# Uma ferramenta para anotação de relações semânticas entre termos

## 1. Introdução

O Processamento de Língua Natural (PLN) adquire importância cada vez maior atualmente. É uma área em evidência pois a quantidade de informação disponível na forma de textos cresce a cada dia; no entanto, os recursos humanos disponíveis para estruturar e extrair conhecimento útil dessa vasta quantidade de dados não acompanham seu ritmo de crescimento, sendo necessário o uso de ferramentas para o processamento automático de textos. Essas ferramentas, por sua vez, precisam de uma quantidade de *corpora* anotados com informações relevantes para treinamento, mas a produção desses recursos é custosa, demandando tempo e anotadores especializados.

Considerando os níveis de processamento linguístico, a morfologia e a sintaxe são níveis em que já existe uma quantidade significativa de *corpora* anotados. Em comparação, a semântica ainda não tem uma quantidade tão grande de recursos, especialmente considerando a língua portuguesa. Essa escassez de dados semânticos é um limitador para diversas aplicações que poderiam se beneficiar de tais recursos, como a tradução automática, recuperação de informação, busca na *web* e outras.

Assim, dado o grande esforço necessário para a construção de recursos linguísticos e a escassez desses recursos em nível semântico, sobretudo para a língua portuguesa, este artigo apresenta uma ferramenta construída com o propósito de auxiliar a tarefa de anotação de relações semânticas. Essa tarefa se propõe a identificar relações semânticas binárias entre termos de um texto. Exemplos clássicos dessas relações incluem a hiponímia (subclasse e superclasse), meronímia (parte e todo) e sinonímia (sinônimos). Este trabalho enfoca sete relações semânticas derivadas da teoria da organização do conhecimento de Minsky [Minsky 1986], descritas na Tabela 1.

**Tabela 1. Relações semânticas consideradas neste trabalho**

| | **Relação semântica** | **Sentença exemplo** | **Relação extraída** |
|---|---|---|---|
| 1 | is-a(subclasse, superclasse) | Maçã é uma fruta | is-a(maçã, fruta) |
| 2 | property-of(algo/alguém, característica) | O prédio é alto | property-of(prédio, alto) |
| 3 | part-of(todo, parte) | Parafuso é uma parte de uma máquina | part-of(máquina, parafuso) |
| 4 | made-of(produto, substância) | Cacau é utilizado para fazer chocolate | made-of(chocolate, cacau) |
| 5 | effect-of(ação/estado, consequência) | Gripe causa febre | effect-of(gripe, febre) |
| 6 | used-for(entidade, função) | Pás são usadas para cavar | used-for(pás, cavar) |
| 7 | location-of(algo/alguém, local) | Uma secretária pode ser encontrada em um escritório | location-of(secretária, escritório) |

Esta ferramenta foi desenvolvida no âmbito de um estudo sobre extração automática de relações semânticas, fornecendo suporte para a construção manual de um corpus anotado, recurso necessário para o treinamento dos métodos computacionais automáticos que serão investigados.

O restante deste artigo se organiza da seguinte forma: a Seção 2 apresenta uma breve revisão da bibliografia sobre ferramentas de anotação semântica; a Seção 3 apresenta a metodologia e as principais decisões tomadas na construção da ferramenta; a Seção 4 apresenta os resultados obtidos com o uso da ferramenta; e a Seção 5 traz as conclusões e considerações para trabalhos futuros.

## 2. Revisão bibliográfica

Algumas das ferramentas já desenvolvidas para a tarefa de anotação semântica de *corpora* são a SALTO [Burchardt et al. 2006] e a GATE [Cunningham et al. 2011].

A SALTO [Burchardt et al. 2006] foi originalmente desenvolvida para a anotação de papéis semânticos no formato da semântica de frames [Baker et al. 1998], embora possa ser utilizada para outros tipos de anotação. Além da funcionalidade da marcação de textos, a ferramenta também pode fazer gerenciamento de corpus e tem uma interface para resolução de conflitos entre anotadores. Os formatos de entrada e saída de dados aceitos pela SALTO são o Tiger XML [Mengel and Lezius 2000] e seu formato próprio, o SALSA/Tiger XML. A SALTO foi desenvolvida utilizando a linguagem Java e a biblioteca Swing para a interface com o usuário. O projeto Propbank-Br [Duran and Aluísio 2011], cujo objetivo é a criação de uma base textual anotada com papéis semânticos em português do Brasil, utiliza esta ferramenta em seu desenvolvimento.

A GATE [Cunningham et al. 2011] é uma plataforma dedicada ao desenvolvimento de qualquer tipo de tarefa de PLN, incluindo etiquetação de *corpora*. Ela fornece um ambiente integrado para o desenvolvimento de aplicações de PLN, possibilitando a integração de todas as etapas do processamento de um texto (p. ex. tokenização, segmentação, análise léxica) em um mesmo local. A GATE aceita diversos formatos de dados de entrada como texto puro, HTML, SGML, XML, PDF, entre outros, mas tem um formato de armazenamento próprio baseado em XML. A linguagem utilizada em seu desenvolvimento foi Java.

Apesar da existência dessas alternativas, optou-se pelo desenvolvimento de uma nova ferramenta pois os trabalhos existentes podem fazer diversos tipos de anotações e possuem muitas funcionalidades que não seriam aproveitadas no escopo da pesquisa em que este trabalho está inserido – a plataforma GATE em especial fornece um ambiente completo de desenvolvimento que seria subutilizado. Assim, preferiu-se o desenvolvimento de uma ferramenta mais simples e alinhada diretamente com a tarefa de marcação de relações semânticas entre termos.

## 3. Metodologia

A ferramenta para anotação de relações semânticas entre termos foi desenvolvida utilizando a linguagem Java versão 1.6, a biblioteca Swing para a interface com o usuário e o ambiente de desenvolvimento NetBeans 7.1. Sua interface principal pode ser vista na Figura 1. O uso da linguagem Java torna a aplicação independente de plataforma e sistema operacional (Windows, Linux, etc.), bastando que o usuário tenha uma versão recente (ao menos 1.6) do Java *Runtime Environment* (JRE) instalado.

O formato de entrada escolhido foi o JSON (`http://www.json.org/`). O motivo dessa escolha sobre outros formatos tradicionais para codificação de corpus como

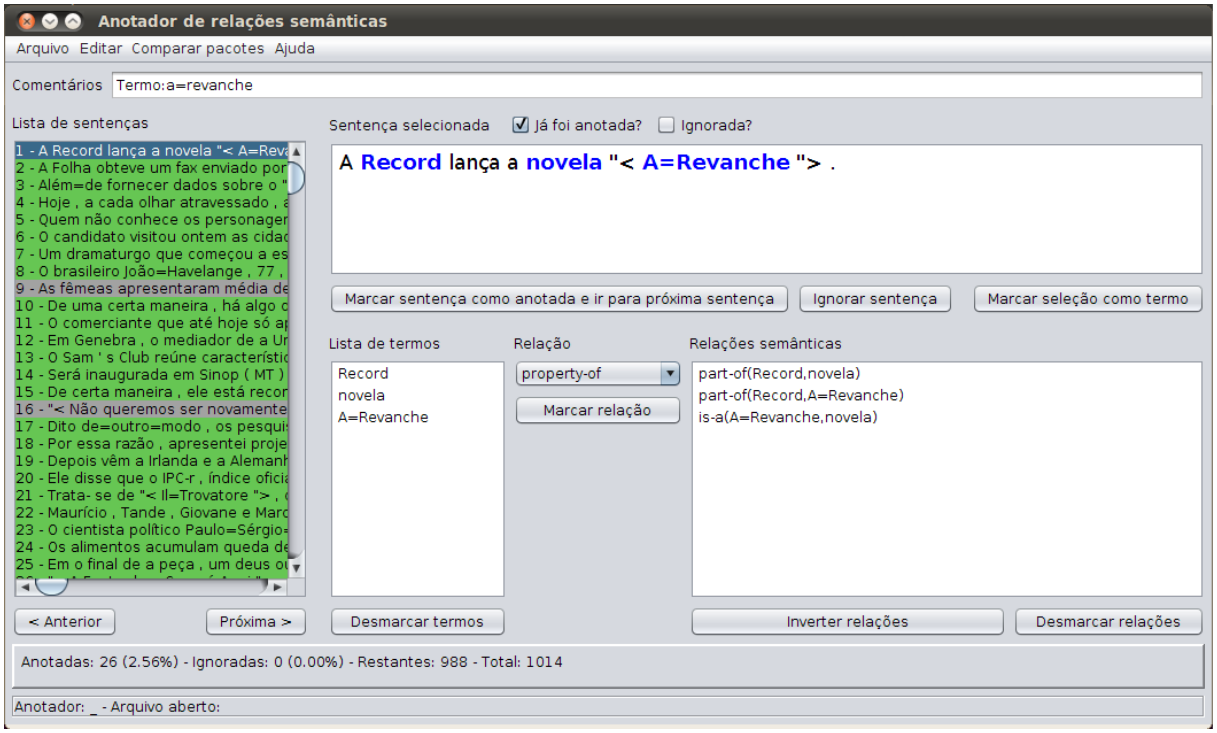


**Figura 1. Interface principal da ferramenta**

o Tiger XML [Mengel and Lezius 2000] e o XCES [Ide et al. 2000] é que o JSON é igualmente portável entre aplicações e fácil de ser processado computacionalmente porém é mais sucinto, resultando em arquivos menores.

A unidade mínima de anotação foi definida como um token, que neste trabalho é considerado uma sequência de quaisquer caracteres exceto o espaço em branco. Um termo é considerado uma sequência de tokens que corresponde a uma entidade no texto, e uma relação é definida como uma tripla <*relação*, *termo1*, *termo2*>, onde *relação* é uma das sete relações descritas na Tabela 1 e *termo1* e *termo2* são os termos que participam da relação. Por exemplo, na Figura 1 temos três relações demarcadas: *part-of(Record, novela)*, *part-of(Record, A=Revanche)* e *is-a(A=Revanche, novela)*.

Outra decisão tomada foi que termos não podem intersectar outros termos, ou seja, um mesmo token não pode fazer parte de dois termos distintos. Essa decisão visa simplificar a marcação dos termos.

O corpus inicial escolhido para anotação foi o CETENFolha (`http://www.linguateca.pt/cetenfolha/`), composto por textos jornalísticos e compreendendo cerca de 25 milhões de palavras, divididas em aproximadamente 1,6 milhões de sentenças anotadas morfossintaticamente pelo *parser* PALAVRAS [Bick 2000]. Uma amostra dessas sentenças foi selecionada de acordo com a frequência de ocorrência de sintagmas nominais (SNs) no corpus – foram escolhidas as sentenças que continham sintagmas que ocorriam de 18 a 51 vezes. Essa faixa foi delimitada pois representa, considerando a lista de frequências ordenada e acumulada de todos os SNs, cerca de 15 mil sintagmas que correspondem a uma faixa de 10% do total de SNs do corpus, cuja frequência está entre 40% e 50% da distribuição acumulada; não são os mais frequentes,

localizados no início da curva de distribuição, e nem os menos frequentes, localizados no fim da curva. O intervalo entre 40% e 50% foi definido pois resulta em uma boa variedade de sintagmas distintos (15 mil) com um número significativo de ocorrências (18 a 51).

As sentenças em que esses 15 mil SNs apareciam foram selecionadas, resultando em uma amostra de aproximadamente 230 mil sentenças a serem anotadas. Essas sentenças foram transformadas do formato de saída do PALAVRAS para o formato JSON aceito pela ferramenta e divididas em pacotes com cerca de 1000 sentenças. Como o processo de anotação do corpus está sendo realizado por dois anotadores, cada par de conjuntos de 1000 sentenças tem por volta de 100 sentenças em comum para o cálculo da concordância entre os anotadores.

A ferramenta possibilita a comparação entre dois conjuntos de dados, mostrando as sentenças comuns aos dois conjuntos e calculando a concordância entre as anotações feitas em cada um (Figura 2). Essa funcionalidade facilita a comparação entre as anotações feitas por dois anotadores distintos.

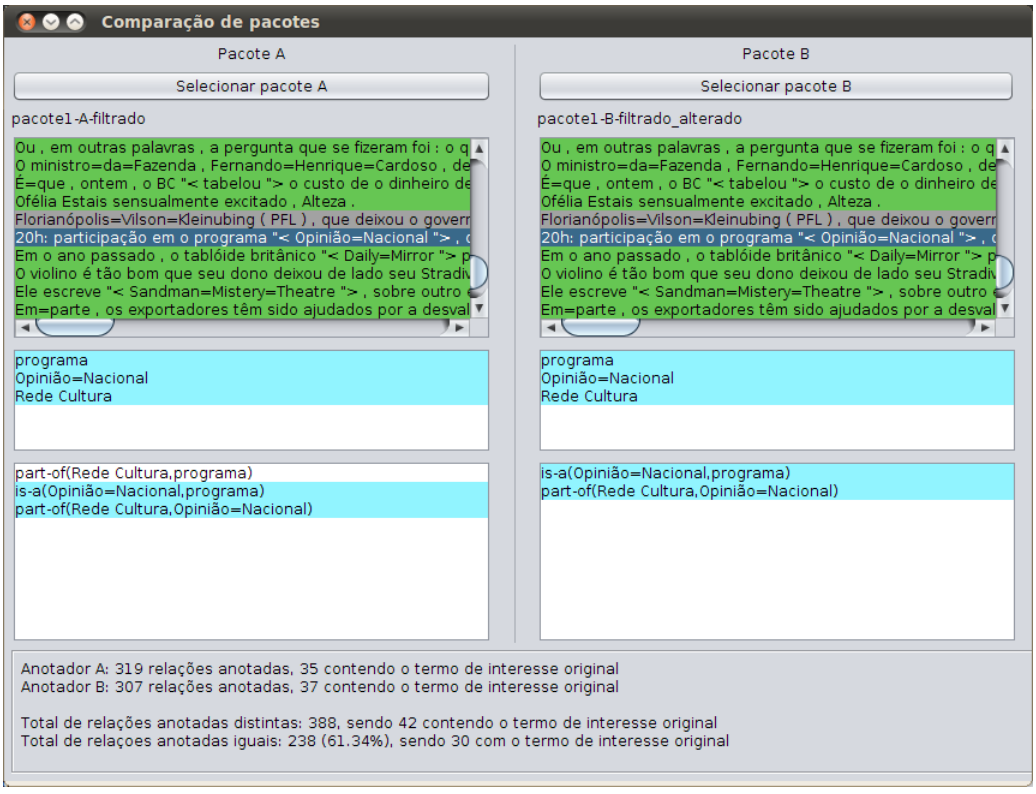


**Figura 2. Interface de comparação de anotações**

## 4. Resultados

Com o uso da ferramenta, um primeiro conjunto de aproximadamente 1900 sentenças foi completamente anotado por dois anotadores: cada anotador etiquetou cerca de 1000 sentenças, sendo que havia por volta de 100 sentenças comuns aos dois anotadores. O número de relações resultante dessa rodada de anotação é apresentado na Tabela 2 a seguir. Percebe-se que as relações mais frequentemente encontradas são a property-of, is-a, part-of e location-of, sendo as demais muito menos frequentes.

**Tabela 2. Número de relações marcadas por cada anotador**

| Relação | Anotador A | Anotador B |
|---|---|---|
| property-of | 1378 | 963 |
| is-a | 680 | 799 |
| part-of | 316 | 429 |
| location-of | 271 | 463 |
| effect-of | 39 | 64 |
| made-of | 14 | 31 |
| used-for | 6 | 45 |
| **Total** | 2704 | 2794 |

A taxa de concordância entre os dois anotadores, verificada sobre o conjunto de cerca de 100 sentenças comuns, é calculada como o número de relações anotadas da mesma forma por ambos dividido pelo número total de relações distintas anotadas. Essa razão, nesta primeira etapa de anotação, foi de 61,34%, sendo que do total de 388 relações semânticas distintas anotadas pelos dois anotadores haviam 238 anotadas da mesma forma. Em paralelo ao trabalho de anotação um manual de anotação está sendo desenvolvido para permitir que outros anotadores possam integrar o processo de anotação futuramente.

## 5. Conclusões e trabalhos futuros

Este artigo apresentou uma ferramenta de anotação de relações semânticas entre termos. Essa ferramenta é específica para a tarefa em questão e utiliza um formato próprio de representação, codificado na estrutura JSON. Ela está sendo utilizada para a anotação semântica de uma amostra do corpus CETENFolha, que por sua vez será usada em estudos sobre extração automática de relações semânticas em *corpora* escritos em português do Brasil, através do treinamento de modelos computacionais automáticos.

Embora a ferramenta já esteja funcional em sua forma atual, novas melhorias estão planejadas para versões futuras. Uma das funcionalidades a serem implantadas é permitir que o usuário customize as relações que queira anotar, possibilitando que outras relações além das 7 definidas possam ser anotadas.

## Referências


Baker, C. F., Fillmore, C. J., and Lowe, J. B. (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, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Bick, E. (2000). *The Parsing System"Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Denmark.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., and Pinkal, M. (2006). Salto – a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, Genoa, Italy.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Duran, M. S. and Aluísio, S. M. (2011). Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology*, Cuiabá/MT, Brazil.

Ide, N., Bonhomme, P., and Romary, L. (2000). An xml-based encoding format for syntactically annotated corpora. In *Proceedings of LREC 2000*, pages 121–126, Atenas, Grécia.

Mengel, A. and Lezius, W. (2000). An xml-based encoding format for syntactically annotated corpora. In *Proceedings of LREC 2000*, pages 121–126, Atenas, Grécia.

Minsky, M. (1986). *The Society of Mind*. Simon and Schuster.