IDENTIFICAÇÃO DE *WORDNET* EM PORTUGUÊS COMO BASE DE DADOS PARA COMPARAÇÃO SEMÂNTICA DE PRODUÇÃO TEXTUAL

Rute Steffany da Silva[1], José Leonardo Oliveira Lima[1]
rute30.10@gmail.com, jjleo@ueg.br (https://orcid.org/0000-0001-8869-3056)

[1] Universidade Estadual de Goiás, Sistemas de Informação, Anápolis, Goiás, Brasil.

RESUMO: A Educação a Distância e a Educação Híbrida trazem desafios para os professores na obtenção de *feedbacks* a respeito da aprendizagem dos estudantes; dentre esses desafios, destaca-se o problema de verificação se o discente, por meio de suas produções textuais, está desenvolvendo conhecimento a partir dos textos fornecidos pelo professor para o estudo de um conteúdo. Face a esse problemática, o objetivo da pesquisa foi identificar a *Wordnet* em Português que melhor se adequa ao processo de comparação semântica – com foco no estudo da base de dados da *Wordnet*, o seu funcionamento e como é organizada no banco de dados – para ser utilizada em um futuro algoritmo de comparação semântica de produção textual. A pesquisa foi metodológica e descritiva, tendo também caráter bibliográfico e documental, esse último em virtude da busca por documentação, código fonte e organização do banco de dados da *Wordnet*. Entre as seis *Wordnets* em português identificadas e analisadas, foi escolhida a OpenWordnet-PT, tendo como critério de escolha prevalente, dentre outros indicadores, o fato de ser livre e de código aberto, o que favorece (ao menos teoricamente) a consulta de sua estrutura e organização e a integração em outros projetos. Outros estudos ainda precisam ser feitos, pois a análise da *OpenWordnet-PT* foi feita utilizando apenas a versão *web* remotamente para consultas e testes, em virtudes de limitações apresentadas na conclusão do artigo.

Palavras-Chave: Comparação semântica, educação a distância, educação híbrida, rede semântica e *Wordnet*.

IDENTIFICATION OF WORDNET IN PORTUGUESE AS DATABASE FOR SEMANTIC COMPARISON OF TEXTUAL PRODUCTION

ABSTRACT: Distance Education and Hybrid Education bring challenges for teachers in obtaining feedback regarding student learning; among these challenges, there is the problem of verifying whether students are developing knowledge from the texts provided by the teacher for content studying, through their writing productions. Delimited to this problem, the main goal was to identify a Wordnet in Portuguese that best suits the semantic comparison process - focusing on the study of the Wordnet database, its operation, and how it is organized in the database system - to be used in a future algorithm for semantic comparison of textual production. The research was methodological and descriptive, also having a bibliographic and documentary characteristic, the latter due to the search for Wordnet's documentation, organization, source code, and database.

Among the six Wordnets in Portuguese identified and analyzed, OpenWordnet-PT was chosen, having as a prevalent choice criterion, among other indicators, the fact that it is free and open-source, which favors, at least theoretically, the consultation of its structure and organization and the integration on other projects. Other studies still need to be done, as the analysis of OpenWordnet-PT was carried out using only the remote web version for consultations and tests, due to limitations presented in the conclusion of the article.

Keywords: Semantic comparison, distance education, hybrid education, semantic network and Wordnet.

1. INTRODUÇÃO

A Educação a Distância (EaD) vem conquistando seu espaço no ensino independentemente da área de estudo ou grau, no mundo todo. A EaD foi introduzida no ensino superior em meados da década de 40, em uma Universidade da África do Sul (PETERS, 2006 apud LIMA, 2016). A *Open University,* na Inglaterra, propôs um modelo pioneiro da educação superior a distância na Europa, oferecendo seus primeiros cursos no começo da década de 70 (LIMA, 2016). Já no Brasil, a implementação da EaD na educação superior foi mais tardia, a partir dos anos 2000, mas atualmente, a EaD é amplamente implementada na graduação nas Instituições de Educação Superior (IES).

O Ensino Híbrido, conhecido também como Educação Híbrida, é outro processo que emerge na atualidade que alia diversos métodos e recursos tecnológicos de aprendizagem (CHAVES FILHO et al., 2006 p. 84 apud RODRIGUES, 2010). No contexto da Educação Híbrida, as tecnologias digitais da informação e comunicação, inclusive as usadas amplamente na EaD, aliadas aos melhores processos e técnicas da modalidade presencial, juntamente com conceitos contemporâneos relacionados à aprendizagem ativa, aprendizagem baseada em problemas, gamificação na educação, simulação virtual, dentre outros, juntam-se para oferecer uma aprendizagem mais efetiva e alinhada com o contexto social e ambiente digital nos quais os estudantes já estão imersos.

Com base na pesquisa de Lima (2016), um dos desafios enfrentados pelos docentes na EaD é a obtenção dos *feedbacks* sobre a aprendizagem dos discentes, que envolve saber se os alunos estão aprendendo com os materiais de estudo que os docentes disponibilizam, e se, por meio desses materiais, os alunos estão conseguindo desenvolver novos conhecimentos. Na EaD (e na Educação Híbrida, dependendo do processo educativo utilizado), uma das formas de verificar a aprendizagem é por intermédio da produção textual realizada pelos estudantes, que sintetiza e permite ao docente verificar e analisar o nível de compreensão e assimilação dos conteúdos estudados pelos estudantes.

Portanto, tem-se como delimitação do presente estudo, a comparação das produções textuais

dos alunos, a fim de identificar se os estudantes estão adquirindo conhecimento com os materiais disponibilizados para leitura e estudo. Para isso, o foco do estudo envolve as comparações semânticas e as *Wordnets*.

Segundo Rosa (2016), a rede semântica é um sistema de representação do conhecimento baseado em grafos, onde os vértices representam as palavras e as arestas os relacionamentos entre elas. Seguindo a mesma lógica, as *Wordnets* também são um tipo de rede semântica, visto que também favorece manipulações de conjuntos de palavras, e de suas semânticas (ROSA, 2016). As *Wordnets* são, portanto, uma base de dados lexical (de palavras) que tem uma estrutura de relação entre as palavras (formado pela rede de grafos e arestas) que envolve a semântica (significado das palavras), favorecendo a identificação de sinônimos, antônimos, frases de conteúdo igual escritas com palavras e termos diferentes, tradução etc.

A similaridade semântica propicia comparar e analisar a aproximação de termos cujos significados se assemelham, utilizando-se, portanto, de redes semânticas como as *Wordnets,* dentre outros.

Face à problemática apresentada e ao crescimento da modalidade de EaD e Educação Híbrida no Brasil, a partir das necessidades identificadas por Lima (2016) e estudo decorrentes realizados por Costa (2017) e Santos (2019), vislumbrou-se o estudo de comparação semântica utilizando as *Wordnets* como recurso para um futuro recurso tecnológico que favoreça um *feedback* para apoio aos professores na avaliação da aprendizagem dos alunos.

Assim, colocou-se como problema de pesquisa: Qual *Wordnet* em Português poderia ser escolhida para ser usada em um algoritmo de comparação semântica de textos, de forma a verificar se o discente, por meio de suas produções textuais, está construindo conhecimento a partir dos textos disponibilizados pelo docente?

Portanto, a pesquisa teve como objetivo geral: Identificar a *Wordnet* em Português que melhor se adequa ao processo de comparação semântica – com foco no estudo da base de dados da *Wordnet,* o seu funcionamento e como ela é organizada no banco de dados – para ser utilizada em um futuro algoritmo de comparação semântica de produção textual.

2. REVISÃO BIBLIOGRÁFICA

2.1 Bases de Dados Lexicais e Processamento de Linguagem Natural

Segundo Greghi (2002) as Bases de Dados Lexicais (BDL) são normalmente utilizadas como repositório central de informações lexicais de uma determinada língua. Tais informações armazenadas podem ser de âmbito sintático, semântico, morfológico, podendo também, expressar relações lexicais da mesma língua, ou entre línguas distintas, sendo que essas bases de dados

lexicais facilitam a manutenção e manipulação das informações (GREGHI, 2002). As bases de dados são disponibilizadas em formato digital por intermédio dos banco de dados.

As informações que são armazenadas em uma BDL podem ser utilizadas em aplicações de Processamento de Linguagem Natural (PLN), por exemplos, os revisores gramaticais e/ou tradutores automáticos. As BDL permitem a manipulação, manutenção e a atualização de dicionários e, também, a troca de informações com outras BDLs (bases de dados com dicionários de outros idiomas, de áreas ou domínios específicos etc.), possibilitando integração e, mesmo, validação da coerência das informações entre diversas BDLs.

O Processamento da Linguagem Natural (PLN), segundo Gonzalez e Lima (2003), envolve fazer com que o computador se comunique na linguagem humana, sendo por meio dos variados aspectos de comunicação, como sons, palavras, frases e referências, entre outros. O PLN pode ser utilizado tanto no entendimento da comunicação dos computadores por meio de linguagem humana, quanto no auxílio de tradução e geração de sínteses, dentre outros. O papel desse processo é dar significado às frases e às palavras.

Existem, portanto, diferentes segmentos que buscam analisar a comunicação humana, como as áreas (e campos) de Comunicação, Letras (linguística, fonologia), Computação (inteligência artificial, similaridade semântica), Medicina (neurociência), Física (acústica) etc., mas que se correlacionam multi e interdisciplinarmente. A PLN é um desses campos interdisciplinares que se vale dessa diversidade e da interação entre ramos do conhecimento.

Convém ressaltar que, para a presente pesquisa, dentro da PLN, o foco foi o estudo da similaridade semântica pela perspectiva da computação, que busca representar os termos e palavras digitalmente e favorecer a comparação levando em consideração os significados.

2.2 Rede Semântica e Similaridade Semântica

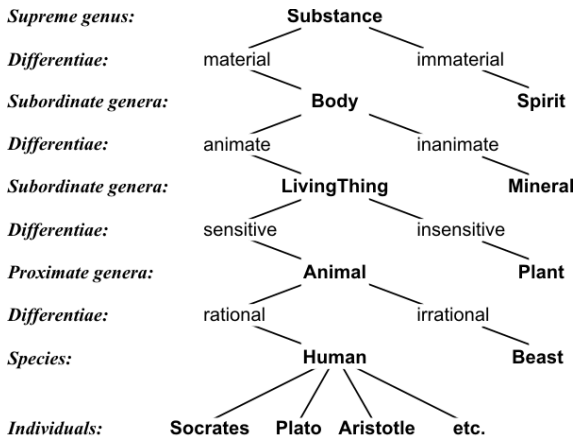
A Rede Semântica, no campo de estudos da inteligência artificial, é definida como uma abordagem para a representação de conhecimento e PLN (ROSA, 2016). Segundo Krippendorff (2004 apud ROSA, 2016, p. 1) muitos pesquisadores da inteligência artificial, na década de 60, analisaram a possibilidade de gerar automaticamente resumo de documentos escritos. As palavras determinadas mais importantes em um texto, são as usadas com mais frequência**;** assim**,** pelo entendimento da época, referidas palavras poderiam definir todo o texto.

Sowa (1992 apud ROSA, 2016) define rede semântica como uma estrutura baseada em grafos representando padrões de vértices e arestas interconectadas, conforme exemplo na ilustração 1.

Para Henrique et al. (2014 apud ROSA, 2016), a rede semântica baseada em palavras é um sistema de representação do conhecimento fundamentado em grafos; os grafos são representados

por vértices (palavras) e arestas (relações entre as palavras). Cunha (2013) define que a rede semântica tem como seu elemento principal a relação entre palavras, adequando-se a variedades de métodos computacionais, sendo, assim, considerado um dicionário legível para máquinas.

Ilustração 1 - Exemplo de grafo e aresta de rede semântica



Fonte: Sowa (1992, p. 2)

A Similaridade Semântica Textual (STS), tem por tarefa verificar a igualdade semântica entre dois textos, considerando que alguns textos são mais semelhantes entre si do que outros (AGIRRE et al., 2013 apud FREIRE et al., 2016). A avaliação de similaridade textual é indispensável nas crescentes tarefas do PLN (LIN; HOVY, 2003 apud FREIRE et al., 2016) para que se possa processar os textos para, por exemplo, identificar o tipo de plágio que é feito alterando palavras do texto original por sinônimos.

De acordo com Wang (2005 apud SILVA, 2008) e Petrakis et al. (2006 apud SILVA, 2008), as medidas de similaridade semântica podem ser divididas em quatro categorias principais, sendo:

- Abordagem baseada em ontologias: Categorizada como todas as abordagens que utilizam recursos e bases de conhecimento, como as ontologias e *Wordnets*, calculando o grau de similaridade;

- Abordagem baseada no índice de informações compartilhadas: Quanto mais um termo tende a ser similar a outro mais se assemelham, ou seja, compartilham o mesmo grau de informação;

- Abordagem baseada em características: Os termos só são semelhantes se compartilharem das mesmas características, exemplo: a palavra “estudante”, é semelhante a aluno, discente, acadêmico, entre outras, que remetem a uma pessoa que estuda;

● Abordagem híbrida: Junção de duas ou mais das abordagens anteriores, tornando o seu processo de similaridade mais seguro, pois faz a comparação dos termos de mais de uma forma possível.

A similaridade semântica tem uma ampla área de aplicação, dentre as quais pode-se destacar: recuperação de informação; integração de dados; integração de base de dados; detecção de ambiguidades e duplicação de informação (COSTA et al.,2014).

Segundo Costa et al. (2014), a *Wordnet* surge como fonte de informação principal para as abordagens descritas. A própria *Wordnet* faz o uso de similaridade semântica, pois classifica e agrupa as palavras conforme os seus significados (FEITOSA; PINHEIRO, 2017).

2.3 *Wordnets* e recursos de representação digital

Aplicando os conceitos previamente citados, as *Wordnets* são BDLs e um tipo de rede semântica, pois manipulam um grande conjunto de palavras, com suas relações semânticas favorecendo os processos de análise de similaridade semântica.

Miller (1995) apresenta que a *Wordnet* é um conjunto de substantivos, adjetivos, verbos e advérbios, organizados em grupos de sinônimos denominados "*synsets*", sendo interligados por relações semânticas. Segundo a definição de Moraes (2008), cada *synset* reúne itens lexicais compartilhando um mesmo conceito. As definições das palavras da *Wordnet* são organizadas de forma conceitual (LEACOCK; CHODOROW, 1998 apud FELLBAUM, 1998), diferenciando-se da organização alfabética de um dicionário comum.

As *Wordnets* se assemelham a dicionários eletrônicos, por fazer manipulação de sinônimos e antônimos de relações lógico-conceitual, essa definição foi introduzida inicialmente para a *Wordnet* de Princeton, conforme apresenta Dias-da-Silva (2005), em seu artigo “A construção da base da Wordnet.Br: conquistas e desafios”.

Quando se fala em *Wordnet* é necessário especificar qual, pois atualmente existem várias *Wordnets* para diversas línguas, cada uma com suas peculiaridades.

A primeira *Wordnet* criada foi a de *Princeton* (WN.Pr), desenvolvida na “*Princeton University*" e está disponível desde a década de 90 (MILLER, 1995 apud FELLBAUM, 1998). A WN.Pr é considerada a mãe de todas as *Wordnets* existentes até a atualidade, servindo de base para tantas *Wordnets* criadas para diversas línguas, tendo se tornado um modelo padrão. A WN.Pr teve grande popularidade e ampla utilização por conta de sua flexibilidade e por ser gratuita.

Dias-da-Silva (2005) explica que a estruturação das unidades sinônimas da WN.Pr é distribuída em quatro categorias lexicais, sendo, os verbos, substantivos, adjetivos e advérbios, ou seja, as quatro categorias básicas e comuns entre todas as *Wordnets*. Essas categorias são

denominadas conjuntos de sinônimos, ou simplesmente, *synsets*.

Por conta da grande popularidade da WN.Pr como base de conhecimento, foi criada uma organização não comercial, a *Global WordNet Association* (GWA), uma plataforma que partilha, discute e faz ligação com as *Wordnets* de todo o globo (BOND; PAIK, 2012; OLIVEIRA et al. 2015). Surgiram também muitas *Wordnets* multilíngues, dentre as quais a EuroWordNet (VOSSEN, 1997), que possibilita criar *Wordnet* para diferentes línguas, todas tendo como base principal a WN.Pr.

A MultiWordNet (PIANTA et al., 2002), é usada para fazer as traduções das *Wordnets* existentes. Além dessas, destacam-se também a BalkaNet (STAMOU et al., 2002 apud OLIVEIRA et al., 2015), para a língua dos Balcãs, a Multilingual Central Repository (MRC) (GONZÁLEZ-AGUIRRE; RIGAU, 2013 apud OLIVEIRA et al., 2015), dedicada às línguas faladas na Espanha. A Open Multilingual WordNet (OMWN) (BOND; FOSTER, 2013 apud OLIVEIRA et al., 2015) é uma iniciativa que tem como objetivo facilitar o acesso às diferentes *Wordnets* de diversas línguas.

Em resumo, o que todas as *Wordnets* aqui apresentadas têm em comum, além da origem a partir da WN.Pr, é que são usadas para organizar informações, fazer relações semânticas entre termos e palavras, por meio dos *synsets*, que são armazenados em suas bases de dados, além de serem usadas na representação do conhecimento.

Dentre tantas *Wordnets* existentes, foi escolhido para o presente trabalho fazer o uso de uma *Wordnet* em português, para isso viu-se então a necessidade de identificar e analisar as que são específicas para o Português.

Até o final da pesquisa, foram identificadas seis *Wordnets* existentes desenvolvidas para o português, umas ainda não finalizadas e outras já prontas. Dentre as *Wordnets* identificadas, há três que é preciso pagar uma licença para usar os recursos que ela oferece, e outras três que são caracterizadas como livres, seu *download* e uso podem ser feitos gratuitamente.

As *Wordnets* pagas são: a WordNet.PT (MARRAFA, 2001, 2002 apud OLIVEIRA et al., 2015), a Wordnet.BR (DIAS DA SILVA et al., 2002 apud OLIVEIRA et al., 2015) e a MultiWordNet.PT (PIANTA et al., 2002 apud OLIVEIRA et al., 2015).

As *Wordnets* livres em Português surgiram a partir do início da década de 2010 e também seguiram o princípio de código aberto, de forma a possibilitar o acesso e uso de recursos léxico-semânticos de forma mais ampla e útil à comunidade de pesquisadores, desenvolvedores e usuários (OLIVEIRA et al., 2015). São elas: a Onto.PT (OLIVEIRA; GOMES, 2010 apud OLIVEIRA et al., 2015), a OpenWordNet-PT (PAIVA et al., 2012; RADEMAKER et al., 2014; OLIVEIRA et al., 2015) e por último a Ufes WordNet (GOMES et al., 2013 apud OLIVEIRA et al., 2015).

As *Wordnets* utilizam de estruturas de representação digital por meio de linguagem ou estrutura de programação, modelos e bancos de dados. Algumas das principais Wordnets analisadas nesse trabalho usam, por exemplo, o RDF/OWL – *Resource Description Framework/Ontology Web Language- e* o SPARQL que detalharemos na sequência.

Segundo Miller (1998 apud FERREIRA; SANTOS, 2013), o RDF é uma estrutura de dados para Web que permite a codificação, o reuso de metadados estruturados e o intercâmbio de informações de uma página Web para outra, sem perda de conteúdo. O RDF permite a utilização de vocabulário legível para humanos e máquinas.

O modelo RDF é representado em forma de grafo por meio de diagramas, possibilitando tanto a facilidade de aprendizagem de sua estrutura quanto a modelagem precisa de um domínio (FERREIRA; SANTOS, 2013). Os grafos do RDF são compostos por nós e arestas, formado por uma tríplice sujeito-predicado-objeto que, respectivamente, são: o recurso, a propriedade e a sentença ou valor, sendo esses os três conceitos básicos do modelo RDF (SANTOS, 2002; FERREIRA; SANTOS, 2013).

A partir do modelo RDF, por ser o modelo base da *web*, muitas outras linguagens foram incorporadas em sua estrutura a fim de aprimoramento de seus recursos, dentre elas a OWL – *Ontology Web Language*, uma linguagem de programação para instanciar ontologias na *web*, considerada também uma linguagem de representação do conhecimento, composta por um conjunto de classes e um conjunto de propriedades, usadas para processamento de informações de algum domínio (REYNOLDS et al., 2005; CHALUB; RADEMAKER, 2016).

Assim, a linguagem OWL é considerada um subconjunto da RDF, tal padrão de linguagem é referenciado como RDF/OWL. Segundo Reynolds et al. (2005), o RDF/OWL modela o mundo em termos de: instância de recursos, classes, propriedades e triplos.

O SPARQL é o protocolo de busca adotado pela W3C, para extração da informação na web semântica, sua estrutura que lembra uma SQL, com o diferencial de ser otimizado para trabalhar com a estrutura de triplas do RDF e sendo considerada uma linguagem de consulta gráfica (PEREZ et al., 2005; NUNES, 2014).

3. METODOLOGIA

Seguindo os critérios de classificação de pesquisas apresentados por Vergara (2013), o **tipo de pesquisa** foi caracterizado em dois tipos básicos, **quanto aos fins e quanto aos meios.** A partir desses critérios, a pesquisa foi caracterizada quanto aos fins como uma pesquisa descritiva e metodológica.

Considerou-se a pesquisa metodológica, pois visou o levantamento teórico das *Wordnets* e

bases de dados para a identificação da *Wordnet* em português que melhor se adequou ao problema de pesquisa; similarmente, foi feito o levantamento metodológico das redes semânticas e das comparações semânticas na intenção de integrá-las à pesquisa.

Descritiva, visto a necessidade de analisar e descrever as características da Wordnet, as metodologias levantadas das similaridades de textos e comparações semânticas, da própria *Wordnet*, visto que as mesmas trabalham com relação entre palavras de um contexto (SANTOS, 2019).

A pesquisa quanto aos meios, foi bibliográfica e documental: Bibliográfica, pois foi necessário, durante o processo de desenvolvimento da pesquisa, estar fazendo análises e pesquisas em fontes secundárias, como livros, dissertações de mestrado, em teses de doutorado e materiais publicados em revistas científicas, para a planificação do projeto.

Documental, pois, segundo Lakatos (2003), foram realizadas coletas de dados em documentos de fontes primárias, que resultaram na explicação do uso do banco de dados, que na pesquisa em Sistemas de Informação é considerado documento em suporte digital, assim como as *Wordnets*.

Como instrumento de coleta de dados foram feitas buscas por meio de palavras-chaves dos assuntos referentes à temática, que foram catalogadas no Sistema Zotero de documentação pessoal. Utilizou-se também as fontes primárias de informação como as bases de dados *Wordnets*.

Os procedimentos de coletas de dados ocorreram por meio das pesquisas bibliográficas em livros, teses de mestrados e doutorados, dissertações e páginas web das *Wordnets* identificadas, encontrados via pesquisa de palavras-chaves com dados pertinentes a temática, fazendo os estudos e levantamento exploratórios e coleta de dados. Nas páginas web encontradas, foram realizadas buscas, com palavras chaves, para a visualização e estudo das comparações semânticas.

Para a análise das *Wordnets*, montou-se um quadro operacional com os indicadores de análise que emergiram do estudo do referencial teórico e que deram origem ao quadro síntese apresentado na sequência (quadro 1).

4. RESULTADOS

Após o levantamento teórico das *Wordnets* existentes, tanto para o português quanto para outras línguas, pôde-se constatar que existem diversas *Wordnets* e que o problema maior seria a escolha da que melhor se adequa ao problema de pesquisa em questão. Com o levantamento das *Wordnets*, especialmente as para o português, foi possível identificar seis *Wordnets* e uma ontologia que se iniciou como uma *Wordnet* e no decorrer do projeto foi adaptada para ontologia, o PULO (SIMÕES; GUINOVART, 2014 apud OLIVEIRA et al., 2015).

Dentre as *Wordnets* para o português identificadas, utilizou-se um processo de classificação para a escolha da que mais se adequa ao presente trabalho. Para essa classificação seguiu-se alguns indicadores de classificação, sendo, a princípio: se a *Wordnet* já está finalizada; a utilização é de uso livre; se possui página *web*; se disponível para *download*; se ela está sendo usada em outros projetos e por último qual a modelagem usada em seu desenvolvimento. O resultado desse levantamento é apresentado no quadro 1.

Quadro 1: Quadro comparativo detalhado das *Wordnets* para o Português

Idicadores	Wordnet.PT	Wordnet.Br	MultiWordnet.PT	Onto.PT	OpenWordnet.PT	Ufes WordNet
Data	1998	2002	2002	2008 - 2010	2010	2013
Responsavel	Palmira Marrafa	Bento Dias da Silva	Antônio Branco e NLX	Projeto de doutoramento de Hugo Gonçalo Oliveira	Valeria de Paiva, Alexandre Rademaker e Gerard de Melo	Projeto de Graduação de Marcelo Gomes
Local	Universidade de Lisboa e CLG	Universidade Estadual Paulista	Universidade de Lisboa	Universidade de Coimbra	Fundação Getulio Vargas	Universidade Federal do Esprito Santo
Versão	WN.PT 1.6, 2006	WN.BR / TeP	MWN.PT v1	Onto.PT 0.6	OpenWN-PT	UfesWN.BR 1.0
Uso	Fechada	*Synsets* livres	Licença paga	Livre	Livre	Livre
Código Aberto	Não	Não	*	Sim	Sim	*
Está sendo atualizada	Sim	*	*	Sim	Sim	*
Já está pronta para uso	Sim	Apenas os *synsets* da primeira fase	Sim	Sim	Sim	Não
Disponivel pra utilização gratuita da página *Web*	Sim	*	Sim	Sim	Sim	Não
Atualização e Criação	Manual	Manual	*	Automatica	semi-automatica / manual	*
Disponivel para *Download* de forma gratuita	Não	Não	Não	Sim	Sim	Não
Qual estrutura de dados utiliza	*	*	*	RDF/OWL	RDF / OWL	Baseada na API do *Google Translate*
Aprimoramento / expandida	WordNet.PT Global	Tep	*	Não	Aprimoramento para a ontologia SUMO	*
Utilização em outros projetos	Não	*	*	Na expansão de sinônimos para recuperação de informação e criação de listas de verbos causais	Projetos FreeLing, OMWN, *Google Translate*	*
E alinhada a *WordNet* de *Princeton*	Sim	Sim	Sim	Não	Sim	Sim
Legenda: * - Não se conseguiu localizar essa informação na documentação analisada						

Fonte: Produzido pelos autores

Após a análise do quadro comparativo, utilizando os critérios citados anteriormente foram descartadas três das *Wordnets:* a Wordnet.PT e a MultiWordnet.PT por terem licença de uso paga; a Wordnet.BR, por ainda não estar disponibilizada na época da pesquisa.

Assim ficaram as três *Wordnets* livres, sendo: a Onto.PT, a OpenWordnet-PT e a *Ufes Wordnet*, essa última que logo também foi eliminada pelo fato de ainda não estar finalizada. A Onto.PT também foi desclassificada, pois em sua própria página *web* é considerada uma ontologia lexical para o português, o que não se adequa a delimitação do trabalho: análise de *Wordnet*.

Com isso, a OpenWordnet-PT foi a escolhida, pois a mesma já se encontra finalizada e pronta para o uso, seus dados podem ser acessados em sua página *web* e o seu código fonte se

encontra disponibilizado no GitHub para *download*, onde se encontra uma breve explicação de como usar, uma breve introdução do RDF utilizado e *link's* para acessar a página *web*, sendo listados os integrantes da equipe de desenvolvimento e os contribuidores do projeto. Tendo também todas as informações de licenças e as datas de atualização.

Em relação ao banco de dados da OpenWordnet-PT , seus desenvolvedores optaram por usar o RDF/OWL por permitir maior amplitude em sua verificação e por permitir consultas semânticas.

Durante o processo de pesquisa e desenvolvimento do trabalho, foram feitas algumas buscas teste na página *online* da OpenWordnet-PT, a primeiro momento foi feito busca somente em português, com a palavra "estudante" e retornou um total de 8 resultados semânticos da palavra, com os seus *synsets* de origem e quantas palavras dentro daquele *synset* são semântica a palavra.

Uma segunda busca foi realizada utilizando a opção de buscar em todos, a busca retornou um total de 144 resultados semânticos para a palavra "*student*", nesta opção retornou-se às palavras semânticas em ambos idiomas, tanto português quanto inglês.

Portanto, a OpenWordnet-PT foi a *Wordnet* em Português, que melhor se qualificou de potencial uso para responder ao problema de pesquisa, porém, é preciso realizar estudos mais detalhados e aprofundados para a compreensão da modelagem de dados da OpenWordnet-PT em virtude das limitações de acesso efetivo ao código da OpenWordnet-PT (vide dificuldades da pesquisa na conclusão), para que, futuramente, possa ser usada em um algoritmo de comparação semântica de texto.

5. CONCLUSÃO

A partir das necessidades levantadas por Lima (2016) e estudo decorrentes realizados por Costa (2017) e Santos (2019), vislumbrou-se a possibilidade de estudo de comparação semântica utilizando as *Wordnets* como base de dados. Assim, surgiu o problema pesquisa que norteou o presente estudo "Qual *Wordnet* em Português poderia ser escolhida para ser usada em um algoritmo de comparação semântica de textos, de forma a verificar se o discente, por meio de suas produções textuais, está construindo conhecimento a partir dos textos disponibilizados pelo docente?".

O objetivo geral[1] proposto para solucionar o problema de pesquisa foi alcançado, com a identificação da *Wordnet* em Português que melhor se adequaria ao processo de comparação semântica, porém com limitações que serão detalhadas no final dessa conclusão. Dentre as *Wordnets* identificadas e analisadas, a OpenWordnet-PT mostrou ser a *Wordnet* com potencial para uso na comparação semântica textual a delimitação do problema de pesquisa. A OpenWordnet-PT

[1] Objetivo geral da pesquisa: Identificar a Wordnet em Português que melhor se adequa ao processo de comparação semântica – com foco no estudo da base de dados da Wordnet, o seu funcionamento e como ela é organizada no banco de dados – para ser utilizada em um futuro algoritmo de comparação semântica de produção textual.

atendeu aos principais critérios: estar finalizada; ser disponibilizada para download (ao menos teoricamente, como será esclarecido mais adiante); ser de uso livre e de código aberto; e poder ser integrada em outros trabalhos para uso na comparação semântica. Segundo a documentação analisada, a OpenWordnet-PT já é utilizada em outros trabalhos, como o Google *Translate*, sendo considerada a representante das *Wordnets* abertas do Português, por conta de sua vasta cobertura e qualidade.

A base de dados da OpenWordnet-PT é o resultado da tradução da base de dados da *Wordnet* de *Princeton*, portanto, contém uma base de dados com grande abrangência. As pesquisas em sua página *web* podem ser realizadas tanto em português quanto em inglês.

No estudo foram identificadas e analisadas seis *Wordnets* em Português, das quais três são pagas e tem o código fechado. As outras três são de uso livre e o seu código é aberto, podendo ser baixadas e integradas em outros projetos, sem necessidade de pagamento.

No que se refere às limitações da pesquisa, algumas dificuldades foram enfrentadas para a análise do código fonte da OpenWordnet-PT, por conta da inconsistência de documentação, não sendo possível realizar testes de busca, utilizando o SPARQL, em sua base de dados disponibilizada online, pois o site encontrava-se fora do ar, problema que perdurou até o momento de finalização da pesquisa, não sendo possível o acesso efetivo e a testagem local. Os testes realizados na OpenWordnet-PT foram feitos de forma limitada e em sua versão web remotamente.

Sugere-se, como futuro estudo, acessar e baixar à OpenWordnet-PT e fazer o estudo de sua estrutura localmente, para compreensão mais aprofundada das suas formas de representação e testagem prática de suas potencialidades e limitações em processamento de similaridade semântica em um algoritmo para análise da produção textual de alunos a partir dos materiais de estudo disponibilizados pelo docente na EaD ou na Educação Híbrida.

REFERÊNCIAS

BOND, Francis; PAIK, Kyonghee. **A survey of wordnets and their licenses**. 2012 p. 64-71. Disponível em: < https://www.researchgate.net/profile/Francis_Bond/publication/267427763_A_Survey_of_WordNets_and_their_Licenses/links/549cccaa0cf2fedbc30fe027 >. Acesso em: 30-05-2020.

CHALUB, Fabricio; RADEMAKER, Alexandre. **Verifying integrity constraints of a rdf-based wordnet**. In: Global WordNet Conference. 2016. p. 309. Disponível em: < https://www.racai.ro/p/gwc2016/Slide-uri/day04/chalub-verifying_integrity_constraints_of_a_rdf-based_wordnet.pdf >. Acesso em: 20-01-2021.

COSTA, Douglas de Jesus. **Análise de algoritmo de comparação semântica de textos:** Uma abordagem sob a ótica cognitivista de Piaget. 2017. 120 p. Monografia (Curso de Sistemas de Informação) - Universidade Estadual de Goiás, Anápolis, 2017.

CUNHA, Marcelo do Vale. **Redes semânticas baseadas em títulos de artigos científicos.** 2013. 127p. Dissertação de Mestrado (Mestrado em Modelagem Computacional e Tecnologia Industrial) – SENAI CIMATEC, Salvador, 2013.

FELLBAUM, Christiane; MILLER, George A. **WordNet:** an electronic lexical database language. 1. ed. MIT Press, 1998. p. 1-423. Disponível em: <https://books.google.com.br/books?id=Rehu8OOzMIMC&lpg=PA265&ots=Irp8HhWXc6&dq=Leacock%20e%20Chodorow%201998&lr&hl=pt-BR&pg=PR3#v=onepage&q&f=false >. Acesso em 30-05-2020.

FEITOSA, David B.; PINHEIRO, Vládia C. **Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual.** In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 1. 2017, Minas Gerais. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2017. p. 161-170. Disponível em: < https://sol.sbc.org.br/index.php/stil/article/view/4012>. Acesso em 23-09-2020.

FERREIRA, Jaider Andrade; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. **O modelo de dados resource description framework (rdf) e o seu papel na descrição de recursos.** Informação & Sociedade: Estudos, p. 13-23, 2013. Disponível em: < http://dx.doi.org/10.6084/m9.figshare.1116375>. Acesso em: 10-01-2021.

FREIRE, J.; PINHEIRO, V.; FEITOSA, D. **FlexSTS:** Um Framework para Similaridade Semântica Textual. Linguamática, v. 8, n. 2, p. 23-31, 31 Dez. 2016. Disponível em: < https://www.linguamatica.com/index.php/linguamatica/article/view/v8n2-3>. Acesso em 20-09-2020.

GREGHI, Juliana Galvani et al. **Projeto e desenvolvimento de uma base de dados lexicais do português.** 2002. Tese de Doutorado. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

LAKATOS, Eva Maria; MARCONI, Maria de Andrade. **Fundamentos de metodologia científica.** 4. ed. São Paulo: Atlas, 2003.

LIMA, J. L. O. **Avaliação discente em cursos de graduação a distância mediados por ambientes virtuais de aprendizagem:** Necessidade de informações dos docentes na visão de especialistas europeus e brasileiros. 2016. 298f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2016.

MILLER, George A. **Wordnet:** a lexical database for english. In:_____. Communications of. 1995. v. 38. p. 39-41. Disponível em: < https://dl.acm.org/doi/10.1145/219717.219748 >. Acesso em: 30-05-2020.

OLIVEIRA ALVES, A.; RODRIGUES, R.; GONÇALO OLIVEIRA, H. **ASAPP:** Alinhamento Semântico Automático de Palavras aplicado ao Português. Linguamática, v. 8, n. 2, p. 43-58, 31 Dez. 2016. Disponível em:< https://www.linguamatica.com/index.php/linguamatica/article/view/v8n2-5>. Acesso em 05-10-2020.

PAIVA, Valeria de; RADEMAKER, Alexandre; MELO, Gerard de. **Openwordnet-pt:** An open brazilian wordnet for reasoning. 2012. Disponível em:< https://hdl.handle.net/10438/10274>.

Acesso em 05-01-2021.

PIANTA, Emanuele; BENTIVOGLI, Luisa; GIRARDI, Christian. **MultiWordNet:** developing an aligned multilingual database. 2002. p. 293-302.

RADEMAKER, Alexandre et al. **OpenWordNet-PT:** a project report. In: Proceedings of the Seventh Global Wordnet Conference. 2014. p. 383-390. Disponível em: < https://www.aclweb.org/anthology/W14-0153.pdf>. Acesso em 20-12-2020.

ROSA, Marcos G. **Modelo empírico para analisar a robustez de redes semânticas.** 2016. 134p. Tese (Doutorado Multi-institucional e Multidisciplinar em Difusão do Conhecimento) - Universidade Federal da Bahia. Faculdade de Educação, Salvador, 2016.

SANTOS, João Vitor Felipe. **Comparação semântica baseada em ontologias de domínio:** Proposição de modelo para compor tecnologia de apoio à avaliação na Educação Superior à Distância. 2019. 78p. Monografia (Curso de Sistemas de Informação) - Universidade Estadual de Goiás, Anápolis, 2019.

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração.** 14. ed. São Paulo: Atlas, 2013.

VOSSEN, Piek; **EuroWordNet:** a multilingual database for information retrieval. 1997. Disponível em: < https://research.vu.nl/ws/files/73708632/Delos97>. Acesso em: 30-05-2020.