VerbNet.Br: construção semiautomática de um léxico verbal para o português do Brasil

Carolina Evaristo Scarton
carol@icmc.usp.br
Orientadora: Sandra Aluísio
sandra@icmc.usp.br

X Encontro de Linguística de Corpus – ELC 2011

Agenda

- Introdução
  - Principais Conceitos e Trabalhos Relacionados
  - VerbNet
- Método de construção da VerbNet.Br
- Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente
- Conclusões e Trabalhos Futuros

Principais Conceitos e Trabalhos Relacionados

- Construção e disponibilização de Recursos Léxicos Computacionais (RLC) → importante para a area de PLN

- RLC → informações:
  - Sintáticas
  - Semânticas
  - ...

Principais Conceitos e Trabalhos Relacionados

- Porém...
  - Construção manual → impraticável
    - Grande carga de trabalho
    - Tempo inaceitável

- Iniciativas para construção automática ou semiautomática
  - Machine Learning
  - **RLCs existentes→ abordagem cross-linguística**

Principais Conceitos e Trabalhos Relacionados

- Inglês → tradição em construir RLCs:

Baker et al. (1998)

Proposition Bank

Palmer et al. (2005)

Fellbaum (1998)

Kipper et al. (2005)

VerbNet

- Informação sintática e semântica de verbos

- Classes Verbais de Levin (Levin, 1993)
  - "Verbs that fall into classes according to shared (syntactic) behavior would be expected to show shared meaning components"

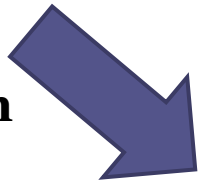
- Comportamento sintático → Alternâncias Sintáticas (Alternâncias de Diátese)
  - "Alternations in the expressions of arguments, sometimes accompanied by changes of meaning" (Levin, 1993)

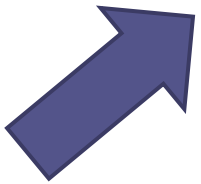
VerbNet

- Alternâncias Sintáticas (comportamento sintático)
  - **to spray**
    - (a) Sharon **sprayed** water **on** the plants.
      Agent **Theme** **Destination**
    - (b) Sharon **sprayed** the plants **with** water.
      Agent **Destination** **Theme**
  - **to load**
    - (a) The farmer **loaded** apples into the cart.
      Agent **Theme** **Destination**
    - (b) The farmer **loaded** the cart **with** apples.
      Agent **Destination** **Theme**

VerbNet

- Alternâncias Sintáticas (comportamento sintático)
  - **to spray**
    - (a) Sharon **sprayed** water **on** the plants.
      Agent **Theme** **Destination**
    - (b) Sharon **sprayed** the plants **with** water.
      Agent **Destination** **Theme**
  - **to load**
    - (a) The farmer **loaded** apples into the cart.
      Agent **Theme** **Destination**
    - (b) The farmer **loaded** the cart **with** apples.
      Agent **Destination** **Theme**





Semântica: putting and covering

Estrutura da VerbNet

Equip-13.4.2		
Thematic Roles and Selectional Restrictions: Agent [+animate | +organization], Theme e Recipient [+animate | +organization]		
Members: charge, invest, ply, arm, equip, rearm, redress, regale, reward, saddle, treat, armor, burden, compensate, encumber, overburden, weight		
Frames:		
NP V NP PP	Brown equipped Jones with a camera.	Agent V Recipient {with} Theme
Semantic Predicates	has_possession(start(E), Agent, ?Theme) has_possession(end(E), Recipient, ?Theme) transfer(during(E), ?Theme) cause(Agent, E)	

- VerbNet tem mapeamentos para a WordNet, PropBank e FrameNet

Trabalhos Relacionados

- Português do Brasil:
  - WordNet.Br (Dias-da-Silva et al., 2002; Dias-da-Silva, 2005; Dias-da-Silva et al., 2008)

- Iniciativas mais recentes:
  - FrameNet Brasil (Salomão, 2009) e FrameCorp (Bertoldi and Chishman, 2009)

  - PropBank.Br (Duran, 2009)

Trabalhos Relacionados

- Português do Brasil:
  - WordNet.Br (Dias-da-Silva et al., 2002; Dias-da-Silva, 200
- Inici
  - Fram
    (Berto
  - PropBank.Br (Duran, 2009)

Não tem informação sobre a interface sintático-semântica dos verbos

Trabalhos Relacionados

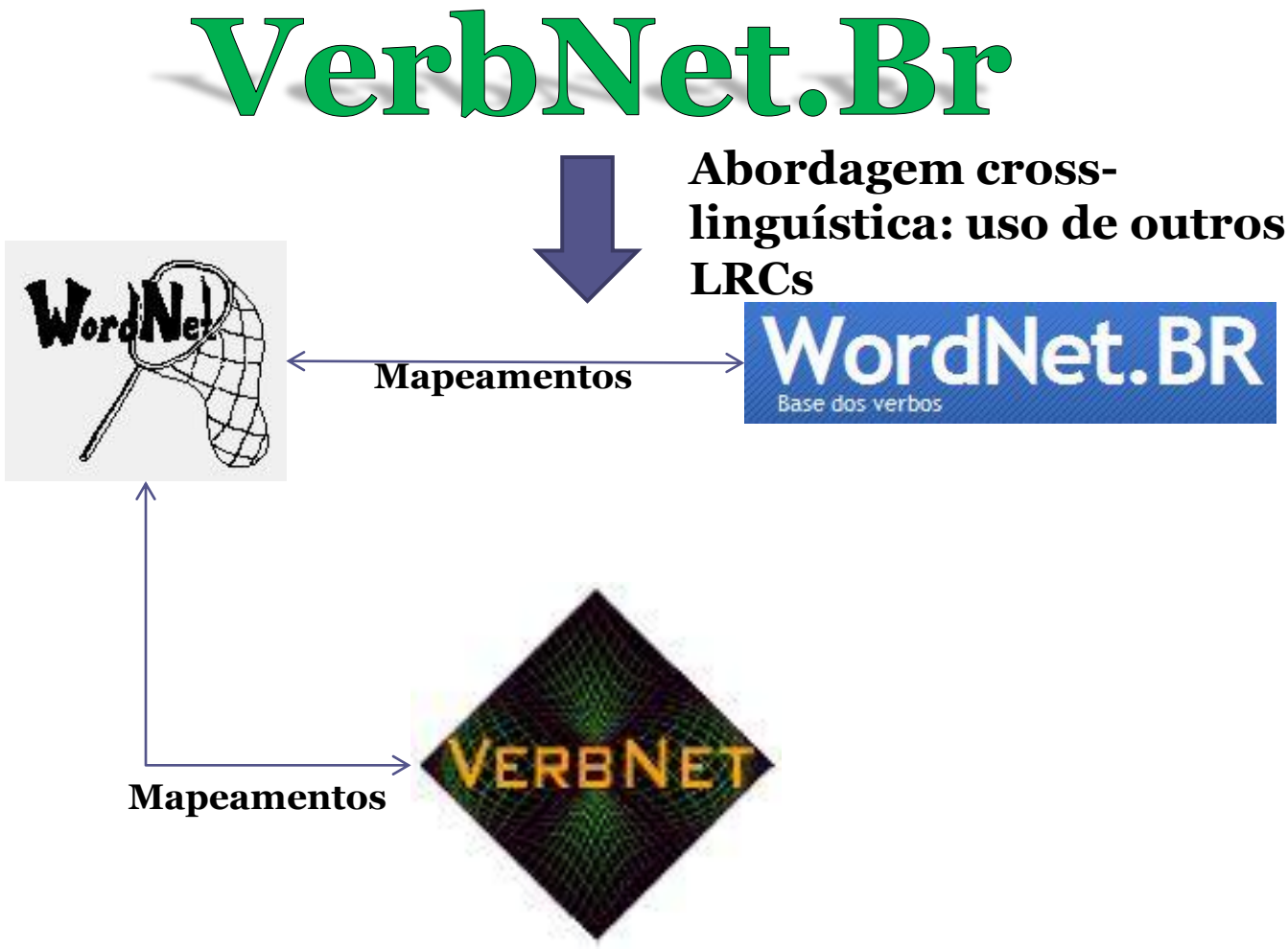
- Descrição do Português:
  - Verbos Psicológicos (Cançado, 1996)
  - Construção Adversativa (Chagas de Souza, 2001)
  - Verbos de Movimento (Moraes, 2008)
  - Verbos de Modo de Movimento (Amaral, 2010)

Trabalhos Relacionados

- Descrição do Português:
  - Verbos Psicológicos (Cançado, 1996)
  - Construçã
  - Verbos
  - Verbos de Modo de Movi (Amaral, 2010)

Resultados não disponíveis computacionalmente

VerbNet.Br


VerbNet.Br
Abordagem cross-linguística: uso de outros LRCs
WordNet
Mapeamentos
WordNet.BR
Base dos verbos
Mapeamentos
VERBNET

Método de Criação da VerbNet.Br

Estágio 1: Definição das alternâncias para o português (manual)

Estágio 2: Busca de alternâncias sintáticas – corpus

Estágio 3: Definição automática de candidatos a membros

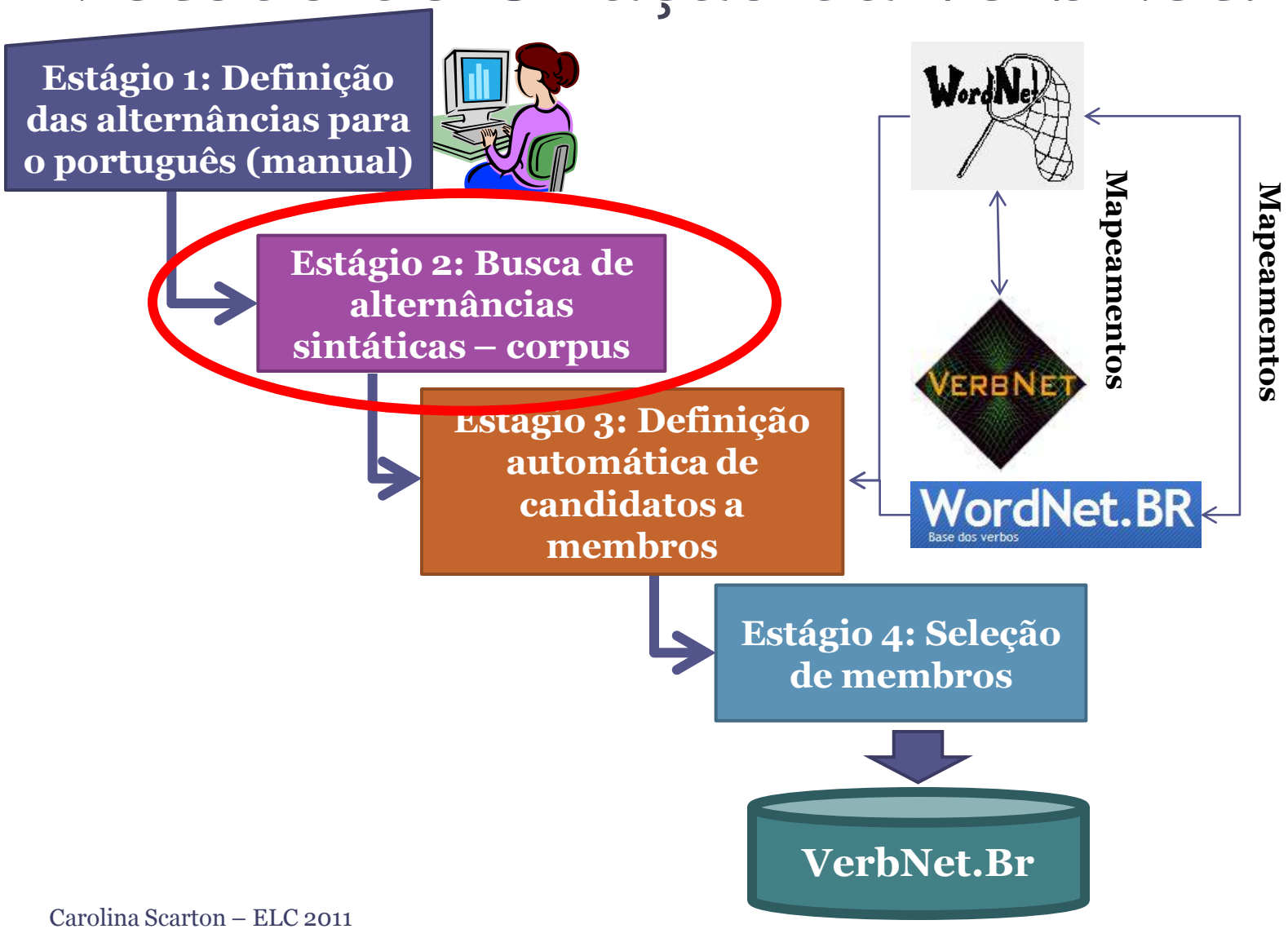
Mapeamentos

Mapeamentos

Estágio 4: Seleção de membros

VerbNet.Br

Método de Criação da VerbNet.Br



Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente

- Busca por padrões:
  - NP V NP
  - NP V de NP
  - NP V NP para NP
  - ...

- Exemplos:
  - Eu quebrei o vaso (NP – V – NP)
  - João gosta de sorvete (NP – V – de – NP)
  - Maria comprou um presente para José (NP – V – NP – para – NP)

Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente

- Mas.... Qual corpus usar?
  - Bosque (Floresta Sintáctica) (Afonso et al., 2002)
    - 186,000 palavras
    - Gênero jornalístico (1994)
  - Mac-Morpho (Aluisio et al., 2003)
    - 1,167,183 palavras
    - Gênero jornalístico (1994)
  - PLN-BR GOLD (Bruckschen et al., 2008)
    - 338,441 palavras
    - Gênero jornalístico (1994 – 2005)

Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente

- Mas.... Qual corpus usar?
  - Corpus NILC (Kuhn et al., 2000; Pinheiro e Aluísio, 2003)
    - 40 milhões de palavras
    - Gêneros: jurídico, didático, literário, técnico e científico, jornalístico e universitário (2000)
  - Lácio-Ref (Aluísio et al., 2004)
    - ~9 milhões de palavras
    - Gêneros: informativo, científico, prosa, poesia e drama (2004)

Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente

- Problemas:
  - Idade do Córpus?
  - Tamanho?
  - Gêneros?

Estágio 2: busca de alternâncias sintáticas em corpus anotado sintaticamente

- Relacionada com a área: **subcategorization frames** (Messiant, 2008; Korhonen et al., 2006 – para o inglês)

- Português do Brasil:
  - Projeto de mestrado de Adriano Zanette (UFRGS) → Busca de alternâncias sintáticas em corpus anotado com parser PALAVRAS (Bick, 2000)
    - Trabalho suportado por linguistas (principalmente o doutorando Leonardo Zilio)

Conclusão e Trabalhos Futuros

- Resultados finais:
  - Papéis temáticos
  - Restrições seletivas
  - Predicados semânticos
- Automaticamente herdados

Conclusão e Trabalhos Futuros

- Trabalhos Futuros
  - Terminar todos os estágios
  - Comparação do resultado da VerbNet.Br com uma abordagem de classificação de verbos totalmente automática
  - Aplicação da VerbNet.Br em uma ferramenta para análise da inteligibilidade em português (Coh-Metrix-Port) (Scarton e Aluísio, 2010)

Conclusão e Trabalhos Futuros

- Este trabalho é importante para a área de PLN→ problemas podem ser resolvidos e projetos podem ser melhorados
- Apoiar a tarefa de frames de subcategorização é importante para as áreas de PLN e Linguística
- Se o método apresentar bons resultados → pode ser aplicados para outras línguas

Obrigada!

Carolina Evaristo Scarton
carol@icmc.usp.br

Mais informações:
http://www2.nilc.icmc.usp.br/portlex/

PortLEX

References

Afonso, S., Bick, E., Haber, R. e Santos, D. (2002): "Floresta sintá(c)tica": a treebank for Portuguese, In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas de Gran Canaria, Espanha, pp.1698-1703.

Allbeck, J., Kipper, K., Adams, C., Schuler, W., Zoubanova, E., Badler, N., Palmer, M. e Joshi, A. (2002): ACUMEN: Amplifying Control and Understanding of Multiple ENtities. In *Proceedings of First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002)*. Bologna, Itália, pp. 191-198.

ALUISIO, S., PINHEIRO, G.M., MANFRIM, A.M.P, OLIVEIRA, L. H. M. de, L. C. GENOVES Jr., TAGNIN, S. E. O. The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In LREC 2004. Proceedings of LREC, 2004, Lisboa, Portugal, p. 1779-1782.

Amaral, L. L. (2010): **O Verbos de Modo de Movimento no Português Brasileiro**. 53f. Trabalho de Conclusão de Curso (Bacharel em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

Baker, C. F., Fillmore, C. J. e Lowe, J. F. (1998): The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, University of Montréal, Canadá, pp. 86-90.

Bertoldi A. e Chishman, R. L. de O. (2009): Desafios para a Criação de um Léxico baseado em Frames para o Português: um estudo dos frames Judgment e Assessing. In *Proceedings of the The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*. São Carlos, SP, Brazil, 1 CD-ROM ISSN 2175-6201.

References

Bick, E. (2000). **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. 505f. Ph.D. Thesis (Philosophy) – University of Aarhus, Dinamarca.

Bruckschen, M., Muniz, F., Souza, J. G. C., Fuchs, J. T., Infante, K., Muniz, M., Gonçalves, P. N., Vieira, R. e Aluísio, S. M. (2008): Anotação Lingüística em XML do Corpus PLN-BR. Série de Relatórios do NILC. NILC-TR-09-08, 39 p.

Cançado, M. (1996): Verbos Psicológicos: Análise Descritiva dos Dados do Português Brasileiro. Revista de Estudos da Linguagem, v. 4, n. 1, pp. 89-114.

Chagas de Souza, P. (2001): Notas Sobre a Construção Adversativa. *Anais do 4º Encontro do Círculo de Estudos Linguísticos do Sul (CELSUL)*, Curitiba, PR, Brasil. Disponível em: http://www.celsul.org.br/Encontros/04/artigos/100.htm. Acessado em: 27 fev. 2011.

Dias-da Silva, B. C., Oliveira, M. F. d., e Moraes, H. R. d. (2002). Groundwork for the development of the brazilian portuguese wordnet. In *Proceedings of the Third International Conference on Advances in Natural Language Processing*. London, UK, pp. 189–196.

Dias-da Silva, B. C. (2005) A construção da base da wordnet.br: conquistas e desafios. In *Proceedings of the Third Workshop in Information and Human Language Technology (TIL 2005), in conjunction with XXV Congresso da Sociedade Brasileira de Computação*. São Leopoldo, RS, Brasil, pp. 2238–2247.

Dias-da-Silva, B. C., Di Felippo, A. e Nunes, M. G. V. (2008). The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 1535-1541.

References

Duran, M. S. (2009). PropBank.BR: Regras Sintático-Semânticas para Mapeamento de Perguntas-Respostas de Verbos do Português e Anotação de Papéis Semânticos em um Corpus do Português do Brasil. Projeto de pós-doutorado aprovado pela FAPESP (processo: 2009/07394-9). ICMC-USP. Aprovado em maio de 2009.

Fellbaum, C. (1998). WordNet: An electronic lexical database. MIT Press. Cambridge, Massachusetts.

Girju, R., Roth, D. e Sammons, M. (2005): Token-level disambiguation of VerbNet classes. In Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes. Saarbruecken, Germany.

Kipper, K. (2005): Verbnet: A broad coverage, comprehensive verb lexicon. 146f. Ph.D. Thesis (Philosophy) - University of Pennsylvania, USA.

Levin, B. (1993): English Verb Classes and Alternation, A Preliminary Investigation. The University of Chicago Press.

Moraes, H. R. (2008): Aspectos sintaticamente relevantes do significado lexical: estudo dos verbos de movimento. 171f. Tese (Doutorado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara.

Palmer, M., Gildea D. e Kingsbury, P. (2005): The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, v. 31, n. 1, pp. 71-106.

Salomão, Maria M. M. (2009): FrameNet Brasil: Um trabalho em progresso. Revista Calidoscópio, v. 7, n. 3, pp. 171-182.

Scarton, C. E. e Aluísio, S. M. (2010): Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. Revista Linguamática (Revista para o Processamento Automático das Línguas Ibéricas - ISSN: 1647-0818), v. 2, n. 1, pp. 45-61.

Shi, L. e Mihalcea, R. (2005): Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In Proceedings of 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005). Cidade do México, México, pp. 99-110.

ZANETTE, Adriano. (2010) *Aquisição de* Subcategorization Frames *para Verbos da Língua Portuguesa*. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio.