que o processo de SA de textos é constituído por etapas, que se assemelham ao de indexação de documentos, em que o grande desafio é a extração de termos que representem e que possuam uma boa cobertura temática dos conteúdos abordados nos documentos. (LANCASTER, 2004).

Santos (2012, p. 7), contribui nesse sentido, ao esclarecer que "A investigação nesta área foi inicialmente motivada pela necessidade de conseguir indexar o crescente número de publicações científicas, uma vez que na altura os recursos tecnológicos não permitiam o armazenamento em grande escala de documentos em formato digital".

Entretanto, é necessário esclarecer as concepções teóricas do que seja um sumário. Para Pardo (2008, p. 2), "[...] em SA, sumário pode tanto se referir a índice quanto a resumo propriamente dito". Já Ribeiro (2016), afirma que existem divergências fundamentais entre resumos, sumários e sumários automáticos. Resumos seriam textos elaborados pelo próprio autor do texto-fonte; sumários, índices; e sumários automáticos, textos de saída produzidos pelos sistemas de sumarização automática.

Já a ABNT NBR 6023:2003, norma que estabelece os requisitos para redação e apresentação de resumos em nível de Brasil, define resumo como uma "Apresentação concisa dos pontos relevantes de um documento". (ASSOCIAÇÃO..., 2003, p. 1). Percebe-se então, as aproximações conceituais entre resumos e sumários que, em tese, possuem o mesmo objetivo: apresentar informações sucintas de um ou mais documentos.

Assim, no que diz respeito ao escopo desta pesquisa, será adotado o conceito de sumários como resumos de documentos, tendo em vista que o ato de sumarizar é atrelado a atividades de condensação de informações, logo, sumários automáticos, serão entendidos como resumos gerados de forma automática com o auxílio de sumarizadores automáticos, que serão abordados de modo mais detalhado na subseção seguinte.

Ainda que os sumários estejam presentes em determinadas realidades, como é o caso dos textos acadêmicos, – em que podemos citar: artigos científicos, monografias, teses, dissertações, relatórios técnicos e outros –, que geralmente são acompanhados de resumos como elementos pré-textuais ou possuem o resumo como metadado de descrição, o que possibilita a introspecção do conteúdo apresentado com o leitor em seu primeiro contato; esse recurso pode ainda alcançar outros materiais textuais, principalmente em especial a representação das informações disponíveis, em se falando do tratamento temático da informação. (SIMONASSI, 2015).

Por conseguinte, a elaboração de resumos dos documentos que estão sendo indexados é uma tarefa importante na determinação do conteúdo temático destes, tendo como pressuposto que o "principal objetivo do resumo é indicar de que trata o documento ou sintetizar seu conteúdo." (LANCASTER, 2004, p. 6).

Acerca desta aplicação dos resumos no processo de indexação, Lancaster (2004, p. 6) afirma ainda que "A indexação de assuntos e a redação de resumos são atividades intimamente relacionadas, pois ambas implicam a preparação de uma representação do conteúdo temático dos documentos.". Além disso, Pardo (2008, p. 2), afirma que, "[...] o uso de sumários pode melhorar certos aspectos da recuperação da informação e da categorização de textos", tendo em vista que

> Com a enorme evolução das ferramentas de Recuperação de Informação (RI) passou a ser fácil o rápido acesso a quantidades enormes de informação, no entanto, o processamento de tal quantidade de informação requer meios automáticos eficazes de condensação de informação, pelo que de outra forma é humanamente impossível filtrar conteúdo relevante de forma eficiente. (SANTOS, 2012, p. 7).

Porém, nos dias atuais, é inviável desenvolver atividades de elaboração de resumos para auxiliar na descrição de documentos, levando em consideração principalmente a quantidade cada vez maior de documentos disponíveis para serem tratados, a urgência do acesso a esses documentos por parte dos usuários, além de outras questões, como a própria falta de recursos disponíveis nas Unidades/Centros de Informação, quanto a recursos humanos, financeiros, materiais, e a disponibilidade de tempo.

E é nesse contexto que o desenvolvimento desta pesquisa se sustenta, no sentido de inferir qualidade e avaliar o grau de contribuição para a Análise Conceitual de documentos, a partir dos resumos gerados por ferramentas de SA, no intuito de entender a possibilidade de esses servirem de auxílio para atividades de representação da informação e do conhecimento, a exemplo da indexação. Assim, a subseção a seguir, tem por objetivo aprofundar as discussões sobre a SA e fazer as caracterizações necessárias.

5.1 Tipos de SA e ferramentas disponíveis

Pardo (2008, p.1) afirma que, "Os sumários comumente chamados resumos, podem ser produzidos por diversas estratégias e pelo uso de conhecimentos de naturezas diversas", o que leva os sumários a serem classificados sob a ótica de diferentes pontos de vista, como por exemplo, quanto à função, audiência e formação, além é claro de levar em consideração, o fim para o qual se destinam.

No quesito classificação, quanto à sua função, os sumários podem ser classificados como informativos, indicativos ou críticos. Sumários informativos possuem características que conferem textualidade aos textos, ou seja, apresentam as principais informações do texto organizadas de forma coerente e coesa, além de terem uma boa progressão temática, e serem gramaticais e legíveis. Sumários indicativos, em contrapartida, possuem uma abordagem mais superficial, no sentido de possibilitarem apenas uma ideia geral do que o texto-fonte trata, não dispensando a consulta ao original. Já os sumários críticos, expressam opiniões para além do conteúdo disposto no texto fonte. (PARDO, 2008).

As caracterizações realizadas por Pardo (2008), no que se diz respeito à classificação dos sumários, referem-se exatamente às características e conceitos apresentados pela ABNT NBR 6023:2003, que estabelece as diretrizes para elaboração de resumos e conceitua os mesmos, em crítico, indicativo e informativo. (ASSOCIAÇÃO..., 2003).

Destaque para Rino e Pardo (2003, p. 6), quando afirmam sobre a funcionalidade dos sumários, em que, “[...] sumários indicativos podem ser utilizados na classificação de documentos bibliográficos, de um modo geral, indicando seu conteúdo e agilizando o acesso às informações relevantes. Nesse caso, eles servem de *indexadores*.”. Enquanto os informativos “[...] por serem autocontidos, servem de meio de informações, porém apresentam uma relação mais complicada com seu texto-fonte, pois de seu objetivo dependerá bastante a avaliação sobre o quanto ele atende às necessidades do usuário”. (RINO; PARDO,

2003, p. 6). Defendendo a ideia de que ambos podem servir para diversas aplicações, porém, destaca-se em especial a área de Recuperação da Informação.

Pardo (2008) os classifica também quanto à sua audiência e diz que estes podem ser genéricos, quando apresentam as informações mais importantes do texto-fonte, sem se preocuparem com os leitores; e focados nos interesses dos leitores, quando customizam informações em função dos conhecimentos dos leitores acerca de determinada temática apresentada no texto-fonte.

Outra classificação dada por Pardo (2008) é quanto à formação dos sumários, que podem ser extratos ou *abstracts*. Extratos são sumários construídos a partir de trechos exatos retirados do texto-fonte, e adaptados através de uma justaposição, em que o Sistema de Sumarização Automática (SSA), faz apenas as adaptações necessárias no sentido de inferir sintaxe e semântica. Já os *abstracts* são sumários que passam por algum tipo de modificação na estrutura dos textos extraídos do texto-fonte, ou seja, são reescritos para compor o texto de saída.

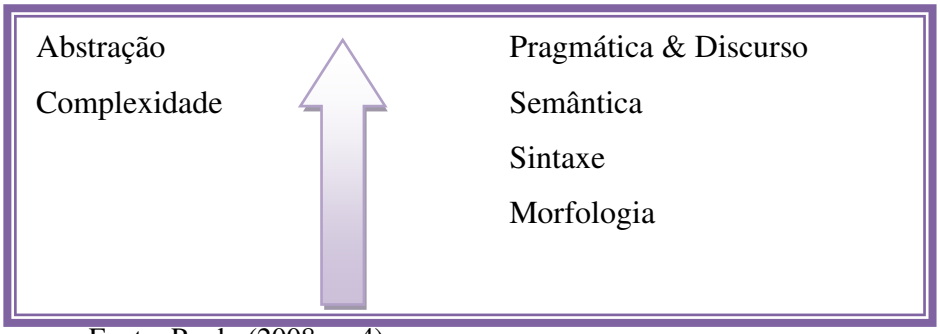
Quanto às abordagens, os sumários podem ser classificados sob a ótica de duas principais abordagens do PLN, a partir da quantidade e do nível de conhecimento linguístico que utilizam, sendo caracterizados quanto à abordagem superficial e abordagem profunda. (Martins *et al,* 2001; Pardo, 2008). Nesse sentido, tais abordagens podem ser definidas por Cardoso (2014, p. 24), da seguinte maneira:

> A abordagem superficial utiliza dados estatísticos ou empíricos e pouco conhecimento linguístico para encontrar a informação principal. Por exemplo, um método que produz um sumário a partir da seleção e justaposição das sentenças que possuem as palavras dos títulos dos textos-fonte é dito superficial.
>
> [...]
>
> A abordagem profunda, por sua vez, caracteriza-se por métodos mais sofisticados, que utilizam modelos linguísticos, regras gramaticais, conhecimento semântico, discursivo e de mundo.

A abordagem profunda conforme Pardo (2008), faz uso de analisadores sintático-semânticos e discursivos, no intuito de identificar as partes mais importantes do texto. Entretanto, existe ainda a possibilidade de uma abordagem híbrida que ocorre quando as duas abordagens, anteriormente citadas, unem-se dando origem a uma nova abordagem.

Em SA, assim como na área maior de PLN, existem níveis de conhecimento utilizado, em termos de abstração linguístico-computacional e sua complexidade para o tratamento computacional. A Figura 15, a seguir, exemplifica os níveis e sua organização.

Figura 15 – Níveis de conhecimento linguístico em SSA



Fonte: Pardo (2008, p. 4).

Assim, quanto mais acentuada é a subida da morfologia em direção à pragmática e ao discurso, maior é a abstração linguístico-computacional, fazendo com que mais difícil se torne obter uma representação formal do nível de conhecimento, e, assim, processar computacionalmente os níveis linguísticos se tornam mais complexos, recorrendo às discussões apresentadas na sessão anterior em que se trata de maneira mais específica sobre o tratamento linguístico computacional em SPLN.

Em se tratando do número de textos de entrada processados, o processo de sumarização pode ser classificado em monodocumento e multidocumento. Logo, a SA monodocumento produz o sumário de um único texto-fonte; enquanto a multidocumento, produz sumário a partir de um conjunto de textos-fonte, e que, por conta desta abordagem, tem ganhado destaque, em vista das possibilidades em torno da quantidade progressiva de informações disponíveis atualmente. (PARDO, 2008).

A esse respeito, Pardo (2008, p. 5), destaca ainda que

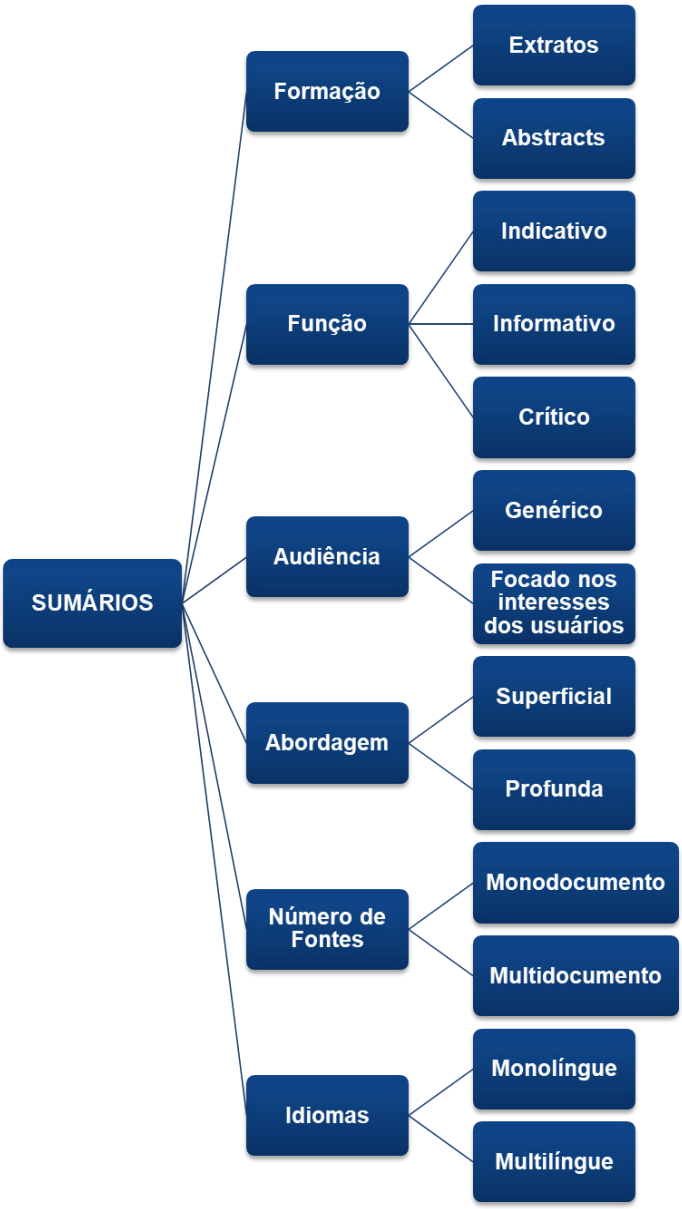
> Na SA multidocumento, além de se identificar o que é informação importante e irrelevante no conjunto de textos, novos desafios surgiram e desafios antigos se tornaram mais complexos, por exemplo, eliminação de informação redundante do sumário, ordenação (temporal ou não) dos segmentos textuais que comporão os sumários, fusão de segmentos textuais com informações complementares, manutenção da coerência do sumário, etc. Deve-se levar em conta também que os textos podem se originar de fontes diferentes e, em geral, são escritos por pessoas diferentes e, portanto, têm estilos diversos.

Cardoso (2014) adiciona, ainda, a essas classificações, uma que é referente à língua do sumário, que, de acordo com ele, podem ser monolíngue ou multilíngue. “A SA monolíngue processa textos-fonte em uma língua e produz um sumário nessa mesma língua. Na SA multilíngue, os textos-fonte podem estar em duas ou mais línguas e o sumário poderá ser em qualquer uma das línguas dos textos de origem.”. (CARDOSO, 2014, p. 24).

Assim, levando em consideração a abordagem de Pardo (2008) e de outros autores apresentados anteriormente, é possível compreender a amplitude e complexidade envolvidas no processo de Sumarização Automática de Texto (SA), no sentido de sistemas baseados

nesse modelo, alcançarem os objetivos propostos. A Figura 16, a seguir, apresenta de forma sintetizada as classificações apresentadas até o momento.

Figura 16 – Esquema-síntese das classificações de sumários



Fonte: Adaptado de Pardo (2008); Martins *et al* (2001); e Cardoso (2014).

É importante destacar que nos dias atuais existem Sistemas de Sumarização Automática capazes de produzir bons sumários, e, que, independente da abordagem utilizada para a construção dos sumários, podem ser empregados para diferentes tarefas, e, a partir disso, atribuem importância para as atividades desenvolvidas.

Assim, destacam-se diversas iniciativas no sentido de oferecer sumarizadores capazes de cumprir com os objetivos que lhes são propostos, em que se destacam principalmente, iniciativas de ordem comercial e acadêmica, como apresentados no Quadro 7, a seguir:

Quadro 7 – Sumarizadores

SUMARIZADOR	DESCRIÇÃO	CATEGORIA
Esummarizer	Criado em 2019, é uma iniciativa comercial que oferece uma ferramenta de sumarização com assinatura mensal.	Proprietário/comercial
FreeSummarizer	Esta ferramenta cria um sumário extrativo baseado nas frequências das palavras. O serviço é gratuito.	Proprietário
Gist Summarizer	Primeiro sistema de SA superficial disponível para o português	Acadêmico
Intellexer Summarizer NE	A ferramenta comercial chamada Intellexer Document Sumarizer (Intellexer Inc., 2014) é uma aplicação desktop em duas versões diferentes: um de uso geral e outra profissional (Pro).	Proprietário/Comercial
Linguakit	Suíte Multilíngue que integra à sua arquitetura o módulo de sumarização automática de textos.	Livre/Gratuito
Neural Summarizer	Sumarizador superficial baseado em uma rede neural artificial, mais especificamente, uma rede de Kohonen.	Acadêmico
Resoomer	Criado em 2020, é um sumarizador automático.	Livre/gratuito
SMMRY	A ferramenta foi desenvolvida em PHP, funciona online e como uma API tendo como entrada arquivos de texto (txt) ou hipertexto (HTML), produzindo uma saída do mesmo tipo de arquivo.	Livre/gratuito
SummarizeThis™	Uma ferramenta de produtividade, oferecida de forma gratuita pela iniciativa Iris Reading.	Livre/Gratuito
Summary Generator	Sumarizador de texto online gratuito baseado em software de sumarização de texto de código aberto.	Livre/gratuito
Text Summarizer	Criada em 2016, é uma API de sumarização de textos, baseada em tecnologias avançadas de PLN e aprendizado de máquina.	Api Livre/gratuito
Turbine Text	Sumarizador online, criado em 2015, possibilita a sumarização automática de textos em outros idiomas, como o Alemão, Português, Inglês e etc.	Livre/Gratuito

Fonte: o autor (2020).

Existe uma vasta oferta de softwares no que diz respeito aos sumarizadores, que variam de acordo com as necessidades dos usuários e finalidades para as quais serão empregados. Assim, os sumários apresentados no quadro anterior foram selecionados, a partir das suas características e das possibilidades que oferecem. Em síntese, a arquitetura geral dos Sistemas de Sumarização Automática possui como etapas do processo de sumarização, como já apresentada anteriormente na Figura 14.

Nessas etapas, a fase de interpretação compreende a análise e interpretação de um texto ou textos originais, no sentido de construir uma representação formal possível de ser

processada computacionalmente/automaticamente. A fase de transformação é responsável por realizar o processo de sumarização, com base na representação realizada pela fase anterior, produzindo a representação interna do sumário, ainda em um formato computacional. E, por fim, a fase de síntese, expressa em língua natural, a versão final do sumário produzido automaticamente. (PARDO, 2008; RINO; PARDO, 2003; SANTOS, 2012).

Com base nos fundamentos teóricos e metodológicos apresentados até então, e a fim de exemplificá-los, a Figura 17 – Ilustração de texto-fonte, mostra um texto original completo e a Figura 18 – Ilustração do sumário da Figura 17, exibe seu sumário gerado automaticamente pelo sistema *Text Compactor*, ferramenta gratuita de resumo on-line, criada em 2016, para ajudar os leitores em dificuldades a processar grandes quantidades de informações.

Figura 17 – Ilustração de texto-fonte

Mulher com 120 anos, 12 filhos e 100 netos vence a Covid-19

Uma mulher de 120 anos venceu a Covid-19 em Sirnak, na Turquia. Mãe de 12 filhos e avó de mais de 100 netos, Menica Encu, ficou duas semanas hospitalizada, mas recebeu alta e já voltou para casa. As informações são do Daily Star.

Menica agradeceu ao hospital infantil, que foi adaptado com uma instalação improvisada para Covid-19. “Eles cuidaram muito bem de mim aqui. Graças aos esforços deles, fiquei boa”, disse a matriarca à Agência Anadolu.

“Minha sogra geralmente só come produtos orgânicos. Ela se recuperou depois de receber atenção médica muito boa. Deus abençoe os médicos e enfermeiras que cuidaram dela”, afirmou Halime Encu, de 65, também à Agência Anadolu.

De acordo com o pneumologista Dr. Bilim Kehya Akcan, Menica respondeu bem ao tratamento e teve alta. Ele acrescentou que a idosa deve continuar a cuidar da saúde durante as próximas semanas, pois ainda está fraca, mas a equipe do hospital avalia que ela está bem o suficiente para ser deixada em casa aos cuidados de sua família.

Fonte: ISTOÉ (2020, on-line).[3]

Assim, o resultado do processamento do texto, se apresenta da seguinte maneira:

Figura 18 – Ilustração do sumário da Figura 17

Mãe de 12 filhos e avó de mais de 100 netos, Menica Encu, ficou duas semanas hospitalizada, mas recebeu alta e já voltou para casa. Deus abençoe os médicos e enfermeiras que cuidaram dela”, afirmou Halime Encu, de 65, também à Agência Anadolu.

De acordo com o pneumologista Dr. Bilim Kehya Akcan, Menica respondeu bem ao tratamento e teve alta. Ele acrescentou que a idosa deve continuar a cuidar da saúde durante as próximas semanas, pois ainda está fraca, mas a equipe do hospital avalia que ela está bem o suficiente para ser deixada em casa aos cuidados de sua família.

Fonte: Text Compactor (2020).

Após as ilustrações anteriores, é possível perceber o potencial uso dos resumos/sumários produzidos a partir de Sistemas de Sumarização Automática, principalmente em se tratando de atividades que visam à condensação de grandes volumes de

[3] Texto disponível em: https://istoe.com.br/mulher-com-120-anos-12-filhos-e-100-netos-vence-a-covid-19/.

informações textuais para fins específicos, como, por exemplo, atividades que envolvem a representação e organização da informação, especificamente no núcleo do tratamento temático da informação.

Nesse sentido, na subseção seguinte, serão detalhados os processos realizados durante o desenvolvimento da parte aplicada da pesquisa. Assim, serão descritos os processos realizados, tal como, os critérios utilizados para as escolhas dos *softwares*; para a criação do corpus; o passo a passo para a submissão do corpus nos SSA, assim como as suas respectivas operacionalizações e demais informações relevantes.

6 USO DA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS NA REPRESENTAÇÃO DA INFORMAÇÃO: apresentação e análise dos resultados

Nessa seção, apresenta-se o uso da sumarização automática de textos aplicados na representação da informação. Assim, a escolha dos *softwares* baseados em modelos de PLN e consequentemente de SA, voltados para o escopo desta pesquisa, se deu a partir dos critérios: quanto à abordagem no quesito formação, em que optamos por sumarizadores extrativos, pois possuem um maior número de ferramentas baseadas nesse modelo.

Sistemas extrativos selecionam as frases mais relevantes de um documento, e, a partir dessa seleção, formam o resumo. O segundo critério foi pautado na iniciativa de *software* livre e aplicação *web*; e outro baseado em uma iniciativa comercial, em sua versão para *desktop.* Levamos em consideração também para a escolha, o fato de selecionar softwares pouco utilizados no âmbito acadêmico, com o intuito de avaliar o desempenho destes.

Inicialmente, a escolha recaiu sobre o Resoomer e o Text Compactor, seguindo os critérios estabelecidos anteriormente. Entretanto, problemas técnicos tornaram impossível a realização do experimento utilizando essas ferramentas.

Optamos então, por utilizar o Turbine Text e o Intellexer Summarizer Network Edition. O Turbine Text é uma ferramenta de sumarização em versão *web*, criado em 2015, que possibilita a sumarização automática de textos em outros idiomas, como o Alemão, Português, Inglês etc. Tais softwares são disponibilizados na versão gratuita e em versões com planos mensais e anuais.

O Intellexer Summarizer Network Edition, é um sumarizador da EffectiveSoft, empresa situada na Rússia e EUA, e atualmente está disponível em três formatos, o Intellexer API; o Intellexer SDK e o aplicativo de desktop. É um software robusto, que apresenta algumas soluções semânticas no intuito de inferir uma melhor qualidade aos resumos gerados automaticamente, este sumarizador é oferecido em uma versão gratuita e em versões pagas.

Nesse sentido, tomou-se como corpus escolhido para realização deste estudo, cinco artigos científicos voltados para a temática da Covid-19, após buscas realizadas no Portal de Periódicos da Capes com a utilização de termos como: COVID-19; Corona Vírus; Pandemia; Novo Corona Vírus, tendo como critério de seleção a ordem de relevância em que foram recuperados.

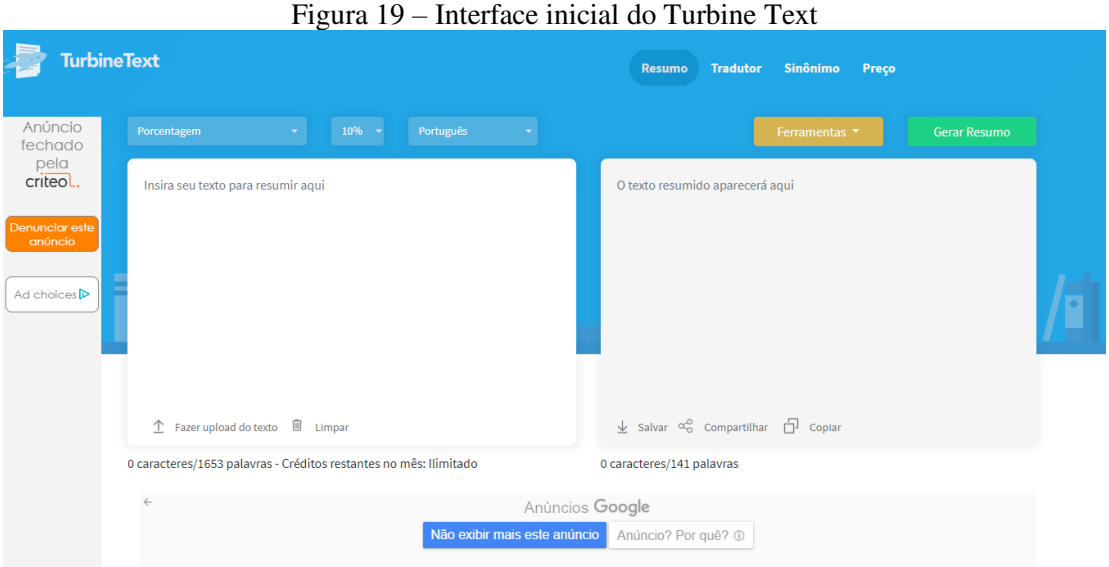
Convém esclarecer a escolha da tipologia documental, pois embora os artigos científicos já possuam resumos como elementos pré-textuais obrigatórios, a escolha por essa tipologia, se deu em torno dos critérios de qualidade já estabelecidos e aceitados pela

comunidade acadêmica, tendo em vista que o texto cientifico possui certo rigor quanto a escrita, e quanto a sua estrutura lógica. E quanto à temática, a escolha se deu em virtude do momento adverso em que o mundo está passando, a quantidade de informações produzidas em todos os âmbitos, quer seja acadêmico, comercial, jornalístico, e outros, com o intuito de satisfazer uma inquietação que se revelou durante o desenvolvimento da pesquisa, que é compreender também o comportamento dos SSA frente a novas terminologias.

Após a montagem do corpus, foi feito um tratamento inicial, no sentido de eliminar quaisquer informações dos textos selecionados que pudessem interferir no processo de sumarização, como, por exemplo, informações referentes a título do periódico em que foram publicados, os resumos já existentes, etc., e salvos em formato .txt, uma vez que é o formato mais aceito pelas ferramentas baseadas em PLN.

Nesse sentido, a fim de tornar conhecida a interface dos sumarizadores utilizados no experimento pesquisa, as figuras a seguir, ilustrarão os *softwares*, assim como algumas operacionalizações importantes.

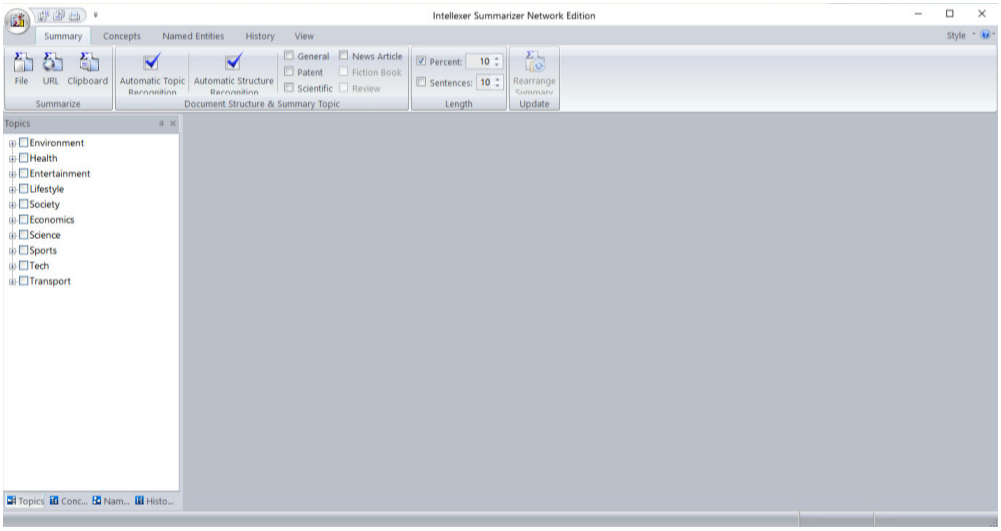
Figura 19 – Interface inicial do Turbine Text



Fonte: Turbine Text (2020).

Como é possível perceber através da ilustração, a interface inicial do Turbine Text é bem intuitiva e simples. A caixa de texto inicial é simples, sem muitas atribuições, contém apenas algumas informações necessárias, e algumas opções de operacionalização. Do lado esquerdo, encontram-se a opção de escolha de porcentagem, idioma e, do lado direito, outras ferramentas disponíveis e o botão “Gerar Resumo”, que é o comando para que o texto possa ser processado e sumarizado. A Figura 20, a seguir, ilustra a página inicial da ferramenta.

Figura 20 - Interface inicial do Intellexer Summarizer Network Edition



Fonte: Intellexer Summarizer Network Edition (2020).

A interface do Intellexer Summarizer Network Edition tambem é simples e intuitiva, embora apresente outras opções. O Texto de entrada pode ser adicionado via upload, url ou clipboard. O sumarizador em questão apresenta opções quanto à estrutura semântica do resumo, a exemplo do reconhecimento automático de tópicos, reconhecimento automático de estrutura e a opção de escolha quanto ao gênero textual do texto de entrada.

Além de permitir ao usuário a escolha quanto à extração do resumo, e pode ser pela porcentagem ou número de setenças. Após o processamento do texto fonte, o Intellexer Summarizer Network Edition apresenta uma saída de dados por módulos, pois apresenta dados do texto fonte relacionados a conceitos e entidades nomeadas.

Ambas as ferramentas permitem ao usuário inserir o texto que deseja resumir por colagem de texto ou via upload, e permite, ainda, ao usuário selecionar o tamanho do resumo, a partir da porcentagem, que varia de 5% a 100%, de acordo com as necessidades e escolhas dos usuários. No Text Sumarizer, em específico, após o processo ter sido concluído, ao usuário é apresentado três opções: a de salvar o resumo em .txt, compartilhar por–e-mail ou twitter; ou copiar o texto.

Após essas considerações acerca dos sumarizadores escolhidos e outras informações relevantes para a realização da pesquisa, em sequência, serão aprensentados e analisados os resultados, caracterizados nos resumos do processo de sumarização a que foram submetidos.

Devido a proposta apresentada pelos sistemas de PLN, em geral, e especificamente em SA, faz-se mais que necessário o uso de instrumentos que possam avaliar a qualidade desses resumos/sumários gerados automaticamente. Necessidades estas que se justificam no intuito de verificar os avanços no desenvolvimento dessas ferramentas; medir o grau de utilidade e

adequação a determinadas tarefas, assim como indicar meios de validação a propostas e metodologias utilizadas na construção e concepção dessas ferramentas. (DIAS DA SILVA *et al*, 2007).

A literatura específica da area de Sumarização Automática de Textos aponta para várias discussões sobre métricas e métodos de avaliação empregados aos sumários/resumos gerados automaticamente. Entretanto, não existe consenso no que diz respeito a melhor forma de avaliar ferramentas desta natureza, suscitando, assim, em diversos desafios. (MANY; MAYBURY, 1999 apud RINO; PARDO, 2003, p. 28).

Desafios dentre os quais, destacam-se a "[...] identificação do que seria um resultado "correto" para um sumarizador automático"; a "[...] identificação de uma taxa de compressão ideal para avaliar adequadamente os sumários automáticos"; a "[...] forma como a qualidade e a informatividade de sumários automáticos podem ser avaliadas automaticamente." e a "[...] identificação da situação e da forma de se utilizar o julgamento humano.". (DIAS DA SILVA *et al*, 2007).

Rino e Pardo (2003, p. 28) destacam que

> Quando se fala em avaliação de um sistema, pode-se ter em mente várias facetas: (a) pode-se avaliar o desempenho computacional do sistema, isto é, o uso que ele faz de memória, seu tempo de execução, a complexidade de seu algoritmo principal, etc.; (b) pode-se considerar a usabilidade do sistema, ou seja, a crítica da clareza de sua interface ou o grau de intuição (dos possíveis usuários) necessário para seu uso ou, ainda, sua consistência e sua flexibilidade para possíveis customizações; (c) pode-se avaliar se os resultados produzidos automaticamente são satisfatórios, isto é, se são os resultados esperados, se são "corretos" ou "adequados".

Os citados autores ponderam, ainda, que "Sparck Jones e Galliers afirmam que o mais importante na avaliação de um sistema de SA é estabelecer claramente o que se quer avaliar.". Posto isto, determina-se o tipo de avaliação que será aplicada, se é intrínseca ou extrínseca, *on-line* ou *off-line*, *black-box* ou *glass-box* e comparativa ou autônoma.

Leite (2010, p. 9), explica que "Na intrínseca, o sumarizador é avaliado de acordo com a qualidade direta dos textos produzidos. Na extrínseca, é mensurado o quanto os sumários são úteis para outra tarefa que os utiliza, como por exemplo QA (*Question Answering*) ou Categorização de Textos.". Assim, a primeira pode envolver medidas de qualidade e informatividade; enquanto a segunda propõe avaliar um sistema em uso, para a realização de tarefas específicas.

E em se tratando da avaliação intrínseca, existem as Metodologias Objetivas de Avaliação Intrínsecas: a exemplo das medidas de precisão, cobertura e *F-Measure*, e o pacote de medidas de informatividade da ferramenta ROUGE; e as Metodologias Subjetivas de

Avaliação Intrínsecas, como as propostas nas DUCs e TACs; o Método da Cobertura de Unidades Elementares (*Basic Elements*) e o Método da Pirâmide. (LEITE, 2010).

Em tese, o tipo de avaliação extrínseca, seria a mais apropriada para o escopo deste experimento, por se tratar de um tipo de pesquisa que avalia o uso de um sistema para a realização de alguma tarefa específica. Entretanto, este trabalho não possui como objetivo avaliar sistemas – embora isso aconteça de uma maneira natural –, e sim, avaliar e inferir qualidade aos sumários/resumos produtos destes sistemas. Portanto, a avaliação intrínseca se harmoniza melhor à proposta desta pesquisa.

Conforme já mencionado, os resumos produzidos pelo Turbine Text e Intellex, foram avaliados por meio do modelo de avaliação proposto pela DUC, a partir de categorias como: gramaticalidade, redundância, clareza referencial, foco, estrutura e coerência. Tais critérios também são indicados por Lancaster (2004) sobre a elaboração de resumos.

No quesito Gramaticalidade, percebemos que os sumários gerados pela ferramenta Turbine Text, tiveram boas avaliações, tendo como variações as notas 4 (Bom) e 5 (Muito bom). Enquanto os mesmos textos, processados pela ferramenta Intellexer Summarizer, tiveram avaliações com maiores variações das notas atribuídas aos sumários, tendo inclusive uma nota 2 (ruim). Os dados podem ser visualizados no Quadro 8.

Quadro 8 – Critério da Gramaticalidade

	TURBINE TEXT	INTELLEX SUMMARIZER NE
Texto 1	4 (Bom)	3 (Aceitável)
Texto 2	3 (Aceitável)	2 (Ruim)
Texto 3	5 (Muito Bom)	3 (Aceitável)
Texto 4	4 (Bom)	4 (Bom)
Texto 5	5 (Muito Bom)	5 (Muito Bom)

Fonte: dados da pesquisa (2020).

Dang (2004), aponta que o aspecto gramatical visa avaliar se o resumo apresenta erros nas sentenças que tornam o texto de difícil leitura, por exemplo, padrões de ortografia, assim como, o emprego incorreto de maiúsculas e minúsculas, erros de pontuação e sintaxe.

Quanto ao parâmetro de Redudância, dos cinco textos sumarizados, foi possível perceber nos sumários resultados do processamento do Turbine Text, que também tiveram boas avaliações, tendo notas variaveis entre 5 (Muito bom), 4 (Bom) e 3 (Aceitável). Os resumos gerados pelo Intellex Summarizer tiveram notas entre 4 (Bom) e 3 (Aceitável), apresentados no Quadro 9.

Quadro 9 – Critério de Redundância

	TURBINE TEXT	INTELLEX SUMMARIZER NE
Texto 1	5 (Muito Bom)	4 (Bom)
Texto 2	4 (Bom)	3 (Aceitável)

Texto 3	4 (Bom)	3 (Aceitável)
Texto 4	3 (Aceitável)	3 (Aceitável)
Texto 5	4 (Bom)	4 (Bom)

Fonte: dados da pesquisa (2020).

Esse parâmetro tem como propósito verificar a inexistência de repetições no resumo, que podem aparecer como frases, sentenças, fatos, etc repetidos ao longo de sua extensão. (DANG, 2005). Lancaster (2004, p. 113), tambem destaca que, "Em suma, as características de um bom resumo são brevidade, exatidão e clareza", trazendo à tona a importância de uma boa estrutura na redação de um resumo, além de não haver repetições desnecessárias, que tornariam o texto menos claro e inexato.

Em se tratando da categoria Clareza Referencial, compreendemos que similarmente os sumários produzidos pelas duas ferramentas tiveram nas avaliações, variantes mais representativas entre 4 (Bom) e 3 (Aceitável), como apresentado no Quadro 10.

Quadro 10 – Critério de Clareza Referencial

	TURBINE TEXT	**INTELLEX SUMMARIZER NE**
Texto 1	4 (Bom)	2 (Ruim)
Texto 2	3 (Aceitável)	3 (Aceitável)
Texto 3	4 (Bom)	4 (Bom)
Texto 4	4 (Bom)	4 (Bom)
Texto 5	4 (Bom)	4 (Bom)

Fonte: dados da pesquisa (2020).

A Clareza Referencial, é "[...] a propriedade de um texto permitir ao leitor identificar a quem ou a que um pronome ou sintagma nominal está se referindo" (LEITE, 2010, p. 2). Logo, as informações no corpo do resumo, devem ser compostas a ponto de ser fácil a identificação de componentes linguísticos que reportam a conceitos, pessoas e etc.

Acerca do parâmetro Foco, constatamos que os sumários produzidos pelas duas ferramentas, tiveram avaliações bem variadas. A exemplo, destacamos o Resumo do Texto 3, em que o sumário produzido a partir do Turbine Text, foi avaliado com nota 5 (Muito boa) e, em contrapartida, o sumário produzido apartir do Intellex Summarizer, teve nota 3 (Aceitável), como exibido no Quadro 11.

Quadro 11 – Critério de Foco

	TURBINE TEXT	**INTELLEX SUMMARIZER NE**
Texto 1	4 (Bom)	2 (Ruim)
Texto 2	3 (Aceitável)	2 (Ruim)
Texto 3	5 (Muito Bom)	3 (Aceitável)
Texto 4	2 (Ruim)	4 (Bom)
Texto 5	4 (Bom)	3 (Aceitável)

Fonte: dados da pesquisa (2020).

Dang (2004), assinala para ao parâmetro Foco, e evidencia que deve avaliar se o resumo é focado, constituído apenas de sentenças que contenham informações relevantes e

relacionadas ao restante do resumo. A respeito disto, Lancaster (2004), salienta sobre a necessidade da clareza e precisão estarem presentes nos resumos.

E, por fim, nos quesitos dos critérios Estrutura e Coerência, que inferem valor quanto a boa estrutura e coerência, percebemos que os sumários produzidos pelo Turbine Text, tiveram três avaliações com notas 5 (Muito bom), enquanto que os gerados pelo Intellex Summarizer, tiveram variáveis das notas 3 (Aceitável) e 2 (Ruim), exposto no Quadro 12.

Quadro 12 – Critério de Estrutura e Coerência

	TURBINE TEXT	INTELLEX SUMMARIZER NE
Texto 1	4 (Bom)	2 (Ruim)
Texto 2	3 (Aceitável)	2 (Ruim)
Texto 3	5 (Muito Bom)	3 (Aceitável)
Texto 4	2 (Ruim)	4 (Bom)
Texto 5	4 (Bom)	3 (Aceitável)

Fonte: dados da pesquisa (2020).

Observamos em geral, que embora alguns sumários produzidos pelas ferramentas apresentados possuam questões a serem estudadas, boa parte dos resumos apresentam boa estruturação e lógica, e as informações extraidas e reorganizadas permitem uma compreensão acerca do que trata o texto-fonte. Entretanto, salienta-se uma obervação acerca da justaposição e contextualização dos resumos automáticos, como ilustrado na Figura 21.

Figura 21 – Resumo do Texto 2 resultado do Intellex Summarizer

Ou seja, 87,5% dos países do globo apresentaram ao menos um caso confirmado (captura dos dados: 12h 01m 27s do dia 24 de março de 2020)6.
Isso porque o acompanhamento gráfico dos casos permite antever o cenário epidemiológico do evento e, com isso, programar políticas públicas e assistenciais próprias ao seu enfrentamento.
É antigo o conhecimento acadêmico sobre este monitoramento, que está pautado em técnica consagrada na literatura, sempre utilizando medidas de incidência (casos novos do evento) para a estimação da velocidade de adoecimento populacional.
Por outro lado, estudo técnico que comparou projeções da epidemia e os casos observados para o mesmo período reflete uma perspectiva otimista para o comportamento do Covid-19 no país, apresentando, em médio prazo, uma tendência ao achatamento da ascensão da curva, ou seja, para a redução da velocidade da epidemia18Corroborando para esta informação, até o dia 16 de março de 2020, momento em que a primeira medida de isolamento físico social foi imposta no país especificamente pelo Governo do Estado do Rio de Janeiro -, a reprodução da doença esteve estimada entre 2,4 a 4,6 pessoas, caindo para uma estimativa entre 2,1 e 3,8 pessoas no dia 24.
Sobre o primeiro aspecto, é importante refletir que a literatura já tem certa robustez teórico-prática sobre o reconhecimento do status de adoecimento e a prevenção de novos casos da doença.
Chama-se atenção que o resultado chinês é anterior a construção recorde (em 10 dias) de dois hospitais para recepção de pacientes com Covid-19, aspecto que pode ter contribuído para uma baixa letalidade quando comparada ao desempenho da Itália e Espanha.
A distribuição proporcional de trabalhadores de enfermagem na Itália e Espanha, nos mesmos anos, é mais próxima à de médicos (58,6 e 55,3 para 10 mil habitantes) e maior do que na China, de 23,0/10.000 habitantes.
Os dados do Brasil informam uma proporção acentuadamente maior da enfermagem, de 97,0/10.000, em 2018.
Deste modo, o terceiro aspecto em análise é a formulação discursiva e prática de atores políticos sobre a experiência com pandemia no Brasil.

Fonte: dados da pesquisa (2020).

A fim de exemplificar, a Figura 21, ilustra o caso em questão. Em alguns casos, percebe-se a inexistência de um ponto de partida do resumo que se caracterize como uma introdução e que confira mais contexto e coerência as primeiras frases do resumo.

Destacamos também o comportamento das ferramentas quanto a temática do corpus utilizado. Em geral, tanto o Turbine Text quanto o Intellex Summarizer, apresentaram bons resultados quanto a apresentação e construção do resumo, tendo o Turbine Text apresentado resultados melhores ainda. Nesse sentido, destacamos o resumo elaborado pela ferramenta Turbine Text do texto 3, texto que possui um caráter de cientificidade bem maior em relação aos demais, tendo em vista que é um texto técnico que trata sobre a Administração de Medicamentos específicos para o combate da Pandemia do Novo Corona Vírus, possuindo uma linguagem bastante técnica e terminologias voltadas a farmácos bem específicas.

É possível, observar na ilustração da figura 21, que os resumos gerados automaticamente, não obdecem as diretrizes de elaboração de resumos no que diz respeito a apresentação gráfica, a exemplo da adoção de parágrafos.

Quanto a avaliação dos resumos obtidos, constatamos que ambos os cinco resumos, tiveram variações nas avaliações. A exemplo dos resumos do texto 4, que no critério de gramaticalidade e clareza referencial, foram atribuidas as notas 4 (bom) tanto para o resumo do Turbine Text quanto do Intellex Summarizer. Já no critério de redundância, observamos que para os resumos dos dois sumarizadores, foram avaliados com a nota 3 (aceitável). E no critério de foco, os resumos do Turbine Text e o Intellex receberam as notas 2 (ruim) e 4 (bom), respectivamente.

Como já pontuado, este trabalho não possui o objetivo de avaliar sistemas, embora isso aconteça de uma maneira natural, tendo em vista que o desenvolvimento da pesquisa, perpassa por esses sistemas e inferências em torno dessas questões. Observamos um bom desempenho computacional e uma boa usabilidade dos sistemas de sumarização, o que se estende, tanto ao Turbine Text, executado em ambiente Web, quanto ao Intellex Summarizer, utilizado em versão *Desktop*.

Sobre o uso e avaliação de sumarizadores, são apresentados por Tabosa *et al* (2020), como protótipos capazes de elaborar resumos automáticos de textos baseado em técnicas de PLN e estatísticas de frequência de palavras, e utilizam para este fim, avaliação humana, a partir da realização de testes cegos. Como resultado, percebermos que há equivalência qualitativa entre os resumos gerados pela ferramenta de SA, com os resumos feitos por humanos, o que reforça as aplicações dessas ferramentas.

Lancaster (2004) acredita que “É pequena a distância entre a etapa de análise conceitual da indexação e a preparação de um resumo aceitável.”, e nesse sentido destaca que a qualidade presente em torno da elaboração de um resumo, interfere na decisão sobre determinadas questões que podem ser incluidas ou omitidas na indexação. Nesse sentido, ponderamos sobre a necessidade de estudos em Biblioteconomia e Ciência da Informação que abarquem a temática da Sumarização Automática no Âmbito da indexação.

Assim, em sequência, faremos as consideracões finais acerca da investigação realizada.

7 CONCLUSÃO

Entendemos que a Organização da Informação e do Conhecimento, assim como diversas áreas do conhecimento humano, passou, passa e passará por mudanças significativas em busca de adaptar-se frente às novas demandas e evoluções advindas do desenvolvimento das sociedades. Essas, em muitos casos, acompanhadas de dificuldades e questões que suscitam novos modos de fazer e realizar tarefas já conhecidas.

Nos dias atuais, grande parte das questões citadas no decorrer da pesquisa, são resultados das Novas Tecnologias Digitais de Comunicação e Informação e, consequentemente, da demasiada produção informacional produzida em larga escala em meio digital.

Nesse sentido, este trabalho teve como intuito, uma questão que problematizava sobre de que forma, a utilização de ferramentas de sumarização automática de textos poderiam contribuir para a análise conceitual de documentos, e quais as implicações do uso dessas ferramentas para a representação temática da informação, no âmbito da indexação de documentos, questão, amparada por objetivos do mesmo escopo, os quais sustentaram a realização desta pesquisa.

A partir disto, depreendemos que a Organização da Informação e do Conhecimento, possui as suas bases no uso de ferramentas, instrumentos e processos que viabilizam o tratamento da informação. Assim, nos dias atuais, a presença simultânea, das informações disponíveis em meios digitais e tradicionais, suscita modos que possibilitem maior celeridade nos processos envolvidos para o fim da representação e organização da informação.

Evidenciamos também, a experiência da pesquisa sob a ótica na interdisciplinaridade e, durante o desenvolvimento da pesquisa, compreendemos de fato a necessidade e a importância da colaboração de diferentes campos do conhecimento, em prol de uma mesma causa. Dessa forma, compreende-se os ganhos para a área da Ciência da Informação, a partir das relações estabelecidas apoiadas com o uso e aplicação de teorias e metodologias de outras áreas do conhecimento.

Nesse sentido, foi possível assimilar que os sistemas baseados em modelos de PLN possuem grandes possibilidades e potenciais usos nos processos de organização, representação para a posterior recuperação da informação. Logo, a partir da apropriação da literatura da área do PLN, houve um entendimento de que tais sistemas baseados nesse modelo, podem e devem ser aplicados no sentido de auxiliar processos deste escopo.

A partir do estudo teórico, podemos nos dar conta de que algumas dessas aplicações de PLN já estão presentes no âmbito da Organização, Representação e Recuperação da Informação em ambientes relacionados à área da Biblioteconomia e Ciência da Informação, a exemplo da Mineração de Textos, Sistemas Inteligentes de Recuperação da Informação, Sintagmas Nominais e outros.

Desse modo, em especial no que diz respeito à Sumarização Automática de Textos, após a avaliação realizada a partir do instrumento escolhido, percebemos que esses sistemas podem potencializar processos que possuem o intuito de tratar e condensar grandes volumes de informações textuais para fins específicos, a exemplo da indexação de assunto.

Entretanto, indicamos que mais estudos nessa área possam ser desenvolvidos no sentido de dar maior sustentação e bases teóricas para o desenvolvimento de técnicas de representação e organização da informação, amparados por sistemas de SA, para permitir uma aproximação mais sólida e concreta entre as áreas em destaque.

Em se tratando do processo de SA, foi possível perceber que é constituído por três etapas: análise, transformação e síntese de textos de entrada/textos-fonte em sistemas de SA, culminando em textos mais sucintos. Sobre as diversas abordagens do funcionamento dos sumarizadores, destacam-se a extrativa e abstrativa. E por fim, percebemos que há uma diversidade de sumarizadores disponíveis nos dias atuais, comerciais, gratuitos, baseados em iniciativas *open source* e etc.

Os softwares baseados em iniciativas *open source*, em sua grande maioria, possuem uma interface e estrutura gráfica básica, e integram à sua arquitetura um número limitado de ferramentas, além de quase todos os softwares citados na presente pesquisa, funcionarem em ambiente web. Em contrapartida, os softwares de iniciativa comercial/proprietários, possuem sistemas mais robustos, e que oferecem ao usuário, mais de uma ferramenta integrada à sua arquitetura, além da disponibilização tanto em versão web quanto desktop, contudo ambos apresentam resultados satisfatórios.

O estudo revelou que as ferramentas de SA podem contribuir significativamente para os processos de representação temática, uma vez que, possibilitam a condensação e a celeridade no tratamento de grandes volumes de informações. Contudo, reforçamos a necessidade de maiores aprofundamentos dos estudos na área, objetivando a aplicação teórica-prática.

Salientamos que o emprego dessas ferramentas de PLN em ambientes de Bibliotecas podem auxiliar e potencializar as atividades desenvolvidas no âmbito do TTI pelos

Bibliotecários. O objetivo é aumentar as capacidades humanas, permitindo a esses profissionais maior efetividade, eficácia e celeridade nas atividades desenvolvidas.

Desse modo, a adoção da sumarização automática de textos nos diversos ambientes informacionais, dadas as suas especificidades, exige profissionais qualificados para a realização da representação temática baseadas nessas ferramentas, a fim de garantir a execução de forma ágil e precisa. Não basta reconhecer que a informação é um insumo estratégico para a manutenção nesses ambientes, faz-se necessário conhecer os diversos métodos, técnicas, instrumentos e ferramentas que podem contribuir para o tratamento da informação, a fim de promover o acesso a elas.

Acreditamos que este trabalho contribua com os fundamentos teóricos-metodológicos na Representação e Organização da Informação, com o intuito de estimular novos estudos sobre aplicação do PLN no processo de indexação e no desenvolvimento de serviços e produtos inerentes à Biblioteconomia e Ciência da Informação.

Como trabalhos futuros, recomendamos investigações sobre os seguintes temas: a avaliação dos resumos gerados automaticamente utilizando outras técnicas de avaliação, com o intuito de comparar os resultados obtidos na presente pesquisa; assim como, o uso de ferramentas de PLN voltadas para indexação de imagens, através de tecnologias de OCR; estudos sobre o reconhecimento de entidades nomeadas; uso de *chatbots* em serviços de referência virtuais; integração de sistemas de conversação texto-fala em bases de dados e repositórios instrucionais, com o intuito de assegurar a acessibilidade à deficientes visuais e outros.

REFERÊNCIAS

AGANETTE, Elisângela Cristina; TEIXEIRA, Livia Marangon Duffles; AGANETTE, Karina de Jesus Pinto. A representação descritiva nas perspectivas do século XXI um estudo evolutivo dos modelos conceituais. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 22, n. 50, p. 176-187, 6 set. 2017. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n50p176. Acesso em: 24 abr. 2020.

ALMEIDA, Carlos Cândido de. Semiologia na organização da informação e do conhecimento. *In*: ALMEIDA, Carlos Cândido de. **Elementos de linguística e semiologia na organização da informação**. São Paulo: Cultura Acadêmica, 2011. Cap. 4. p. 169-181. Disponível em: http://www.culturaacademica.com.br/catalogo/elementos-de-linguistica-e-semiologia-na-organizacao-da-informacao/. Acesso em: 29 ago. 2020.

ALMEIDA, Gladis Maria de Barcellos. A teoria comunicativa da terminologia e a sua prática. **Alfa**, São Paulo, v. 50, n. 2, p. 85-101, 2006. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/viewFile/1413/1114. Acesso em: 31 out. 2020.

ALVARENGA, Lidia. Organização da informação nas bibliotecas digitais. *In*: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Org.). **Organização da informação**: princípios e tendências. Brasília: Briquet de Lemos, 2006. p 76-98.

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da Ciência da Informação em tempo e espaço digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 8, n. 15, p. 18-40, jan. 2003. ISSN 1518-2924. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18. Acesso em: 31 mar. 2020.

ALVES, Rachel Cristina Vesú. Da catalogação aos metadados: o tratamento descritivo da informação e as tecnologias de informática. *In:* ALVES, Rachel Cristina Vesú. **Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Cap. 2, p. 25-42. Disponível em: http://hdl.handle.net/11449/103361. Acesso em 30 mar. 2020.

ANNA, Jorge Santa. A (r)evolução digital e os dilemas para a catalogação: os cibertecários em atuação. **Rdbci**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 13, n. 2, p. 312, maio/ago. 2015. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8634632/3388. Acesso em: 27 maio 2020.

ARAÚJO, Carlos Alberto Ávila. Teorias e tendências contemporâneas da Ciência da Informação. **Informação em Pauta**, Fortaleza, v. 2, n. 2, p. 9-34, dez. 2017. Disponível em: http://periodicos.ufc.br/informacaoempauta/article/view/20162. Acesso em: 30 mar. 2020.

ARAUJO, Eliany Alvarenga de. A construção social da informação: dinâmicas e contextos. **DataGramaZero** - Revista de Ciência da Informação, [Rio de Janeiro], v.2, n.5, out 2001. Disponível em: http://www.brapci.inf.br/index.php/article/view/0000001246/d11daa9de3ea05fb4652e9cde6bef943/. Acesso em: 30 abr. 2020.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**. Métodos para análise de documentos: determinação de seus assuntos e seleção de termos de indexação: procedimento. Rio de Janeiro, 1992. 4 p.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6028**. Informação e documentação: resumos: apresentação. Rio de Janeiro, 2003. 2 p.

BARBOSA, Eduardo Caetano; KOBASHI, Nair Yumiko. Extroversão e descoberta: visualização de dados no auxílio a buscas e recuperação de informações. **Revista Brasileira de Biblioteconomia e Documentação**, v. 13, p. 115-120, 2017. Disponível em: http://hdl.handle.net/20.500.11959/brapci/1157. Acesso em: 05 out. 2020.

BARROS, Flávia de Almeida; ROBIN, Jacques. Processamento de Linguagem Natural. 1997. Disponível em: https://www.cin.ufpe.br/~fab/cursos/jai96/ProcessamentoDeLinguagemNatural.pdf. Acesso em: 29 jul. 2020.

BORKO, H. Information science: what is it?. **American Documentation**, v. 19, n. 1, p. 3-5, jan. 1968. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/k---artigo-01.pdf. Acesso em 23 maio. 2020.

BRASCHER, Marisa; CAFÉ, Lígia. Organização da Informação ou Organização do Conhecimento? In: Encontro Nacional de Pesquisa em Ciência da Informação, 9., 2008, São Paulo. **Anais[...]** . São Paulo: Ancib, 2014. p. 1 - 14. Disponível em: http://enancib.ibict.br/index.php/enancib/ixenancib/paper/view/3016. Acesso em: 2 jul. 2019.

CABRAL, Luciano de Souza. **Uma Plataforma para Sumarização Automática de Textos Independente de Idioma**. 2015. 139 f. Tese (Doutorado) - Curso de Doutorado em Engenharia Elétrica, Universidade Federal de Pernambuco, Recife, 2015. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/14968/1/lsc_tese_corrigida_rdl_versaoDigital.pdf. Acesso em: 16 abr. 2020.

CAFÉ, Lígia; SALES, R. Organização da informação: Conceitos básicos e breve fundamentação teórica. *In*: Robredo, Jaime; Bräscher, Marisa. (orgs.). **Passeios no Bosque da Informação**: Estudos sobre Representação e Organização da Informação e do Conhecimento. Brasília DF: IBICT, 2010. 335 p. Cap. 6, p. 115-129. Disponível em: http://www.ibict.br/publicacoes/eroic.pdf. Acesso em 30 mar 2020.

CAMARA JUNIOR, Auto Tavares da. **Processamento de linguagem natural para indexação automática semântico-ontológica**. 2013. 181 f. Tese (Doutorado) - Curso de Doutorado em Ciência da Informação, Universidade de Brasília, Brasília, Df, 2013.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. **Perspectivas em Ciência da Informação**, [S.l.], v. 12, n. 1, nov. 2007. ISSN 19815344. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54/47. Acesso em: 30 mar. 2020.

CARDOSO, Paula Christina Figueira. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. 2014. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2014. doi:10.11606/T.55.2014.tde-16032015-161912. Acesso em: 03 abr. 2020.

CARMO, Juliana Rabelo do; CONCEIÇÃO, Valdirene Pereira da. Processamento da linguagem natural do domínio musical: do sentido à gestão terminológica no ambiente Etermos. **Informação & Informação**, Londrina, n. 3, v. 23, p.314-341, set./dez. 2018. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/29622/pdf. Acesso em: 22 abr. 2019.

CHAUMIER, Jacques. Indexação: conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. 1/2, p. 63-79, 1988. Disponível em: https://www.brapci.inf.br/index.php/article/download/19202. Acesso em: 05 maio 2020.

DANG, Hoa Trang. Overview of DUC 2005. *In*: **Proceedings of the Document Understanding Conference**, 2005. Disponível em: https://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf. Acesso em: 16 nov. 2020.

DAVENPORT, Thomas H. Informação e seus dissabores: uma introdução. *In*: DAVENPORT, Thomas H. **Ecologia da informação**. São Paulo: Futura, 1998. Cap. 1. p. 11-25. Disponível em: https://ppgic.files.wordpress.com/2018/07/davenport-t-h-2002.pdf. Acesso em: 01 abr. 2020.

DIAS DA SILVA, Bento Carlos Dias da. O estudo Lingüístico-Computacional da Linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, jun. 2006. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/597. Acesso em: 29 jun. 2020.

DIAS DA SILVA, Bento Carlos *et al*. **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. São Carlos: NILC - ICMC-USP, 2007. 121 p. (Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional). Disponível em: http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0710-DiasDaSilvaEtAl.pdf. Acesso em: 25 jun. 2019.

DIAS, Eduardo Wense; NAVES, Madalena Lopes. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007.

DODEBEI, Vera Lucia Doyle. Representação, Memória e Linguagem. *In:* DODEBEI, Vera Lucia Doyle. **Tesauro:** linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002. p. 19-59.
em: 25 jun. 2019.

EVERS, Aline. Processamento de Língua Natural. *In*: EVERS, Aline. **Processamento de língua natural e níveis de proficiência do português**: um estudo de produções textuais do exame celpe-bras. 2013. 174 f. Dissertação (Mestrado em Letras) – Programa de Pós Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: http://www.ufrgs.br/acervocelpebras/arquivos/textos-publicados/evers-aline. Acesso em: 26 jul. 2020. p. 86-106.

FAVARÃO, Neide Rodrigues Lago; ARAÚJO, Cíntia de Souza Alferes. Importância da Interdisciplinaridade no Ensino Superior. **EDUCERE**. Umuarama, v.4, n.2, p.103-115, jul./dez., 2004. Disponível em: https://www.revistas.unipar.br/index.php/educere/article/viewFile/173/147. Acesso em: 22 set. 2020.

FERNEDA, Edberto. **Recuperação de informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 2003. 147 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/pt-br.php. Acesso em 26 maio 2020.

FUJITA, Mariângela Spotti. RUBI, Lopes Milena Polsinelli. BOCCATO, Vera Regina Casari. As diferentes perspectivas teóricas e metodológicas sobre indexação e catalogação de assuntos. *In*: FUJITA, Mariângela Spotti Lopes et al (org.). **A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias: um estudo de observação do contexto sociocognitivo com protocolos verbais**. São Paulo: Cultura Acadêmica, 2009. Cap. 1. p. 19-42. Disponível em: https://repositorio.unesp.br/handle/11449/109109. Acesso em: 30 abr. 2020.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 5. ed. São Paulo: Atlas, 2010. 183p.

GONZALEZ, Marco; LIMA, Vera Lúcia Strube de. Recuperação de Informação e Processamento da Linguagem Natural. 2003. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-06---gonzales-e-lima-2003.pdf. Acesso em 24 jul. 2020.

GRAEL, Felipe Fink. **Processamento de linguagem natural e cases de aplicação**. [S. l.: s. n.], 2019. 1 vídeo (50 min). Publicado pelo canal Instituto Infnet. Disponível em: http://www.youtube.com/watch?v=iwPj0qgvfIs. Acesso em: 25 jul. 2020.

GUIMARAES, José Augusto Chaves. A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knowledge Organization (ISKO). Revista Ibero-Americana de Ciência da Informação, v. 1, p. 77-99, 2008. Disponível em: https://www.brapci.inf.br/index.php/res/v/70663. Acesso em: 02 maio 2020.

GUIMARÃES, José Augusto Chaves. Abordagens teóricas de tratamento temático da informação (TTI): catalogação de assunto, indexação e análise documental. **Ibersid: revista de sistemas de información y documentación**, v. 3, p. 105-117, 15 set. 2009. Disponível em: https://www.ibersid.eu/ojs/index.php/ibersid/article/view/3730. Acesso em: 21 abr. 2020.

GUINCHAT, Claire; MENOU, Michael. A indexação. *In*: **Introdução geral às técnicas da informação e documentação**. 2.ed. Brasília: IBICT, 1994. p. 175-185.

IVO, Amanda. **Algumas noções básicas de lingüística**: os níveis descritivos da linguagem. Disponível em:

https://grad.letras.ufmg.br/arquivos/monitoria/ApostilaConceitos%20b%c3%a1sicos_Aula1.pdf. Acesso em 15 jul. 2020.

KURAMOTO, Hélio. Sintagmas nominais: uma nova abordagem no processo de indexação. *In*: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Org.). **Organização da informação**: princípios e tendências. Brasília: Briquet de Lemos, 2006. p. 117-137.

LADEIRA, Ana Paula. **Processamento de linguagem natural**: caracterizacao da produção científica dos pesquisadores brasileiros. 2010. 262 f. Tese (Doutorado) - Curso de Doutorado em Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010. Disponível em: https://repositorio.ufmg.br/handle/1843/ECID-8B3Q6C. Acesso em: 21 jul. 2020.

LADEIRA, Ana Paula; ALVARENGA, Lídia. PROCESSAMENTO DE LINGUAGEM NATURAL: em busca de evidências temáticas nas publicações nacionais e contemporâneas. 2012. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/69/GT%202%20Txt%2019-%20LADEIRA%2c%20Ana%20Paula%3b%20ALVARENGA%2c%20L%c3%addia.pdf?sequence=1. Acesso em: 21 jul. 2020.

LANCASTER, F. W. **Indexação e Resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos/livros, 2004.

LEITE, Daniel Saraiva. **Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizagem de máquina para sumarização automática de textos em português**. 2010. 213 f. Dissertação (Mestrado) - Curso de Mestrado em Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2010. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/459/3512.pdf?sequence=1. Acesso em 27 set. 2020.

LEIVA, Isidoro Gil. Aspectos conceituais da indexação. *In*: LEIVA, Isidoro Gil Leiva; FUJITA, Mariângela Spotti Lopes (Ed.). **Política de Indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 31-106. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 30 abr. 2020.

LIMA, Gercina Ângela Borém. Interfaces entre a ciência da informação e a ciência cognitiva. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 77-87, jan./abr. 2003. Disponível em: https://www.brapci.inf.br/_repositorio/2010/02/pdf_77053b8355_0008068.pdf. Acesso em: 19 maio 2020.

LIMA, José Leonardo de Oliveira; ALVARES, Lillian. Organização e representação da informação e do conhecimento. *In*: ALVARES, Lillian (org.). **Organização da informação e do conhecimento:** conceitos, subsídios interdisciplinares e aplicações. São Paulo: B4 Editores, 2012. Cap. 1. p. 21-48. Disponível em: https://www.researchgate.net/profile/Jose_Leonardo_Lima/publication/281969932_Organizacao_e_representacao_da_informacao_e_do_conhecimento/links/5600067308ae07629e522ad1/Organizacao-e-representacao-da-informacao-e-do-conhecimento.pdf. Acesso em: 20 abr. 2020.

LIMA, Vera Lúcia Strube de; NUNES, Maria das Graças Volpe; VIEIRA, Renata. Desafios do processamento de línguas naturais. *In*: **Anais do XXVII Congresso da Sociedade Brasileira de Computação**, 2007. p. 2202–2216. Disponível em: https://www.inf.pucrs.br/linatural/Recursos/Desafios.pdf. Acesso em: 23 jul. 2020. linguagem. 2019. Disponível em: https://grad.letras.ufmg.br/arquivos/monitoria/ApostilaConceitos%20b%c3%a1sicos_Aula1.pdf. Acesso em 15 jul. 2020.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. Fundamentos de metodologia científica. São Paulo: Atlas, 2003

MARTINS, Agnaldo Lopes. Potenciais aplicações da Inteligência Artificial na Ciência da Informação. **Informação & Informação**, Londrina, v. 15, n. 1, p. 1-16, 19 set. 2010. Universidade Estadual de Londrina. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/3882. Acesso em: 20 maio 2020.

MARTINS, Camilla Brandel. **Introdução à sumarização automática**. 2001. Disponível em: https://sites.icmc.usp.br/taspardo/RTDC00201-CMartinsEtAl.pdf. Acesso em: 19 set. 2020.

MARTINS, Carlos Wellington Soares. **Plantando bibliotecas para colher desenvolvimento**: análise do programa de bibliotecas rurais "Arca das Letras", no município de Codó-MA. Curitiba: Editora Crv, 2014. 155 p.

MARTINS, Paulo George Miranda. **Evolução das tecnologias de representação**: das linguagens de marcação aos dados interligados. 2018. 102 f. Dissertação (Mestrado) - Curso de Mestrado em Ciência da Informação, Universidade Federal de São Carlos, São Carlos, 2018. Disponível em: https://repositorio.ufscar.br/handle/ufscar/10491?show=full. Acesso em: 30 abr. 2020.

MEADOWS, Arthur Jack. **A comunicação cientifica**. Brasilia: Briquet de Lemos/Livros, 1999. 268 p.

MEY, Eliane Serrão Alves. **Não brigue com a catalogação**. Brasília: Briquet, 2003. 186 p.

MEY, Eliane Serrao Alves; SILVEIRA, Naira Christofoletti. **Catalogação no plural**. Brasilia: Briquet de Lemos/Livros, 2009. 217 p.

MEY, Eliane. **Introdução à catalogação**. Brasília: Briquet de Lemos, 1995.

MORESI, Eduardo. **Metodologia da Pesquisa**. Brasília: UCB, 2003. Disponível em: http://www.inf.ufes.br/~pdcosta/ensino/2010-2-metodologia-de-pesquisa/MetodologiaPesquisa-Moresi2003.pdf. Acesso em 22 set. 2020.

NEVES, Bárbara Coelho. As perspectivas e aplicações da computação cognitiva em Unidades de Informação. **ENANCIB**, Brasil, out. 2019. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1421. Data de acesso: 23 maio 2020.

NHACUONGUE, Januário Albino; DUTRA, Moisés Lima. A terminologia em Sistemas de Recuperação da Informação baseada na WORDNET.PT. **Informação & Sociedade: Estudos**, v. 30, n. 2, 26 maio 2020. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/50756. Acesso em: 21 set. 2020.

NOVELLINO, Maria Salet Ferreira. Instrumentos e metodologias de representação da informação. **Informação & Informação**, Londrina, v. 1, n. 2, p. 37-45, jul./dez. 1996. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/1603. Acesso em: 20 maio 2020.

NUNES, Maria das Graças Volpe. **Processamento de línguas naturais**: para quê e para quem. Notas Didáticas ICMC-USP, n. 73. São Carlos, 2008.

OLIVEIRA NETO, João Mendes de; TONIN, Sávio Duarte; PRIETCH, Soraia Silva. **Processamento de Linguagem Natural e suas Aplicações Computacionais**. 2010. Disponível em: https://www.inpa.gov.br/erin2010/Artigo/Artigo9.pdf. Acesso em 16 jun. 2020.

PARDO, Thiago Alexandre Salgueiro. **Sumarização Automática:** Principais Conceitos e Sistemas para o Português Brasileiro. São Carlos: NILC - ICMC-USP, 2008. 15 p. (Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional). Disponível em: http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0804-Pardo.pdf. Acesso em: 22 abr. 2019.

PINHEIRO, Lena Vania Ribeiro. Campo interdisciplinar da ciência da informação: fronteiras remotas e recentes. *In*: PINHEIRO, Lena Vania Ribeiro (org.). **Ciência da informação, ciências sociais e interdisciplinaridade**. Brasília; Rio de Janeiro: Ibict, 1999. p. 155-183. Disponível em: https://livroaberto.ibict.br/bitstream/1/1000/1/PINHEIRO.%20Ci%C3%AAncia%20da%20Informa%C3%A7%C3%A3o,%20Ci%C3%AAncias%20Sociais%20e%20Interdisciplinariedade.pdf. Acesso em: 20 maio 2020.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013. Disponível em: http://www.feevale.br/Comum/midias/8807f05a-14d0-4d5b-b1ad-1538f3aef538/E-book%20Metodologia%20do%20Trabalho%20Cientifico.pdf. Acesso em: 22 set. 2020.

RAMPÃO, Talita de Souza; TSUNODA, Denise Fukumi. Mineração de dados em bases jurídicas: um estudo de caso. **Revista Brasileira de Educação em Ciência da Informação**, v. 6, n. Especial, p. 61-76, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/114081. Acesso em: 05 out. 2020.

RINO, Lúcia Helena Machado; PARDO, Thiago Alexandre Salgueiro. A Sumarização Automática de Textos: principais características e metodologias. *In:* **Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA)**, 2003, p. 203-245. Campinas-SP.

ROBREDO, Jaime. Da Ciência da informação revisitada aos sistemas humanos de informação. Brasília DF: Thesaurus; SSRR Informações, 2003, 262 p.

RODRIGUES, Iraci Oliveira. **A organização da informação e a organização do conhecimento na produção científica em ciência da informação**. 2015. Dissertação (Mestrado em Cultura e Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2015. doi:10.11606/D.27.2016.tde-01022016-144902. Acesso em: 1 ABR. 2020.

ROSA, João Luís Garcia. Processamento de Línguas Naturais. *In*: ROSA, João Luís Garcia. **Fundamentos da Inteligência Artificial**. Rio de Janeiro: Ltc, 2011. p. 136-169. Disponível em: http://walderson.com/2011-2/IA/FIA.pdf. Acesso em 02 jun. 2020.

ROWLEY, Jennifer. Introdução à biblioteca eletrônica. *In:* ROWLEY, Jennifer. **A biblioteca eletrônica**. Brasília: Briquet de Lemos Livros, 2002. p. 3-23.

RUBI, Milena Polsinelli. Os princípios da política de indexação na análise de assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários. *In*: FUJITA, Mariângela Spotti Lopes et al (org.). **A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias: um estudo de observação do contexto sociocognitivo com protocolos verbais**. São Paulo: Cultura Acadêmica, 2009. Cap. 4. p. 81-95. Disponível em: https://repositorio.unesp.br/handle/11449/109109. Acesso em: 30 abr. 2020.

SALES, Adriana Silva. **O processo de indexação da informação jurídica como base para elaboração da Política de Indexação do Tribunal de Justiça do Maranhão**. 2020. 91 f. Monografia (Graduação) - Curso de Biblioteconomia, Universidade Federal do Maranhão, São Luís, 2020.

SANTOS, Alessa Fabíola dos. **A atividade indexação apoiada por plataformas computacionais**. 2017. 153 f. Dissertação (Mestrado) - Curso de Mestrado Profissional em Gestão de Unidades de Informação, Centro de Ciências Humanas e da Educação, Universidade do Estado de Santa Catarina, Florianópolis, 2017. Disponível em: http://www.repositorio.mar.mil.br/handle/ripcmb/844421?mode=full. Acesso em: 04 maio 2020.

SANTOS, Ângelo Filipe da Silva dos. **Sumarização automática de texto**. 2012. 76 f. Dissertação (Mestrado) - Curso de Mestrado em Engenharia Informática, Universidade da Beira Interior, Covilhã, 2012. Disponível em: https://ubibliorum.ubi.pt/handle/10400.6/3738. Acesso em: 20 mar. 2020.

SARACEVIC, Teiko. Ciência da Informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**. Belo Horizonte, v. 1, n. 1, p.41-62, jan./jun. 1996. Disponível em: https://www.brapci.inf.br/_repositorio/2010/08/pdf_fd9fd572cc_0011621.pdf. Acesso em: 26 abr. 2020.

SIMONASSI, Rafael. **Uma abordagem de sumarização automática de textos aplicadas a debates online**. 2015. 113 f. Dissertação (Mestrado) - Curso de Mestrado em Gestão do Conhecimento e da Tecnologia da Informação, Universidade Católica de Brasília, Brasília, Df, 2015. Disponível em: https://bdtd.ucb.br:8443/jspui/bitstream/123456789/1458/1/Rafael%20Simonassi.pdf. Acesso em: 20 mar. 2020.

SOUSA, Aline Rocha de. **Processamento automático de línguas naturais**: um estudo sobre a localização do IBM Watson™ para o português do Brasil. 2015. 76 f., il. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas)—Universidade de Brasília, Brasília, 2015. Disponível em: https://bdm.unb.br/bitstream/10483/18730/1/2015_AlineRochaDeSousa_tcc.pdf. Acesso em 28 jun. 2020.

SOUZA, Fernanda Possenti de; HILLESHEIM, Araci Isaltina de Andrade. Tratamento da informação e o uso das tecnologias da informação e comunicação. **Biblionline**, João Pessoa, v. 2, n. 10, p. 81-96, jan. 2014. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/biblio/article/view/16748/12483. Acesso em: 22 abr. 2020.

SOUZA, Osvaldo de *et al*. Um método de sumarização automática de textos através de dados estatísticos e Processamento de Linguagem Natural. **Informação & Sociedade: Estudos**, v. 27, n. 3, 24 dez. 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/32571. Acesso em: 09 ago. 2020.

TABOSA, Hamilton Rodrigues et al. Avaliação do desempenho de um software de sumarização automática de textos. **Informação & Informação**, Londrina, v. 25, n. 1, p. 189-210, jan./mar. 2020. Disponível em: http://repositorio.ufc.br/handle/riufc/51441. Acesso em: 20 ago. 2020.

UNESCO. Guidelines for the establishment and development of monolingual thesauri. [s.n.t]. 37p.

UNISIST. Princípios de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 10, n. 1, p. 83-94, mar. 1981. Disponível em: https://www.brapci.inf.br/index.php/res/download/87644. Acesso em: 28 abr. 2020.

VIEIRA, Renata; LIMA, Vera L. S. Lingüística computacional: princípios e aplicações. *In*: **Anais do Encontro Nacional de Inteligência Artificial**, 2001. Disponível em: https://www.inf.pucrs.br/linatural/Recursos/jaia-2001.pdf. Acesso em: 20 jul. 2020.

VIEIRA, Renata; LOPES, Lucelene. Processamento de Linguagem Natural e o Tratamento Computacional de Linguagens Científicas. In: PERNA, Cristina Lopes; DELGADO, Heloísa Koch; FINATTO, Maria José (Org.). **Linguagens especializadas em corpora**: modos de dizer e interfaces de pesquisa. Porto Alegre: Edipucrs, 2010. p. 183-201. Disponível em: http://bibliodigital.unijui.edu.br:8080/xmlui/handle/123456789/1496. Acesso em: 22 jul. 2020.

APÊNDICE – ARTIGOS SUMARIZADOS

FARIAS, Luis Arthur Brasil Gadelha et al. O papel da atenção primária no combate ao Covid-19: impacto na saúde pública e perspectivas futuras. **Revista Brasileira de Medicina de Família e Comunidade**, Rio de Janeiro, v. 15, n. 42, p. 1-8, 19 maio 2020. Disponível em: https://rbmfc.org.br/rbmfc/article/view/2455/1539. Acesso em: 09 jun. 2020.

TEXTO 1 – RESULTADO DA SUMARIZAÇÃO DO TURBINE TEXT

RESUMO: Covid-19 é o acrônimo oficial da doença causada pelo agente SARS-CoV-2, isolado e genotipado a partir de amostras respiratórias de pacientes da cidade de Wuhan, cidade da China, após um surto de pneumonia de origem não identificada em dezembro de 2019.1 O SARS-CoV-2, também conhecido como novo corona vírus, é um beta corona vírus de origem zoonótica cuja origem específica ainda não foi elucidada e tem sido associado a uma mutação do corona vírus presente em morcegos ou pangolim (animal comum na Ásia), ambos estando relacionados aos hábitos alimentares chineses e que podem ter transmitido aos seres humanos.

A propagação do vírus se dá por gotas ou aerossóis, bem como pelo contato da mucosa com partículas nas superfícies pelas mãos.2 A doença apresenta espectro clínico variado, desde sintomas leves como uma síndrome gripal (febre, coriza, tosse) até quadros de síndrome do desconforto respiratório agudo (SDRA) com evolução fatal.

Com o intuito de fornecer subsídios para as populações de maior risco e fortalecer a igualdade de acesso à saúde, o Grupo de Trabalho de Saúde da População Negra da Sociedade Brasileira de Medicina de Família e Comunidade (SBMFC) e Associação de Medicina de Família e Comunidade do Rio de janeiro (AMFaC-RJ), disponibilizaram um manual de orientações para populações vulneráveis, com destaque para orientações para os indivíduos que necessitam trabalhar durante a quarentena como proteção ao usar transporte público, circular nas ruas ou andar de mototáxi, bem como orientações acerca da limpeza da casa, convívio com caso suspeito, uso de proteção individual e como entreter as crianças durante o período de quarentena, com soluções adequadas para tarefas que fazem parte do cotidiano do brasileiro.

TEXTO 1 – RESULTADO DA SUMARIZAÇÃO DO INTELLEX SUMARIZER

RESUMO: Segundo dados do Instituto Superior de Sanidade (ISS), órgão de controle de saúde pública da Itália, cerca de 10% dos casos infectados são profissionais da saúde, que estiveram expostos ao alto potencial de transmissibilidade do vírus no âmbito assistencial.

Com o intuito de fornecer subsídios para as populações de maior risco e fortalecer a igualdade de acesso à saúde, o Grupo de Trabalho de Saúde da População Negra da Sociedade Brasileira de Medicina de Família e Comunidade (SBMFC) e Associação de Medicina de Família e Comunidade do Rio de janeiro (AMFaC-RJ), disponibilizaram um manual de orientações para populações vulneráveis, com destaque para orientações para os indivíduos que necessitam trabalhar durante a quarentena como proteção ao usar transporte público, circular nas ruas ou andar de mototáxi, bem como orientações acerca da limpeza da casa, convívio com caso suspeito, uso de proteção individual e como entreter as crianças durante o período de quarentena, com soluções adequadas para tarefas que fazem parte do cotidiano

Ações como cancelamento de consultas eletivas não essenciais, agendamento por horário e montagem de consultórios ao ar livre têm sido experimentadas por secretarias de saúde e pelas próprias equipes de ESF, as quais podem individualizar as ações de acordo com as particularidades de seus territórios e de suas UBS.

O Ministério da Saúde, por meio da sexta versão do protocolo supracitado, busca proteger profissio-nais da saúde em grupo de risco, sugerindo que estes devam ser afastados ou remanejados para tarefas em que não estejam diretamente em contato com pacientes suspeitos de Covid-19.

Com a possibilidade de triagem dos pacientes suspeitos a partir da metodologia fast-track , os profissionais da saúde em grupo de risco podem atender exclusivamente pacientes sem sintomas gripais, fincando mais protegidos.

Esse tipo de atendimento é recomendado no seguimento dos pacientes suspeitos e/ou confirmados, a fim de identificar sinais de piora clínica no paciente ou surgimento de novos casos relacionados a este.

Além disso, países com atenção primária fortalecida geralmente proporcionam dados epidemiológicos mais fidedignos, principalmente devido à maior capilaridade para notificação de casos e, por conseguinte, no fim da linha do cuidado, dados mais confiáveis no que tange a letalidade e a taxa de cura.

Neste momento, a melhor ferramenta de controle existente para o Covid-19 é a prevenção e não existe melhor lugar para desenvolvê-la do que na APS.

RAFAEL, Ricardo de Mattos Russo et al. Epidemiologia, políticas públicas e pandemia de Covid-19: o que esperar no Brasil?. **Revista Enfermagem UERJ**, [S.l.], v. 28, p. 1-6, abr.

2020. ISSN 0104-3552. Disponível em: https://www.e-publicacoes.uerj.br/index.php/enfermagemuerj/article/view/49570/33134. Acesso em: 09 jun. 2020.

TEXTO 2 – RESULTADO DA SUMARIZAÇÃO DO TURBINE TEXT

RESUMO: Este cenário complexo impõe desafios adicionais à vigilância epidemiológica, às relações internacionais e à programação de políticas públicas, sobretudo por meio de medidas que reduzam as desigualdades de acesso aos sistemas de saúde e a condições estruturais para o autocuidado.

No Brasil, a situação ainda inicial da epidemia já aponta para uma característica ascensional e veloz da curva epidêmica, saindo, em 2 de março de 2020, de dois casos alóctones confirmados para 2201 casos autóctones no dia 24 de março, e já com a expectativa de atingir 6981 casos (IC 95%: 6243 – 7807) no dia 29 deste mês.

Por outro lado, estudo técnico que comparou projeções da epidemia e os casos observados para o mesmo período reflete uma perspectiva otimista para o comportamento do Covid-19 no país, apresentando, em médio prazo, uma tendência ao achatamento da ascensão da curva, ou seja, para a redução da velocidade da epidemia18Corroborando para esta informação, até o dia 16 de março de 2020, momento em que a primeira medida de isolamento físico social foi imposta no país – especificamente pelo Governo do Estado do Rio de Janeiro -, a reprodução da doença esteve estimada entre 2,4 a 4,6 pessoas, caindo para uma estimativa entre 2,1 e 3,8 pessoas no dia 24.

Dados do Observatório global de Saúde da Organização Mundial da Saúde informam que a proporção da força médica é maior em países europeus como Itália (2017) e Espanha (2016) - de 40,9 e 40,6 para cada 10 mil habitantes, respectivamente -, enquanto no Brasil, o informado para o ano de 2018 era de 21,4 e na China (2015) era de 17,8 por 10 mil habitantes.

TEXTO 2 – RESULTADO DA SUMARIZAÇÃO DO INTELLEX SUMARIZER

RESUMO: Ou seja, 87,5% dos países do globo apresentaram ao menos um caso confirmado (captura dos dados: 12h 01m 27s do dia 24 de março de 2020)6.

Isso porque o acompanhamento gráfico dos casos permite antever o cenário epidemiológico do evento e, com isso, programar políticas públicas e assistenciais próprias ao seu enfrentamento.

É antigo o conhecimento acadêmico sobre este monitoramento, que está pautado em técnica consagrada na literatura, sempre utilizando medidas de incidência (casos novos do evento) para a estimação da velocidade de adoecimento populacional.

Por outro lado, estudo técnico que comparou projeções da epidemia e os casos observados para o mesmo período reflete uma perspectiva otimista para o comportamento do Covid-19 no país, apresentando, em médio prazo, uma tendência ao achatamento da ascensão da curva, ou seja, para a redução da velocidade da epidemia18Corroborando para esta informação, até o dia 16 de março de 2020, momento em que a primeira medida de isolamento físico social foi imposta no país especificamente pelo Governo do Estado do Rio de Janeiro -, a reprodução da doença esteve estimada entre 2,4 a 4,6 pessoas, caindo para uma estimativa entre 2,1 e 3,8 pessoas no dia 24.

Sobre o primeiro aspecto, é importante refletir que a literatura já tem certa robustez teórico-prática sobre o reconhecimento do status de adoecimento e a prevenção de novos casos da doença.

Chama-se atenção que o resultado chinês é anterior a construção recorde (em 10 dias) de dois hospitais para recepção de pacientes com Covid-19, aspecto que pode ter contribuído para uma baixa letalidade quando comparada ao desempenho da Itália e Espanha.

A distribuição proporcional de trabalhadores de enfermagem na Itália e Espanha, nos mesmos anos, é mais próxima à de médicos (58,6 e 55,3 para 10 mil habitantes) e maior do que na China, de 23,0/10.000 habitantes.

Os dados do Brasil informam uma proporção acentuadamente maior da enfermagem, de 97,0/10.000, em 2018.

Deste modo, o terceiro aspecto em análise é a formulação discursiva e prática de atores políticos sobre a experiência com pandemia no Brasil.

MENEZES, Carolline Rodrigues; SANCHES, Cristina; CHEQUER, Farah Maria Drumond. Efetividade e toxicidade da cloroquina e da hidroxicloroquina associada (ou não) à azitromicina para tratamento da COVID-19: o que sabemos até o momento?. **J. Health Biol Sci**, [s.l], v. 1, n. 8, p. 1-9, abr. 2020. Disponível em: https://pesquisa.bvsalud.org/portal/resource/pt/biblio-1095354. Acesso em: 09 jun. 2020.

TEXTO 3 – RESULTADO DA SUMARIZAÇÃO DO TURBINE TEXT

RESUMO: A pandemia de Coronavírus (COVID-19) é um quadro de grave crise global de saúde e representa uma situação de grandes incertezas pelo desconhecimento acerca do vírus e sobre o manejo dos pacientes que vem crescendo exponencialmente.

A Anvisa também realizou divulgação pública em 27 de março acerca da liberação de pesquisas com uso de

hidroxicloroquina e azitromicina para prevenção de complicações em pacientes com infecção pelo novo coronavírus (Covid-19) com casos leves e moderados, e para avaliação da segurança e eficácia clínica desses fármacos em pacientes com pneumonia causada por infecção pelo vírus Sars-CoV-2 (pacientes graves).

Trata-se de uma revisão narrativa com a seguinte pergunta norteadora: "Qual a efetividade dos tratamentos do COVID-19 com o uso da cloroquina ou hidroxicloroquina associa-da (ou não) à azitromicina no tratamento do COVID-19 e os seus possíveis efeitos adversos e tóxicos?" De acordo com a pergunta norteadora, foi estabelecido o "PICOS": "P"(population): pacientes que fizeram uso de terapia medicamentosa com cloroquina ou hidroxicloroquina associada (ou não) à azitromicina.

O estudo de Borba et al dividiu seus pacientes em dois grupos: no primeiro grupo (n=41) usou alta dose de cloroquina: 600mg 2 vezes ao dia por 10 dias e no segundo grupo (n=40) utilizou baixa dose de cloroquina: 450mg 2 vezes ao dia no primeiro dia e 450mg 1 vez ao dia por 4 dias.

tratamento com hidroxicloroquina mostrou-se eficaz na redução da carga viral em 70% (n= 18,2) dos pacientes do estudo de Gautret et al, e sendo que 100% (n=6) do grupo tratado com associação de hidroxicloroquina com azitromicina apresentaram cura do ponto de vista de carga viral detectada dos pacientes se comparada com cura do controle de 12,5% (n=2), indicando um possível efeito sinérgico com utilização dos dois fármacos.

No entanto, é importante salientar as limitações que perpassam todos os sete estudos ao utilizarem os fármacos em concentrações maiores do que as recomendados - nos dois estudos de Gautret em que são utilizadas doses diárias de 600mg, no estudo de Barbosa37 600mg 2 vezes ao dia e no estudo de Borba, 450 mg 2 vezes ao dia ao invés da dose máxima preconizada de 400mg/dia no caso de tratamento de afecções reumatológicas, sujeitando os pacientes ao desenvolvimento de efeitos adversos severos, ou mesmo a toxicidade.

Também devem ser consideradas as limitações individuais dos estudos, as quais incluem: baixa população avaliada, falta de um grupo controle adequado, curto período de avaliação, uso de outros terapias medicamentosas juntamente com hidroxicloroquina (como a azitromicina, por exemplo), falta de avaliação a longo prazo, limitação nos parâmetros de avaliação dos resultados e principalmente falta de seguimento clínico dos pacientes para estudar os efeitos adversos e tóxicos desta terapêutica.

TEXTO 3 – RESULTADO DA SUMARIZAÇÃO DO INTELLEX SUMARIZER

RESUMO: É necessário aprofundamento dos mecanismos de ação e dos efeitos no aparelho auditivo, para que os pacientes usuários de tais medicamentos possam tomar medidas preventivas e, portanto, evitar maiores complicações.

Em face do exposto, o presente estudo tem como objetivos abordar as evidências científicas existentes até o presente momento sobre a efetividade do uso da cloroquina, da hidroxicloroquina associada (ou não) à azitromicina para tratamento da afecção pelo coronavírus e seus possíveis efeitos adversos e tóxicos aos seres humanos.

Foram considerados elegíveis artigos disponíveis na íntegra em inglês ou português, e que descreveram sobre a efetividade ou sobre os efeitos adversos em seres humanos associados ao uso dos fármacos: cloroquina, hidroxicloroquina associada (ou não) à azitromicina no tratamento do COVID-19.

As variáveis para esta revisão foram: autor, ano de publicação, localidade, tamanho da amostra, população, sexo, idade, tempo de seguimento, dose diária, grupo controle, descrição dos efeitos adversos ao medicamento, efetividade (efeito curativo ou redução da carga viral) e limitações do estudo.

O estudo de Gautret e colaboradores33 já foi encerrado testando a eficácia do tratamento com hidroxicloroquina (neste caso, uma parcela do grupo de tratamento foi testado utilizando apenas hidroxicloroquina e outra parcela testou hidroxicloroquina em associação com a azitromicina) no tratamento de pacientes com COVID-19 (n=42).

O grupo estudou a eficácia do tratamento com hidroxicloroquina versus hidroxicloroquina em associação com a azitromicina no tratamento de pacientes com COVID-19.

Quanto à dose diária, o estudo Gautret e colaboradores usou a dose de 600 mg de hidroxicloroquina (n=20), com variação de 600 mg de hidroxicloroquina associada a 500 mg de azitromicina no primeiro dia e depois nos seguintes dias 250mg (n=6); o segundo estudo de Gautret et al. utilizou 600mg de sulfato de hidroxicloroquina durante os seis dias de tratamento associada a 500 mg de azitromicina no primeiro dia e depois 250 mg de azitromicina para os cinco dias restantes de tratamento para todos os pacientes (n=80).

O estudo de Barbosa et al37 utilizou hidroxicloroquina off-label de acordo com o seguinte esquema terapêutico: 400mg 2 vezes ao dia pelos 2 primeiros dias e sequencialmente, 200 a 400mg por dia, sendo 1 vez ao dia. O estudo de Borba et al dividiu seus pacientes em dois grupos: no primeiro grupo (n=41) usou alta dose de cloroquina: 600mg 2 vezes ao dia por 10 dias e no segundo grupo (n=40) utilizou baixa dose de cloroquina: 450mg 2 vezes ao dia no primeiro dia e 450mg 1 vez ao dia por 4 dias.

O estudo de Barbosa et al observou como efeito adverso da hidroxicloroquina o desenvolvimento de arritmia do tipo torsade de pointes e o estudo de Borba et al38 demonstrou desenvolvimento de miocardite, rabdmiólise e prolongamento do segmento QT (não especificando se houve arritmia) com o uso da

cloroquina.
O tratamento com hidroxicloroquina mostrou-se eficaz na redução da carga viral em 70% (n= 18,2) dos pacientes do estudo de Gautret et al, e sendo que 100% (n=6) do grupo tratado com associação de hidroxicloroquina com azitromicina apresentaram cura do ponto de vista de carga viral detectada dos pacientes se comparada com cura do controle de 12,5% (n=2), indicando um possível efeito sinérgico com utilização dos dois fármacos.
Essa hipótese foi novamente testada em Gautret et al, utilizando hidroxicloroquina e azitromicina na totalidade dos pacientes e foi observada cura da doença em 97,5% dos casos (n=78).
O estudo de Gao et al relatou 100% de cura (n=100) em pacientes com pneumonia associada com a COVID-19 mas não descreveu seus resultados devido ao caráter em andamento do estudo.
No entanto, é importante salientar as limitações que perpassam todos os sete estudos ao utilizarem os fármacos em concentrações maiores do que as recomendados - nos dois estudos de Gautret em que são utilizadas doses diárias de 600mg, no estudo de Barbosa37 600mg 2 vezes ao dia e no estudo de Borba, 450 mg 2 vezes ao dia ao invés da dose máxima preconizada de 400mg/dia no caso de tratamento de afecções reumatológicas, sujeitando os pacientes ao desenvolvimento de efeitos adversos severos, ou mesmo a toxicidade.
Os estudos que analisaram a efetividade da hidroxicloroquina no tratamento da COVID-19 apresentaram, em comum, a seguinte limitação: quantidade insuficiente de ensaios realizados devido à emergência da utilização da hidroxicloroquina no contexto da pandemia do SARS-CoV-2.
No cenário nacional, o Ministério da Saúde disponibilizará para uso, a critério médico, os medicamentos cloroquina e hidroxicloroquina como terapia adjuvante no tratamento do COVID-19 de formas graves, em pacientes hospitalizados, sem que outras medidas de suporte sejam preteridas em seu favor, embora exista necessidade de maior investigação.

NETO, Mercedes et al. Fake news no cenário da pandemia de COVID-19. **Cogitare Enfermagem**, [S.l.], v. 25, apr. 2020. ISSN 2176-9133. Disponível em: https://revistas.ufpr.br/cogitare/article/view/72627. Acesso em: 09 jun. 2020.

TEXTO 4 – RESULTADO DA SUMARIZAÇÃO DO TURBINE TEXT

RESUMO: Anos mais tarde (2012), o MERS-Cov, outro tipo de coronavírus, provocou surto na Arábia Saudita, propagando o contágio por meio de perdigotos e deficiência na higiene, especialmente das mãos, dentre outros hábitos hodiernos da cultura dos cuidados, para além das lacunas nas políticas públicas, quando 27 países foram atingidos.
Em tempos de avanços tecnológicos, estas notícias falsas são veiculadas nas redes sociais, de forma rápida e multiplicada entre a população, que, em linguagem metafórica, pode-se entender como um vírus que contamina a comunicação e promove ações e comportamentos contrários às orientações das autoridades técnicas no campo da saúde.
Estas, com termos técnicos, próprios dos centros de pesquisa, precisam ser decodificadas à população para melhor entendimento, o que remete à aplicação da técnica da comunicação denominada de AIDA — Atenção, Interesse, Desejo e Atitude — utilizada pelos jornalistas para a imprensa social.
Elas atraem aos que as disseminam quando há interesse, mas possuem sobrevida curta, bem como fazem os operadores das comunicações até que outro assunto seja mais interessante para a indústria e/ou comércio das informações, visando aos leitores na formação de opinião pública e as redes sociais como seus consumidores.

TEXTO 4 – RESULTADO DA SUMARIZAÇÃO DO INTELLEX SUMARIZER

RESUMO: A busca das notícias Fake News ocorreu no banco de dados do Ministério da Saúde, no cenário da pandemia de COVID-19, no período de 29 de janeiro a 31 de março de 2020, quando foram identificados 70 registros.
Refletir sobre as Fake News na contemporaneidade é pensar nas publicações com base nas evidências científicas.
Isto se articula com a intencionalidade que a indústria da informação e seus operadores aplicam às matérias jornalísticas, no intuito do efeito de real, no sentido de poder fazer e crer, o que se faz ver.
Por um lado, não se pode negar que a tipificação não foi apresentada à análise, o que permite lacunas nesta comunicação; por outro lado, instiga o aprofundamento das tipificações para análise em estudos futuros.
Ademais, destaca-se que a literatura brasileira é escassa sobre a pandemia de COVID-19 e a velocidade desta produção do conhecimento vai de encontro com a produção das Fake News.
Não obstante, enfatiza-se a necessidade de a população conhecer o site do Ministério da Saúde brasileiro, o qual aponta as Fake News, para que ocorra educação em saúde com informações corretas e seguras.

FARIAS, Heitor Soares de. O avanço da Covid-19 e o isolamento social como estratégia para redução da vulnerabilidade. **Espaço e Economia [Online]**, [S.l], v. 17, abr. 2020. Disponível em: http://journals.openedition.org/espacoeconomia/11357. Acesso em: 15 jun.

2020.
TEXTO 5 – RESULTADO DA SUMARIZAÇÃO DO TURBINE TEXT
RESUMO: Dia 26 de março de 2020, um mês após ser registrado o primeiro caso de COVID-19 no Brasil, são 2433 casos confirmados, 59 mortes e surgiram 284 novos casos nas últimas 24 horas. Embora comecem a aparecer evidências de que a China omitiu informações iniciais sobre a nova doença, e com isso perderam tempo precioso na divulgação dos fatos, o que poderia impedir que mais pessoas se contaminassem, em um momento seguinte de maior transparência, o governo conseguiu que a epidemia ficasse restrita à província de Hubei, onde está Wuhan3. Além disso, a Coreia do Sul, país vizinho atingido na sequência, obteve sucesso em controlar a epidemia, baseando-se na multiplicação de testes diagnóstico realizados massivamente para identificação dos indivíduos infectados, colocando-os em quarentena para interromper o avanço da doença. Trump justificou dizendo que o aumento é reflexo do grande número de testes realizados recentemente, entretanto no país foram 1480 mortes nas últimas 24 horas, o maior número diário em um país desde o início da epidemia. Diante desse fato e da inércia do governo federal, no dia 13 de março, o governador do Rio de Janeiro, Wilson Witzel, decretou o fechamento de escolas, teatros e cinemas, além de atividades que proporcionassem aglomerações como a visitação a presos e a realização de eventos esportivos, eventos científicos, comícios e passeatas por 15 dias. A crise econômica é certa, e não ocorrerá pela imobilidade do cidadão brasileiro, impedido de ir ao trabalho, mas pela dimensão mundial dos estragos econômicos causados pela pandemia do coronavírus em diferentes países. Mesmo na China, onde foram tomadas medidas muito duras com restrições severas à circulação de pessoas, e se conseguiu zerar os casos de transmissão local da Covid-19, o retorno de chineses que estavam fora do país no início da epidemia fez aparecer novamente casos da doença.
TEXTO 5 – RESULTADO DA SUMARIZAÇÃO DO INTELLEX SUMARIZER
RESUMO: Um ritmo mais intenso do que a China e a Itália alcançaram. Em 11 de março a OMS declarou a pandemia do novo coronavírus, porque nas últimas duas semanas o número de casos de Covid-19 fora da China aumentou 13 vezes e a quantidade de países afetados triplicou. Já era esperado que São Paulo e Rio de Janeiro apresentassem os primeiros casos no Brasil, pois possuem as cidades mais ricas do país, mais populosas e concentram maior número de voos internacionais. Diante desse fato e da inércia do governo federal, no dia 13 de março, o governador do Rio de Janeiro, Wilson Witzel, decretou o fechamento de escolas, teatros e cinemas, além de atividades que proporcionassem aglomerações como a visitação a presos e a realização de eventos esportivos, eventos científicos, comícios e passeatas por 15 dias. Na mesma semana, no dia 19 de março, ocorreu a primeira morte em consequência do coronavírus no estado do Rio de Janeiro. Uma mulher de 63 anos residente no interior do estado, município de Miguel Pereira, que trabalhava como doméstica no Leblon, Zona Sul do Rio de Janeiro. Decretou o fechamento de bares, restaurantes e lanchonetes, pontos turísticos, sendo mantidas somente as atividades essenciais. Sendo que a competência dessas atividades é do governo federal, através de suas agências reguladoras, e dependia da anuência de ambas para serem confirmadas. Uma vitória da oposição, tendo em vista que a proposta inicial do governo federal era de R$200,00. Em São Paulo, no dia 1 de abril havia 201 corpos aguardando o resultado do exame. Nesse sentido, a informação divulgada nos meios de comunicação reforçando a necessidade da população ficar em casa, junto ao programa de ajuda financeira do governo, são fundamentais para reduzir a exposição dos grupos mais vulneráveis. No estado do Rio de Janeiro são 1074 casos confirmados e 47 mortes, enquanto no estado de São Paulo são 4048 casos e 219 mortes, seis vezes mais mortes do que a China, considerando os 13 dias após o primeiro registro de óbito por Covid-19.