Rumo a um Recurso Lexical para a Linguagem Jurídica Brasileira

Anderson Bertoldi[1], Rove Chishman[1]

[1]Programa de Pós-Graduação em Linguística Aplicada - Universidade do Vale do Rio dos Sinos (UNISINOS)
Caixa Postal 275 – 93.022-000 – São Leopoldo – RS – Brasil
andersonbertoldi@yahoo.com, rove@unisinos.br

Abstract. *This paper describes the initial steps to create a lexical resource for the Brazilian legal language. It is presented the methodology adopted to create the* `Processo_penal` *frame of the Brazilian legal system.*

Resumo. *Este artigo descreve as etapas iniciais do desenvolvimento de um recurso lexical para a linguagem jurídica brasileira. É apresentada aqui a metodologia adotada para a criação do* frame `Processo_penal` *para o sistema jurídico brasileiro.*

1. Introdução

Este artigo descreve as etapas iniciais do desenvolvimento de um recurso lexical para a linguagem jurídica brasileira. Trata-se de um projeto lexicográfico baseado na Semântica de Frames, sendo desenvolvido no âmbito do projeto *Tecnologias Semânticas e Sistemas de Recuperação de Informação Jurídica*[1]. Esse projeto prevê a construção de recursos lexicais e recursos baseados em conhecimento jurídico, como léxicos e ontologias, para uso em recuperação de informação jurídica. O projeto lexicográfico que se apresenta neste artigo é parte integrante desse conjunto de esforços que vem sendo realizado na descrição do conhecimento e linguagem jurídicos brasileiros.

Neste trabalho apresentam-se: (i) a metodologia adotada para a criação do *frame* `Processo_penal` e (ii) os *subframes* que compõem o *frame* `Processo_penal`, bem como as unidades lexicais evocadoras desses *subframes* e os exemplos de sentenças anotadas para o português.

2. Léxicos Jurídicos

Este trabalho, que aplica o paradigma FrameNet para a construção de uma base de dados lexicais para a linguagem jurídica brasileira, inspira-se em trabalhos anteriores de criação de bases de dados de linguagem jurídica: JurWordNet (Sagri et al., 2003) e LOIS (Dini et al., 2005). Tanto a JurWordNet como a LOIS são bases de dados lexicais terminológicas baseadas na estrutura da WordNet (Miller, 1995).

[1] O projeto *Tecnologias Semânticas e Sistemas de Recuperação de Informação Jurídica* conta com o apoio da CAPES e do Conselho Nacional de Justiça (Edital CNJ – Acadêmico nº. 020/2010/CAPES/CNJ) e é coordenado pela professora Dra. Rove Chishman.

Anais da II Jornada de Descrição do Português, páginas 13–20, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

WordNets terminológicas como a JurWordNet têm como objetivo melhorar a recuperação de informação jurídica pela conexão de termos por meio da relação semântica de sinonímia. Como a sinonímia não é muito produtiva entre os termos especializados, as relações de sinonímia ligam os termos especializados usados pelos operadores do Direito às palavras não especializadas usadas pelos cidadãos. Um exemplo é a palavra *affitto* (aluguel), utilizada pelo não especialista, e o termo jurídico *locazioni di immobili* (locação de imóvel), preferido pelos especialistas. Essa organização por rede de sinônimos reduz a barreira entre a linguagem especializada utilizada pelo especialista e a linguagem não especializada utilizada pelo cidadão não especialista.

A LOIS (*Lexical Ontologies for Legal Information Sharing*) é uma extensão multilíngue da JurWordnet. Relações semânticas conectam termos jurídicos em diferentes línguas. A arquitetura da LOIS é baseada na EuroWordNet (Vossen, 1998). As *wordnets* de diferentes línguas são conectadas através de um índice de interlíngua.

Tanto a JurWordNet quanto a LOIS seguem o paradigma WordNet (Miller, 1995). Neste trabalho, propõe-se o uso do paradigma FrameNet (Fillmore et al., 2003) para a construção de recursos lexicais jurídicos. O objetivo é descreverem-se os participantes dos atos jurídicos, como, por exemplo, o *juiz* e o *réu*.

3. O FrameNet e a Metodologia de Criação de Frames

O FrameNet é uma base de dados lexicais que descreve o significado das palavras de acordo com os princípios da semântica de *Frames*. No FrameNet, os itens lexicais são concebidos como unidades lexicais. A unidade lexical é a junção de uma palavra a um significado. Assim, cada novo significado de uma palavra representará uma nova unidade lexical. Nos termos da FrameNet, cada nova unidade lexical evoca um frame semântico diferente.

Segundo Fillmore e Baker (2010), o método de análise lexical da FrameNet segue cinco passos:

Caracterização do *frame*. Caracteriza-se a situação descrita pelas unidades lexicais, por exemplo, a prisão de um suspeito, como no caso do *frame* `Arrest`.

Descrição e nomeação dos elementos de *frame*. Após a caracterização de um *frame* específico, identificam-se todos os possíveis participantes da situação e criam-se nomes para cada participante, por exemplo, AUTORIDADES, SUSPEITO, OFENSA e ACUSAÇÕES.

Seleção das unidades lexicais. Após a descrição da situação e da identificação e nomeação dos elementos de *frame*, as unidades lexicais e expressões evocadoras do *frame* são identificadas: *apprehend.v, apprehension.n, arrest.n, arrest.v, book.v, bust.n, bust.v, collar.v, cop.v, nab.v, summons.v*

Anotação de sentenças. Sentenças selecionadas para exemplificar os padrões sintáticos e semânticos de cada unidade lexical são anotadas com elementos de *frame*.

Geração automática de entradas lexicais. Os exemplos anotados para cada unidade lexical são transformados automaticamente em uma entrada lexical contendo a definição da unidade lexical, as realizações sintáticas de cada elemento de *frame* e os padrões valências.

Conforme Fillmore e Baker (2010), os elementos de *frame* representam propriedades ou entidades que podem ou devem estar presentes em qualquer instância de um *frame*. A FrameNet diferencia os elementos de *frame* em **centrais**, **periféricos** e **extratemáticos**. Segundo Fillmore e Baker (2010), a distinção entre esses tipos nem sempre é clara. De uma forma geral, elementos de *frame* que são obrigatoriamente expressos são centrais. Os elementos de *frame* periféricos expressam em geral funções de adjuntos, expressando tempo, lugar ou modo. A diferença entre elementos centrais e periféricos depende da necessidade de complementação da unidade lexical. Os elementos de *frame* extratemáticos introduzem informação referente a outro *frame*, como o propósito motivador de algum evento ou ação. Os elementos de *frame* periféricos e extratemáticos são agrupados na FrameNet sob a denominação de elementos **não-centrais**.

4. O Desenvolvimento do *Frame* `Processo_penal`

Esta seção apresenta a metodologia utilizada para a descrição do *frame* `Processo_penal`, a organização dos *subframes* que compõem esse *frame* e as unidades lexicais evocadoras de *frames*. O foco da atenção nesta seção será o agrupamento das unidades lexicais segundo o *frame* por elas evocado, a definição dos *frames* jurídicos, o reconhecimento dos elementos de *frame* centrais de cada *frame* jurídico e anotação de exemplos de sentenças.

4.1. Metodologia

Este trabalho representa a primeira etapa de um projeto lexicográfico que objetiva a aplicação do paradigma FrameNet para a criação de um recurso lexical da linguagem jurídica brasileira. Nessa fase do projeto, utilizou-se a metodologia de expansão (Vossen, 1999). A metodologia de expansão é utilizada por projetos como o Spanish FrameNet (Subirats, 2009). A metodologia de expansão, aplicada ao desenvolvimento de recursos lexicais baseados em *frames*, consiste em utilizar os mesmos *frames* semânticos da FrameNet para o desenvolvimento da base de dados lexicais da nova língua, substituindo as unidades lexicais do inglês pelos seus equivalentes de tradução na outra língua e adaptando os *frames* semânticos quando necessário.

Assim, a metodologia deste trabalho seguiu quatro passos:

(i) Primeiramente, identificaram-se equivalentes de tradução em português para as unidades lexicais evocadoras de *frame* em inglês. Nessa etapa, utilizou-se o Dicionário Jurídico Bilíngue Noronha (Goyos Junior, 1992) para não contar apenas com a intuição dos pesquisadores sobre as línguas em comparação.

(ii) Em segundo lugar, identificou-se o evento jurídico evocado por cada uma das unidades lexicais em português. Essa etapa envolveu a análise do conhecimento jurídico vinculado pela unidade lexical em português, que nem sempre era compatível com o contexto jurídico evocado pelas unidades lexicais em inglês.

(iii) Em terceiro lugar, criaram-se os *subframes* que compõem o *frame* `Processo_penal`. Dois passos metodológicos distintos foram seguidos nessa etapa. No caso de o evento jurídico evocado pela unidade lexical em português ser correspondente ao evento jurídico evocado pela unidade lexical em inglês, adotou-se a

metodologia de expansão. O *frame* semântico foi mantido o mesmo, apenas substituindo-se as unidades lexicais do inglês pelas unidades lexicais em português e anotando-se exemplos de sentenças em português para cada *frame* semântico. No caso de o evento jurídico evocado pela unidade lexical em português não ser correspondente ao evento evocado pela unidade lexical em inglês, adotaram-se os passos metodológicos utilizados pela FrameNet (Fillmore e Baker 2010), descritos na seção 3.

(iv) O último passo envolveu a seleção de exemplos de sentenças e anotação manual com elementos de *frame*. Para esta etapa, utilizou-se o corpus NILC. A anotação de sentenças apresentada aqui não é extensiva, e sim seletiva. Foram selecionadas apenas aquelas sentenças que melhor exemplificam a anotação de sentenças com os elementos de *frame*.

A metodologia apresentada aqui apresenta diversas limitações. Primeiramente, a metodologia de expansão não se presta bem para o domínio jurídico. Os *frames* jurídicos tendem a apresentar mais incompatibilidades que compatibilidades, pois são dois sistemas jurídicos que estão em contraste, o americano e o brasileiro. A metodologia de expansão também limita a criação de novos *frames* na base de dados lexicais da nova língua, uma vez que se parte dos *frames* e das unidades lexicais já descritos pela FrameNet. Outra limitação da metodologia utilizada diz respeito ao *corpus*. O *corpus* NILC, apesar de ser representativo para a pesquisa realizada, não é um *corpus* especializado da linguagem jurídica. Assim, em futuras etapas do projeto lexicográfico descrito neste artigo, serão necessárias a análise de documentos jurídicos e a compilação de um *corpus* especializado. Através dos documentos jurídicos é possível se identificar as fraseologias típicas do Direito, o que não é possível de se identificar com um dicionário jurídico bilíngue.

4.1. O *Frame* `Processo_penal`

O *frame* `Processo_penal` está dividido cinco *subframes*: `Prisão`, `Denúncia`, `Audiência_de_instrução`, `Pronúnica`, e `Julgamento`. O *frame* `Julgamento` está dividido em três *subframes* que representam os passos de um julgamento pelo procedimento do Tribunal do Júri: `Instrução`, `Veredito` e `Sentença`. O *frame* `Julgamento` está em relação de perspectiva com o *frame* `Julgar_acusado`. O *frame* `Julgar_acusado` especifica o evento legal geral representado pelo *frame* `Julgamento`. Enquanto o *frame* `Julgamento` representa os principais passos de um Tribunal do Júri, o *frame* `Julgar_acusado` representa o evento de julgar um réu. Ambos os *frames* descrevem o mesmo evento, mas de pontos de vista diferentes. Como representam pontos de vista diferentes, as unidades lexicais evocadoras de *frame* são diferentes para cada *frame*. A figura 1 apresenta a organização do *frame* `Processo_penal`.

O *frame* `Prisão` descreve um ato em que AUTORIDADES privam um SUSPEITO da liberdade por ACUSAÇÕES contra ele. Os elementos de *frame* centrais a este *frame* são: AUTORIDADES, SUSPEITO, OFENSA, ACUSAÇÕES. As unidades lexicais evocadoras deste *frame* são: *prender*, *prisão*, *fichar*, *deter*, *capturar*, *em cana*. As sentenças anotadas a seguir exemplificam as ocorrências de cada unidade lexical e os elementos de *frame* que ocorrem com cada uma delas.

(1) [França $_{\text{AUTORIDADES}}$] **prende** [95 suspeitos $_{\text{SUSPEITO}}$] [de colaboração com terror argelino. $_{\text{OFENSA}}$]

O *frame* `Denúncia` representa um evento jurídico em que o promotor, AUTORIDADE_DE_ACUSAÇÃO, denuncia o ACUSADO por ACUSAÇÕES contra ele. Os elementos de *frame* centrais deste *frame* são ACUSADO, AUTORIDADE_DE_ACUSAÇÃO e ACUSAÇÕES. As unidades lexicais evocadoras de *frame* são *acusar*, *acusação*, *denunciar* e *denúncia*.

(2) A partir desses documentos, [o Ministério Público $_{\text{AUTORIDADE\_DE\_ACUSAÇÃO}}$] **denunciou** [os bicheiros $_{\text{ACUSADO}}$] novamente e ficou comprovado que eles mantinham suas atividades mesmo de trás das grades

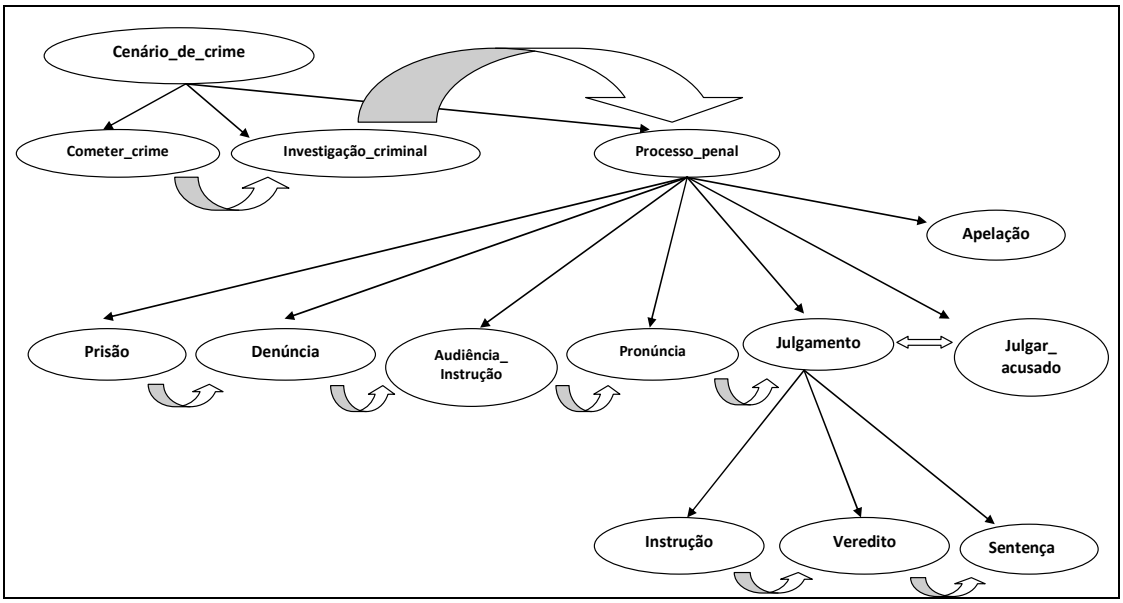

Figura 1. O *frame* `Processo_penal`

O *frame* `Audiência_de_instrução` representa a audiência preliminar em que um JUIZ ouve o ACUSADO de um crime com o objetivo de decidir sobre o prosseguimento do processo. São elementos de *frame* centrais JUIZ, ACUSADO, TESTEMUNHAS e ACUSAÇÕES. As unidades lexicais evocadoras deste *frame* são *interrogar* e *depor*.

(3) [Principal testemunha da chacina $_{\text{TESTEMUNHA}}$] **depõe** no II Tribunal do Júri reafirma denúncias e diz que Emanuel mentiu ao inocentar Côrtes.

O *frame* `Pronúncia` descreve o evento jurídico em que o JUIZ,presidente do Tribunal do Júri, faz a apreciação preliminar das provas, em sentença, para submeter o

RÉU posteriormente a julgamento. Os elementos de *frame* centrais são JUIZ e RÉU. As unidades lexicais evocadoras do *frame* `Pronúncia` são *pronúncia* e *pronunciar*.

(4) [O juiz JUIZ] deve **pronunciar** [o réu RÉU] (TJSP, RCrim 71.325, RT 648 / 275).

O *frame* `Julgamento` descreve o evento jurídico em que um JUIZ, presidente do Tribunal do Júri, e um corpo de jurados, o JÚRI, devem decidir sobre a culpa ou inocência de um RÉU. A PROMOTORIA tenta provar a culpa e a DEFESA tenta provar a inocência do RÉU. Os elementos de *frame* centrais a este *frame* são JUIZ, JÚRI, PROCURADOR, RÉU, DEFESA, ACUSAÇÕES, TRIBUNAL e AÇÃO. As unidades lexicais evocadoras de *frame* são *julgamento*, *processo* e *ação penal*.

(5) O recurso pode provocar, em 95, um novo **julgamento** [dos acusados RÉU] [pelos desembargadores do Tribunal de Justiça do Estado. JUIZ]

O *frame* `Instrução` descreve a fase de instrução em plenário, em que o JUIZ interroga o RÉU e as TESTEMUNHAS da defesa e da acusação depõem. Os elementos de *frame* são RÉU, TESTEMUNHA e JUIZ. As unidades evocadoras de *frame* são *depor*, *interrogar* e *testemunhar*.

(6) [Principal testemunha da chacina TESTEMUNHA] **depõe** no II Tribunal do Júri reafirma denúncias e diz que Emanuel mentiu ao inocentar Côrtes.

O *frame* `Veredito` descreve a fase de votação em que o JÚRI decide sobre a culpa ou inocência do RÉU. Os elementos de *frame* são JUIZ, DECISÃO, ACUSAÇÕES. As unidades evocadoras de *frame* são *decidir*, *considerar*, *absolver*, *inocentar*, *condenar*, *condenação* e *veredito*.

(7) Quanto a essa acusação, o [júri JUIZ] **decidiu** [absolver DECISÃO] [o réu Alexandre Cardoso, o Topeira, RÉU] e [condenar DECISÃO] [Sandro Baggi e André Rodrigues da Silva, o Gargamel. RÉU]

O *frame* `Sentença` descreve a fase do *frame* julgamento em que o JUIZ profere a sentença ao RÉU. Os elementos de *frame* centrais são CONDENADO, TRIBUNAL, OFENSA e PENA. A unidade lexical evocadora deste *frame* é *condenar*.

(8) [Ubirajara CONDENADO] foi **condenado** [a 19 anos PENA] [para cada homicídio OFENSA] e [a 12 anos PENA] [pela tentativa de homicídio de Orlando OFENSA]

O *frame* `Julgar_acusado` descreve o julgamento de um RÉU, que é acusado de um crime. Um corpo de jurados, o JÚRI, é responsável por avaliar as ACUSAÇÕES e decidir se o RÉU é culpado pelo crime, a OFENSA. Os elementos de *frame* centrais são JUIZ, JÚRI, RÉU, OFENSA e ACUSAÇÕES. A unidade lexical evocadora deste *frame* é *julgar*.

(9) Para o governador, o fato de [os acusados $_{\text{RÉU}}$] serem **julgados** [por um júri popular $_{\text{JÚRI}}$] é muito positivo.

Apesar de se utilizar a metodologia de expansão para a criação do *frame* `Processo_penal`, há muitas diferenças conceituais entre os frames da FrameNet e os *frames* jurídicos brasileiros. Essas diferenças levam a certas conclusões sobre a continuidade do projeto lexicográfico descrito neste artigo.

5. Direções Futuras

Este artigo descreveu o primeiro estágio de criação de um recurso lexical baseado em *frames* para a linguagem jurídica brasileira. Iniciou-se o desenvolvimento desse recurso lexical através do estudo do *frame* `Criminal_process` e sua expansão para o sistema jurídico brasileiro, resultando no *frame* `Processo_penal`. Como uma primeira conclusão, é possível dizer que *frames* complexos são difíceis de serem expandidos para outras línguas, por causa das diferenças entre os sistemas jurídicos e as leis de cada país. Agora é necessário testar *frames* que representam nódulos menores, ou seja, *frames* menos complexos, como `Law` e `Legality`.

O desenvolvimento de um recurso lexical baseado em *frames* da linguagem jurídica brasileira é apenas parte de um projeto maior que objetiva a criação de recursos lexicais e bases de dados de conhecimento, como léxicos e ontologias, para uso em recuperação de informação jurídica. O projeto lexicográfico apresentado aqui tem dois objetivos. O primeiro é o desenvolvimento de uma base de dados lexical de grande porte da linguagem jurídica brasileira. O segundo é o uso das etiquetas semânticas desenvolvidas no âmbito desse projeto para anotação semântica de um *corpus* jurídico para ser utilizado como *corpus* de treinamento em processamento de linguagem natural.

A pressuposição aqui é que as etiquetas semânticas da FrameNet não são completamente aplicáveis para outras línguas. Considerando-se que o Direito é uma criação socialmente orientada, o evento jurídico descrito por alguns *frames* da FrameNet podem não ser equivalentes em diferentes línguas/sistemas jurídicos. Por essa razão, decidiu-se expandir os *frames* quando possível, adaptar aqueles *frames* que possuíssem alguma similaridade e criar novos *frames* sempre que necessário. Diferentemente das bases de dados baseadas no paradigma WordNet, as relações semânticas entre palavras não são o foco de uma base de dados baseada em *frames*. Portanto, uma base de dados baseada em *frames* tem diferentes aplicações em processamento de linguagem natural e recuperação de informação.

As etiquetas semânticas poderiam ser utilizadas em uma serie de aplicações em processamento de linguagem natural, como sumarização automática, recuperação de

informação jurídica e extração de informação jurídica. A anotação automática de decisões judiciais pode permitir a sumarização automática e a geração automática de ementas. Essas ementas são resumos do teor das decisões dos tribunais e permitem aos advogados conhecerem o tema das decisões sem ter que ler o documento na íntegra. Outra possibilidade de uso dos *frames* semânticos para recuperação de informação jurídica é a anotação dos participantes nas decisões judiciais, como o *réu*, o *juiz*, o *promotor*, o *advogado*, e o resultado dos eventos legais, como o *veredito* e a *pena*.

O trabalho de descrição da linguagem jurídica brasileira está apenas no início. Há ainda importantes procedimentos a serem feitos. Primeiro, expandir o número de *frames* para melhor representar o universo da linguagem jurídica brasileira. Segundo, compilar um *corpus* jurídico para ser utilizado em anotação semântica. Esse corpus poderia ser utilizado como fonte de exemplos para a base de dados lexicais ou como corpus de treino para aplicações automáticas. Terceiro, programar uma interface amigável para disponibilizar a base de dados gratuitamente. O estudo do *frame* `Processo_penal` representa apenas o primeiro estágio deste projeto lexicográfico que está focado na inovação tecnológica das bases de dados dos tribunais brasileiros.

Referências

Dini, L. et al. (2005). Cross-lingual legal information retrieval using a WordNet architecture. In: Proceedings of the 10th International Conference on Artificial Intelligence and Law, Bologna. ACM Press: New York, p.163-167.

Fillmore, C.J.; Baker, C. (2010). A Frames Approach to Semantic Analysis. *In*: *The Oxford Handbook of Linguistic Analysis*. Oxford: OUP, p. 313-339.

Fillmore, C. J.; Johnson, C. R.; Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*. Vol.16, Nº.3, p.235-250.

Goyos Júnior, D. N. (1992). *Noronha's Legal Dictionary – Noronha Dicionário Jurídico:* English-Portuguese, Portuguese-English – Inglês-Português, Português-Inglês. 1.ed. São Paulo: Observador Legal.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*. New York: ACM Press. Vol.38, Nº.11, p.39-41.

Sagri, M. T.; Tiscornia, D.; Bertagna, F. (2003). Jur-WorNet. In: Sojka, P. et al. (Eds.) Second International Wordnet Conference. Brno: Masaryk University, p.305-310.

Subirats, C. (2009). Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In: Boas, H. C. (Ed.) *Multilingual FrameNets in computational lexicography:* Methods and applications. Berlin/New York: Mouton de Gruyter, p.136-162.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*. Vol.32, Nº.2-3, p.73-89.

Vossen, P. (1999). EuroWordNet General Document. Version 3. Technical report, University of Amsterdam.