

FrameCorp: Reflexões Preliminares sobre a Anotação do Corpus SUMMIT

Rove Chishman¹, João Gabriel Padilha², Carla ten Kathen³

^{1,2,3}Programa de Pós-Graduação em Linguística Aplicada - Universidade do Vale do Rio dos Sinos (UNISINOS)

rove@unisinis.br, joaogabriels1@hotmail.com,
carlatenkathen@yahoo.com.br

Resumo. *O objetivo desta comunicação é apresentar os resultados preliminares dos estudos empreendidos no âmbito do projeto FrameCorp¹, cujo objetivo geral é a investigação semântico-computacional do léxico do Português do Brasil a partir da utilização de corpus eletrônico. É nosso propósito verificar em que medida a teoria dos frames de Fillmore (1982, 1985) se presta à descrição semântica de verbos. Justifica-se esta proposta ressaltando a escassez de investigações voltadas para a construção de recursos computacionais com base em estudos semânticos. Em termos teóricos, pretende-se não apenas validar a Semântica dos Frames, verificando sua capacidade descritiva e explanatória, como também contribuir para seu avanço, haja vista que ocupamo-nos de dados lingüísticos do Português. Em termos aplicados, interessa a esta investigação a construção de corpus com anotação semântica a partir de evidência empírica.*

Abstract. *The objective of this communication is to present the preliminary results of the studies undertaken in the ambit of the FrameCorp project, which general aim is the semantic-computational investigation of the Brazilian Portuguese lexicon based on electronic corpus. It is our purpose to verify to some extent the Charles Fillmore's frame semantics theory (1982, 1985) can be applied to the semantic description of the verbs. Such proposal is justified by emphasizing the lack of investigations concerning the construction of computational resources related to semantic studies. In theoretical terms, it is intended not only to validate the Frame Semantics, verifying its descriptive and explanatory capability, but also contribute to its advancement, since We are dealing with Portuguese linguistic data. In applied terms, it concerns to this investigation the construction of a semantic-annotated corpus from empirical evidence.*

Palavras-chave: Anotação Semântica, Corpus Anotado, Frames

¹ O Projeto FrameCorp conta com suporte do CNPq através do Edital 03/2008. João Gabriel Padilha realizou este estudo com bolsa de IC da FAPERGS e Carla ten Kathen, com bolsa PIBIC/CNPq.

1. Introdução

Neste trabalho, concentramo-nos em apresentar os resultados concernentes à tarefa de descrição semântica tendo em vista o processo de anotação de corpus. O foco nesta aplicação justifica-se pela necessidade de se desenvolverem recursos para processamento computacional do Português, ainda pouco expressivo em se tratando principalmente de conteúdo semântico. Um *corpus* com anotação semântica, neste sentido, vem contribuir para o desenvolvimento de diferentes aplicações computacionais para o Português, tais como os *parsers* semânticos e os sistemas de recuperação de informação.

Em termos aplicados, tomamos como ponto de partida um *corpus* formado por 50 textos do domínio *ciência*, extraídos do *corpus* PLN-BR GOLD, ocupando-nos aqui em apresentar os resultados obtidos através da anotação semântica realizada por meio da ferramenta SALTO. Em termos teóricos, seguimos a Semântica de *Frames* (Fillmore et al., 2003), haja vista o potencial descritivo da teoria em representar a semântica dos diferentes domínios.

Para tratarmos das questões apontadas anteriormente, o presente artigo está organizado em quatro seções. Na Introdução, apresentamos o tema deste estudo, assim como a estrutura a ser seguida. Na seção 2, apresentamos uma reflexão sobre anotação semântica de *corpora*. Na seção 3, discutimos a abordagem teórica adotada neste trabalho, a Semântica de *Frames*. Já na seção 4, apresentamos os resultados da anotação semântica a partir de evidência empírica. Por fim, apresentamos as considerações finais.

2. Anotação Semântica

A anotação semântica é uma área de investigação promissora, mas ainda carente, principalmente em se tratando de recursos para a Língua Portuguesa. Por um lado carente porque iniciativas de construção de corpus com anotação semântica são raras. Uma das razões pode estar relacionada à complexidade e subjetividade desse tipo de conhecimento lingüístico e a conseqüente dificuldade de realizar essa tarefa de maneira automática. Por outro lado promissora por estar associada ao conceito de *web* semântica, cuja implementação depende da estruturação do conteúdo das páginas da web, de maneira a garantir o acesso tanto por usuários como por máquinas.

Nesse novo contexto, a *web* será capaz de representar associações entre coisas que, em princípio, poderiam não estar relacionadas, facilitando a recuperação e extração de informação. Segundo Berners-Lee et al (2001), o idealizador desse novo conceito de web, através do acesso a coleções estruturadas de informações (dados e metadados) e de regras de inferência, computadores e usuários poderão trabalhar de forma cooperativa através da composição de um grande número de pequenos componentes ontológicos que apontam entre si. Anotação semântica, nesse novo cenário, remete a um objeto acrescido a um documento e a atividade que produz este objeto.

Em obra clássica sobre o tema, Garside et al. (1997) ressaltam que, nos últimos trinta anos, o termo *corpus* vem sendo associado a um corpo de material lingüístico que existe em formato eletrônico e que pode ser processado pelo computador para atender a demandas relacionadas tanto à pesquisa lingüística como à engenharia da linguagem. A anotação de *corpus* é apontada pelos autores como um dos fatores que agregam valor a

um *corpus* à medida que este passa a ser enriquecido com informação lingüística ou interpretativa que pode ser útil para a extração de informação. Padó et al. (2006), por sua vez, referindo-se aos *corpora* com múltiplas camadas de anotação lingüística, destacam a possibilidade de investigar empiricamente as interações entre os diferentes níveis de análise lingüística. Os autores lembram também da utilidade de tais *corpora* para a aquisição de modelos estatísticos para anotação automática.

Leech (1991), referindo-se ao uso de *corpora* para a pesquisa lingüística, enfatiza o seu valor como fonte de evidência empírica sobre a linguagem, o que pode valer não apenas como elemento complementar aos estudos teóricos, como também como elemento que vem suplantando outras fontes e a própria perspectiva introspectiva. Esta linha de pensamento é seguida à risca pelos pesquisadores vinculados ao UCREL (Universidade de Lancaster), centro de excelência e pioneirismo em diversas áreas que se beneficiam da metodologia baseada em *corpus*. É importante mencionar a *expertise* do grupo em desenvolver recursos ligados à anotação de *corpora*. O grupo tem propostas de anotação que envolvem desde a anotação gramatical, com as chamadas POS tags (ou *part of speech tags*), até a proposição de esquemas de anotação anafórica e semântica, com a identificação de campos semânticos.

Convém lembrar que as etiquetas gramaticais estão entre os casos mais típicos de anotação. Merecem destaque também os *corpora* com anotação sintática, os chamados *treebanks*, cuja utilização tem possibilitado a criação e o aperfeiçoamento de vários recursos lingüísticos, como léxicos e gramáticas, além do treinamento automático de *parsers*. O *corpus* do projeto Penn Treebank, desenvolvido pelo grupo da Pennsylvania (Marcus et al., 1993), é outra iniciativa importante ligada a esse nível de anotação, contribuindo para o desenvolvimento de ferramentas computacionais de extração de informações de língua inglesa.

Para o Português, há diversas propostas de criação de *corpus*. Em se tratando de *corpora* com anotação, contudo, os exemplos de que se tem conhecimento dizem respeito à anotação gramatical ou sintática. É o caso dos seguintes projetos: (i) AC/DC (Acesso Corpora/Disponibilização Corpora) (Santos & Bick, 2000), (ii) Corpus do NILC (Núcleo Interinstitucional de Lingüística Computacional) e (iii) Lácio-Web (Aluísio et al., 2003).

A dificuldade de automatização do trabalho de anotação semântica dificulta a sua expansão, porém a inclusão de um nível semântico a *corpora* agregaria valor à anotação gramatical e sintática. Gildea & Juravsky (2002) chamam a atenção para o fato de as análises sintáticas produzidas pelos *parsers* construídos a partir de tais recursos não contarem com a representação do significado das sentenças. Os autores acreditam que a inclusão de um nível de representação semântica é importante para uma série de aplicações, tais como sistemas de extração de informação e de pergunta e resposta, assim como sistemas de tradução automática. Entre os fenômenos semânticos abarcados por tais sistemas, eles destacam a quantificação, a correferência anafórica e os conceitos de aspecto e modalidade.

Uma das poucas iniciativas de anotação semântica de *corpus* é a do projeto Proposition Bank (PropBank), também da Pennsylvania, que se propõe a adicionar informação sobre predicados e argumentos ou rótulos de papel semântico às estruturas sintáticas do Penn Treebank. Outra proposta de peso é o *corpus* SemCor (Miller et al,

1993), pela Princeton University, que é composto por 352 textos, contendo 41.497 verbos com anotação semântica. Assim, devido à importância do tema e a carência de recursos para o Português, justifica-se o presente experimento e o seu futuro desdobramento em um corpus com anotação semântica.

3. A Semântica de Frames

A Semântica dos *Frames* (Fillmore, 1977, 1982) é uma tentativa de explicitar o significado das palavras evitando as formas tradicionais de estudo semântico realizados até então, como as listas de traços, de orientação estruturalista, e as condições de verdade, de orientação formal. O termo *frame* é um empréstimo da psicologia cognitiva. Assim, destacamos, desde já, a forte orientação cognitiva desta teoria.

Para compreendermos a Semântica dos *Frames* é importante distinguirmos dois conceitos inter-relacionados: *cena* e *frame* (Fillmore, 1977). A *cena* trata-se de uma experiência vivenciada, moldando a cognição do ser humano e lhe propiciando conhecimento de mundo. O *frame* é a contraparte lingüística desse conhecimento, ou seja, as palavras relacionadas a uma determinada *cena*. Assim sendo, podemos afirmar que o conhecimento de um *frame* é um conhecimento “experencial”, visto que, o conhecimento de um *frame* pressupõe o conhecimento completo de evento que esse *frame* representa (Fillmore, 1977).

Como podemos ver, o conceito de *contexto* desempenha um papel fundamental para a descrição de um *frame*. Uma palavra pode ter significados diferentes conforme os contextos em que ela é utilizada, ou seja, uma palavra pode integrar diferentes *frames*. Considerando isso, os *frames* exercem um papel fundamental nos processos de compreensão e comunicação. Segundo Fillmore (1977), cenas e *frames* se associam na mente das pessoas, um ativando o outro. Assim, uma cena traz à mente um *frame* e um *frame* pode trazer à mente outros *frames* relacionados a esse. Isso explica a sua importância para a comunicação, bem como para a compreensão de textos.

Tomando por base a Semântica dos *Frames*, em 1997 teve início o projeto FrameNet (Baker et al, 1998). O projeto FrameNet aplica a Semântica dos *Frames* à descrição lexicográfica. Um dos objetivos do projeto é a descrição da valência sintático-semântica das palavras.

Segundo Fillmore e Petruck (2003), *frame* é uma representação esquemática de uma situação envolvendo vários participantes, os chamados **elementos frame**. Os **elementos frame** são os papéis semânticos dos participantes de um evento, sendo específicos a cada *frame* e por isso também chamados de papéis situacionais. Enquanto o *frame* é uma categoria conceitual, os **elementos frame** são instanciados no texto através das **unidades lexicais**. Uma **unidade lexical** é a combinação entre um lema e um *frame*, ou seja, é uma palavra considerada segundo um *frame* específico. Nesses termos, as palavras no FrameNet não são polissêmicas, pois cada novo sentido está associado a um *frame* diverso, caracterizando, assim, uma nova unidade lexical.

4. Análise e Discussão dos Dados

A metodologia empregada neste trabalho segue os princípios propostos pelo grupo de Berkeley. Uma descrição semântica de *frames* implica associar uma unidade lexical a

um determinado *frame*, que, por sua vez, é constituído por vários papéis conceituais, ou *elementos frames* na metalinguagem do FrameNet. Consideremos o *frame Statement* e seus componentes conceituais a partir de uma das sentenças de nosso estudo:

O presidente disse que a ministra Marina Silva criou um grupo de trabalho para analisar as propostas.

O *frame statement* descreve uma situação em que um Falante (*Speaker*) fala sobre um Tópico (*Topic*) ou envia uma Mensagem (*Message*) através de um Meio (*Medium*). Além da descrição do *frame*, a base de dados traz os elementos evocadores de *frames* e os elementos *frames*. Os elementos evocadores de *frames* são, em essência, palavras predicadoras, como verbos, nomes e adjetivos. Também chamados de palavras *target*, os elementos evocadores de *frame* são o ponto de partida para a anotação. Na sentença acima, o verbo *dizer* funciona como elemento evocador, e *o presidente* e toda a oração completiva (*que a ministra Marina Silva criou um grupo de trabalho para analisar as propostas*) são elementos *frame*, mais especificamente *Speaker* e *Message*. Podemos dizer que o trabalho do anotador consiste em: (i) identificar o elemento evocador da sentença, (ii) identificar um equivalente de tradução em Inglês, (iii) localizar o *frame* associado a essa unidade lexical e (iv) descrever a sentença, selecionando os elementos *frames* relacionados.

É importante considerar que tal processo se complexifica nos casos de anotação de sentenças em línguas que não sejam a língua inglesa, como é o caso de nosso projeto e dos demais projetos bilíngües – German FrameNet, Spanish FrameNet e Japanese FrameNet. Isso acontece porque as descrições oferecidas pela base de dados para consulta são propostas para o Inglês. A grande questão a ser respondida por todas estas iniciativas é se os *frames* propostos para o Inglês são válidos para as descrições semânticas de outras línguas.

O trabalho de anotação pode seguir duas orientações, dependendo da forma como as sentenças são escolhidas: a anotação lexicográfica e a anotação *running-text*. A primeira orientação parte de uma determinada unidade lexical e busca sentenças contendo a unidade a ser analisada. Trata-se do método utilizado pelos projetos que objetivam construir léxicos em que as palavras aparecem associadas a *frames*. Obtém-se um *corpus* de sentenças representativas do *frame* em análise. A segunda orientação metodológica, ou anotação *running-text*, prevê que o anotador respeite a seqüência do texto, identificando a cada nova sentença um elemento evocador de *frame*. Como consequência deste procedimento, selecionam-se novos *frames* à medida que surgem unidades lexicais distintas.

Feita esta breve exposição, podemos sistematizar as principais etapas do processo de anotação de *frames* do Summ-it:

Segmentação dos textos em sentenças: Esta etapa se justifica pela necessidade de identificar previamente os elementos evocadores de *frames* ou elementos *target*. Este conjunto corresponde às sentenças que contêm verbos plenos ou típicos predicadores. *Verbos plenos* são aqueles que semanticamente têm significação lexical e sintaticamente ocupam o núcleo do predicado num sintagma verbal (Borba, 2000). Excluíram-se os verbos do tipo não-pleno, que são os copulativos, os modalizadores, os verbos-suporte e os verbos auxiliares de tempo e modo.

Extração das palavras mais frequentes: Através da ferramenta X-Tractor (Gasperin, 2003), foi gerada uma *wordlist*, que nos permitiu perceber que um número considerável de verbos pertencia ao *frame* Statement. Das 774 sentenças, 135 contêm verbos relacionados ao *frame* Statement. Este resultado nos levou a iniciar a anotação por estes verbos, caracterizando o modo lexicográfico, conforme esclarecemos acima.

Anotação semântica: Tendo iniciado o trabalho pelos verbos do *frame* Statement – *dizer, afirmar, alegar, anunciar, escrever e explicar*, – partimos para a metodologia *running-text*. Este procedimento, necessário haja vista as próprias exigências da ferramenta SALTO, requer que o anotador analise uma frase de cada vez, deparando-se com novos *frames* à medida que se diversificam os evocadores de *frames*. Esta segunda fase envolveu uma equipe de 4 anotadores, que editaram suas anotações diretamente na ferramenta SALTO.

Consulta a dicionários e thesaurus: O processo de anotação *running-text*, por não permitir a escolha de itens lexicais a serem anotados, nos obrigou a tratar de uma grande variedade de palavras. Assim, a consulta a dicionários e thesaurus, principalmente em Inglês, nos ajudou a clarificar o significado de palavras e a encontrar equivalentes de tradução para os itens lexicais em anotação.

Confronto entre as anotações: Um dos recursos da ferramenta SALTO é a possibilidade de dois anotadores anotarem o mesmo *corpus* e, depois, confrontar as anotações e corrigir os desencontros, conforme abordaremos na seção seguinte. Assim, após pares de anotadores anotarem o *corpus*, a anotação das duplas foi confrontada no SALTO e corrigida. Na primeira anotação, das 512 sentenças anotadas, 91 apresentaram discordância de anotação entre anotadores e 89 não foram anotadas por pelo menos um dos anotadores (por provocarem dúvidas). Esses dados serão discutidos na seção seguinte.

Após a primeira anotação, o confronto e o ajuste das sentenças com informações incompatíveis, chegamos à versão final do *corpus*. De um total de 774 sentenças do *corpus*, 512 foram anotadas (66,14% do *corpus*). 142 *frames* foram utilizados na anotação, sendo que *frames statement* e *evidence* foram os mais produtivos, totalizando 135 e 32 sentenças, respectivamente. O paradigma FrameNet apresenta uma vantagem sobre os tradicionais padrões de anotação de papéis semânticos: a possibilidade de descrição de uma estrutura conceitual e uma maior especificidade de papéis semânticos. Esses papéis semânticos, chamados de elementos *frame*, por estarem sempre associados a um *frame*, possibilitam uma especificidade de descrição maior que os tradicionais papéis semânticos (agente, paciente etc). A sentença abaixo ilustra a análise de uma sentença utilizando as etiquetas semânticas do FrameNet. O verbo *mostrar* é o elemento evocador do *frame evidence*, os conchetes demarcam a estrutura sintática da sentença e os respectivos elementos *frame*.

[**evidence**]

[*Mas novos cálculos support*] **mostram** [*que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio. Proposition*]

A anotação do *corpus* teve que gerenciar problemas de sutilezas semânticas entre *frames*, como no caso dos *frames remembering_information* e *remembering_experience*,

ambos atribuídos ao verbo *esquecer*. *Frames* como *activity_start* e *process_start*, apesar de semelhantes, o primeiro descreve uma *atividade* iniciada por um *agente*, enquanto o segundo descreve o início de um *evento*. A estrutura sintático-semântica da sentença ajuda na diferenciação de ambos *frames*, mas requer um conhecimento gramatical mais aprofundado por parte dos anotadores.

Também a polissemia gerou divergências na anotação. Como uma mesma palavra pode estar relacionada a diferentes *frames*, e como nem sempre é fácil identificar o *frame* mais adequado, ocorrem diferenças na anotação de diferentes anotadores. Não há uma escolha única e segura em se tratando de anotação de *frames*, pois não é fácil identificar e separar os diferentes sentidos de uma palavra polissêmica. Esse é o caso, por exemplo, do verbo *vir*, na sentença abaixo, em que um anotador identificou como *arriving* e outro como *origin*. Na verdade, ambas as leituras estão corretas, pois a origem de algo indica a sua proveniência, ou seja, o lugar de onde essa entidade veio. Por fim, no confronto optou-se por *origin*.

E que, apesar da sua longa permanência nas Américas, a grande maioria dos cães do continente veio mesmo da Europa depois da "descoberta" de Cristóvão Colombo, em 1492.

Outra dificuldade de anotação, ao empregar o paradigma FrameNet, foi o nível de especificidade de algumas Unidades Lexicais. Em casos em que os anotadores não encontraram um *frame* mais específico para anotar a sentença, a orientação foi utilizar um *frame* mais genérico, para não deixar a sentença sem anotação. Esse é o caso de *frames* como *causation* e *intentionally_act*.

A larva [induz causation] quimicamente a aranha a modificar o formato da própria teia para que o casulo da vespa possa se desenvolver.

O trabalho [foi coordenado intentionally_act] pelo geocientista Charles Vörösmarty, da Universidade de New Hampshire, nos Estados Unidos.

Demais divergências de anotação estavam relacionadas à imprecisão lexical. Tais casos suscitam dúvidas nos anotadores por não serem facilmente compreendidos. Esse é o caso do verbo *envolver*, na sentença abaixo:

*Os estudos **envolveram** análise do DNA de uma estrutura das células, as mitocôndrias, que é transmitido apenas pelas mães.*

Nessa sentença o verbo *envolver* está sendo aplicado como sinônimo de *incluir*. Foi somente através de uma paráfrase que os anotadores conseguiram chegar ao *frame inclusion*, e assim anotar a sentença. O problema da imprecisão lexical na anotação semântica é intensificada pela falta de paralelismo entre as línguas. Em casos em que não se conseguiu chegar a um equivalente de tradução razoável, como no caso de *creditar*, a sentença ficou sem anotação.

*Ele **credita** à sua colega Hannah Faye Chua a idéia de testar de forma visual manner um dado já verificado verbalmente.*

Uma das possibilidades da ferramenta SALTO é a criação de novos *frames*. Neste primeiro trabalho, tratamos de conhecer a base de dados de *frames* do projeto FrameNet; dessa forma, não propusemos novos *frames*, atividade que certamente

teremos que enfrentar em um próximo trabalho. Com base nas observações feitas aqui, chegamos às conclusões expostas na seção seguinte.

5. Considerações Finais

Este trabalho descreve os resultados finais do trabalho de anotação semântica do *corpus* Summ-it. Para tanto, optou-se pelo paradigma FrameNet por ser uma semântica de orientação pragmática, descrevendo estruturas conceituais de situações. Para a anotação do *corpus* com as etiquetas semânticas do FrameNet foi utilizada a ferramenta SALTO, já empregada por outro projeto de anotação semântica, o SALSA. Trataremos aqui da avaliação dos resultados, dos futuros desdobramentos deste trabalho e da sua aplicação.

A base de dados FrameNet é bastante extensa. Isso, por um lado, enriquece a anotação, mas dificulta o trabalho dos anotadores, que devem se familiarizar com mais de 800 *frames* para realizar uma anotação *running-text*. Essa riqueza de dados pode provocar dúvidas nos anotadores, resultando em incompatibilidade entre as anotações. A vagueza, a polissemia e os diferentes padrões de lexicalização também interferiram na anotação, tendo em vista o trabalho bilíngüe que os anotadores devem realizar, encontrando para o *target* um equivalente de tradução em Inglês. Casos em que a Língua Inglesa e a Portuguesa apresentam maior paralelismo, como o *frame statement* (*state-affirmar, announce-anunciar*), por exemplo, foram mais fáceis de anotar, porém casos de polissemia, como o verbo *vir*, e de vagueza, como o verbo *envolver*, resultaram em anotações divergentes.

Neste trabalho, não enfrentamos casos complexos, como os verbos-suporte, verbos copulativos e modalizadores, pois ainda não há uma orientação metodológica clara por parte do FrameNet sobre o que fazer com tais verbos. Tal opção nos rende um índice razoavelmente baixo de sentenças anotadas, apenas 66,1%. Ainda temos dúvidas sobre as sentenças negativas, uma vez que a negação tem escopo sobre toda a estrutura conceitual do *frame*, negando todo o conteúdo. Assim, como etapas futuras deste trabalho, estão a anotação dos verbos suporte, copulativos e modalizadores e o aprofundamento dos estudos sobre as implicações da negação em uma anotação baseada em *frames*.

Quanto às aplicações tecnológicas deste trabalho, ele é o primeiro passo rumo a um *corpus* maior com anotação semântica *running-text*. A partir de um *corpus* manualmente anotado será possível a criação de ferramentas de anotação semântica automática, tais como o Shalmanaser (Erk e Padó, 2006) para o Alemão.

6. Referências

ALUÍSIO, S. et al. (2003). The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: Macnery, T. et al. (eds.), CORPUS LINGUISTICS 2003, Lancaster. *Proceedings of the Corpus Linguistics 2003*, UCREL Technical Papers, v.16. p.14-21.

ALUÍSIO, S. et al. (2007). Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: *Corpus Linguistics 2007 Conference*, Birmingham.

- BAKER, C. et al. (1998). The Berkeley FrameNet Project. In: ANNUAL MEETING OF THE ACL, 1998, Montréal. *Proceedings of the 36th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, v. 1. p. 86-90.
- BICK, E. (2000). *The Parsing Sistem PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PHD thesis, Arhus University.
- BOAS, H. (2002). Bilingual FrameNet Dictionaries for Machine Translation. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain. Vol. IV: 1364-1371.
- BORBA, F. (2000). *Uma gramática de valências para o português*. São Paulo: Ática.
- BURCHARDT, A., Erk, K., Frank, A., Kowalski, A., Padó, S. and Pinkal, M. (2006). The SALSA Corpus: a German Corpus Resource for Lexical Semantics. *Proceedings of LREC 2006*, Genoa, Italy.
- BURCHARDT, A., Erk, K., Frank, A., Kowalski, A. and Padó, S. (2006). SALTO - A Versatile Multi-Level Annotation Tool. *Proceedings of LREC 2006*, Genoa, Italy.
- ERK, K., Kowalski, A., Padó, S., Pinkal, M. (2003). Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. *Proceedings of ACL 2003*, Sapporo.
- FILLMORE, C. J., Johnson, C. R and Petruck, M. (2003). Background to FrameNet. *International Journal of Lexicography*. Vol.16, no.3, pp.235-250.
- GASPERIN, C. et al. (2003). Uma ferramenta para resolução automática de correferência. In: IV Encontro Nacional de Inteligência Artificial, 2003, Campinas. *Anais do XXIII Congresso da Sociedade Brasileira de Computação, IV ENIA*. Campinas: SBC, 2003
- MARCUS, M. (1994). The Penn TreeBank: A revised corpus design for extracting predicate-argument structure. In: THE ARPA HUMAN LANGUAGE TECHNOLOGY WORKSHOP, Princeton, 1994. *Proceedings of the ARPA Human Language Technology Workshop*.
- PADÓ, S. et al. (2006) The SALSA Corpus: a German corpus resource for lexical semantics. In: THE FIFTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, Genoa. *Proceedings of the 5th LREC*.
- PALMER, M. et al. (2001). Automatic Predicate Argument Analysis of the Penn TreeBank, In: FIRST INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2001, San Francisco. *Proceedings of HLT 2001*.
- RINO, L. et al. (2007). Summ-it: um corpus anotado com informações discursivas visando à sumarização automática. In: Workshop de Tecnologia da Informação e da

Linguagem Humana, 2007, Rio de Janeiro. *Proceedings do Congresso Nacional da SBC*.

RUPPENHOFER, J., Ellsworth, M, Petruck, M., Johnson, C. and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. ICSI.

SUBIRATS, C. & Petruck, M. (2003). Surprise: Spanish FrameNet. International Congress of Linguists. *Workshop on Frame Semantics*, Prague (Czech Republic).

VIEIRA, R. et al. (2007). An Agent-Oriented Programming Language for Computing in Context. In: *IFIP 19th World Computer Congress*, Santiago do Chile. Professional Practice in Artificial Intelligence - IFIP 19th World Computer Congress, TC-12 Professional Practice Stream. Berlin : Springer, 2006.