A TERMINOLOGIA EM SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO BASEADA NA WORDNET.PT

A WORDNET.PT-BASED TERMINOLOGY FOR INFORMATION RETRIEVAL SYSTEMS

Januário Albino Nhacuongue[1]
Moisés Lima Dutra[2]

RESUMO
O artigo resulta da pesquisa de pós-doutorado realizada na Universidade Federal de Santa Catarina. O objetivo é propor estratégias de recuperação da informação baseadas no processamento da linguagem natural, para extrair relações semânticas da WordNet.Pt, e utilizá-las na representação de documentos e de expressões de busca dos usuários. O enfoque é qualitativo, exploratório e aplicado aos problemas de ambiguidade na recuperação da informação. Quanto aos procedimentos utilizados, trata-se de uma pesquisa bibliográfica. A discussão é motivada pelo problema da baixa precisão e alta revocação em buscas do usuário, influenciado tanto pela ausência da correspondência semântica entre expressões de busca e termos utilizados na indexação como pela falta da determinação da similaridade semântica entre termos de documentos que, mesmo sendo lexicograficamente diferentes, possuem o mesmo significado. O núcleo de pesquisa justifica-se pela vantagem de desenvolvimento de sistemas que combinam a linguagem natural e a linguagem controlada orientada, para uma busca interativa. Embora de forma parcial, a pesquisa aponta para resultados importantes na solução da ambiguidade lexical, por meio de relacionamentos semânticos na representação de documentos e busca do usuário. Por um lado, esse sucesso garante a restrição do espaço da busca e, consequentemente, a precisão. Por outro lado, a expansão de consultas por meio de sugestão de termos equivalentes de vocabulários controlados e da língua natural e suas variantes.
Palavras-chave: WordNet.Pt. Processamento da linguagem natural. Sistemas de recuperação da informação. Ambiguidade. Relevância informacional.

ABSTRACT
The article results from post-doctoral research conducted in Universidade Federal de Santa Catarina. The goal is to propose information retrieval strategies based on natural language processing, to extract semantic relations from WordNet.Pt, and use them to represent documents and users' search expressions. The approach is qualitative, exploratory and applied to ambiguity problems in information retrieval. As for the procedures used, it is a bibliographic search. The discussion is motivated by the problem of low precision and high recall in user searches, influenced both by the absence of semantic correspondence between search expressions and terms used in indexing and by the lack of determination of the semantic similarity between document terms that, even being lexicographically different, have the same meaning. The research core is justified by the advantage of developing systems that combine natural language and controlled language, for an interactive search. Although in a partial way, the research points to important results in the solution of lexical ambiguity, through semantic relationships in the representation of documents and user search. On the one hand, this success guarantees the restriction of the search space and, consequently, precision. On the other hand, the expansion of consultations by suggesting equivalent terms from controlled vocabularies and the natural language and its variants.
Keywords: WordNet.Pt. Natural Language Processing. Information Retrieval Systems. Ambiguity. Informational relevance.

Artigo submetido em 19/02/2020 e aceito para publicação em 12/03/2020

1 Professor da Universidade Federal de São Carlos, Brasil. ORCID: http://orcid.org/0000-0002-6679-1306. E-mail: januario80@gmail.com
2 Professor daUniversidade Federal de Santa Catarina, Brasil. ORCID: https://orcid.org/0000-0003-1000-5553. E-mail: moises.dutra@ufsc.br

1 INTRODUÇÃO

O processo de recuperação da informação, na maioria das vezes, envolve operações seletivas sobre o *corpus* documental no qual o universo do conhecimento se encontra representado. Trata-se de determinar, por meio de processos cognitivos e de linguagem, a relevância e a pertinência da informação, quando o sujeito interage com o objeto.

Atualmente, existem vários Sistemas de Recuperação da Informação – SRI, que se distinguem pela escala da sua operacionalidade. Os pessoais utilizam processos manuais e automáticos, por exemplo, de classificação e pesquisa de e-mails pessoais ou de documentos armazenados em computadores pessoais. Os sistemas de pesquisa corporativa permitem a recuperação de domínios específicos de documentos internos da empresa, armazenados em arquivos centralizados ou bancos de dados. Os sistemas de recuperação na Web baseiam-se no processamento da linguagem, para determinar a precisão e revocação.[3] (MANNING; RAGHAVAN; SCHUTZE, 2008). O nosso trabalho concentra-se nos sistemas de recuperação na Web.

Na fase inicial, a maioria das pesquisas em Processamento da Linguagem Natural – PLN concentrou-se na necessidade de resolver a ambiguidade sintática e semântica com base em entradas gramaticais de contexto local ou individual. Esse fato deveu-se principalmente a três razões: (i) foco na tradução automática de línguas específicas, por exemplo, de russo ao inglês; (ii) limitação tecnológica de processamento para a análise de sentenças longas; e (iii) a resolução da ambiguidade centrada na sintaxe por meio de equivalentes de dicionários de sinônimos. Não houve abordagens profundas sobre o contexto linguístico complexo da comunicação humana. Logo, faltou a caracterização da estrutura inteira de frases, para que fossem ordenadas de forma geral, e de acordo com as regras e contextos de aplicação de cada língua. Muitos SRI não satisfaziam às necessidades dos usuários porque se baseavam em frases autônomas ou em situações de línguas específicas. (JONES, 2001).

A língua é essencial na comunicação, pois constitui a herança, a cultura, a identidade, e a história de cada comunidade, nação ou grupo étnico. Ela reflete os mecanismos de produção e apropriação do seu conhecimento. Chomsky (2009, p.13) nos mostra que, nos estudos de linguagem, a língua é "[...]um sistema integrado de regras e princípios, a partir do qual podem ser derivadas as expressões da língua, cada um dos quais uma coleção de instruções para o pensamento e para a ação."

3 Enquanto a precisão consiste em determinar a relevância de documentos recuperados, a revocação determina a relevância em relação aos documentos da coleção ou *corpus*. Por isso, quanto maior for a precisão, menor será a revocação (SANDERSON, 1996; STEWART, 2008, MANNING; RAGHAVAN; SCHUTZE, 2008).

O nosso trabalho incide sobre a rede léxico-conceitual WordNet.Pt, que constitui a base do universo do conhecimento da língua portuguesa.

As reflexões sobre importância da língua se inserem no poder e na ação das sociedades pós-modernas, proporcionados tanto pelas tecnologias de comunicação como pela concomitância do conhecimento científico e social[4]. O fundamento para esta afirmação é encontrado em Lyotard (1991), quando define o pós-modernismo como estado das transformações que afetaram as regras de jogo da ciência, da literatura e das artes a partir do final do século XIX. Essas transformações modificaram a natureza da ciência pelo impacto dos desenvolvimentos tecnológicos sobre o saber, provocando a incredulidade em relação aos seus metarrelatos ou características universais. Assim, o saber passou a depender de formas de organização, armazenamento, distribuição, acesso e apropriação da informação, exigindo uma concepção operacional da ciência.

A tecnologia afetou as principais funções do saber: a pesquisa e a transmissão de conhecimentos. O saber deixou de ser para si mesmo seu próprio fim (valor informativo) e virou mercadoria. Esse novo cenário caracteriza-se por uma heterogeneidade de jogos de linguagem[5], em que, além de narrativos, os elementos de linguagem também se tornam denotativos, prescritivos, descritivos, performativos, etc., conforme a necessidade e o poder de ação de cada ator. O conhecimento científico passou a sofrer um processo de dupla legitimação - o direito de decidir sobre sua veracidade dentro do discurso científico, e de decidir sobre sua validade no discurso pragmático. Por outras palavras, com o acesso ubiquitário às informações, cada ator exerce o poder de decisão como remetente, destinatário ou referente, numa rede de relações de comunicação. Além de denotativas, prescritivas, avaliativas, performativas, etc., as mensagens comportam a função agonística relacionada ao pragmatismo ou aos usos sociais, enriquecendo a contribuição da linguagem. (LYOTARD, 1991).

O estudo da linguagem ganhou novos contornos de complexidade quando, em vez do comportamento e seus produtos, passou-se a concentrar-se na cognição para entender os mecanismos de ação e interpretação na construção do conhecimento. Como nos mostra Chomsky (2009), cada pessoa possui uma competência linguística, através da qual faz uso real e criativo da linguagem. Essa

4 O saber científico não é todo o saber; ele sempre esteve ligado a seu conceito, em competição com uma outra espécie de saber que, para simplificar, chamaremos de narrativo[...]. Não se trata de dizer que este último possa prevalecer sobre ele, mas seu modelo está relacionado às ideias de equilíbrio interior e de convivialidade, comparadas às quais o saber contemporâneo empalidece, sobretudo se tiver que sofrer uma exteriorização em relação àquele que sabe (*sachant*) e uma alienação em relação a seus usuários bem maiores do que antes (LYOTARD, 1991, P.12).

5 O saber é uma espécie de discurso. Nos estudos de linguagem, os discursos produzem efeitos (enunciados ou jogos de linguagem), que possuem regras específicas sobre suas propriedades e seu uso. Tais regras são legitimadas pelo acordo entre os jogadores, e a sua modificação implica alteração da natureza do jogo. Por isso, toda a ação social do indivíduo ou grupo no discurso é um jogo.

competência, por um lado, envolve a teoria da gramática gerativa, um sistema de regras que especifica a relação som-significado numa dada língua. Parte das regras gramaticais incluem descrições estruturais[6], que fazem representação fonética das sentenças e especificação do seu significado. Por outro lado, o uso real da linguagem envolve crenças extralinguísticas ou princípios da estrutura cognitiva que não são aspectos da linguagem.

Outros problemas complexos da semântica contemporânea começaram no século XX, com o enigma das atitudes proposicionais ou sentenças de crença, de Frege e Russell. Partindo da crítica de Kripke, Ibaños (2009) mostra a problemática de crença do sujeito no uso de nomes próprios. Isso acontece porque o critério de identificação ao seu referente, que é associado aos nomes na linguagem, apenas fixa a referência do nome e não o seu sentido. Nós entendemos que a interpretação semântica de nomes próprios é construída dentro de processos cognitivos e representacionais de cada sujeito.

Os SRI incorporam a complexidade linguística. Enquanto a gramática se torna fundamental pelo conjunto de regras de uma certa língua na representação e interpretação semântica dos fonemas, o modelo perceptivo no todo envolve processos mentais tanto de representações gramaticais como de aspectos cognitivos para representação sintática, semântica e fonética. O poder cognitivo de cada indivíduo, tomado a partir da dimensão espacial e das instâncias de mediação política, econômica e social, para a transformação de dados operacionais em informações úteis pode ser percebido nas considerações de Santos (2013). As formas de representação externas são relativas ao meio e facilitam o seu entendimento, enquanto as internas resultam da exteriorização do comportamento do indivíduo. Esse fenômeno é característico da linguagem, ou seja, enquanto o léxico representa os objetos linguísticos gerais, a cultura, a cognição, o espaço, o tempo, etc., de cada sujeito interferem no seu modelo perceptivo em espaços de mediação informacional.

Por exemplo, na comunicação corriqueira usam-se expressões idiomáticas cuja semântica não pode ser deduzida a partir das palavras constituintes. As expressões “bater as botas” e “esticar o pernil” resultam de uma gramaticalização e lexicalização que dependem da relação e associação para produzir o significado comum “morte ou morrer”. Esse significado pode ser expresso através do léxico e relacionado à uma ontologia ou à um tesauro para controlar o vocabulário. Desse modo, a sua semântica pode ser recuperada, quando utilizadas como termos de indexação ou em expressões

6 As transformações gramaticais correspondem ao conjunto de regras que exprimem a relação entre as estruturas profunda e superficial das sentenças. Enquanto a estrutura profunda representa frases chaves na interpretação semântica de uma sentença, as estruturas superficiais representam a expressão linguística e as categorias a que pertencem as frases. A estrutura profunda fornece dados importantes nos processos gramaticais que permitem operações mentais de representação, interpretação e apropriação semântica (CHOMSKY, 2009).

de busca. Uma busca pelo assunto "morte" poderá recuperar documentos que contenham os termos "morte" e "bater as botas" ou "esticar o pernil", como conceitos lexicalizados por combinação de palavras nas relações de equivalência.

O objetivo deste trabalho é propor estratégias de recuperação da informação baseadas no PLN, para extrair relações semânticas do léxico WordNet.Pt, e utilizá-las na representação dos documentos e das expressões de busca dos usuários.

Vossen (1998) aponta para a importância do entendimento da diversidade léxica, do conhecimento social, e do caráter subjetivo da representação no desenvolvimento de SRI. Rijsbergen (1979), por sua vez, acrescenta que o uso de computadores na recuperação resolveu grande parte dos problemas sobre o armazenamento, mas deixou sem solução o problema intelectual de caracterizar o conteúdo do documento. Esse problema não consiste apenas em extrair informações importantes sobre documentos, mas, principalmente, em associá-las e usá-las para decidir sobre a relevância. Aliás, a representação no tratamento documentário, dentro do contexto da economia da informação, conforme aponta Marcondes (2001 *apud* SANTOS, 2013), serve como um dispositivo de inferência sobre relevância do recurso informacional para as necessidades de informação do usuário que a interpreta. Aqui se encontra uma das bases do nosso trabalho, ao propor estratégias de processamento linguístico de relações semânticas para reduzir ambiguidades na representação e busca de informações.

A ambiguidade é um fenômeno semântico que se manifesta pelo uso de uma palavra ou grupo de palavras associadas a mais de um significado. Na literatura semântica, existem vários tipos de ambiguidade: lexical (homonímia e polissemia), sintática, de escopo, semântica, atribuição de papéis temáticos e construções com gerúndios. (CANÇADO, 2008). Embora o processamento da linguagem envolva todos os tipos de ambiguidade, o nosso trabalho concentra-se principalmente na ambiguidade lexical e semântica.

O aspecto social da Ciência da Informação (CAPURRO, 2003) alarga o âmbito do seu objeto (informação), segundo a tripartição de Buckland (1991), para aspectos sobre cognição, cultura e outros modelos de representação que permeiam o conhecimento do mundo. Ingwersen (1992) enfatiza esse comprometimento social com base nas dimensões da Ciência da Informação, preconizando uma visão holística que enaltece, entre outros aspectos, a ecologia da informação ou informação na estrutura social, o conhecimento em ação como razão da transferência da informação e a facilidade de comunicação entre geradores e usuários da informação. É dentro dessa dimensão social que desenvolvemos este trabalho.

2 TRABALHOS RELACIONADOS

Em relação ao léxico, muitos trabalhos já foram desenvolvidos para fundamentar pesquisas em Linguística, PLN, Inteligência Artificial, Ciência da Informação, Ciência da Computação, Recuperação da Informação, entre outras áreas. Marrafa, Amaro e Mendes (2011) mostraram a importância da WordNet.Pt global, uma extensão da WordNet.Pt, para o desenvolvimento de aplicações linguísticas abrangendo as variantes da língua portuguesa. Essas variantes resultam da comunicação nos oito países cuja língua oficial é portuguesa (Angola, Brasil, Cabo Verde, Timor-Leste, Guiné-Bissau, Moçambique, Portugal e São Tomé e Príncipe). O trabalho consistiu na extração de 10.000 conceitos da WordNet.Pt, associando-os às expressões léxicas que os denotam em cada língua dos oito países, exceto Portugal, cujas expressões já estão codificadas no modelo. As autoras consultaram os falantes nativos nas respectivas comunidades de origem, identificando as variantes dos conceitos. O trabalho das autoras traz relações semânticas importantes para o PLN e a recuperação da informação.

Marrafa *et al.* (2012) analisaram estratégias para evitar a ambiguidade léxica, induzida pela polissemia ou por efeitos de função sintática, nos sistemas de linguagem controlada do português. Para isso, partiram da noção de Linguagens Naturais Controladas como conjuntos de restrições linguísticas a serem aplicadas aos textos escritos em uma determinada língua. Um dos objetivos da pesquisa destes autores era garantir a usabilidade dos sistemas de tradução automática (português – inglês), fornecendo aos usuários não especialistas ferramentas para melhorar os resultados. A análise dos autores foi baseada no projeto PorSimples[7], para reduzir a complexidade lexical ou sintática de um texto usando um dicionário de sinônimos e uma ontologia lexical.

O trabalho de Mendes, Necsulescu e Bel (2012), por um lado, é extremamente importante porque prioriza a produção automática de recursos, concretamente a extração de substantivos sinônimos, com base nos dados. Assim, contrapõe-se à maioria de abordagens que recaem sobre recursos lexicais estruturados pré-existentes de linguagens produzidas manualmente, acarretando mais custos e maior grau de arquitetura. Por outro lado, está relacionado ao nosso trabalho, já que que a sinonímia pode ser usada em aplicações do PLN, para produção de resumos, respostas às perguntas nos formulários de busca, entre outras.

7 Desenvolvido pelo Núcleo Interinstitucional de Linguística Computacional, da Universidade de São Paulo, voltado para tecnologias de Processamento de Linguagem Natural, para promover a inclusão digital e acessibilidade de pessoas com baixos níveis de alfabetização. A iniciativa envolve duas abordagens: simplificação e elaboração de texto. Na simplificação procura-se reduzir a complexidade lexical ou sintática de um texto, enquanto se preserva o significado e a informação. A simplificação pode ser lexical e sintática, resumo automático, etc. A elaboração de texto visa esclarecer e explicar as informações, tornando explícitas as conexões em um texto (por exemplo, através de sinônimos).

Sussna (1993), reconhecendo os problemas da baixa precisão (polissemia) e da baixa revocação (sinonímia) quando termos são usados sem desambiguação semântica, propõe o uso da WordNet na indexação de documentos, para que a busca se faça a partir do significado pretendido e não de termos relacionados. A estratégia utilizada se baseia no conceito de distância semântica entre nós da rede, para desambiguar termos polissémicos de entrada. Com isso, melhora-se a precisão e revocação nos processos de recuperação da informação.

O trabalho de Varelas et al. (2005) demonstra algumas vantagens na recuperação de informações, a partir do mapeamento de termos para calcular a sua similaridade semântica ou relação na WordNet. Comparado aos modelos de espaço vetorial, probabilístico, booleano, entre outros que se baseiam na correspondência de termos lexicográficos, o modelo de similaridade semântica dos autores destaca-se pela possibilidade de recuperação de documentos cujos termos, embora lexicograficamente diferentes, têm mesmo significado ou são sinônimos.

A tese de Sanderson (1996) investiga a relação entre recuperação de informações e ambiguidade lexical, mostrando as vantagens da desambiguação do sentido de palavras na melhoria da eficácia da recuperação, através do uso de pseudo-palavras que simulam palavras ambíguas.

Reconhecendo a ausência de relacionamentos semânticos entre termos nos tradicionais algoritmos de *clusters*, Wei et al. (2015) propõem integrar a WordNet às cadeias lexicais, para desambiguar o sentido das palavras e representar com precisão o significado de documentos.

A pesquisa de Ferneda e Dias (2015) não é voltada para o léxico. Mesmo assim, insere-se no PNL aplicado ao modelo espaço vetorial, para padronização terminológica na recuperação da informação. Para tal, os autores desenvolveram o *OntoSmart*, um modelo de recuperação baseado em ontologias. No modelo proposto, a ontologia é formada por elementos linguísticos que constituem termos do vocabulário da área de domínio, usados para padronização terminológica tanto na representação de documentos como em buscas. Esse sistema melhora a precisão, já que as necessidades de usuários são expressas através de termos de domínio específico, também usados na indexação. Além disso, permite a expansão de consultas, agregando terminologias de áreas específicas.

Os estudos citados mostram a importância do processamento da linguagem para melhorar a comunicação em vários sistemas. Dentro do mesmo objetivo, o nosso trabalho tem a particularidade de inserir algumas estratégias estudadas no contexto da Ciência da Informação, para ambientes de representação e busca informacional, e de construção de linguagens controladas ou de vocabulários.

3 PROCESSAMENTO DA LINGUAGEM E TERMINOLOGIA

Segundo a revisão histórica de Jones (2001), a primeira fase do PLN (final de 1940 até a década de 1960) centrou-se no emprego de máquinas para a tradução automática (principalmente, do russo para o inglês). Como reflexo, foi realizada a primeira Conferência Internacional sobre Tradução Automática em 1952. Também se realizou a Conferência Internacional de Washington sobre Informação Científica em 1958, abordando aspectos de processamento de linguagem em relação à recuperação de informação, como o uso de dicionário de sinônimos. A Conferência Internacional em Tradução Automática e Análise Aplicada da Língua de Teddington, em 1961, teve destaque tanto pela referência a pesquisas feitas em outros países no âmbito do PLN como por enfoques na morfologia, na sintaxe e semântica.

As primeiras investigações sobre terminologia no concernente ao processamento da linguagem podem ser relacionadas à busca de soluções para o problema do aumento de informações e consequente especialização científica necessária para o progresso. Por isso, a maioria das pesquisas em campos análogos envolvia o uso de computadores e visava à recuperação da informação. Na recuperação da informação, por exemplo, Chu (2007) considera que o período de 1950 a 1980 foi marcado pela introdução de computadores e pelos estudos de Hans Peter Luhn, para a indexação automática, resumos e combinação de palavras-chave nos formulários de busca. Esse esforço mostra a preocupação em relação ao uso da terminologia para reduzir a ambiguidade e garantir um acesso rápido e preciso aos documentos.

A preocupação em torno da terminologia também é mostrada por Lewis e Jones (1996), ao situarem a recuperação da informação nas tarefas do PLN. Conforme os autores apontam, a relação entre a necessidade do usuário e o documento que a atenda não é necessariamente óbvia. Logo, a indexação se torna essencial para fixar uma linguagem, um vocabulário de termos e um método para relacionar buscas de usuários às descrições de documentos. Ao solucionar as necessidades de usuários, a indexação lida com vários problemas, alguns dos quais consistem na dificuldade de representação semântica da expressão de busca do usuário. Outros, na variabilidade de formas em que um conceito pode ser expresso. Ambos são questões de linguagem. (LEWIS; JONES, 1996).

Jones (2001) lembra que, mesmo limitado pela tecnologia da época, a fase inicial do processamento de linguagem contribuiu para a construção da gramática e do léxico para algumas línguas envolvidas na época como Russo, Inglês e Chinês. Também permitiu abordar questões gerais

de arquitetura de sistemas e estratégias de processamento como a tradução direta e interlingual. O outro destaque apontado pela autora consistiu na concepção do processo da tradução como pesquisa, através da análise da estrutura de frases para solucionar os problemas de ambiguidade sintática e semântica. Assim, muitos esforços foram concentrados na sintaxe, por exemplo, resolvendo problemas de anáforas e polissemias.

A segunda fase das pesquisas sobre o PLN vai do final dos anos 1960 até o final dos anos 1970, e foi influenciada pela Inteligência Artificial na construção de bases de conhecimento para manipulação e representações do significado, isto é, para semântica orientada. Essas bases influenciaram estudos sobre inferências na interpretação de discursos na década 1970, a partir de relatórios e de diálogos. Assim, retomou-se a preocupação com a anáfora e com a construção de sistemas baseados nas manifestações culturais e crenças gerais de cada léxico. (JONES, 2001).

O desenvolvimento de pesquisas sobre PLN foi impulsionado tanto pela revolução cognitiva dos anos 1950 e 1960 nos estudos da linguagem, que passou a priorizar a mente como sistema computacional[8], como pela psicologia evolucionária das décadas de 1960 e 1970, que se concentrou na análise das faculdades da mente para solução de problemas adaptativos. É sobre esses aspectos que se desenvolve o modelo perceptivo chomskiano, analisando as regras da gramática universal para línguas e o processamento baseado em outras faculdades mentais para representações sintáticas, semânticas e fonéticas. (PINKER, 2009).

A terceira fase na historiografia do processamento linguístico, conforme Jones (2001), vai do final dos anos 1970 até o final dos anos 1980 e foi caracterizada pelo enfoque gramatical-lógico, ou seja, pelo desenvolvimento da teoria gramatical e pela transição para o uso da lógica na representação do conhecimento e raciocínio em Inteligência Artificial. Nesse caso, as novas gramáticas orientadas para computação tinham uma base contextual para apoiar os algoritmos de análise. O processamento, por sua vez, era orientado à sintaxe em formas lógicas de sentido e representação do conhecimento, de acordo com as crenças e intenções do usuário.

Como se pode depreender, a preocupação com a ambiguidade foi estendida ao contexto específico em que cada usuário interpreta e se apropria do signo na sua relação com objetos. É nesse contexto que os tesauros, dentro da Ciência da Informação, procuram atender domínios específicos em termos semânticos.

8 Computação como conceito de racionalidade mecânica, através da qual a mente processa a informação com base em crenças e exerce o *feedback* para reduzir a diferença entre um estado atual e um estado almejado (PINKER, 2009, P.12).

Ao referirmo-nos à ambiguidade no contexto de apropriação da cultura e da necessidade do usuário consideramos, em grande medida, a ambiguidade lexical, limitada principalmente à polissemia contrastiva e complementar. Na polissemia contrastiva, dois ou mais sentidos não-relacionados estão associados a uma única expressão. Por exemplo, o termo "casa" denota três conceitos (construção, habitação, botoeira) cujas lexicalizações dependem do sentido pretendido pelo usuário. Na polissemia complementar, os sentidos estão relacionados à única expressão. Por exemplo, o termo "cordeiro" significa tanto o próprio animal como a carne de cordeiro. Importante realçar que, mesmo tratados de forma semelhante através de codificação em diferentes nós em WordNet.Pt, os conceitos de cordeiro ainda se mostram problemáticos dentro do escopo do nosso trabalho. Para garantir a precisão na recuperação, importa determinar em que sentido a expressão "cordeiro" é empregue. (MARRAFA *et al.* 2012).

A preocupação em relação à sintaxe e às diversidades de representação pode ser inserida no âmbito da organização do conhecimento na Ciência da Informação. Ora, a concepção dos sistemas de organização de conhecimento impõe um plano multidimensional que transponha fronteiras culturais e geográficas de acesso e representação no controle semântico para sanar a ambiguidade (BOCCATO, 2009). É o que acontece, por exemplo, com as listas de cabeçalhos de assunto, tesauros e ontologias. Esse princípio reforça cada vez mais a importância da WordNet.Pt como base de estruturas semânticas da linguagem natural e artificial em diferentes ambientes controlados. Aliás, a tendência crescente dos trabalhos sobre o léxico no final dos anos 1980 foi justificada pela sua importância na abordagem gramatical-lógica e na tradução automática multilíngue. (JONES, 2001).

A quarta fase do PLN, dos anos 1990, foi marcada pela crescente influência do léxico que nasceu na década 1980. Esta tendência foi verificada na análise de grandes volumes de dados legíveis para computadores dentro da classificação semântica, envolvendo a probabilidade na aprendizagem de máquinas. Essas técnicas foram aplicadas em textos de notícias e na modelagem em reconhecimento da fala. No geral, esta fase remete-nos à tecnologia da linguagem, associada à necessidade de projeção de sistemas de tarefas genéricas e multilíngues cujos recursos linguísticos são fornecidos, por exemplo, pela WordNet. (JONES, 2001).

O léxico é importante para pesquisa do texto completo. Lewis e Jones (1996) demonstram esta importância em três linhas: a relação entre palavras, locuções e frases, bem como relações sintagmáticas que são objetos do processamento de documentos; a estrutura de classificação de documentos que permite a substituição de termos controlados na indexação e busca, baseada no

PLN; mecanismos baseados no processamento para busca e correspondência. Por exemplo, o PLN determina como o termo composto (casa protocolar) é selecionado e ponderado a partir dos seus constituintes (casa e protocolo). Desse modo, melhora a precisão e revocação, usando recursos de normalização, na indexação e busca, para reduzir as várias formas a uma única forma canônica. Por exemplo, a derivação, que é uma normalização com base na morfologia, dos termos "protocolar", "protocolarmente" e "protocolo" para protocolo.

Com o atual volume excessivo de informações, o PLN se torna essencial na análise da sintaxe, da semântica e do discurso, para reduzir ambiguidades e melhorar a precisão em ambientes de recuperação da informação baseados em métodos estatísticos. Isso acontece em esferas globais e transcendentais do valor humano, manifestados pelo universo do conhecimento. O léxico constitui a base do conhecimento para essas manifestações e a Ciência da Informação, no seu caráter interdisciplinar e social, contribui em questões cognitivas do usuário, de facilitação da sua interação com os geradores de informação, dos modelos e tecnologias de construção, transferência e transformação informacional.

4 LÉXICO-CONCEITUAL DA LÍNGUA PORTUGUESA WORDNET.PT

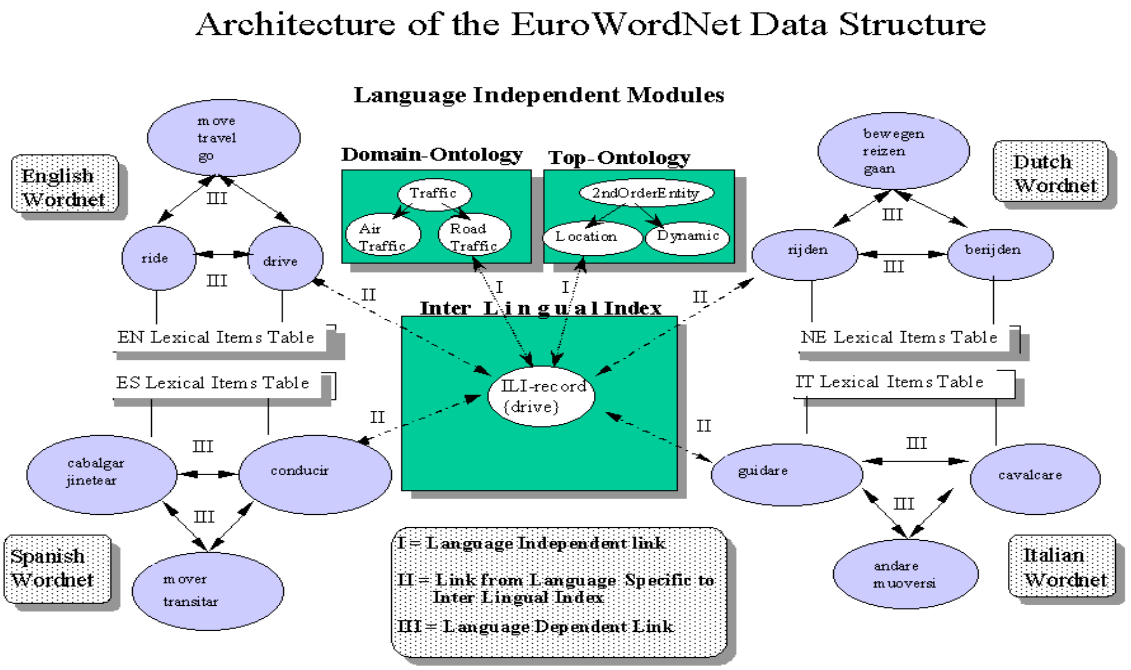
WordNet.Pt, segundo Marrafa *et al.* (2005), é uma base de dados de conhecimento da língua portuguesa, desenvolvida pelo Grupo de Computação de Conhecimentos Léxicos e Gramáticos do Centro de Linguística da Universidade de Lisboa, com base no modelo da EuroWordNet para línguas europeias e da WordNet de Princeton para o inglês. Trata-se de uma ontologia lexical que se aplica à várias áreas, desde a comunicação até a produção de termos e dos respectivos conceitos, para construção de linguagens artificiais em ambientes controlados.

Marrafa, Amaro e Mendes (2011) apontam que a primeira fase do projeto da WordNet.Pt (1999-2003) consistiu na seleção de um conjunto de dados semânticos abrangendo conceitos com alta produtividade na comunicação cotidiana. Esses conceitos fornecem subsídios para a projeção de sistemas de recuperação consistentes, por meio de um conjunto de relações de lexicalização em diferentes domínios semânticos e em outras línguas.

O léxico WordNet.Pt representa a base do conhecimento da língua portuguesa, com as respectivas variações linguísticas. Pode ser visto como resposta às dificuldades da fase inicial do PLN em incorporar o conhecimento do mundo através de modelos universais. Por isso, os sistemas baseavam-se na tradução de textos tratados como frases independentes ou no discurso de uma única fonte, dificultando

a resolução de anáforas, da terminologia e de outros problemas sintáticos e semânticos. Também pode ser considerado produto da teoria semântica lexical cujos alguns desenvolvimentos, segundo Cançado (2008), buscam viabilidade de aplicação da teoria linguística às práticas da Ciência da Computação.

Figura 1 - Arquitetura da linguagem no modelo da EuroWordNet

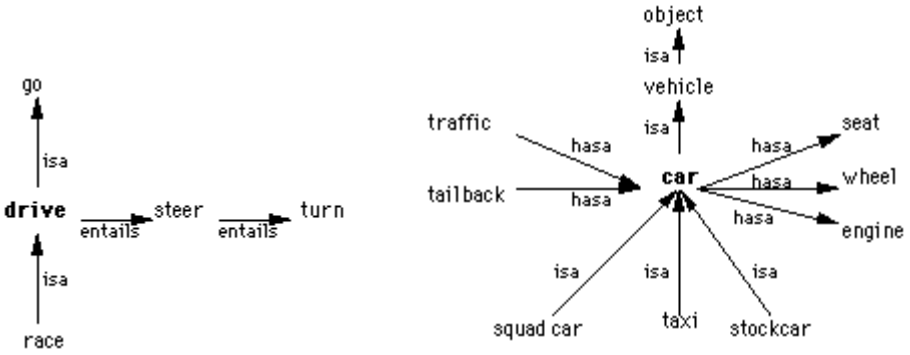


Fonte: Vossen (1998, p.7)

EuroWordNet é um banco de dados multilíngue com *wordnets* de várias línguas europeias (holandês, italiano, espanhol, alemão, francês, tcheco e estoniano). Os *wordnets* são estruturados da mesma forma que o WordNet americano para inglês, em termos de *synsets* ou conceitos (conjuntos de palavras sinônimas) com relações semânticas básicas entre eles. Os conceitos do WordNet são integrados à rede multilíngue do modelo do EuroWordNet, através do *Inter-Lingual-Index – ILI*, que permite relacionar as diversas *wordnets* entre si. Desse modo, cada conceito em uma língua é relacionado aos conceitos de outras línguas por sinonímia, hiperonímia, meronímia e relações funcionais. O ILI também permite o acesso ao topo da ontologia compartilhada de 63 distinções semânticas. O topo da ontologia fornece uma estrutura semântica comum para todas as línguas, enquanto as propriedades específicas de cada língua são mantidas nas *wordnets* individuais. (VOSSEN, 1998).

De acordo com a Figura 1, é possível criar o domínio ontológico para padronizar a terminologia nas áreas de trânsito, tráfego aéreo e tráfego rodoviário. Conforme se referiu anteriormente, o topo da ontologia fornece o primeiro nível ou uma estrutura semântica comum à todas as línguas, para entidades de segunda ordem, no âmbito da localização e do processo dinâmico. No nível II o ILI relaciona as *wordnets* de várias línguas (por exemplo, o termo *"drive"* em inglês corresponde aos termos *"conducir"* em espanhol, *"guidare"* em italiano e *"rigden"* em holandês). No nível III relacionam-se conceitos dentro de um léxico específico (por exemplo, em inglês o termo *"drive"* tem relação com os termos *"ride", "move", "travel" e "go"*).

Figura 2 - Relação entre os termos "*drive*" e "*car*"



Fonte: EuroWordNet (1999).

Cada léxico pode utilizar o modelo EuroWordNet para produzir relações semânticas que podem ser interligadas às *wordnets* de outras línguas. O léxico da língua portuguesa WordNet.Pt usa o mesmo princípio na definição dos conceitos e suas lexicalizações. Esses relacionamentos são necessários para padronizar a terminologia usada na indexação e na busca para usuários não especialistas.

A Figura 2 ilustra as relações semânticas entre termos e o agrupamento de termos com significados relacionados que podem ser acessados como uma espécie de enciclopédia. Por exemplo, o substantivo *"car"* (carro) está ligado, entre outras, a todas as palavras que têm com ele relações de sinonímia "tem mesmo significado" (*vehicle* é sinônimo de *car*); ou meronímia "contém parte ou substância" (*car* contém *wheel*). O verbo *"drive"* (conduzir) está ligado, entre outras, a todas as palavras que têm relações de hiponímia. Desse modo, é possível expandir a consulta usando termos

de relações lexicais, para melhorar a revocação nos SRI. Por exemplo, uma consulta com os termos *"drive"* e *"car"* será expandida para combinações como: *go + car*, *race + vehicle*, *steer + car*, *turn + wheel* e *race + engine*.

O processamento da linguagem baseado no WordNet.Pt é importante para o relacionamento semântico entre a linguagem natural e de especialidade. Os tesauros, por exemplo, consistem em vocabulários controlados e dinâmicos, isto é, numa linguagem especializada para o relacionamento sintático e semântico entre termos, baseado na linguagem natural usada em documentos ou pelos indexadores e usuários. (CURRÁS, 2010; CUNHA e CAVALCANTI, 2008). Com a WordNet.Pt, pode-se expandir ou integrar os relacionamentos semânticos nos tesauros e em outras linguagens documentárias. As relações léxico-semânticas são princípios organizadores do léxico, e a sua codificação em recursos de linguagem tornou-se um mecanismo importante no desempenho de aplicações do PLN. (MENDES; NECSULESCU; BEL, 2012).

Segundo Marrafa *et al.* (2005), a sinonímia é uma relação entre termos que correspondem ao mesmo conceito, por exemplo, "automóvel" e "carro". No mesmo entendimento, Mendes, Necsulescu e Bel (2012) apontam que a sinonímia é relativa às palavras que compartilham o mesmo significado, sendo geralmente definida em relação ao impacto que a substituição de uma expressão por outra em uma frase tem em termos do seu valor de verdade. Desse modo, A e B são sinônimos se ao se substituir A por B ou B por A nunca muda o valor da verdade da sentença em que ocorrem. No tesauro, os dois termos representam o mesmo conceito ou a relação de equivalência. Por isso, um pode ser usado como descritor ou termo preferencial e o outro como não-descritor.

Neste trabalho propomos o uso do léxico WordNet.Pt no PLN para aplicações em SRI baseados na Web, com vista à melhoria dos problemas de ambiguidade lexical e sintática. Para isso, adotamos o contexto de recursos baseados em conceitos, no qual a sinonímia é ligada a um dado contexto ou domínio, para produzir a semântica. (MENDES; NECSULESCU; BEL, 2012). O fundamento para esta abordagem está na principal característica das linguagens documentárias, que consiste na padronização terminológica dentro de determinados domínios específicos do conhecimento. Além de sinônimos intercambiáveis no léxico comum, duas ou mais palavras podem ser consideradas sinônimas em um contexto linguístico se representarem o mesmo valor dentro das variantes da língua ou de comunicação.

Na Figura 2 o termo *"taxi"* é um hipônimo, tipo ou espécie de *"car"*, enquanto que *"car"* é um hiperônimo, gênero de *"taxi"*. Os termos "carro" e "táxi" representam uma relação hierárquica genérica. Além da relação hierárquica do tipo gênero/espécie, os tesauros incluem relações hierárquicas do tipo

todo/parte e relações associativas. (CINTRA *et al*., 2005). No léxico, as relações hierárquicas partitivas são denominadas meronímia/holonímia e incluem partes propriamente ditas, membros, porções, matéria e localização. (MARRAFA *et al*., 2005).

O léxico utiliza relações semânticas de correspondência, superordenação/subordinação, contraste, ordenação e ambiguidade. Por exemplo, a sinonímia é relação de correspondência baseada na similaridade. Sussna (1993) afirma que as relações hiperonímia/hiponímia e holonímia/meronímia são consideradas relações verticais assimétricas e de ordem, enquanto as relações sinonímia e antonímia são consideradas relações horizontais, simétricas e sem ordem.

No PLN, as relações semânticas ajudam a orientar o usuário através do assunto, criando associações entre termos e conceitos que podem ser usados tanto para redefinir a necessidade de informação como para aprimorar o conhecimento.

5 SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO NA WEB BASEADOS EM RELAÇÕES SEMÂNTICAS

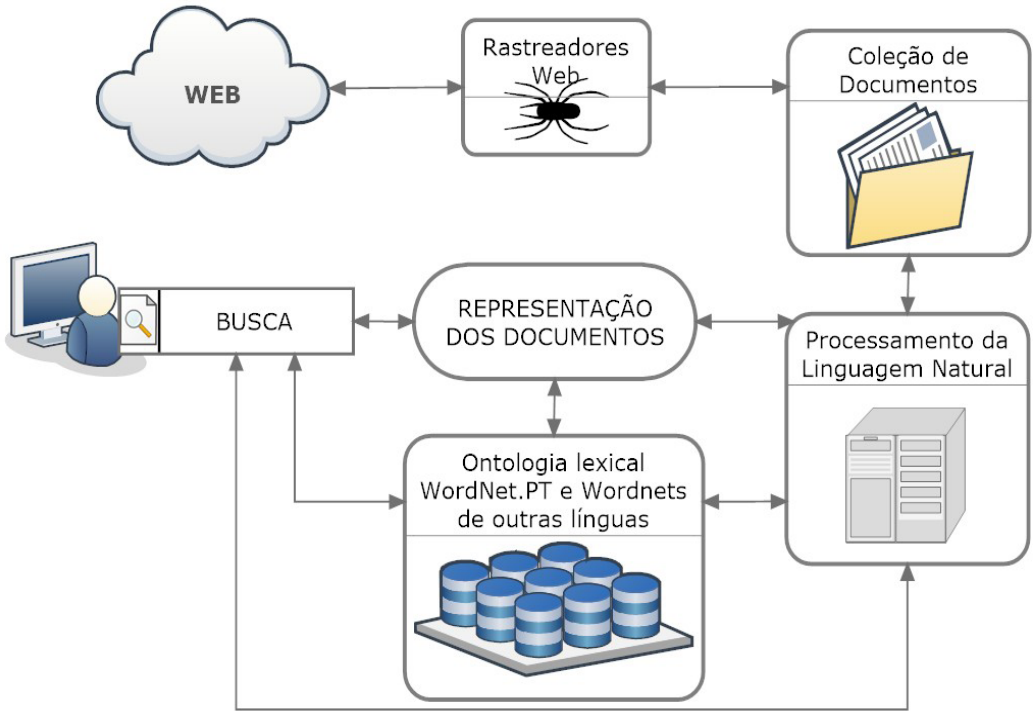
A determinação da relevância na maioria dos SRI consiste no grau de similaridade entre termos que compõem as expressões de buscas de usuários e a ocorrência em documentos da coleção ou nos termos de indexação. Embora esse processo produza bons resultados, muitas vezes contém ruídos ou *false drops*[9]. Por exemplo, a busca por documentos sobre "computador de mesa" poderia retornar tanto os documentos relacionados ao termo "computador de mesa", como os relacionados à "mesa de computador" ou até que tivessem os termos "computador" e "mesa". Como Sanderson (1996) argumenta, a ambiguidade afeta os sistemas de recuperação porque, na maioria das vezes, é difícil determinar em que sentido a ocorrência da palavra ou do termo depende do contexto em que aparece.

Além do problema referenciado, outra questão se prende à dificuldade de determinar a semântica dos termos que compõem a expressão da busca do usuário a partir da respectiva necessidade informacional. Ou seja, a representação e recuperação da informação são processos eminentemente subjetivos e cognitivos. Às vezes, a necessidade e a relevância da informação do usuário mudam em função da interação e dos resultados de busca.

9 O termo *false drop* refere-se à recuperação de documentos que, mesmo contendo os termos da busca, não são verdadeiramente relevantes para a pesquisa, ou seja, só se alcança a revocação perfeita se houver a queda na proporção dos documentos considerados relevantes para o tema da busca.

Para diminuir a ambiguidade na recuperação da informação, propomos estratégias baseadas no PLN, para extrair relações semânticas da WordNet.Pt, e utilizá-las na representação dos documentos e das expressões de busca dos usuários.

Figura 3 - Modelo de sistema de recuperação na Web baseado em relações semânticas



Fonte: Elaborado pelos autores

Mendes, Necsulescu e Bel (2012) apontam que as bases de dados lexicais são valiosas fontes de conhecimento sobre palavras e seus significados. A sua aplicação envolve diversas áreas, das quais destacamos o PLN e a recuperação da informação. O PLN incide sobre línguas naturais, para subsidiar aplicações que lidam com a desambiguação do sentido das palavras, recuperação da informação, tradução automática, interação humano – computador, etc.

O modelo da Figura 3 apresenta rastreadores de conteúdo na Web para indexação de documentos (textos, imagens, vídeos, etc.) na coleção que será exibida no mecanismo de busca do usuário. Os termos de entrada podem ser indexados com base no próprio conteúdo do documento utilizando a semântica fornecida pelos *synsets* ou sinônimos da WordNet.Pt. Esse processo do PLN envolve: a) tokenização – identificação das palavras do documento; b) remoção de *stopwords* ou palavras frequentes (preposições, conjunções e artigos); c) *stemming* – redução das palavras a uma forma normalizada ou ao seu radical (por exemplo, “caule” e “caules” se tornam “caule”). As palavras que

restam formam os nós da rede semântica do léxico: sinonímia (tem mesmo significado), hiperonímia (é um), hiponímia (tem instância), holonímia (é parte de, é substância em), meronímia (tem parte, contêm substância), antonímia (inverso). (SUSSNA, 1993; MANNING, RAGHAVAN, SCHUTZE, 2008).

Várias ferramentas de PNL podem ser utilizadas antes da desambiguação. Por exemplo, o marcador gramatical – as palavras a serem desambiguadas seriam marcadas com sua categoria gramatical (substantivo, verbo, adjetivo, pronome, etc.), reduzindo, assim, o número de sentidos a serem tratados pelo desambiguador. (SANDERSON, 1996).

Na indexação dos documentos pode-se utilizar o processo de desambiguação semântica, considerando a distância semântica[10] entre *synsets* na WordNet.Pt. Para isso, pode-se atribuir pesos conforme a semelhança semântica expressa em cada relação. Relações de sinonímia recebem o peso mais alto, enquanto relações de antonímia recebem o peso mais baixo. (SUSSNA, 1993).

A indexação e recuperação, baseadas em relações semânticas do léxico WordNet.Pt, podem reduzir a ambiguidade da sinonímia e polissemia.

> [...] Uma única forma de palavra pode ter vários significados, e um único significado pode ser expresso por várias formas de palavra. Ambas as multiplicidades causam problemas para qualquer abordagem da pesquisa de conteúdo com base em formas de palavras. Acreditamos que, para fazer uma recuperação no nível quase humano, precisamos ir além das palavras e obter significados. A desambiguação do texto durante a indexação deve melhorar a precisão, combatendo a polissemia. (SUSSNA, 1993, p.67).

Ainda no PLN, propomos a estratégia da determinação da similaridade semântica entre termos de documentos que, mesmo sendo lexicograficamente diferentes, possuam o mesmo significado. Os substantivos e verbos resultantes da tokenização, remoção de *stopwords* e *stemming* podem ser relacionados como nós no WordNet.PT. Para isso, métodos como diferença no conteúdo de informação de dois termos em função da probabilidade de ocorrência na coleção, comparação cruzada de termos de ontologias diferentes (por exemplo, WordNet.Pt e a ontologia de termos médicos MeSH), etc., podem ser usados na determinação da similaridade semântica. A ponderação de termos por meio de relacionamentos com outros termos semanticamente semelhantes e a expansão de termos através de hipônimos e hiperônimos semanticamente semelhantes aos termos da consulta favorece a expansão de consultas do usuário. (VARELAS et al. 2005).

10 Quanto menor a distância, maior a relação. Para a desambiguação, a hipótese é que, dado um conjunto de termos ocorrendo próximos um do outro no documento, cada um dos quais com múltiplos significados, escolhendo os sentidos que minimizam a distância, selecionamos os sentidos corretos (SUSSNA, p.69).

Além do PLN, da recuperação de informações, do resumo e da categorização de texto, a similaridade semântica desempenha um papel importante no agrupamento de documentos. Desse modo, também propomos a desambiguação semântica para construção de algoritmos de agrupamento de documentos. Inicialmente, constroem-se as cadeias lexicais para documentos desambiguados, ou seja, constroem-se os termos relacionados que fornecem o conteúdo semântico do documento e permitem a identificação dos principais tópicos. Em seguida, adicionam-se os pesos das relações envolvendo as cadeias lexicais com base no princípio de distância semântica. O uso do WordNet.Pt para desambiguar o sentido de palavra pode reduzir os problemas de sinonímia e polissemia. (WEI et al. 2015).

Os algoritmos de *cluster* agrupam conjuntos de documentos segundo critérios determinados, para permitir que se comportem de maneira semelhante em relação à relevância das necessidades de informações. Esses documentos compartilham termos que permitem a sua recuperação associada nas buscas; por isso, para evitar o *false drops*, os termos precisam ter similaridade semântica. Os algoritmos de agrupamento são usados em processos de buscas para fornecer informações relevantes ao usuário, oferecer alternativas de pesquisas, garantir resultados mais rápidos, etc. (MANNING, RAGHAVAN, SCHUTZE, 2008).

O PLN pode assegurar uma busca interativa, em que a linguagem de indexação seja acessível ao usuário para formulação da busca, combinando a linguagem natural e a linguagem controlada orientada. Em vez de utilizar a linguagem fortemente controlada, os relacionamentos semânticos do léxico e das linguagens artificiais podem permitir buscas em linguagem natural. Além disso, com base na análise gramatical, é possível cobrir expressões idiomáticas e outras cujas lexicalizações não se encontram no WordNet.Pt. (LEWIS; JONES, 1996).

A interface de busca do usuário também pode disponibilizar termos de ontologias a serem selecionados pelo usuário para compor a sua consulta, garantindo resultados precisos e familiaridade com a terminologia do domínio de interesse. (FERNEDA; DIAS, 2015). Do mesmo modo, pode-se utilizar termos de outros vocabulários controlados como tesauros, de variantes de língua, desde que estejam relacionados semanticamente.

A redução da ambiguidade é importante tanto para restringir o espaço de busca, garantindo a precisão, como para expandir o universo de documentos sobre o mesmo assunto, mas que estejam indexados por outros termos. Na nossa proposta, uma busca por “autocarro” (“ônibus” em Portugal) resulta em “veículo pesado concebido e usado como meio de transporte de passageiros” (WORDNET.PT,

2017), e poderia recuperar outros documentos. Mesmo tratando do mesmo assunto, tais documentos poderiam usar termos de variantes da língua, por exemplo, “machibombo” ou “machimbombo”, que são mais comuns no português de Moçambique. (MARRAFA; AMARO; MENDES, 2011).

Como Logan (2012) nos coloca, palavras são atratores estranhos, pois o seu significado muda em cada contexto de uso ou em cada contexto de representação/apropriação. Isso se torna complexo na simbolosfera, isto é, no universo formado pela mente humana e seus produtos como o pensamento simbólico abstrato, a linguagem e a cultura. Por isso, para o autor, a linguagem é um organismo vivo que propaga sua organização. Compreender as mutações que surgem no seu uso idiossincrático pode ser essencial para a projeção de melhores SRI.

6 CONSIDERAÇÕES

Ao tratar do problema de ambiguidade semântica nos ambientes de produção, tratamento, difusão, apropriação e uso, este trabalho esmiúça parte da natureza, manifestações, e efeitos da informação e do conhecimento que são objetos da Ciência da Informação. Por isso, contribui para pesquisas que são desenvolvidas dentro do campo, em relação à linguagem, aos SRI, à relevância, à semântica, entre outros aspectos.

O PLN envolvendo estratégias de similaridade semântica com base nos *synsets* ou nós da rede WordNet melhora a precisão e revocação na recuperação da informação conforme demonstrado pelos trabalhos citados. O uso do léxico WordNet.PT no processamento de linguagem também pode reduzir a ambiguidade na representação de documentos e de buscas de usuários em ambientes de recuperação da informação.

As estratégias apresentadas no trabalho são do nível conceitual, o que pode caracterizar uma limitação da pesquisa. Além disso, na sua maioria, dependem de conceitos matemáticos, computacionais e estatísticos. Mesmo assim, a discussão trazida no nosso trabalho vai contribuir nas abordagens que vêm sendo desenvolvidas na Ciência da Informação, sobre complexidade da língua e da linguagem em ambientes de produção, organização, disseminação e acesso da informação e do conhecimento. Entendemos que, mesmo sem aprofundamento matemático e computacional, as estratégias propostas podem ser utilizadas por bibliotecários e outros profissionais de informação nos vários espaços de representação e busca informacional, bem como de construção de linguagens ou vocabulários controlados.

Qualquer reflexão sobre problemas da língua e linguagem poderá ser insuficiente diante da complexidade cognitiva, discursiva, argumentativa e representacional humana. Por isso, trabalhos futuros se mostram necessários para aprofundar o debate.

REFERÊNCIAS

BOCCATO, V. R. C. A linguagem documentária vista pelo conteúdo, forma e uso na perspectiva de catalogadores e usuários. In: FUGITA, M. S. L. (org.) **A indexação de livros:** a percepção de catalogadores e usuários de bibliotecas universitárias. São Paulo: Cultura Acadêmica, 2009.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science** (JASIS), v.45, n.5, p.351-360, 1991.

CANÇADO, M. **Manual de semântica:** noções básicas e exercícios. 2ed. Belo Horizonte: Editora UFMG, 2008.

CAPURRO, R. Epistemologia e Ciência da informação. In: Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), 5., Belo Horizonte, 2003. **Anais eletrônicos...** Belo Horizonte: UFMG, 2003. Disponível em: http://www.capurro.de/enancib_p.htm. Acesso em: 31 mar. 2011.

CHOMSKY, N. **Linguagem e mente**. São Paulo: Editora UNESP, 2009.

CHU, H. **Information representation and retrieval in the digital age**. 3ª Tiragem, New Jersey: Asist&T, 2007.

CUNHA, M. B.; CAVALCANTI, C. R. O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, 2008.

CURRÁS, E. **Ontologias, taxonomias e thesaurus em teoria de sistemas e sistemática**. Tradução de ROBREDO, J. Brasília: Thesaurus, 2010.

FERNEDA, E.; DIAS, G. A. OntoSmart: proposta de um modelo de recuperação de informação baseado em ontologia. In: **XII Congreso ISKO España y II Congreso ISKO España-Portugal**: organización del conocimiento para sistemas de información abiertos. Murcia: Universidad de Murcia, 2015. Disponível em: http://www.iskoiberico.org/wp-content/uploads/2015/11/80_Ferneda.pdf. Acesso em: 28 set. 2016.

IBAÑOS, A. M. T. Kripke sobre sentenças de crença. In: CAMPOS DA COSTA, J. PEREIRA, V. W. (org.) **Linguagem e cognição:** relações interdisciplinares. Porto Alegre: EDIPUCRS, 2009, p.143-152.

INGWERSEN, P. Conceptions of Information Science. In: VAKKARI, P. CRONIN, B. (ed.) **Conceptions of Library and Information Science:** historical, empirical and theoretical perspectives. London: Taylor Graham, 1992. p.299-312.

INTERAÇÃO PESSOA-MÁQUINA EM LINGUAGEM NATURAL – INQUER, 2003. Disponível em: http://www.clul.ulisboa.pt/clg/inquer/proj_inquer.htm. Acesso em: 12 jan. 2017.

JONES, K. S. **Natural language processing:** a historical review, 2001. Disponível em: https://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf. Acesso em: 8 set. 2016

LEWIS, D. D.; JONES, K. S. Natural language processing for information retrieval. **Association for Computing Machinery – ACM**, v.39, n.1, 1996. Disponível em: http://dl.acm.org/citation.cfm?id=234173.234210. Acesso em: 7 set. 2016

LOGAN, R. K. **Que é informação?**: a propagação da informação na biosfera, na simbolosfera, na tecnosfera e na econosfera. Rio de Janeiro: Contraponto: PUC-Rio, 2012.

LYOTARD, J. F. **O pós-moderno**. 3.ed. Tradução de BARBOSA, R. C. Rio de Janeiro: José Olympio J.O. Editora, 1991.

MANNING, C. D.; RAGHAVAN, P.; SCHU˙ TZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MARRAFA, P.; AMARO, R.; MENDES, S. WordNet.PTglobal – Extending WordNet.PT to Portuguese Varieties. **Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics**, Edimburgo, p. 70-74, 2011. Disponível em: http://www.aclweb.org/anthology/W/W11/W11-2609.pdf. Acesso em: 22 dez. 2016.

MARRAFA, P. *et al.* Portuguese Controlled Language: coping with ambiguity. In: KUHN, T.; FUCHS, N. E. (eds.) **Controlled Natural Languages - CNL**, Berlin Heidelberg: Springer-Verlag, p. 152-166, 2012. Disponível em: http://www.clul.ul.pt/clg/files/CNL2012.pdf. Acesso em: 10 jan. 2017.

MARRAFA, P. *et al.* **WordNet.Pt – uma rede léxico-conceptual do Português on-line**. Lisboa, 2005. Disponível em: http://www.clul.ul.pt/clg/files/WordnetPT_APL.pdf. Acesso em: 10 ago. 2015.

MENDES, S.; NECSULESCU, S.; BEL, N. Synonym extraction using a language graph model. In: MITITELU, V. B.; POPESCU, O.; PEKAR, V. (orgs.) **Semantic Relations II:** enhancing resources and applications - Language Resources and Evaluation Conference, Istambul, p. 7-15, 2012. Disponível em: http://www.lrecconf.org/proceedings/lrec2012/workshops/10.Semantic%20Relations%20II%20Proceedings.pdf. Acesso em: 10 jan. 2017.

PINKER, S. Afinal, como de fato funciona a mente? In: CAMPOS DA COSTA, J. PEREIRA, V. W. (org.) **Linguagem e cognição:** relações interdisciplinares. Porto Alegre: EDIPUCRS, 2009, p.11-46.

RIJSBERGEN, C. J. V. **Information Retrieval**. 2.ed. Londres: Butterworths, 1979.

SANDERSON, M. **Word sense disambiguation and information retrieval**. 1996. 136f. Tese (Doutorado) – Departamento da Ciência da Computação, Universidade de Glasgow, 1996. Disponível em: http://theses.gla.ac.uk/4463/. Acesso em: 14 jun. 2019.

SANTOS, P. L. V. A. C. Catalogação, formas de representação e construções mentais. **Tendências da Pesquisa Brasileira em Ciência da Informação** (ANCIB), v. 6, n. 1, 2013. Disponível em: http://inseer.ibict.br/ancib/index.php/tpbci/article/view/100/140. Acesso em: 9 jan. 2017.

STEWART, D.L. **Building Enterprise Taxonomies**. Mokita Press, 2008.

SUSSNA, M. Word sense disambiguation for free-text indexing using a massive semantic network. **International Workshop on Web Information and Data Management** (WIDM): Anais da segunda conferência internacional sobre gestão da informação e conhecimento. Nova Iorque: Association for Computing Machinary, 1993. Disponível em: https://dl.acm.org/doi/10.1145/170088.170106. Acesso em: 13 jun. 2019.

VARELAS, G. et al. Semantic similarity methods in WordNet and their application to information retrieval on the Web. **International Workshop on Web Information and Data Management** (WIDM): Anais do 7º workshop internacional da Association for Computing MACHINARY (ACM) sobre gerenciamento de dados e informações da Web. Nova Iorque: Association for Computing Machinary, 2005. Disponível em: https://dl.acm.org/doi/10.1145/1097047.1097051. Acesso em: 13 jun. 2019.

VOSSEN, P. EuroWordNet: building a multilingual database with wordnets for European languages. **The ELRA Newsletter**, v.3, n.1. Paris, 1998, p. 7-10. Disponível em: https://research.vu.nl/ws/portalfiles/portal/74104472/elra. Acesso em: 14 fev. 2017.

WEI, T. et al. A semantic approach for text clustering using WordNet and lexical chains. **Expert Systems with Applications**, v.42, n.4, 2015. P.2264-2275. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417414006472. Acesso em 3 set. 2019.

WERSIG, G. Information Science: the study of postmodern knowledge usage. **Information Processing and Management**: An International Journal, Tarrytown-Nova Iorque, v.29, n.2, p.229-239, Mar./Abr. 1993

WORDNET.PT – REDE LÉXICO-CONCEPTUAL DO PORTUGUÊS, 2017. Disponível em: http://www.clul.ul.pt/clg/wordnetpt/index.html. Acesso em: 13 jan. 2017.