Verbos do domínio jurídico: uma proposta de organização ontológica com vistas ao PLN

Isa Mara da Rosa Alves (UNESP)
Rove Luiza de Oliveira Chishman (UNISINOS)
Paulo Miguel Torres Duarte Quaresma (Universidade de Évora)

Resumo

Apresenta-se, neste artigo, uma proposta de estruturação ontológica de verbos do domínio jurídico com vistas ao aperfeiçoamento de sistemas de Processamento Automático da Língua Natural (PLN), mais especificamente, ao sistema de busca e extração de informações da Procuradoria Geral da República de Portugal. As relações lógico-semânticas, os papéis semânticos e os *frames* foram as abordagens semânticas que se mostraram mais produtivas para a construção da ontologia proposta.
Palavras-chave: Semântica Verbal; Relações Lógico-Semânticas; Papéis semânticos; Frames; Ontologia.

Introdução

Esta é uma pesquisa interdisciplinar, comprometida fundamentalmente com a Semântica Lexical Computacional em interface com a área de estudo

dedicada ao Processamento Automático da Língua Natural (PLN), subárea da Lingüística Computacional e, por sua vez, da Inteligência Artificial (IA). O PLN, área de estudo que surgiu da necessidade de uma melhor interação homem-máquina, trata do desenvolvimento de sistemas computacionais que buscam processar a língua humana. Nas investigações sobre o léxico com vistas ao PLN, convergem abordagens e metodologias provenientes de diferentes disciplinas, tais como a Psicolingüística, a Antropologia, a Lingüística (Formal e Aplicada), bem como a Lingüística Computacional. Sistemas como tradutores automáticos, corretores gramaticais e ortográficos, sumarizadores textuais, thesaurus eletrônicos, ontologias, base de dados lexicais, processadores de fala e sistemas de busca e extração de informações são alguns exemplos de recursos computacionais que fazem uso do PLN. O foco aqui é a construção de uma representação ontológica de verbos do domínio jurídico que sirva para o aperfeiçoamento de sistemas *on-line* de busca e extração de informações.

Sabe-se da necessidade e da vontade comercial e científica de criar sistemas de busca cada vez mais eficientes (ágeis e precisos). Há consenso na comunidade científica de que a forma de melhorar a qualidade dos sistemas de PLN é ampliar a sua capacidade de processar informações semânticas. Essa, no entanto, é uma tarefa bastante custosa que não pode ser realizada de forma totalmente automática e exige o trabalho de Lingüistas em cooperação com Lingüistas Computacionais.

Para a melhoria do desempenho de sistemas de busca na Web, surge a proposta de reorganização da Web atual em uma Web Semântica (do Inglês: *Semantic Web*). De maneira bastante simplificada, a idéia é incluir anotações semânticas nos documentos disponibilizados na Web de maneira que os sistemas de busca possam realizar inferências sobre o conteúdo dos textos. Em dezembro de 2003, a World Wide Web Consortium (W3C) estabeleceu como linguagem padrão dessa revolucionária concepção de Internet a OWL (Ontology Web Language), em lugar da linguagem de marcação de hipertexto HTML (Hypertext Markup Language). Nota-se que a Semântica, em especial a Semântica Lexical Computacional, tem papel central para a construção de sistemas de busca no contexto da Web Semântica.

Apresentam-se aqui abordagens semânticas próprias para a descrição verbal de forma a possibilitar a construção baseada em córpus de uma ontologia de domínio jurídico que tem a finalidade específica de contribuir para a melhoria do desempenho de sistemas de busca e extração de informações na Web, em especial, o sistema da Procuradoria Geral da República de Portugal (PGR). Este estudo reúne pesquisadores da Lingüística do Brasil e da Informática de Portugal que colaboram no âmbito de projetos maiores que unem alguns países da União Européia no intuito de construir uma *wordnet* multilingüe especializada para o domínio jurídico que facilite a busca de informações desse domínio em sites jurídicos oficiais desses países em suas próprias línguas (português, italiano, inglês, checo e holandês).

O presente trabalho objetiva responder fundamentalmente a seguinte questão: como representar ontologicamente a semântica de verbos de domínio jurídico de forma a possibilitar a melhoria do desempenho de sistemas *on-line* de busca e extração de informações? Para tanto, o artigo foi organizado como segue: em (1), apresenta-se uma sistematização de aspectos da semântica verbal que servirão para a descrição ontológica dos verbos, bem como alguns resultados da análise do córpus; em (2), trata-se da descrição ontológica dos verbos, caracterizando o corpus e descrevendo os procedimentos e questões

metodológicas da análise; em (3), descreve-se como as informações semânticas representadas na estrutura ontológica proposta foram inseridas em uma ferramenta própria para a edição de ontologias, o Protégé.

1. Aspectos da Semântica Verbal

Lingüistas, psicolingüistas, antropologistas e cientistas da computação têm criado (ou utilizado) diferentes representações do léxico, dependendo dos aspectos da língua que desejam focalizar. Nesta seção, serão apresentadas abordagens lingüísticas que se prestam à descrição ontológica do significado de verbos do domínio jurídico. A partir da análise das informações lingüísticas empregadas por outros projetos de construção de ontologias, léxicos computacionais e bases de dados lexicais (Mikrokosmos, Cyc, Wordnet e *Frame*net), o critério de seleção das abordagens lingüísticas apresentadas aqui visa aos seguintes fins: (i) descrever de forma completa o significado dos verbos em questão, portanto, percorrendo diferentes níveis de análise (semântico, sintático e pragmático) sem perder de vista o rigor teórico; e (ii) formalizar essas descrições em linguagem computacionalmente compatível, no caso, a OWL.

Borba (1996) considera que, para descrever exaustivamente os verbos de uma língua, o importante é percorrer todos eles e identificar particularidades e idiossincrasias, mas isso só se faz com eficiência depois de se identificar uma base comum entre eles, mesmo porque essa operação tem interesse prático: economia, uniformidade e simplicidade descritiva. Nesse caso, uma solução (ou sugestão, ou alternativa) apresentada pelo autor é procurar *áreas temáticas* que permitam estabelecer traços gerais comuns e, dentro de cada área, descobrir equivalências parciais (já que totais são raras). Esse critério, auxiliado pela taxonomia[1], fornecerá várias pistas para uma descrição adequada e completa.

Serão apresentadas, nesta seção, abordagens da semântica verbal que contribuem para a descrição dos verbos de uma maneira completa. Trata-se de perspectivas teóricas que contemplam papéis semânticos e *frames* são os níveis de análise.

1.1. Aspectos lógico-semânticos do significado verbal

Lehrer (1974) explica que uma análise de *campos semânticos* – ou *áreas temáticas*, conforme Borba (1996) – é baseada na concepção de que o significado de uma palavra em um dado campo surge das *relações semânticas* – de similaridade e de contraste – que se estabelecem entre ela e as outras. Uma abordagem relacional aceita a existência de domínios e descreve como os elementos de um domínio estão relacionados a outros. Os nós que os concectam são chamados relações lexicais ou semânticas.

Saint Dizier e Viegas (1995) ressaltam que as relações semânticas desempenham um papel essencial na semântica lexical e interferem em muitos níveis de compreensão e produção em linguagem natural. Elas são também consideradas elementos centrais na organização de bases de conhecimento lexicais ou ontologias. Uma abordagem relacional do léxico é útil ainda para o estudo das diferenças entre as línguas, detectando padrões distintos de lexicalização; portanto, especialmente interessante para este trabalho.

Tomando como base as relações propostas pelas *wordnets*, foram selecionadas as seguintes relações que permitem estruturar um léxico do domínio jurídico: (a) antonímia (b) acarretamento, (c) causa, (d) hiponímia e (e) sinonímia. Apresentam-se abaixo algumas características fundamentais de cada relação. Por falta de espaço aqui não serão tratadas certas sutilezas ou particularidades da definição das relações, tais como a apresentação dos subtipos e dos testes aplicados para garantir a identificação das relações.

A *antonímia*, assim como a sinonímia, é uma relação semântica bastante familiar. Interessa-nos aqui vê-la simplesmente como uma relação de exclusão entre conceitos opostos. Fellbaum (1990) destaca a antonímia verbal entre elementos de um mesmo campo semântico referindo-se a uma mesma atividade, mas envolvendo participantes diferentes, pois estes desempenham funções opostas na atividade (predicação), é o caso de *absolver* e *condenar*.

O *acarretamento* é uma relação que a semântica lexical importou da lógica de proposições. Referimo-nos aqui a um tipo de acarretamento geral que se dá entre dois verbos V1 e V2: quando uma sentença em que *Alguém* V1 acarreta *Alguém V2* for verdadeira. O acarretamento é facilmente confundido com a hiponímia. Para que isso não ocorra, é importante que se tenha em mente que, na relação de acarretamento, V1 e V2 ligam-se por uma relação de inclusão temporal, ou seja, V1 ocorre antes de V2, por exemplo. Na hiponímia, por outro lado, trata-se de uma relação entre dois verbos que co-ocorrem temporalmente. Como exemplo de acarretamento, tem-se *julgar* e *condenar*. Um juiz precisa primeiro *julgar* para depois *condenar*, mas também pode *julgar* sem *condenar*; conseqüentemente, *condenar* acarreta *julgar*.

As relações de hiponímia e sinonímia assumiram papel de destaque na descrição semântica dos verbos que compõem a ontologia. Bem como ocorre em léxicos de nominais, conforme descreve Fellbaum (1998), a hiponímia é a base da organização ontológica dos verbos.

Em linhas gerais, pode-se dizer que a hiponímia é uma relação lexical correspondente à inclusão de uma classe em outra caracterizada pelo acarretamento unilateral, ou seja, V1 *é um tipo de* V2, mas V2 *não é um tipo de* V1. Há autores, como Miller e Fellbaum (1991), que contestam a aplicabilidade dessa frase lógica para os verbos e preferem V1 *é uma forma de/ é uma maneira de* V2, preferindo, por essa razão, denominar a relação de troponímia ou relação de maneira. Contudo, bem como Cruse (2000) e Vossen (1997), não será adotada essa distinção de nomenclatura aqui.

Algumas vezes, a hiponímia pode ser confundida com a sinonímia. A característica fundamental de distinção entre elas é a questão do acarretamento unilateral na hiponímia/hipernímia e bilateral na sinonímia. Outra distinção entre elas é o fato de que os sinônimos são mutuamente exclusivos quando ocorrem em um mesmo contexto/sentença por causarem redundância, ao contrário dos hipônimos, que são parcialmente intercambiáveis.

A identificação de sinônimos não é algo simples nem consensual na literatura, talvez o único consenso seja sobre a inexistência de sinonímia absoluta. De acordo com Cruse (1986; 2000), sinônimos são palavras cujas similaridades são mais salientes do que as diferenças. Percebe-se, em sua posição, que ele desconsidera a existência de palavras com total identidade semântica, isso porque há aspectos relativos ao *uso* que devem ser levados em conta na identificação dos pares de sinônimos. Os sinônimos, de acordo com essa perspectiva, são itens lexicais cujos sentidos são idênticos no que se refere aos traços semânticos centrais, mas diferem no que diz respeito a traços periféricos (VIEIRA et al., 2003).

No caso da representação ontológica de verbos proposta aqui, a hiponímia facilitará a realização de inferências por parte do sistema, possibilitando, por exemplo, a interpretação de perguntas como: *que veículos automotores mais se acidentam?* O sistema, para responder a esse tipo de pergunta, deve ser capaz de associar *carro*, *motocicleta*, *motociclo*, *autocarro* ("ônibus" em PE) com *veículos automotores*.

A relevância da sinonímia na ontologia aqui proposta deve-se à possibilidade de auxiliar o usuário no momento da busca, na medida em que ele não precisará se limitar a empregar um termo jurídico específico, podendo realizar suas consultas através de termos equivalentes.

1.2. Aspectos gramaticais do significado verbal

Há informações fundamentais para a descrição da semântica verbal que só podem ser expressas com um estudo das relações entre as entidades verbais e os elementos que co-ocorrem com elas nas sentenças. Abandona-se o nível de análise exclusivamente lexical e passa-se a analisar o nível sentencial, na interface sintaxe e semântica.

As sentenças descrevem situações – estados, eventos, ações – em que várias entidades estão envolvidas de diferentes formas. Tais entidades serão chamadas de participantes. Para classificar os tipos de situações em que os verbos do corpus ocorreram, seguiu-se exclusivamente Borba (1996) por descrever os verbos do Português a partir de uma análise de corpus bastante completa, tomando como base a Gramática de Valência.

Observou-se uma predominância dos verbos de ação e ação-processo, seguidos dos verbos de processo. Os verbos estativos não constituíram objeto dessa análise.

Para classificar os participantes das situações, ou seja, os argumentos dos verbos, diferentes autores contribuíram para as conclusões (ex.: FILLMORE, 1968; CHAFE, 1970; JACKENDOFF, 1975; DOWTY, 1979; FRAWLEY, 1992; BORBA, 1996; SAEED, 1997; KEARNS, 2000). A tarefa de classificação dos argumentos de acordo com o papel semântico (ou temático) aos quais eles remetem não é nada simples dado o alto grau de subjetividade da tarefa. O estudo realizado mostrou que há muitas sobreposições nas classificações dos referidos autores e que há diferenças muito sutis entre um papel e outro dos argumentos. Tendo em vista os fins práticos de construção de uma ontologia, não serão considerados certos detalhamentos de certos tipos de argumentos dos verbos. Optou-se por agrupar certos papéis temáticos em uma só denominação. Como exemplo, tem-se o papel de *paciente,* que é o receptor lógico típico que refere-se ao elemento modificado no processo ou na ação. Sob essa noção incluem-se também os papéis *tema* e *objetivo*, no sentido empregado por Chafe (1970). A denominação *objetivo* foi adotada para caracterizar outro tipo de participante, aquele também chamado de *meta* (o destino/alvo da predicação).

Abaixo se apresenta uma figura que sistematiza os papéis semânticos os quais farão parte deste estudo organizados de forma hierárquica. A estrutura geral é inspirada na proposta de Frawley (1992), que divide os papéis temáticos em papéis semânticos participantes, os quais se subdividem em atores lógicos, receptores lógicos e papéis espaciais, e papéis semânticos não-participantes, que dizem respeito aos papéis opcionais para a predicação.

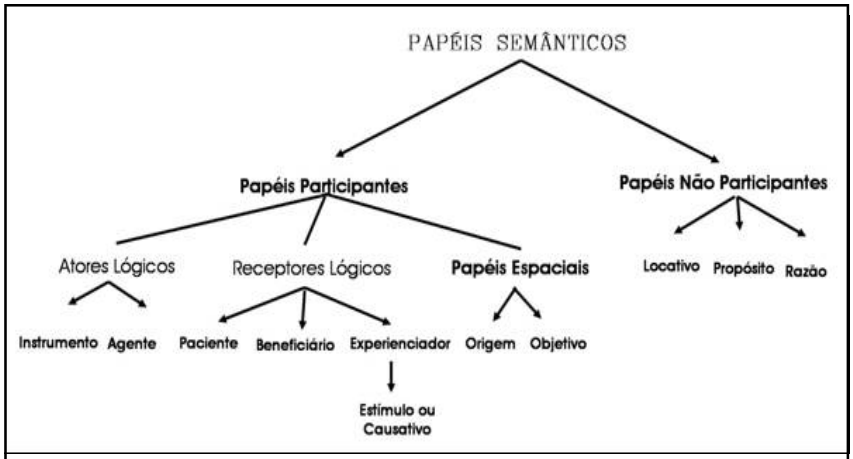

Figura 1. Papéis Semânticos

Os papéis semânticos identificados no corpus foram: (a) agente, (b) instrumento, (c) beneficiário, (d) paciente, (e) objetivo, (f) origem, (g) locativo, (h) propósito e (i) razão.

1.3. Aspectos contextuais do significado verbal

Nesta seção será discutida a concepção de *frame* que fornece subsídio para a inclusão de aspectos contextuais na descrição ontológica proposta para os verbos jurídicos em questão.

A noção de *frame* recebe interpretações variadas tanto no escopo da Lingüística quanto da Computação. Algumas abordagens mantêm-se mais ligadas à estrutura da sentença (ex.: *wordnets*), outras menos (ex.: *Framenet*). Quem originalmente tratou da questão foi Minsky (1974), que propôs que o significado fosse organizado através de um conjunto de dados estruturados para representar formalmente situações estereotipadas em uma só estrutura de dados contendo todo o conhecimento relevante acerca dessa dada entidade. Em termos formais, um *frame* é uma estrutura composta por atributos, valores, relações e restrições sobre os elementos participantes de uma situação (ex. julgamento). Uma estrutura *frame*, portanto, pode englobar diferentes níveis de informações sobre um item lexical, dependendo da orientação teórica – semântica, sintática, textual, pragmática – e da aplicação proposta.

Tomou-se como base para a identificação dos *frame*s ativados pelos verbos do corpus a base de dados *Framen*et (FN), de Charles Fillmore. Para representar os elementos *frame* dos dez verbos do corpus (condenar, absolver, julgar, acordar, conceder, revogar, concluir, alegar provar e recorrer), foram selecionados 9 *frame*s semânticos do FN, sendo, algumas vezes, necessário combinar diferentes *frame*s para descrever de maneira mais completa os participantes da situação em questão. Os elementos *frame* identificados no corpus são: (a) avaliador (juiz, magistrado, tribunal); (b) avaliado (réu, representante, autor); (c) argumentador (réu, representante, autor); (d) reconhecedor (réu, representante, autor // juiz, magistrado, tribunal); (e) meio; (f) base legal; (g) razão; (h) propósito; (i) evidência; (j) tópico; (k) conteúdo; (l) mensagem; (m) pedido. Propriedades como *tempo, condição, local e maneira* nem sempre são lexicalizados na sentença ou no corpo do texto, mas sempre estão presentes no cabeçalho de todos os Acórdãos. Fazer referência a eles é

fundamental, pois veiculam informações que situam o documento em relação a quando o processo foi julgado, as condições do julgamento, onde o processo foi julgado e o tipo de processo.

Passa-se agora para a discussão dos aspectos de descrição ontológica dos verbos do corpus.

2. A descrição ontológica dos verbos

As discussões de base fundamentalmente teóricas empreendidas nas seções anteriores convergem aqui para a descrição semântica dos verbos selecionados no corpus e para a construção de representações formais desses dados semânticos, o que dará origem à ontologia de domínio jurídico aqui proposta – a UNIVERBUE.

2.1 Corpus, procedimentos e questões metodológicas

O corpus da pesquisa é constituído de documentos disponibilizados eletronicamente[1] nas bases de dados jurídicos do Instituto das Tecnologias de Informação na Justiça de Portugal.

A metodologia adotada pode ser dividida em três grandes fases: (i) a fase de pré-análise, (ii) a fase de análise semântica do corpus; e (iii) a fase de construção da representação ontológica de verbos do domínio jurídico, a qual será apresentada na seção seguinte.

A fase de pré-análise do corpus tem como objetivo geral prepará-lo para a análise propriamente dita. Para tanto, foram realizados procedimentos que possibilitaram sua definição, seleção e contextualização. Foram escolhidos aleatoriamente 6 acórdãos homologados no período 2002-2003 sobre o tema "acidentes rodoviários" (no sistema português: "acidentes de viação") e, após a análise dos textos, foi feita a extração e contagem automática das ocorrências verbais e suas respectivas concordâncias com o intuito de selecionar os verbos que comporão a ontologia. Os 10 mais freqüentes foram selecionados (21 Recorrer, 14 Condenar, 12 Julgar, 10 Provar, 10 Revogar, 9 Concluir, 8 Acordar, 7 Alegar, 6 Absolver e 4 Conceder), totalizando 101 concordâncias analisadas. Na etapa de contextualização do corpus, com base em abordagens discursivas, investigou-se como se organizam textualmente os Acórdãos, bem como seu papel no contrato de comunicação em que se inserem. Essa compreensão global do corpus auxiliou na identificação dos papéis semânticos dos argumentos, bem como nas outras questões semânticas específicas da análise.

De posse das informações descritas acima, iniciou-se a fase de análise propriamente dita. O primeiro passo foi, com auxílio do concordanceador do Wordsmith**,** analisar as sentenças nas quais os verbos em questão ocorrem a fim de proceder a uma representação ontológica em quatro níveis para cada um dos verbos. Trata-se da seleção de (i) uma definição[2], (ii) das relações lógico-semânticas, (iii) dos papéis temáticos e, (iv) dos elementos *frame*. Abaixo se apresenta o resultado da análise do verbo *condenar*, como uma ilustração do estudo feito.

O verbo *condenar* apresentou 14 ocorrências em 6 acórdãos. *Condenar* é um verbo polissêmico que remete a diferentes tipos de situações de acordo com seus complementos. Conforme Borba (2002), condenar pode expressar dois tipos de situação: *ação-processo* e *ação*. Interessa para o domínio jurídico

a situação do tipo *ação-processo* de sentido "declarar culpado; pronunciar uma sentença a alguém em um tribunal, reconhecendo-o responsável pelo delito, crime ou falta e atribuindo-lhe uma pena".

A figura abaixo mostra as entidades verbais que possuem relações lógico-semânticas com o verbo *condenar* no sentido focalizado. A base para essa identificação, conforme já foi dito, foi a Wordnet; no entanto, não se seguiu rigorosamente a descrição ali apresentada. Adaptações foram feitas de acordo com a relevância para domínio e aplicação.

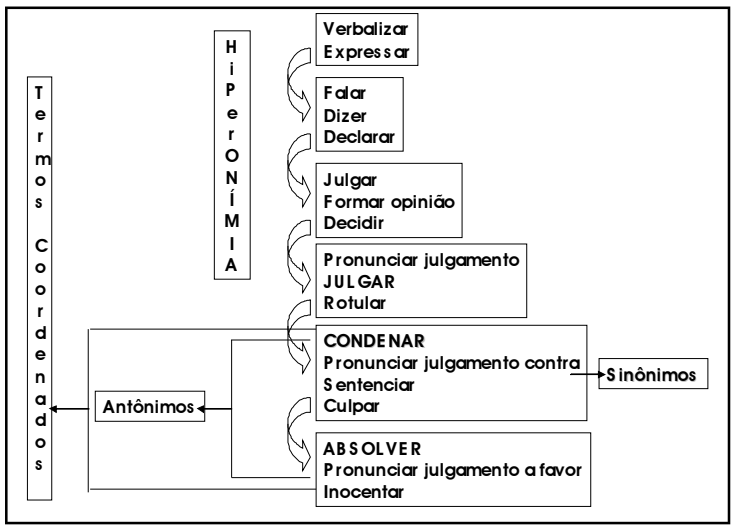

Figura 2. Relações lógico-semânticas *condenar*

A estrutura argumental de *condenar* (VTDI) prevê que *alguém* (argumento externo – ARG 0) condene um outro *alguém* (argumento interno – ARG 1) a *algo* (argumento interno – ARG 2). Entre as 14 sentenças analisadas, em apenas uma delas o ARG 0 aparece explícito. Esse fato não compromete a descrição dos argumentos, uma vez que na estrutura do verbo ele está previsto. Abaixo se apresenta uma sentença que mostra dois dos papéis semânticos atribuídos ao verbo *condenar*.

Exemplo 1 (a):

"Foi proferida nova sentença (fls. 241 a 252). E, no essencial, com o mesmo conteúdo da anterior, sendo a Ré [paciente] condenada a pagar aos autores, a mesma quantia global, de 6151000 escudos [objetivo]". (Fonte: Acórdão 02B2159)

O mesmo exemplo (1) será utilizado abaixo para mostrar alguns elementos *frame* ativados por *condenar*.

Exemplo 1 (b)

"Foi proferida nova sentença [meio] (fls. 241 a 252) [base_legal]. E, no essencial, com o mesmo conteúdo da anterior, sendo a Ré [avaliado] condenada a pagar aos autores, a mesma quantia global, de 6151000 escudos [tópico]". (Fonte: Acórdão 02B2159)

No sentido específico do domínio jurídico, *condenar* pode ser considerado um verbo de *comunicação_de_julgamento*, pois há um *comunicador* (o qual está implícito na sentença acima) que comunica um julgamento sobre um *avaliado* (a autora) a um sujeito alvo (a autora e os réus não explícitos na sentença). Essas informações semânticas sobre o papel situacional do ARG 0 e do ARG 1 serão de grande valia para o funcionamento da ontologia, pois possibilitam a inserção de restrições como *agente* do tipo *comunicador* e *paciente* do tipo *avaliado*. Vê-se a presença ainda de uma entidade *meio*[1] (sentença) que identifica o local abstrato através do qual a condenação é preferida. O *tópico* identifica a penalidade sofrida pelo *avaliado* (pagar pensão anual).

Para encerrar a apresentação ilustrativa da análise feita nos 10 verbos do corpus, abaixo se apresenta um quadro que sistematiza a descrição ontológica do verbo *condenar*.

ENTIDADE: *condenar*
Definição*: declarar culpado; pronunciar uma sentença a alguém em um tribunal, reconhecendo-o responsável pelo delito, crime ou falta e atribuindo-lhe uma pena (WN - tradução minha; BORBA, 2002 e DLPC, 2001)*
Relações lógicas-semânticas: *sinonímia, antonímia, hiperonímia, termos coordenados.*
Papéis semânticos*: agente, paciente, objetivo, instrumento, razão.*
Frame semântico: *avaliador, avaliado, meio, base legal, razão, local, condições, tempo, maneira, tópico.*

Quadro 1. Estrutura Ontológica *Condenar*

Ressalta-se que a descrição da semântica dos verbos aqui apresentada partiu das informações fornecidas pelo corpus, mas não se limitou às possibilidades ali expressas. Dessa forma, estabeleceu-se relação semântica dos dez verbos de base com outros independentemente de sua ocorrência ou não nos acórdãos, considerando apenas as restrições do domínio. No caso dos papéis semânticos, seguiu-se o mesmo caminho: foram representados participantes das situações que não ocorreram com os verbos em questão na amostragem de textos do corpus (ex.: agentes implícitos nas sentenças).

3. A edição da ontologia UNIVERBUE

Nesta seção será apresentado um exercício de construção de uma ontologia, a UNIVERBUE, no intuito de propor uma organização para os diferentes níveis de análise lingüística em linguagem computacionalmente legível.

É importante lembrar que uma ontologia é uma representação estruturada do conhecimento (informações lingüísticas). Sem o *conhecimento* não há, portanto, ontologia. No entanto, sem estruturação desse conhecimento, sua utilização em PLN torna-se mais complicada. De posse da descrição ontológica dos verbos, a tarefa agora é sistematizar essas informações no editor de ontologias Protégé que possibilitará a conversão dos dados para a OWL.

Com o auxílio do Protégé, as descrições semânticas aqui propostas passam a ter um alcance muito maior do que apenas a proposição de uma metodologia para a descrição semântica de verbos. Será possível atingir, dessa forma, os principais objetivos da construção de uma ontologia, são eles: (i) compartilhar conhecimento estruturado de informações comuns entre pessoas e máquinas (sistemas computacionais); (ii) possibilitar o reuso do conhecimento de determinado domínio; (iii) tornar explícito o conhecimento sobre determinado domínio; (iv) separar o conhecimento de um domínio do conhecimento operacional de construção de um sistema; (v) analisar o conhecimento de um domínio.

As vantagens da editoração de uma ontologia em uma ferramenta desse tipo são indiscutíveis, mas não se pode negar que esse exercício de organização do conhecimento verbal revelou também algumas limitações do Protégé. Não serão discutidas a fundo aqui tais questões, apenas será apresentada a organização proposta para a representação ontológica dos verbos do domínio jurídico.

A figura apresenta uma visão geral de como o conhecimento lingüístico descrito na seção anterior foi inserido na ferramenta.

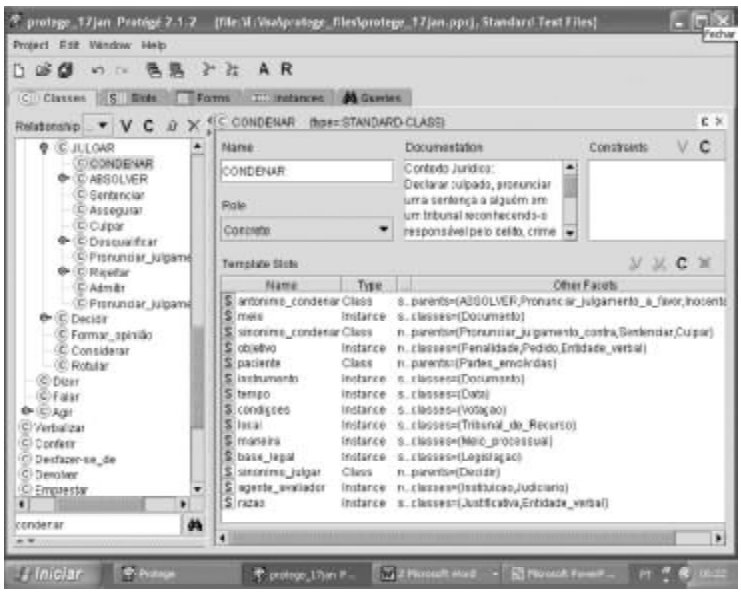

Figura 3. Protégé - Visão geral

A *definição* foi inserida no campo *documentação*. A relação semântica de hiponímia, a qual é a base da ontologia, foi inserida no campo *relationship*. A especificação das demais relações lógico-conceituais entre as classes expressas no campo *relationship* foram inseridas nos *template slots*, bem como os papéis semânticos e alguns dos elementos *frame*.

A ferramenta Protégé encontra-se organizada em sintonia com uma abordagem de *frames* sob a ótica de Minsky (1974). O que foi dito na seção referente aos *frames* vai se tornar verdade absoluta nesta etapa da ontologia: foi criada uma *estrutura frame* (*template slot*) para cada verbo que inclui as informações ontológicas descritas na seção anterior de maneira integrada.

Os papéis semânticos e os elementos *frame* referem-se a entidades semelhantes em uma situação de forma mais ou menos independente da estrutura sintática. A diferença reside no fato de que os papéis semânticos representam os participantes da predicação e os elementos *frame* representam os participantes da situação evocada pela unidade lexical em questão (nesse caso, os verbos). Assim, o desafio, nesta etapa de construção da ontologia, foi encontrar uma forma de não inserir informações sobrepostas nem perder preciosas informações. A solução encontrada foi incluir todos os papéis semânticos na estrutura *frame* (*template slot*) da ferramenta e apenas os elementos *frame* que representarem informações ainda não descritas pelos papéis semânticos foram inseridos entre os *slots*, caso de *agente_avaliador* e de *razão* (figura 3). Os elementos *frame* que não foram inseridos entre os *slots* serviram para organizar a estrutura hierárquica de entidades referentes a outras classes gramaticais, tais como *avaliado*, *avaliador*, *tribunal*, conforme figura 4, abaixo:

Figura 4. Protégé – Hierarquia Nominais

Ressalta-se que algumas vezes foi necessário substituir o nome padrão dado pelo Framenet e apresentado na estrutura ontológica por algum equivalente que fosse mais genérico. Esse foi o caso de noções como *avaliado*, *réu*, *autor, representante* que foram agrupadas em *avaliado*; *juiz, magistrado, falante, argumentador,* que foram agrupadas em *avaliador*; o papel de *meio* foi substituído por *documento* para referir *petição*, *acórdãos*, etc.

Esses elementos *avaliador* e *avaliado* serviram também para subespecificar a noção semântica expressa pelos papéis semânticos *agente* e *paciente*. Isso foi feito para contornar a limitação já comentada na seção sobre as relações lógico-semânticas: o sistema não permite a inserção de um mesmo *slot* com facetas diferentes. O que deu origem a *slots* do tipo: *agente_avaliador* e *agente_avaliado*.

Vimos, na análise do corpus, que muitas vezes os verbos recebem complementos oracionais. A ferramenta Protégé, como um editor de ontologias, não permite a inserção de orações como classes. A solução encontrada para a representação desse tipo de complemento verbal foi indicar que o verbo tem como complemento alguma subclasse de outra *entidade_verbal*. A análise da estrutura interna desse complemento oracional será feita quando o verbo núcleo dessa oração for descrito.

A partir da inserção dos dados lingüísticos no editor de ontologias Protégé, as informações foram exportadas para a linguagem implementacional própria para a codificação de ontologias na era da Web Semântica, a OWL. Somente a partir desses dados formais é que a ferramenta de busca e extração de informações na Web fará uso do conhecimento descrito no Protégé. Tendo a ontologia jurídica representada em OWL, é possível aos sistemas de busca usarem esta informação, de modo a melhorarem o seu desempenho e responderem de um modo mais exato às questões dos usuários, interagindo em língua natural. Um exemplo de interação usuário-sistema jurídico é o seguinte:

Usuário: Qual foi a condenação da ré A do processo 03B3338?

Na base de documentos existe uma frase relacionada com essa interrogação:

Documento: (...) a Relação de Lisboa julgou-a procedente em parte e, conseqüentemente, procedente a reconvenção condenando a ré a pagar, a uma e a outra, 50% dos prejuízos que ambos sofreram sendo de 973.720$00 os do A e de 330.997$00 (...) (Fonte: 03B3338).

Como se pode observar, a resposta para a pergunta do usuário não está explícita na sentença e, portanto, um sistema de busca de informação que não possua conhecimento do domínio com organização semântica dificilmente conseguirá extrair a informação desejada. Com a utilização de ferramentas de PLN (análise sintática e semântica), é possível identificar na interrogação a intenção do usuário em ser informado sobre uma dada condenação, ou seja, sobre o argumento classificado aqui como *objetivo*. Com o uso da ontologia UNIVERBUE, o sistema consegue recuperar que a ação "condenar" possui como atributo (slot) o *objetivo* da ação, sendo capaz de responder adequadamente: "pagar 50% dos prejuízos que ambos sofreram sendo de 973.720$00 os do A e de 330.997$00 (...)".

Conclusão

O objetivo maior deste trabalho foi apresentar uma metodologia de estruturação ontológica de verbos do domínio jurídico de maneira teoricamente rigorosa, de forma a possibilitar a construção de uma ontologia que possa contribuir para o funcionamento mais eficiente de um sistema jurídico *on-line* de busca e extração de informações que interaja com o usuário através de pergunta e resposta em língua natural.

A posição adotada aqui – conforme Cruse (2000) – é a de que, apesar de o significado de uma palavra poder ser visto como infinitamente variável e sensível ao contexto, há regiões de "alta densidade significativa" que formam "blocos de significações" com maior ou menor estabilidade em relação às

mudanças contextuais. Foram essas propriedades do significado verbal que procurou-se representar formalmente na ontologia UNIVERBUE. Contudo, somente a partir de um estudo visando à contextualização do córpus de forma a conhecer a situação de comunicação em que ele está inserido é que se pode estabelecer critérios e fazer opções teóricas para a descrição do conteúdo semântico de cada verbo na constituição do todo significativo dos textos em questão, Acórdãos Judiciais da PGR-PT.

A base teórica fundamental deste estudo foi a Semântica, em especial, a Semântica Lexical. A decisão por incluir abordagens que fazem interface com a sintaxe e com a pragmática deve-se ao compromisso de descrição do conhecimento ontológico dos verbos. Semanticistas que se dedicam à descrição da língua com vistas ao PLN recebem uma espécie de "licença poética" para incluir no léxico o máximo possível de informações relacionadas à significação." Enfatiza-se que os objetivos aplicados deste trabalho serviram como pretexto para muitas das opções teóricas, mas procurou-se ter o máximo de cuidado para não tratar os fenômenos da língua levianamente devido à necessidade de formalização de tais informações.

A análise do córpus evidenciou, abordagens baseadas unicamente em relações semânticas, papéis temáticos e *frames* não são suficientes para representar todo o conteúdo de um texto, como se podia prever. Entretanto, ficou evidente que é possível recuperar informações chave tanto para a construção das sentenças quanto para a construção do sentido do texto a partir de tais abordagens.

A metodologia adotada para a representação do conhecimento ontológico dos verbos permitiu a ampliação dos limites previstos para a UNIVERBUE. As relações lógico-semânticas se confirmaram como um recurso bastante produtivo e ágil para a inclusão de classes em uma ontologia. Foram inseridas inicialmente 10 classes verbais as quais geraram um total de 120, isso porque foi possível expandir a análise relacional para além dos 10 verbos base. A atribuição de papéis temáticos e de elementos *frame* possibilitaram ampliar o escopo de análise para os nominais e alguns adjetivos. A construção da UNIVERBUE partiu de seis Acórdãos Judiciais da PGR-PT; daí foram extraídas 359 ocorrências verbais diferentes e selecionadas as 120 referentes aos verbos em questão. Ao final da etapa apresentada aqui, a ontologia possui com 120 entidades verbais e 74 entidades não verbais.

Além das referidas contribuições teóricas, este trabalho possui especial valor social. Isso porque uma ontologia de domínio legal aplicada a sistemas de busca de informação possibilita o acesso e a compreensão dos conteúdos digitais de bases legais públicas tanto por meio de palavras-chave quanto de pergunta-resposta sem a necessidade de linguagem especializada (jargão jurídico). Esta pesquisa tem sua aplicação direta entre parceiros da Comunidade Européia, uma vez que a ontologia criada será parte de um sistema de busca multilíngüe de consulta a bases jurídicas de países europeus. O Brasil, contudo, pode beneficiar-se desta pesquisa não só pela possibilidade de consultar esse sistema de busca inteligente, como também, pela utilização dos resultados teóricos e aplicados para descrições semânticas de bases ontológicas aplicadas a sistemas brasileiros de PLN.

Enfatiza-se, por fim, que este estudo retrata um trabalho cooperativo entre profissionais da Lingüística e da Computação. Com isso, pretende-se ressaltar a importância dessa postura colaborativa entre profissionais dessas áreas para o aperfeiçoamento de sistemas de PLN.

ABSTRACT

This paper brings a proposal of ontological structure of judging verbs aiming the to contribute towards the enhancement of Natural Language Processing (NLP) systems, particularly to the system of Procuradoria Gerald a República de Portugal. The logical-semantic relationships, semantic roles, end *frame*s are the approaches witch have been more useful for the description of verbal semantics.
Keywords: Verbal semantics; Logical-semantic relationships; Semantic roles; Frames; Ontology.

Notas

1 O rótulo de relações *taxonômicas* refere-se aqui às relações semânticas que envolvem alguma hierarquia de classes.

2 Disponível em http://www.dgsi.pt

3 A definição foi selecionada a partir da Wordnet, além dos dicionários do PB, Borba (2002) e, do PE, Dicionário da Língua Portuguesa Contemporânea (2001). Entre os verbos selecionados, dentro do domínio jurídico, não foram observadas diferenças de significado entre o PB e o PE.

4 O *elemento frame meio* é uma noção também adotada por uma abordagem baseada em papéis semânticos , porém, neste trabalho ele é consideraremos somente um *elemento frame*. No nível temático, a posição indicada pelo *meio* é denominada *instrumento*; contudo, não se pode esquecer que a noção de *meio* como um *elemento frame* não está restrita à estrutura argumental da predicação.

Referências bibliográficas

BASE DE DADOS LEXICAIS PARA A LÍNGUA INGLESA – FrameNet. Universidade da Califórnia em Berkeley: The International Computer Science Institute (ICSI) Disponível em: <http://www.icsi.berkeley.edu/framenet/>. Acesso ao longo do desenvolvimento da dissertação.

BASE DE DADOS LEXICAIS PARA A LÍNGUA INGLESA – WORDNET 2.0. Universidade de Princeton: Laboratório de Ciências Cognitivas. Disponível em: <http://wordnet.princeton.edu/>. Acesso ao longo do desenvolvimento da dissertação.

BORBA, F. S. *Dicionários de Usos do Português do Brasil*. São Paulo: Ed. Ática, 2002.

_____. *Dicionário da Língua Portuguesa Contemporânea – Verbo*. Vol. 1 e 2, Editorial Verbo e Academia das Ciências de Lisboa, 2001.

BORBA, F. S. *Uma Gramática de Valências para o Português*. São Paulo: Editora Ática, 1996.

CHAFE, W. L. *Significado e Estrutura Lingüística*. Chicago: The University of Chicago Press, 1970.

CRUSE, D. *Meaning in Language: an Introduction to Semantics and Pragmatics*. New York: Oxford University Press, 2000.

CRUSE, D. A. *Lexical Semantics*. Cambridge: Cambridge University Press, 1986.

DOWTY, D. *Word Meaning and Montague Grammar*. Dordrecht: D. Reidel, 1979.

EVENS, M. W. *Relational Models of the Lexicon*. Cambridge: Cambridge University Press, 1988.

FELLBAUM, C. A Semantic Network of English Verbs. In: Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.

FELLBAUM, C. English Verbs as a Semantic Net. In.: *International Journal of Lexicography*, 3, pp. 278-301, 1990.

FILLMORE, C.J. The Case for Case. In.: Bach and Harms (Ed.): *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, New York, 1968.

FRAWLEY, W. *Linguistic Semantic*. London: Lawrence Erlbaum Associates, Publishers, 1992.

JACKENDOFF, R. *Morphological and Semantics Regularities in the Lexicon*. Language, Baltimore: v.51, 3, 1975.

KEARNS, K. *Semantics*. New York: St. Martin Press, 2000.

LEHRER, A. *Semantic fields and lexical structure*. Amsterdam: North-Holland, 1974.

MINSKY, M. A. *A Framework for Representing Knowledge*. Artificial Intelligence Memo 306, MIT AI Lab, 1974.

MILLER, G. A. E; FELLBAUM, C. Semantic Networks of English. In:. Levin e S. Pinker (eds), *Lexical and Conceptual Semantics*. Cambridge, MA: Blackwell, 1991.

SAEED, J. *Semantics*. Oxford: Brasil Blackwell, 1997.

SAINT-DIZIER, P. e VIEGAS, E. *Computational Lexical Semantics*. Cambridge: Cambridge University Press, 1995.

CHISHMAN, R. L. de O. ; VIEIRA, R. ; ALVES, I. M. da R. ; RIGO, S. Synonymy for Query Expansion in Information Search. In: *11th Portuguese Conference on Artificial Intelligence*, 2003, Beja. Lecture Notes in Artificial Intelligence. Berlin: Springer, 2003. p. 445-449.

VOSSEN, P. EuroWordNet: A Multilingual Database for Information Retrieval. (1997). In: THIRD DELOS WORKSHOP Cross-Language Information Retrieval, pp. 85-94. European Research Consortium For Informatics and Mathematics. Disponível em: <http://www.ercim.org/publication/ws-proceedings/DELOS3/Vossen.pdf>. Acesso em 01/03/2003.