INTERFACE TECNOLÓGICA

LINGUÍSTICA COMPUTACIONAL: um mapeamento bibliográfico de 2000 a 2020

COMPUTATIONAL LINGUISTICS: a bibliographic mapping from 2000 to 2020

Anne Cleres da Silva Oliveira – annecleres.ac@gmail.com
Faculdade de Tecnologia (Fatec) – Catanduva – SP – Brasil

Ana Carolina Freschi – anafreschi@gmail.com
Faculdade de Tecnologia (Fatec) – Catanduva – SP – Brasil

DOI: 10.31510/infa.v17i2.1056
Data de publicação: 18/12/2020

RESUMO

A Linguística Computacional é um campo multidisciplinar que envolve a Inteligência Artificial, a Informática e a Linguística. Tal área busca desenvolver, por meio de um modelamento lógico-formal, sistemas com capacidade de reconhecer e de produzir informações da linguagem natural. O objetivo deste trabalho, pautado na pesquisa bibliográfica exploratória, é apresentar um mapeamento bibliográfico de estudos que envolvem a Linguística Computacional entre os anos de 2000 e de 2020 a fim de reconhecer suas diversas aplicações. Esta investigação se justifica pela própria importância desse campo de pesquisa para o desenvolvimento de novas tecnologias e pelo conhecimento de possíveis áreas em que ele possa ser usado. Na análise, foram encontradas outras áreas relacionadas à Linguística Computacional e como elas estão interligadas na construção de novas ferramentas. Dessa maneira, com esta investigação, é possível ter um claro entendimento da sua origem e da sua usabilidade nos processos de criação de sistemas relacionados à manipulação da linguagem humana.

Palavras-chave: Linguística computacional. Revisão bibliográfica. Informática.

ABSTRACT

Computational Linguistics is a multidisciplinary field that involves Artificial Intelligence, Informatics, and Linguistics. This area seeks to develop, through logical-formal modeling, systems with the capacity to recognize and produce information presented by natural language. The objective of this work, based on exploratory bibliographic research, is to present a bibliographic mapping of studies involving Computational Linguistics between 2000 and 2020 to recognize its diverse applications. This investigation is justified by the very importance of this research field for new technologies development and by the knowledge of possible areas in which it can be used. In the analysis, it was found other areas related to Computational Linguistics and how they are interconnected in the construction of new tools. Therefore, this

investigation shows a clear understanding of its origin and usability in the processes of systems creation related to human language manipulation.

Keywords: Computational linguistics. Literature review. Computing.

1 INTRODUÇÃO

De acordo com a *Stanford Encyclopedia of Philosophy* (2014), a Linguística Computacional possibilita (i) a formulação de estruturas gramaticais e semânticas para caracterizar as línguas de forma a permitir implementações computacionalmente tratáveis de análise sintática e semântica e (ii) o desenvolvimento de modelos computacionais cognitiva e neuro cientificamente plausíveis de como o processamento e a aprendizagem da linguagem podem ocorrer no cérebro. Em outras palavras, a Linguística Computacional busca uma maneira de transformar a linguagem natural em um artefato que possa ser produzido e processado de maneira computacional de maneira útil e simples. Um exemplo disso é o reconhecimento que um tradutor faz da fala natural para gerar um resultado claro de uma palavra falada de maneira diferente da qual se aprende em ambiente escolar.

Ainda conforme a *Stanford Encyclopedia of Philosophy* (2014), alguns dos objetivos desse campo variado e amplo são a recuperação de textos sobre algum tópico desejado; a tradução automática (*machine translation*); a resposta a perguntas (*question answering*); a sumarização de textos; a análise de textos escritos ou falados por meio do tópico, do sentimento ou de outros atributos psicológicos; os agentes de diálogo para a realização de certas tarefas (compras, resolução de problemas técnicos, planejamento de viagens, manutenção de horários, aconselhamento médico, etc.); e, finalmente, a criação de sistemas computacionais com competência semelhante à humana no diálogo, na aquisição da linguagem e na obtenção de conhecimento de textos.

Tendo em vista a importância dessa área sem, muitas vezes, ser percebida, este trabalho busca explorar em quais investigações ela foi abordada nos últimos anos. Desse modo, este artigo tem o objetivo de apresentar um mapeamento bibliográfico de estudos que envolvem a Linguística Computacional entre 2000 e 2020 a fim de buscar reconhecer suas diversas aplicações e indicar possíveis áreas de atuação.

Tal estudo se justifica pela própria importância desse campo teórico para o desenvolvimento de novas tecnologias e pelo conhecimento de possíveis áreas em que ele possa ser usado. Além disso, espera-se, com esta investigação, ter um claro entendimento da origem

e da aplicação dos processos que são utilizados na criação dos sistemas relacionados à manipulação da linguagem humana.

Este artigo está organizado da seguinte forma: esta introdução, que apresenta a delimitação do assunto, os objetivos e a justificativa, uma seção sobre Linguística Computacional, que contém os princípios teóricos dessa área de estudo, os Métodos, em que são discutidos a forma como o levantamento bibliográfico foi feito e categorizado, os Resultados e Discussões, que abordam as categorias encontradas e suas características, as Considerações Finais, que retomam os objetivos e discutem limitações e futuros estudos, e, por fim, as Referências.

2 LINGUÍSTICA COMPUTACIONAL

A Linguística Computacional é a área de pesquisa multidisciplinar que estuda o processamento sintático, semântico e lógico da linguagem natural. De acordo com Vieira e Lima (2001, p. 1), essa disciplina pode ser entendida como "a área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural". Nessa mesma trilha, para *Stanford Encyclopedia of Philosophy* (2014), a Linguística Computacional é a disciplina científica de engenharia preocupada com a compreensão da linguagem escrita e falada de uma perspectiva computacional e com a construção de artefatos que processam e produzem linguagem de maneira útil, seja em massa ou em um ambiente de diálogo.

De acordo com Impacta (2019), tal campo de estudo se iniciou na década de 1950, especificamente nos Estados Unidos, que utilizavam computadores para traduzir de maneira automática e rápida os documentos escritos em outras línguas durante a Guerra Fria. Esses textos eram escritos em russo e traduzidos para inglês. Na época, apesar de a tradução não ser perfeita, já era possível conseguir um resultado de qualidade razoável.

Ainda sobre sua origem, Othero (2016) discute que sua história está intrinsecamente ligada ao desenvolvimento na área da Inteligência Artificial. Daí percebe-se sua forte relação com a Informática.

Sobre isso, Grisham (1992, p. 1) aponta que

> o potencial [dos computadores] para o processamento de linguagem natural foi reconhecido bem cedo no desenvolvimento de computadores, e trabalhos em Linguística Computacional – basicamente para tradução automática –

> começaram na década de 1950 em diversos centros de pesquisa. O rápido crescimento na área, no entanto, aconteceu principalmente a partir do final dos anos 1970.

Conforme apontam Vieira e Lima (2001), ao longo dos anos, pode-se observar a consolidação desse campo de pesquisa em diferentes áreas que inclui o desenvolvimento de algoritmos, métodos e softwares que dispõem aos computadores formas de conseguirem lidar com uma língua natural de forma disciplinada e útil às nossas necessidades.

Nos dias de hoje, de acordo com a Impacta (2019) e a Infopédia (2020), entre as principais aplicações da Linguística Computacional, podem-se citar:

- a criação de ferramentas que auxiliam na produção e na escrita de um texto, como corretores gramaticais e ortográficos, dicionários eletrônicos e de sinônimos, analisadores sintáticos e morfológicos, sistemas especializados na composição de cartas comerciais, entre outros;
- o desenvolvimento de ferramentas que auxiliam na leitura e no folheamento de páginas eletrônicas, como sistemas que leem e-mails, dispositivos de procura em base de dados especializados e recursos de busca em material falado utilizando técnicas de reconhecimento de voz;
- a simplificação da tradução automática entre diferentes línguas;
- o desenvolvimento de modelos com interface de voz capazes de questionar o computador por meio de técnicas de reconhecimento de fala para obter respostas aplicáveis em sistemas inteligentes para a casa ou o carro e em jogos lúdicos e didáticos;
- a facilitação de tarefas diárias, como compras em *e-commerce*;
- a criação de ferramentas computacionais que facilitam e estimulam o processo de ensino aprendizagem em geral e, em particular, o ensino-aprendizagem de línguas estrangeiras e maternas, como jogos interativos e didáticos e programas que ajudam a aperfeiçoar a pronúncia;
- o acesso de pessoas com deficiências articulatórias ou visuais à sociedade de informação por meio de programas de reconhecimento e síntese que os ajudem a realizar tarefas de leitura e escrita;
- a indexação de bases de dados utilizadas em bibliotecas digitais e coleções de materiais diversos;

- e a segurança do usuário com base em características individuais da voz e que apenas reajam às suas ordens.

Como pode-se observar, são variados os usos e as aplicações dessa área multidisciplinar. Tendo isso em vista, enfoca-se em conhecer quais são as áreas estudadas entre os anos de 2010 e 2020 a fim de reconhecer o que e como essas pesquisas são realizadas e apontar futuros estudos que ainda podem ser feitos.

3 MÉTODOS

Conforme já mencionado, a metodologia de pesquisa deste artigo é baseada na pesquisa bibliográfica com o intuito de demostrar um claro entendimento da origem, da usabilidade e dos processos que são utilizados na criação dos sistemas relacionados a manipulação da linguagem humana. Macedo (1996) afirma que esse tipo de pesquisa envolve a busca de informações bibliográficas, a seleção de documentos que se relacionam com o problema de pesquisa (em livros, em verbetes de enciclopédia, em artigos de revistas, em trabalhos de congressos, em teses etc.) e o respectivo fichamento das referências para que sejam posteriormente utilizadas.

Tendo isso em vista, a abordagem desta pesquisa é qualitativa, em que se é demonstrado o estudo e entendimento da Linguística Computacional por meio de um ponto de vista amplo sobre o assunto abordado. Segundo Denzin e Lincoln (2006), essa abordagem está relacionada a uma abordagem interpretativa do mundo, o que significa que seus pesquisadores estudam seus objetos em seus cenários naturais, tentando entender os fenômenos em termos dos significados que as pessoas a eles conferem.

A coleta de dados foi realizada por meio de pesquisa em bancos de dados que contém trabalhos científicos, como o Google Acadêmico[1], o SciELo[2] e o Periódicos da Capes[3] por conta da facilidade de acesso[4]. Foram buscados os termos Linguística Computacional e Processamento de Linguagem Natural (que foi considerado por ser uma subárea da Ciência da

[1] https://scholar.google.com.br/?hl=pt
[2] https://scielo.org/
[3] http://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php?
[4] A coleta de dados deste artigo foi realizada durante o segundo semestre de 2020, quando as bibliotecas físicas estavam fechadas por conta da necessidade de isolamento social causada pela pandemia de covid-19.

Computação, da Inteligência Artificial e da Linguística que estuda os problemas da geração e da compreensão automática de línguas humanas naturais) nos títulos, nos resumos ou nas palavras-chave.

No total, foram encontrados 17 trabalhos no Google Acadêmico, 23 trabalhos no SciELo e 16 no Periódicos da Capes, totalizando 56 artigos encontrados. Desses, foram lidos apenas vinte artigos, pois (i) excluíram-se os que estavam em outras línguas que não fossem em português, em espanhol ou em inglês, por questão de proficiência de leitura das autoras e (ii) os que não estavam entre os anos de 2000 e de 2020. Tal período foi selecionado por conter dados mais atuais que podem revelar tendências futuras.

4 RESULTADOS E DISCUSSÃO

A análise permitiu a identificação de quatro categorias bastante citadas nos trabalhos. A primeira categoria é a que contém o termo "Linguagem Natural", que enfoca a necessidade de compreender a forma como a comunicação ocorre por meio da fala para a construção de sistemas. Os estudos que mais abordaram esse tema são "Linguistas e computadores: que relação é essa?" (BRITO 2004) e "*Análisis de contenido y lingüística computacional: su rapidez, confiabilidad y perspectivas*" (CHÁVEZ; YAMAMOTO 2014).

Em seguida, o mais explorado foi "Inteligência Artificial", que mostra a junção da linguística e da informática para criar processos computacionais para o controle da linguagem humana. Os artigos que mais citam a inteligência artificial são "*Computational Linguistics*" (*Stanford Encyclopedia of Philosophy,* 2014), "*Modelo para detección automática de errores léxico-sintácticos en textos escritos en español*" (RODRÍGUES; OSPINA; VELÁSQUEZ, 2018) e "Conheça tudo sobre área de Linguística Computacional!" (IMPACTA, 2019).

Outro tema bastante abordado foi "Linguística de Corpus". Tal relação se dá pelo fato de a Linguística Computacional facilitar o planejamento, a construção e a análise de corpus, que é uma base de dados de textos escritos e de registros orais que servem para análise linguística. Esse assunto é abordado nos seguintes trabalhos: "Linguística computacional: uma intersecção de áreas" (FROMM, 2006), "A preparação de material terminológico em língua inglesa por meio de ferramentas linguístico-computacionais" (SILVA; BABINI 2011), "Sobre a construção de um léxico da afetividade para o processamento computacional do português" (FREITAS, 2013), "Corpus, Linguística Computacional e as Humanidades Digitais"

(FREITAS, 2015) e "*Los Corpus Del Español Clásico y Moderno: entre la filología y la Linguística Computacional*" (CAMPOS, 2019).

E, por último, um tema recorrente foi "Processamento de Linguagem Natural", que se preocupa diretamente com o estudo da linguagem voltado para a construção de *softwares*, aplicativos e sistemas computacionais específicos. Ele foi abordado nos estudos "Linguística Computacional: uma breve introdução" (OTHERO, 2006), "O estudo Linguístico-Computacional da Linguagem" (SILVA, 2006), "*Generating abstracts from genre structure through lexicogrammar: Modelling of feature selection and mapping*" (CASTEL, 2006), "Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade" (NUNES, ALUISIO; PARDO, 2010) , "O pecado original da linguística computacional" (MARTINS, 2011), "*La generación de lenguaje natural: análisis del estado actual*" (VICENTE, BARROS, PEREGRINO, AGULLÓ, LIORET, 2015) e "*Modelo para detección automática de errores léxico-sintácticos en textos escritos en español*" (RODRÍGUES; OSPINA; VELÁSQUEZ, 2018),

Entre os trabalhos lidos um se encaixou em duas categorias, sendo elas "Inteligência Artificial" e "Processamento da Linguagem Natural". Esses temas foram abordados no artigo "*Modelo para detección automática de errores léxico-sintácticos en textos escritos en español*" (RODRÍGUES; OSPINA; VELÁSQUEZ, 2018).

Dentre os vinte trabalhos lidos, quatro não se encaixaram nas categorias mais citadas acima. Eles tratavam sobre a Linguística Computacional, mas não abordavam em si as categorias que mais se observaram em outros trabalhos lidos. Os que não se encaixaram foram "Donatus: uma interface amigável para o estudo da sintaxe formal utilizando a biblioteca em Python do NLTK" (ALENCAR, 2012), "Categoria: Linguística Computacional" (IMPORT LINGUISTICS, 2016), "Uma implementação computacional de construções verbais perifrásticas em francês" (ALENCAR, 2017) e "Linguística computacional" (GARCIA, 2020).

5 CONSIDERAÇÕES FINAIS

Este trabalho visou contribuir para o entendimento da Linguística Computacional com o intuito de demostrar um claro entendimento não só da sua origem, mas também da sua usabilidade nos processos de criação de sistemas relacionados à manipulação da linguagem

humana. Além disso, com esta pesquisa, foi possível encontrar outras áreas pertinentes à Linguística Computacional.

Pode-se ainda observar que esse campo de pesquisa está intimamente ligado à Inteligência Artificial, que segundo Brito (2004), constitui-se em um conjunto de técnicas de programação para resolver determinados tipos de problemas em informática. Ela procura imitar, por meio dos programas que comandam máquinas, as formas de resolução de problemas do mesmo modo que o homem o faz. Assim, a Inteligência Artificial e a Linguística Computacional são motivadas pelo desejo de fazer uma tecnologia melhor e de entender melhor os processos humanos de comunicação.

Os trabalhos mais encontrados foram do ano 2010 e, apesar de ser um campo teórico muito usado, existe uma escassez de trabalhos dos anos atuais. Uma das principais usabilidades citadas da Linguística Computacional é a questão de tradução, pois ainda se buscam mecanismos eficientes para uma perfeita compreensão da fala natural.

Dentre as limitações para a realização desta pesquisa, destaca-se a necessidade de usar apenas trabalhos disponíveis nas bases de dados de amplo acesso que são gratuitas. Algumas bases que poderiam conter mais estudos são pagas. Além disso, muitas teses acessadas nas bases de amplo acesso não foram encontradas, porque seus arquivos não estavam disponíveis. Para futuras pesquisas, ressalta-se a necessidade de incluir esses dados que podem revelar outras relações dessa área multidisciplinar.

REFERÊNCIAS

ALENCAR, L. F**.** Uma implementação computacional de construções verbais perifrásticas em francês. **Alfa: Revista de Linguística**, v. 61, n. 2, p. 351-380, 2017. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-57942017000200351&lang=pt. Acesso 30 out. 2020.

_____. Donatus: uma interface amigável para o estudo da sintaxe formal utilizando a biblioteca em Python do NLTK. **Alfa: Revista de Linguística**, v. 56, n. 2, p. 523-555, 2012. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-57942012000200008&lang=pt. Acesso em: 30 out. 2020.

BRITO, G. S. Linguistas e computadores: que relação é essa**? Working papers em Linguística**, n. 4, p. 7-23, 2004. Disponível em: https://periodicos.ufsc.br/index.php/workingpapers/article/viewFile/4711/3951. Acesso em: 17 set. 2020.

CAMPOS, M. C. Los corpus del español clásico y moderno: entre la filología y la lingüística computacional. **RLA**, v. 57, n. 2, p. 41-64, 2019. Disponível em: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-48832019000200041&lang=pt. Acesso em: 29 out. 2020.

CASTEL, V. Generating abstracts from genre structure through lexicogrammar: modelling of feature selection and mapping. **Revista Signos**, v.39, n. 62, p.327-356, 2006. Disponível em: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-09342006000300001&lang=pt. Acesso em: 30 out. 2020.

CHÁVEZ, B. L.; YAMAMOTO, J. M. Análisis de contenido y lingüística computacional: su rapidez, confiabilidad y perspectivas. **Anal. Psicol**., v. 30, n.3, p. 1146-1150, 2014. Disponível em: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-97282014000300039&lang=pt. Acesso em: 29 out. 2020.

COMPUTATIONAL LINGUISTICS. *In*: STANFORD ENCYCLOPEDIA OF PHILOSOPHY. Stanford: Stanford University, 2014. Disponível em: https://plato.stanford.edu/entries/computational-linguistics/. Acesso em: 16 set. 2020.

DENZIN, N. K.; LINCOLN, Y. S. Introdução: a disciplina e a prática da pesquisa qualitativa. *In*: DENZIN, N. K.; LINCOLN, Y. S. (org). **O planejamento da pesquisa qualitativa**: teorias e abordagens. Porto Alegre: Artmed, 2006. p. 15-41.

FREITAS, C. Corpus, Linguística Computacional e as Humanidades Digitais. *In*: LEITE, M.; GABRIEL, C. T. (org). **Linguagem, Discurso, Pesquisa e Educação**. Rio de Janeiro: De Petrus, 2015, p. 18-46.

_____. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **RBLA**, v. 13, n. 4, p.1031-1059, 2013. Disponível em: https://www.scielo.br/scielo.php?pid=S1984-63982013000400004&script=sci_abstract&tlng=pt. Acesso em: 29 out. 2020.

FROMM, G. Linguística Computacional: uma intersecção de áreas. **Revista Factus**, n. 5, p. 135-140, 2006. Disponível em: http://www.ileel.ufu.br/guifromm/wp-content/uploads/2014/05/linguisticacomputacional.pdf. Acesso em: 22 set. 2020.

GARCIA, A. I. C. **Linguística Computacional**. Disponível em: https://www.usc.gal/export9/sites/webinstitucional/gl/centros/filoloxia/guiacentros/arquivos/LINGCA/LINGSTICA_COMPUTACIONAL.pdf. Acesso em: 29 out. 2020.

IMPACTA. **Conheça tudo sobre área de Linguística Computacional!** [*S.l.*], 2020. Disponível em: https://www.impacta.edu.br/blog/conheca-tudo-area-linguistica-computacional/. Acesso em: 22 set. 2020.

IMPORT LINGUISTICS. **Categoria**: Linguística Computacional. [*S.l.*], 4 fev. 2016. Disponível em: https://importlinguistics.com/category/linguistica-computacional/. Acesso em: 22 set. 2020.

LINGUÍSTICA COMPUTACIONAL. *In*: INFOPÉDIA. Porto: Porto Editora, 2003-2020. Disponível em: https://www.infopedia.pt/$linguistica-computacional. Acesso em: 27 out 2020.

MACEDO, N. D. **Iniciação à pesquisa bibliográfica**. São Paulo: Unimarco Editora, 1996.

MARTINS, R. O pecado original da linguística computacional. **Alfa: Revista de Linguística**, v. 5, n. 1, p. 1-22, 2011. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4178/3776. Acesso em: 29 out. 2020.

NUNES, M. G. V.; ALUISIO, S. M.; PARDO, T. A. S. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. **LinguaMática**, v. 2, n. 2, p. 13-27, 2010. Disponível em: https://www.linguamatica.com/index.php/linguamatica/article/view/66. Acesso em: 27 out. 2020.

OTHERO, G. A. Linguística Computacional: uma breve introdução**. Letras de Hoje**, v. 41, n. 2, p. 341-351, 2006. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/605/436. Acesso em: 26 out. 2020.

RODRÍGUEZ, M. D. B.; OSPINA, A. A. P.; VELÁSQUEZ, I. M. R. Modelo para detección automática de errores léxico-sintácticos en textos escritos en español. **TecnoLógicas**, v. 21, n.42, p. 199-209, 2018. Disponível em: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-77992018000200199&lang=pt. Acesso em: 29 out. 2020.

SILVA, B. C. D. O estudo Linguístico-Computacional da Linguagem. **Letras de Hoje**, v. 41, n. 2, p. 103-138, 2006. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/597/428. Acesso em:26 out. 2020.

SILVA, E. B.; BABINI, M. A preparação de material terminológico em língua inglesa por meio de ferramentas linguístico-computacionais. **Trabalhos em Linguística Aplicada**, v. 50, n. 1, p. 119-132, 2011. Disponível em: https://www.scielo.br/scielo.php?pid=S0103-18132011000100007&script=sci_arttext. Acesso em: 26 out. 2020.

VICENTE, M.; BARROS, C.; PEREGRINO, F. S.; AGULLÓ, F.; LIORET, E. La generación de lenguaje natural: análisis del estado actual. **Comp. y Sist.**, v. 19, n. 4, p. 721-756, 2015. Disponível em: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462015000400721&lang=pt. Acesso em: 29 out. 2020.

VIEIRA, R.; LIMA, V. L. S. Linguística computacional: princípios e aplicações. *In*: NEDEL, L. P. (ed.). **IX Escola de Informática da SBC-Sul**. Passo Fundo, Maringá, São José. SBC-Sul, 2001.