

## Tesouros Distribucionais para o Português: avaliação de metodologias

Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira,  
Gabriel Gonçalves, Aline Villavicencio

<sup>1</sup> Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)

{rswilkens,lzilio,eduardo.ferreira}@inf.ufrgs.br

{gcgoncalves,avillavicencio}@inf.ufrgs.br

**Abstract.** *In recent decades there has been an increase in interest on methods for the automatic construction of distributional thesauri from corpora. Efforts to systematically evaluate and improve the resulting thesauri have been made for languages like English and French, but for Portuguese there is an urgent need for such initiatives. This paper presents a comparative investigation of the two main approaches for thesaurus generation: count-based and predictive methods, focusing on Portuguese. For the evaluation we propose a TOEFL-like test for Portuguese which was automatically generated from BabelNet, using nouns and verbs.*

**Resumo.** *Nas últimas décadas, houve um crescente interesse em métodos para a construção automática de tesouros distribucionais a partir de corpora. Esforços para a avaliação e aprimoramento sistemáticos dos recursos resultantes têm sido feitos para línguas como o inglês e o francês, mas, para o português, há ainda uma necessidade de tais iniciativas. Este artigo apresenta uma investigação comparativa entre dois métodos para construção de tesouros: baseados em contagens e preditivos, com foco no português. Para avaliação, é proposto um teste similar ao TOEFL para o português, o Brazilian BabelNet-based Semantic Gold Standard (B<sup>2</sup>SG), que contém questões automaticamente geradas a partir do BabelNet, com foco em substantivos e verbos.*

### 1. Introdução

A importância de recursos como a WordNet [Fellbaum 1998], que explicitam relações entre palavras, pode ser medida pelo número de iniciativas dedicadas a (re)produzi-los para outras línguas, tais como a EuroWordNet<sup>1</sup> [Vossen 1998] e a Global WordNet Association<sup>2</sup> [Bond and Paik 2012]. Tais recursos têm sido utilizados em inúmeras aplicações de tecnologia de linguagem, como sistemas de perguntas e respostas, de simplificação de texto e de análise de sentimentos. Para o português, estão disponíveis o Onto.PT<sup>3</sup> [Gonçalo Oliveira and Gomes 2010], OpenWN-PT<sup>4</sup> [de Paiva et al. 2012], MultiWordnet of Portuguese<sup>5</sup>, o WordNet.PT<sup>6</sup> [Marrafa 2002], WordNet.Br<sup>7</sup> [Dias-da-Silva et al. 2008],

<sup>1</sup><http://www.illc.uva.nl/EuroWordNet/>

<sup>2</sup><http://globalwordnet.org/wordnets-in-the-world/>

<sup>3</sup><http://ontopt.dei.uc.pt>

<sup>4</sup><https://github.com/arademaker/openWordnet-PT>

<sup>5</sup><http://mwnpt.di.fc.ul.pt/>

<sup>6</sup><http://www.clul.ul.pt/clg/wordnetpt/index.html>

<sup>7</sup><http://143.107.183.175:21380/wordnetbr>

entre outros.

A construção manual desse tipo de recurso requer conhecimento especializado, além de ser cara e demorada. Além disso, o recurso resultante é estático, tem cobertura limitada e se aplica a um domínio geral. Por isso, como alternativa, investigam-se métodos baseados em corpora para a construção automática de tesouros distribucionais com associações semânticas entre palavras. Esses métodos são independentes de linguagem e aplicáveis a qualquer domínio [Lin 1998], e os recursos gerados podem complementar a informação de recursos lexicais e ontológicos como a WordNet.

Assim, muita atenção tem sido devotada para construção, avaliação e aprimoramento sistemáticos de tesouros distribucionais, principalmente para o inglês, mas também para outras línguas, como o francês. Para essas línguas, o desenvolvimento de conjuntos de testes e gold standards disponíveis para a comunidade, tais como o English Lexical Substitution Task<sup>8</sup> [McCarthy and Navigli 2009], o TOEFL [Landauer and Dumais 1997] e o teste derivado do TOEFL, o WordNet-Based Synonymy Test (WBST) [Freitag et al. 2005], tem permitido a comparação direta de técnicas diferentes e a quantificação precisa de melhorias na qualidade dos recursos gerados. Questões como a influência do método usado (baseado em contagem ou preditivo) [Baroni et al. 2014, Lebrecht and Collobert 2015], da medida de associação e medida de similaridade [Lin 1998, Padró et al. 2014], do tipo de contexto (bag-of-words ou dependências sintáticas) e de seu tamanho ( $1 \times 2 \times 5 \times n$  palavras em torno de cada palavra-alvo) [Freitag et al. 2005] têm sido cuidadosamente analisadas para determinar a melhor estratégia para se obter um tesouro de qualidade de acordo com língua, tamanho e tipo de corpus.

Para o português, ainda faltam estudos comparativos e conjuntos de dados e gold standards. Este trabalho tem por objetivo contribuir na criação de gold standards que possam ser usados para avaliações comparativas desses métodos, através da construção de um teste similar ao TOEFL para o português. O Brazilian BabelNet-based Semantic Gold Standard ( $B^2SG$ ) foi automaticamente gerado a partir do BabelNet [Navigli and Ponzetto 2010], contendo questões que envolvem o cálculo de similaridade entre uma determinada palavra e candidatos a palavras semanticamente relacionadas. Este artigo também visa a responder parte das questões sobre a qualidade dos tesouros gerados com foco no português, através de uma investigação comparativa entre dois métodos para a construção de tesouros (baseado em contagem e preditivo).

Esses tópicos são discutidos no artigo da seguinte forma: em §2, são apresentados os trabalhos relacionados sobre tesouros distribucionais e, em §3, os materiais e métodos empregados. A avaliação comparativa e os resultados são discutidos em §4, e as conclusões e trabalhos futuros são expostos em §5.

## 2. Tesouros Distribucionais

A palavra tesouro surgiu na Lexicografia com Peter Mark Roget, em 1852, para designar seu *Thesaurus of English Words and Phrases* [Moreira and Moura 2006]. O nome foi usado para designar o seu dicionário, em que as palavras se organizavam “de acordo

---

<sup>8</sup>Disponível em <http://nlp.cs.swarthmore.edu/semeval/tasks/task10/summary.shtml>.

com as ideias que exprimiam” [Gomes et al. 1990]. Assim, surgiram dicionários que exprimiam a similaridade entre as palavras por meio de relações.

No português, existe o *Dicionário analógico da língua portuguesa* [Santos Azevedo 1990], que divide as palavras em seis classes primárias: relações abstratas, espaço, matéria, entendimento, vontade e afeições. Em formato eletrônico, temos como exemplo o TEP [Dias-Da-Silva and Moraes 2003], o BabelNet [Navigli and Ponzetto 2010] e o Onto.PT [Oliveira and Gomes 2014]. Dentre esses, como veremos mais adiante, optamos por usar o BabelNet como comparação devido principalmente à sua cobertura e à distinção de polissemia.

Para a construção automática de tesouros distribucionais a partir de corpora, tradicionalmente, utiliza-se como base a hipótese distribucional de Harris de que se pode conhecer uma palavra pelas palavras que costumam ocorrer com ela [Lin 1998]. Há duas principais abordagens para a construção de tesouros: uma, mais tradicional, baseada em contagem [Lin 1998, Baroni and Lenci 2010] e outra, mais recente, baseada em redes neurais [Mikolov et al. 2010]. Avaliações sobre a qualidade dos recursos gerados por cada abordagem existem para algumas línguas e domínios [Padró et al. 2014]; porém, avaliações comparativas das duas abordagens ainda são raras [Baroni et al. 2014, Lebret and Collobert 2015] e reportam resultados divergentes. Por exemplo, comparando modelos tradicionais e modelos preditivos em 14 tarefas diferentes, os modelos preditivos obtiveram os melhores resultados [Baroni et al. 2014], mas, em outras tarefas, ambos os modelos obtiveram resultados comparáveis [Lebret and Collobert 2015]. Neste artigo, apresentamos os dois modelos e uma avaliação comparativa para o português.

## 2.1. Modelos baseados em contagem

Os modelos tradicionais baseados em contagem foram propostos para a construção automática de tesouros distribucionais, variando principalmente em termos de (a) tipo e tamanho do contexto a ser utilizado, (b) medidas utilizadas para calcular a associação de uma palavra-alvo com o contexto em que ocorre e (c) medidas para calcular a similaridade entre palavras a partir de seus contextos.

Em (a), o contexto usado para representar o perfil distribucional da palavra-alvo pode envolver relações sintáticas (por exemplo, para verbos, pode-se usar sujeito e objeto) ou uma *bag-of-words* (BoW) contendo as  $n$  palavras de conteúdo à sua volta [Freitag et al. 2005]. Em (b), são utilizadas medidas estatísticas para determinar um valor de associação entre cada palavra do contexto e o alvo. Para calcular a associação entre a palavra-alvo e cada palavra nos seus contextos de ocorrência, são usadas várias medidas estatísticas de associação, tais como *pointwise mutual information* (PMI),  $\chi^2$ , *log likelihood*, entre outras [Lin 1998]. A similaridade entre duas palavras (c) é então calculada com base na semelhança de seus contextos, usando medidas de proximidade (p.ex., cosseno), de distância (p.ex., Manhattan ou Euclidiana) ou de divergência (p.ex., Kulback-Leibler) [Lin 1998, Freitag et al. 2005].

## 2.2. Modelos baseados em predição

Redes neurais têm sido utilizadas com bastante sucesso para o problema clássico da construção de modelos de linguagem: a predição da probabilidade de uma sequência de palavras. Em particular, o trabalho de Mikolov em redes neurais recorrentes

para modelar a linguagem gerou modelos que, ao serem treinados para predizerem sequências de palavras, as distribuem num espaço que captura propriedades linguísticas [Mikolov et al. 2010]. A arquitetura típica dessas redes consiste em uma camada de entrada e uma de saída, uma camada oculta com conexões recorrentes e uma matriz de pesos. Os vetores de entrada e saída codificam as palavras e a camada oculta mantém o histórico de representação. Nesse modelo, não são utilizados conhecimentos sintáticos, morfológicos ou semânticos explicitamente. Ele apenas recebe como entrada um texto simples, sem qualquer anotação ou pré-processamento.

### 2.3. Avaliação

A avaliação de tesouros distribucionais é uma tarefa complexa, pois faltam recursos que meçam a similaridade entre palavras. Para realizar a avaliação, pode-se utilizar uma validação por juízes; contudo, essa forma de avaliação é lenta e custosa. Uma alternativa é a utilização de ontologias lexicais, como a WordNet, comparando a ontologia e o tesouro. A avaliação também pode ser indireta, através de tarefas que necessitam da quantificação da similaridade, tais como:

**Deteção de relações semânticas:** objetiva agrupar as palavras segundo uma relação predeterminada. Datasets incluem o BLESS [Baroni and Lenci 2011], com 200 substantivos agrupados em 17 classes; o ESSLLI, [Baroni et al. 2008] com 44 conceitos em 6 classes; o Strudel, com 83 conceitos e 10 classes [Baroni et al. 2010]<sup>9</sup>; e o SemEval 2010 Task 8 [Hendrickx et al. 2010], baseado em 9 relações semânticas profundas.

**Identificação da preferência seletional de verbos:** objetiva identificar qual a relação sintática mais adequada entre um verbo e um substantivo. Existem conjuntos de 211 verbos [Padó 2007] e de 100 verbos [McRae et al. 1998].

**Identificação de analogia:** usa exemplos da relação para fazer inferência de analogias morfológicas, sintáticas e semânticas [Mikolov et al. 2013b], tais como *man está para woman assim como king está para queen* [Mikolov et al. 2013a].

**Identificação de itens relacionados semanticamente:** objetiva identificar palavras que são relacionadas por alguma relação semântica (p.ex., *tigre* e *animal*, *areia* e *praia*). Datasets incluem 65 pares de substantivos [Rubenstein and Goodenough 1965], 80 substantivos (TOEFL [Landauer and Dumais 1997]), 353 pares (WordSim353 [Finkelstein et al. 2001]), 2003 pares com sentenças de contexto (SCWS [Huang et al. 2012]) e 3000 pares (MEN [Bruni et al. 2014]). No TOEFL, para cada uma das 80 palavras-alvo, há quatro alternativas, dentre as quais se deve identificar a palavra mais próxima semanticamente. Já o WordNet-Based Synonymy Test (WBST) [Freitag et al. 2005] é uma extensão gerada automaticamente a partir da WordNet.

## 3. Materiais e métodos

Nesta seção, apresentamos a metodologia de criação do recurso de avaliação utilizado, o Brazilian BabelNet-based Semantic Gold Standard ( $B^2SG$ ), o corpus utilizado para o treino dos modelos e, por fim, o desenvolvimento dos tesouros distribucionais.

---

<sup>9</sup>As tarefas de agrupamento são avaliadas com base na pureza [Baroni and Lenci 2011, Baroni et al. 2008, Baroni et al. 2010].

### 3.1. Gold standard para português

A fim de avaliar a performance das diferentes abordagens no português, criamos um gold standard para o português baseado no WBST do inglês [Freitag et al. 2005]. Diferente do WBST, que explora apenas a relação de sinonímia, o recurso desenvolvido explora as relações de sinonímia, antonímia, hiperonímia, hiponímia e outras<sup>10</sup>. Outro ponto de diferença é nosso foco em substantivos e verbos, por causa de sua relevância entre as classes gramaticais. Como no WBST, para cada palavra-alvo, há 4 alternativas: 1 semanticamente relacionada e 3 não relacionadas.

O Brazilian BabelNet-based Semantic Gold Standard ( $B^2SG$ ) foi gerado em 3 etapas: (1) criação da lista de palavras-alvo, (2) seleção das palavras relacionadas e (3) seleção das palavras não relacionadas. Para a lista de alvos, utilizou-se uma lista de palavras (substantivos e verbos) anotadas com a frequência de um corpus de referência do projeto AC/DC<sup>11</sup>. A anotação do grau de polissemia das palavras foi feita com base no BabelNet [Navigli and Ponzetto 2010], e as palavras não contidas nele foram removidas. Para a geração da lista de palavras semanticamente relacionadas, foi utilizado o BabelNet para identificar sinônimos, antônimos, hiperônimos etc. de cada alvo. A escolha de palavras não relacionadas utilizou como base a mesma lista de palavras relacionadas, porém, para cada palavra-alvo, selecionaram-se palavras sem relação com ela de acordo com o BabelNet. As palavras selecionadas como relacionadas e não relacionadas tiveram frequência e polissemia uniformizadas por meio de filtros baseados, respectivamente, no AC/DC e no BabelNet.

**Frequência:** com base na anotação de frequência, as alternativas foram ordenadas pela menor distância em relação à frequência da palavra-alvo, e a média da frequência das palavras não relacionadas. Selecionamos os 10.000 substantivos e os 5.000 verbos com menor distância por relação<sup>12</sup>.

**Polissemia:** parecida com a filtragem por frequência, a filtragem por polissemia usou a ordenação pela distância entre o valor de polissemia da palavra relacionada e a média dos valores das palavras não relacionadas, sendo que a seleção dos substantivos e verbos foi baseada na menor distância por relação.

Como resultado, foram geradas 5 listas de verbos e 5 de substantivos, uma para cada tipo de relação, num total de 11.235 perguntas (2.700 para verbos e 8.535 para substantivos), como na Tabela 1, considerando relação e classe gramatical.

### 3.2. Corpus

Como o tamanho do corpus tem impacto na performance de muitas tarefas em PLN, procuramos utilizar o maior corpus possível. Nesse sentido, a metodologia WaC (Web as Corpus kool yinitiative) provê uma forma rápida e simples de criação de grandes corpora [Baroni et al. 2009]. Para o português, foi utilizado o brWaC [Boos et al. 2014] com anotação de rótulo morfossintático do TreeTagger [Schmid 1994]. Esse processo resultou em um corpus com 52 milhões de tokens e 875 mil types.

<sup>10</sup>Neste trabalho, incluímos quatro classes de relações que são explicitamente distintas (sinonímia, antonímia, hiperonímia e hiponímia) e uma classe que inclui outros tipos de relações (por exemplo, meronímia).

<sup>11</sup>Disponível em <http://www.linguateca.pt/ACDC>.

<sup>12</sup>A relação de antônimo em ambas as classes gramaticais não alcançou o número mínimo (10.000 substantivos e 5.000 verbos), então usamos todos os valores como saída do processo de filtragem de frequência.

**Tabela 1. Tamanho em número de palavras-alvo nas diferentes relações semânticas no gold standard proposto**

|              | Sinônimos   | Hiperônimos | Hipônimos   | Antônimos  | Outros      | Total        |
|--------------|-------------|-------------|-------------|------------|-------------|--------------|
| Verbos       | 500         | 500         | 500         | 200        | 1000        | 2700         |
| Subst        | 1667        | 1667        | 1667        | 200        | 3334        | 8535         |
| <i>Total</i> | <i>2167</i> | <i>2167</i> | <i>2167</i> | <i>400</i> | <i>4334</i> | <i>11235</i> |

### 3.3. Tesauros de contagem

O tesauro baseado em contagem foi construído seguindo Padró et. al [Padró et al. 2014]: descartaram-se palavras com ocorrência de bigrama menor que 5 no corpus e utilizou-se a implementação *DISSECT* [Dinu et al. 2013]. O tipo de contexto usado são as palavras em torno de substantivo ou verbo como uma bag-of-words, isto é, uma janela de  $n$  palavras de contexto antes e depois da palavra-alvo. Foram gerados tesauros com dois tamanhos de janela: 5 e 10. Assim, um contexto  $(s, c, t)$  é a ocorrência de substantivo  $s$ , contexto  $c$  e marcação  $t$ , e o número de ocorrências de um contexto em um corpus é representado por  $||s,c,t||$ . Por exemplo, a frase “O cão comeu o osso” gera duas triplas ( $s = \text{“cão”}$ ,  $c = \text{“comer”}$ ,  $t = \text{“verbo”}$ ) e ( $s = \text{“cão”}$ ,  $c = \text{“osso”}$ ,  $t = \text{“substantivo”}$ ). Utilizaram-se apenas contextos com mais de 5 ocorrências e com PMI maior do que zero.

### 3.4. Tesauros de predição

Com os modelos baseados em predição, foram criados dois tesauros distintos a partir do brWaC: um com apenas os lemas das palavras (corpus normalizado) e outro com os lemas e sua anotação morfossintática (corpus anotado). Esses modelos foram construídos a partir do pacote *word2vec* [Mikolov et al. 2010], que possui diversos parâmetros, tais como: (1) tamanho desejado do vetor (a quantidade de nós que são passados para a rede neural); (2) janela de contexto que será analisada pelo algoritmo; (3) *downsampling* (limiar para que palavras de alta frequência sejam aleatoriamente ignoradas); e (4) quanto de ruído será extraído por técnicas de suavização.

Cada tesauro utilizou um algoritmo de bag-of-words para a geração do modelo, com um vetor de palavras de tamanho 500, uma janela de contexto de tamanho 10, um limiar de *downsampling* de  $1e-5$ , uma amostragem de 25 para o algoritmo de treinamento negativo e uma frequência mínima de 5 ocorrências no corpus para que a palavra fosse utilizada no tesauro resultante<sup>13</sup>.

## 4. Avaliação

A avaliação e a comparação de métodos de criação de tesauros distribucionais usam o recurso descrito na Seção 3.1 para verificar a capacidade dos modelos para identificarem a resposta adequada em nível de relação.

Na Tabela 2, são apresentados os acertos dos dois modelos (contagem e predição). No modelo de contagem, os resultados são divididos entre as bag-of-words com janela de tamanho 5 e 10. No modelo de predição, o corpus normalizado apresenta os dados

<sup>13</sup>Os valores utilizados na parametrização do pacote são os sugeridos para grandes corpora no site do pacote (<https://code.google.com/p/word2vec/>).

sem distinção morfossintática, enquanto o corpus anotado considera a anotação morfossintática. Os acertos foram calculados com base no número de perguntas em que a resposta certa se encontrava no vocabulário (ou seja, consideramos apenas os casos em que o modelo poderia acertar). Observou-se que o modelo preditivo sem morfossintaxe tem um resultado superior aos demais para a maioria das relações. Consideramos que o resultado inferior do modelo preditivo usando anotação morfossintática se deve à maior esparsidade nos dados.

Analisando a diferença entre o corpus usado para gerar o teste (AC/DC) e o corpus usado para gerar os modelos percebemos uma diferença quanto à distribuição da frequência<sup>14</sup>, o que pode afetar a cobertura de vocabulário. A fim de evitar o impacto das palavras fora de vocabulário sobre a performance dos modelos, rodamos também um teste restrito em que foram consideradas apenas as palavras-alvo em que todas as 4 alternativas eram conhecidas pelo modelo (Tabela 3). Nesse teste mais restrito, o modelo preditivo sem anotação morfossintática continuou sendo superior na maioria das relações.

**Tabela 2. Porcentagem de acertos obtidos nos quatro modelos**

|            | Contagem     |           | Predição       |                    |
|------------|--------------|-----------|----------------|--------------------|
|            | Janela 5     | Janela 10 | Corpus anotado | Corpus normalizado |
| Antônimo   | 64.5%        | 62.5%     | 55.7%          | <b>67.3%</b>       |
| Hiperônimo | 60.7%        | 56.2%     | 56.2%          | <b>64.3%</b>       |
| Hipônimo   | 54.2%        | 53.6%     | 56.3%          | <b>59.4%</b>       |
| Sinônimo   | 65.8%        | 64.5%     | 61.4%          | <b>68.4%</b>       |
| Outras     | <b>57.0%</b> | 55.7%     | 55.5%          | 55.3%              |

**Tabela 3. Porcentagem de acertos obtidos com as quatro alternativas conhecidas**

|            | Contagem     |              | Predição       |                    |
|------------|--------------|--------------|----------------|--------------------|
|            | Janela 5     | Janela 10    | Corpus Anotado | Corpus Normalizado |
| Antônimo   | <b>64.0%</b> | 61.1%        | 50.6%          | 62.2%              |
| Hiperônimo | 55.1%        | 52.8%        | 33.8%          | <b>56.0%</b>       |
| Hipônimo   | 47.8%        | 46.5%        | 38.5%          | <b>50.0%</b>       |
| Sinônimo   | 61.8%        | 60.7%        | 46.7%          | <b>62.5%</b>       |
| Outras     | <b>49.9%</b> | <b>49.9%</b> | 39.5%          | 49.5%              |

Esses resultados para o português são compatíveis com os obtidos para o inglês por Baroni et al. [Baroni et al. 2014], já que o método preditivo teve um resultado superior na identificação de itens semanticamente relacionados. Desse modo, o modelo parece capturar aspectos da representação semântica das palavras nessas línguas.

## 5. Conclusões

Recentemente houve um aumento no interesse pela construção automática de tesauros distribucionais a partir de corpora. Para línguas como o inglês e o francês, já existem avaliação e melhora dos recursos, mas, para o português, há ainda muito espaço para

<sup>14</sup>Analisando as palavras com frequência superior a 5 nos corpora brWaC e AC/DC observamos uma correlação 0,5298 (com 99,99% de confiança).

desenvolvimento. Nesse sentido, este artigo apresentou uma investigação comparativa entre dois métodos para construção de tesouros: baseado em contagem e preditivo, com foco no português.

Para avaliação, foi proposto um teste similar ao TOEFL, um dos principais testes utilizados na língua inglesa. Esse teste (denominado Brazilian BabelNet-based Semantic Gold Standard –  $B^2SG$ )<sup>15</sup> contém questões automaticamente geradas a partir de um recurso lexical similar à WordNet, o BabelNet.

A comparação apresentada neste trabalho aponta que a utilização de um método preditivo sem uso de anotação morfossintática tem um resultado superior para a criação de tesouros. Um ponto importante a ser considerado é que o BabelNet foi construído automaticamente, de modo que pode haver erros nas relações, impactando o teste gerado neste trabalho.

Como trabalhos futuros, pretendemos avaliar manualmente as perguntas e alternativas do teste, além de estender os testes avaliados, incluindo testes de preferência lexical. Com isso, poderemos delimitar melhor a qualidade dos tesouros do português para uma tarefa em particular: a simplificação textual, e, em um âmbito maior, avaliar a qualidade dos modelos preditivos de maneira interlinguística.

### Agradecimento

Agradecemos ao Instituto de Informática da UFRGS pelo apoio à pesquisa. Parte dos resultados apresentados neste trabalho foram obtidos no projeto *Simplificação Textual de Expressões Complexas* patrocinado pela Samsung Eletrônica da Amazônia Ltda., através da lei 8.248/91, e também contou com apoio do CNPq (482520/2012-4, 312184/2012-3).

### Referências

- Baroni, M., Barbu, E., Murphy, B., and Poesio, M. (2010). Strudel: A distributional semantic model based on properties and types.
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.
- Baroni, M., Evert, S., and Lenci, A. (2008). Bridging the gap between semantic theory and computational simulations: Proceedings of the esslli workshop on distributional lexical semantics. *Hamburg, Germany: FOLLI*.
- Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.
- Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

---

<sup>15</sup>O Brazilian BabelNet-based Semantic Gold Standard ( $B^2SG$ ) será disponibilizado para a comunidade em <http://www.inf.ufrgs.br/pln/explaintext/index.php?title=Publications>



- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference*, pages 64—71.
- Boos, R., Prestes, K., Villavicencio, A., and Padró, M. (2014). brwac: a wacky corpus for brazilian portuguese. In *Computational Processing of the Portuguese Language*, pages 201–206. Springer.
- Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.
- de Paiva, V., Rademaker, A., and de Melo, G. (2012). Openwordnet-pt: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*. See at <http://www.coling2012-iitb.org> (Demonstration Paper). Published also as Techreport <http://hdl.handle.net/10438/10274>.
- Dias-da-Silva, B. C., Felippo, A. D., and das Graças Volpe Nunes, M. (2008). The automatic mapping of princeton wordnet lexical-conceptual relations onto the brazilian portuguese wordnet database. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.
- Dias-Da-Silva, B. C. and Moraes, H. R. d. (2003). A construção de um thesaurus eletrônico para o português do brasil. *ALFA: Revista de Linguística*.
- Dinu, G., N. P., and M. B. (2013). Dissect-distributional semantics composition toolkit. In *System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*.
- Fellbaum, C. (1998). *WordNet*. Wiley Online Library.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.
- Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., and Wang, Z. (2005). New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32. Association for Computational Linguistics.
- Gomes, H. E., da Educação, M., de Pessoal, B. C. d. A., de Estudos, F., et al. (1990). *Manual de elaboração de tesouros monolíngües*. Programa Nacional de Bibliotecas das Instituições de Ensino Superior.
- Gonçalo Oliveira, H. and Gomes, P. (2010). Towards the automatic creation of a wordnet from a term-based lexical network. In *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pages 10–18. ACL Press.
- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.
- Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Lebret, R. and Collobert, R. (2015). Rehabilitation of count-based models for word vector representations. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing - 16th International Conference*, volume 9041 of *Lecture Notes in Computer Science*, pages 417–429. Springer.

- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774. Association for Computational Linguistics.
- Marrafa, P. (2002). *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto de Camões, Lisboa.
- McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.
- McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.
- Moreira, M. P. and Moura, M. A. (2006). Construindo tesauros a partir de tesauros existentes: a experiência do tci-tesauro em ciência da informação. *DataGramaZero-Revista de Ciência da Informação*, 7(4).
- Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.
- Oliveira, H. G. and Gomes, P. (2014). Eco and onto. pt: A flexible approach for creating a portuguese wordnet automatically. *Language resources and evaluation*, 48(2):373–393.
- Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. PhD thesis, Universitätsbibliothek.
- Padró, M., Idiart, M., Villavicencio, A., and Ramisch, C. (2014). Comparing similarity measures for distributional thesauri. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA)*.
- Padró, M., Idiart, M., Villavicencio, A., and Ramisch, C. (2014). Nothing like good old frequency: Studying context filters for distributional thesauri. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 419–424. ACL.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
- Santos Azevedo, F. F. d. (1990). *Dicionário analógico da língua portuguesa*. Lexikon.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.
- Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.