Buscador Online do CINTIL-Treebank

Patricia Nunes Gonçalves, António Branco

Faculdade de Ciências da Universidade de Lisboa

Abstract

This paper describes the CINTIL-Treebank Online Searcher, a freely available online service to search and view the parse and dependency trees of the CINTIL-Treebank.

Keywords/Palavras-chave: parse tree; dependency tree; annotated corpus treebank; treebank search; árvores sintácticas; árvores de dependência; corpus anotado; busca em corpora.

1. Introdução

Recentemente vêm crescendo os estudos de pesquisa sobre corpora anotados. Em particular, tem crescido o interesse em corpora anotados com árvores que expressam relações de constituência e de dependência gramatical. Para optimizar esses estudos e para melhor tirar proveito desses recursos, tem sido disponibilizado, nas línguas mais utilizadas, ferramentas de buscas sobre corpora. Essas ferramentas permitem pesquisas de nível não trivial no sentido de encontrar no corpus anotado árvores que se conformam em um padrão de busca especificado por um utilizador. Esta tarefa seria de grande esforço caso fosse realizada manualmente sem a ajuda de uma ferramenta computacional.

Este trabalho tem como finalidade apresentar o CINTIL-Treebank e uma ferramenta de consulta online a árvores sintácticas deste treebank. Esta ferramenta se apresenta em forma de serviço online e está disponível sem custo a qualquer utilizador.

2. CINTIL-Treebank

O CINTIL-Treebank é um corpus anotado em que as frases estão associadas às suas árvores sintácticas e suas relações de constituência e dependência gramatical. O treebank é composto por um recorte do CINTIL-Corpus Internacional do Português (Barreto et al 2006), desenvolvido pela Universidade de Lisboa pelo Grupo REPORT do CLUL-Centro de Linguística[1] e pelo NLX-Natural Language and Speech[2] do Departamento de Informática.

O Corpus CINTIL está anotado com categorias morfossintácticas, lemas, informação de flexão e indicação de entidades nomeadas. Para a construção do treebank, a anotação realizada no CINTIL foi herdada e algumas frases desse corpus foi

[1] http://www.clul.ul.pt/index.php

[2] http://nlx.di.fc.ul.pt

alargada com árvores sintácticas de constituência e dependência, além de etiquetas de papéis semânticos.

2.1. Processo de anotação

O trabalho de anotação foi realizado por linguistas de acordo com o método de múltipla anotação independente seguida de adjudicação. A anotação foi realizada com o apoio da gramática computacional LXGram (Branco e Costa, 2008) que realiza processamento linguístico profundo de frases em Português.

O processo de anotação funciona da seguinte forma: Para cada frase, a gramática é usada para gerar todas as análises possíveis. O facto de recorrer a uma gramática computacional garante que os diferentes níveis de anotação são consistentes entre si.

Após este processamento automático, cada anotador humano tem que escolher a análise que considera correcta. Em caso de divergência entre anotadores na selecção manual, o adjudicador decidirá pela árvore correcta. Este processo de anotação garante grande confiabilidade nas informações geradas.

Neste momento, o CINTIL-Treebank está disponível com 1.204 frases, contendo um total de 10.387 tokens. O processo de anotação continua a ser realizado e novas frases estão a ser anotadas para aumentar o tamanho do treebank.

2.2 Árvores de Constituência

As árvores de constituência registam as habituais relações entre constituintes sintácticos segundo um esquema X-barra básico.

No CINTIL-Treebank, as árvores de constituência encontram-se anotados com três conjuntos de etiquetas: (i) categorias lexicais e sintagmáticas, (ii) funções gramaticais e (iii) papéis semânticos.

As etiquetas lexicais e sintagmáticas são acrónimos das designações em inglês das categorias. As etiquetas utilizadas para marcação de papéis semânticos foram inspiradas no trabalho de (Palmer et al 2005).

A tabela 1 mostra o conjunto das etiquetas de categorias lexicais e sintagmáticas.

Etiqueta	Descrição
A	Adjectivo
AP	Sintagma Adjectival
ADV	Advérbio
ADVP	Sintagma Adverbial
C	Complementador
CP	Sintagma Complementador

Etiqueta	Descrição
DEM	Demonstrativo
N	Nome
NP	Sintagma Nominal
P	Preposição
PP	Sintagma Preposicional
POSS	Possessivo

CARD	Cardinal			QNT	Quantificador
CONJ	Conjunção			S	Frase
CONJP	Sintagma Conjuncional			V	Verbo
D	Determinante			VP	Sintagma verbal

Tabela 1: Etiquetas de categorias lexicais e sintagmáticas

A tabela 2 mostra o conjunto das etiquetas de funções gramaticais usadas.

Etiqueta	**Descrição**
C	Complemento
DO	Objecto Directo
IO	Objecto Indirecto
M	Modificador
N	Relação de palavras de nome próprio
OBL	Complemento Oblíquo
PRD	Predicador
SJ	Sujeito
SP	Especificador

Tabela 2: Etiquetas de funções gramaticais

A tabela 3 mostra o conjunto das etiquetas de papel semântico utilizadas.

Etiqueta	**Descrição**		**Etiqueta**	**Descrição**
ADV	Adverbial		M	Modificador
ARG1	Primeiro Argumento		MNR	Modo/Maneira
ARG2	Segundo Argumento		NULL	Nulo
ARGA	Agente causativo de verbos com alternância causativa		PNC	Objectivo/Propósito
CAU	Causa		POV	Ponto de Vista
DIR	Direcção		PRD	Predicação secundária
EXT	Extensão		TMP	Tempo
LOC	Localização			

Tabela 3: Etiquetas de papel semântico

A figura 1 mostra um exemplo de árvore de constituência da frase “A nova ponte já tem nome” retirada no CINTIL-Treebank.

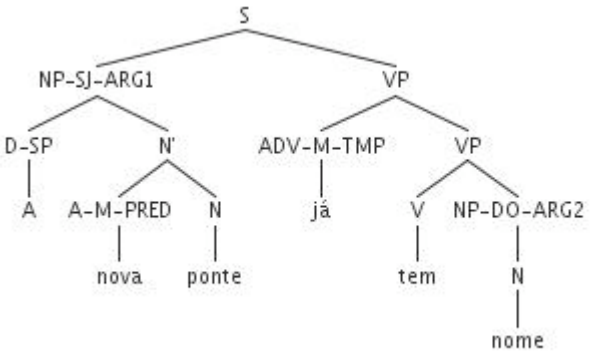

Figura 1: Árvore de Constituência

Em cada um dos nós da árvore de constituência, em muitos casos, a etiqueta é dividida em três partes, separadas por hífen “-”. A primeira parte indica a categoria sintagmática, a segunda parte indica a função gramatical e a terceira parte o papel semântico. Por exemplo, o nó marcado com `NP-SJ-ARG1` está associado como sendo um sintagma nominal (`NP`) que é sujeito (`SJ`) com papel semântico de primeiro argumento de predicação verbal (`ARG1`).

2.3 Árvores de Dependência

As árvores de dependência codificam as relações entre palavras de acordo com as funções gramaticais relevantes. Estas árvores são constituídas por arcos dirigidos que conectam palavras. Estes arcos estão decorados com funções gramaticais (Hellwig, 1986). A figura 2 mostra um exemplo de árvore de dependência para a frase “A nova ponte já tem nome” retirada do CINTIL-Treebank.

Figura 2: Árvore de Dependência

3. Buscador Online do CINTIL-Treebank

O Buscador Online do CINTIL-Treebank é um serviço online para a busca e visualização das árvores sintácticas e de dependência do CINTIL-Treebank.

Este serviço está a ser desenvolvido e mantido pelo NLX-Grupo de Fala e Linguagem Natural da Universidade de Lisboa e encontra-se disponível no endereço electrónico (`http://cintiltreebank.di.fc.ul.pt/`). Tem como objectivo servir de apoio para estudantes de computação e linguistas interessados em pesquisa baseada em corpus anotado ou de qualquer outra área que envolve o estudo gramatical da língua Portuguesa.

3.1. Consulta no Buscador Online CINTIL-Teebank

Para realizar uma busca por árvores de constituência, o Buscador recebe como entrada a descrição da estrutura com base no padrão de consulta da Tregex (Levy e Andrew, 2006), um motor de busca para árvores sintácticas.

A sintaxe de consulta apesar de ser bastante simples segue algumas condições de boa formação que devem ser respeitadas. Na tabela abaixo é apresentado a sintaxe e os símbolos usados para pesquisa nas árvores sintácticas. Esse tipo de sintaxe é usada combinando símbolos específicos usados na busca e as etiquetas descritas na secção 2.2.

Símbolo	**Significado**	**Exemplo**
A << B	A domina B	NP << N
A >> B	A é dominado por B	V >> VP
A < B	A domina imediatamente B	PP < P
A > B	A é imediatamente dominado por B	CONJ > NP
A $ B	A é irmão de B	NP $ CONJ
A .. B	A precede B	P .. POSS-M
A . B	A precede imediatamente B	CONJ . VP
A ,, B	A segue B	CARD ,, VP
A , B	A segue imediatamente B	D-SP , NP-C
A >>, B	A é o descendente mais à esquerda de B	VP >>, P
A >>- B	A é o descendente mais à direita de B	PP >>- N
A >, B	A é o primeiro filho de B	PP >, P
A >- B	A é o último filho de B	PP >- NP-C
A >i B	A é o *i-ésimo* filho de B	ADV >1 ADVP
A >: B	A é o único filho de A	N >: NP
@A	Etiquetas com a categoria sintagmática A	@NP

Tabela 4: Sintaxe e símbolos para pesquisa

Os símbolos podem ser combinados para aumentar a expressividade da consulta. Para ilustrar a sintaxe de consulta usando a combinação de símbolos, considere o exemplo: `S<VP<<NP-DO-ARG2`. Esta consulta realiza a busca por árvores sintácticas que contêm um nó `(S)` que domina imediatamente um sintagma preposicional `(VP)` e que também domina (não imediatamente) um sintagma nominal `(NP)` com função gramatical de objecto directo `(NP-DO)`. Quando a consulta é realizada, as frases que correspondem o padrão solicitado são mostradas ao utilizador como resposta, como mostra a figura 3.

Figura 3: Buscador Online CINTIL-Treebank

O utilizador pode parametrizar o número de frases retornadas. No exemplo acima são mostradas apenas as 5 primeiras frases que respeitam esse padrão. Para este padrão foram encontradas um total de 455 frases em todo o corpus. As próximas frases podem ser pesquisadas com a ajuda dos botões com setas que encontram-se acima das frases.

A árvores de constituência e de dependência podem ser visualizadas quando o utilizador escolhe a frase desejada. As imagens com as árvores aparecerão logo abaixo e a frase seleccionada fica indicada na tela. A figura 4 mostra as árvores de constituência e de dependência da primeira frase que foi seleccionada.

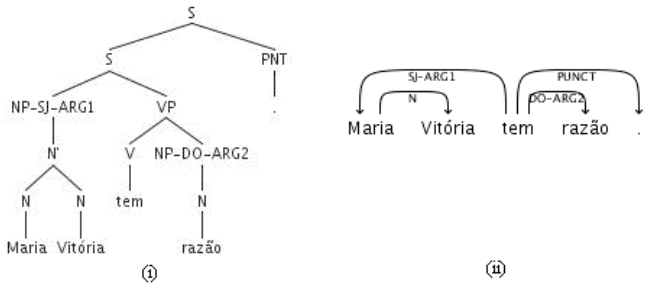

Figura 4: (i) Árvore de constituência – (ii) Árvore de dependência

3.1.1 Consulta avançada

O sistema permite a expansão da consulta, para isto, o utilizador pode fazer uso de expressões regulares, bastando para isso colocar a expressão a procurar entre barras. A pesquisa usando expressões regulares no Buscador Online do CINTIL-Treebank segue os símbolos usuais estabelecidos para este tipo de pesquisa:

- Alternância: as pesquisas alternativas são marcadas com uma barra vertical: "|". Por exemplo: **`/NP|VP/`** dá como resultado todas as árvores sintácticas com sintagmas verbais ou sintagmas nominais.

- Iteração: Para uma busca por iteração, os operadores ".*" (ponto e asterisco) faz com que o caracter ou expressão que o precede seja realizado zero ou mais vezes. Por exemplo: **`/NP.*/`** dá como resultado todas as árvores sintácticas em que a etiqueta inicie por NP, por exemplo: NP, NP-C, NP-M e NP-SJ.

- Delimitadores: Para delimitar o início e o fim de alguma etiqueta podemos usar os caracteres especiais "^" e "$". Esse tipo de pesquisa é útil quando se deseja procurar árvores sintácticas com uma composição de etiquetas gramaticais e papéis semânticos. Por exemplo: **`/^NP.*.ARG1$/`** dá como resultado todas as árvores sintácticas em que a etiqueta se inicia por NP e que tenha qualquer outra etiqueta no meio mas obrigatoriamente termine com a etiqueta ARG1, indicando o papel semântico de primeiro argumento, por exemplo: NP-DO-ARG1 e NP-SJ-ARG1.

3.1.2 Outras formas de consulta

Outras formas de consultas foram desenvolvidas com base na necessidade dos utilizadores. Uma delas é a consulta por palavras. As palavras encontram-se nas folhas das árvores sintácticas. Para realizar a pesquisa por palavras basta digitá-la na caixa de texto de pesquisa. Por exemplo:

A pesquisa por palavras está associada ao padrão em que ela se encontra no treebank, podendo estar escrita usando letra minúscula, maiúscula ou usando maiúscula e minúscula. Para cobrir tais casos na busca de resultados, a pesquisa deve explicitamente contemplar as diferentes formas de escrita usando o operador de alternância, como demonstra a imagem abaixo:

Todas as frases do CINTIL-Treebank possuem um identificador único. O identificador é mostrado ao utilizador juntamente com as respostas, após a pesquisa. Esse identificador serve para uma procura rápida e posterior quando alguma frase anteriormente seleccionada na pesquisa venha a servir de exemplo. Para realizar a pesquisa pelo identificador da frase, é necessário que o número correspondente devolvido na resposta seja anotado. A pesquisa é então feita usando a palavra *"ID:"*, como mostra o exemplo a seguir:

A consulta no Buscador Online do CINTIL-Treebank disponibiliza uma opção para busca por árvores que não contenham determinado padrão, o que é chamado de pesquisa invertida. Para realizar a pesquisa invertida é necessário acrescentar a palavra *"INV"* seguida de dois pontos ":" e logo em seguida serão devolvidas frases onde o padrão solicitado não foi encontrado.

Com o exemplo acima, todas as frases que não contenham um sintagma verbal serão seleccionadas e mostradas ao utilizador. Esse tipo de pesquisa é muito útil quando é necessário pesquisar algum fenómeno pouco frequente na língua.

3. Conclusão

Neste artigo apresentamos o CINTIL-Treebank. Este treebank foi construído com o apoio de uma gramática computacional para o Português. Esta gramática é usada para gerar todas as análises possíveis de uma frase, cabendo então aos anotadores humanos seleccionar a análise correcta entre todas as que foram geradas. Actualmente o treebank contém 1204 frases, sendo que o processo de anotação continua a ser desenvolvido com o objectivo de se alargar o corpus anotado.

Apresentamos também o Buscador Online do CINTIL-Treebank, um serviço online e disponível gratuitamente para pesquisa e estudo da língua portuguesa. A ferramenta de busca possui uma linguagem rica e poderosa que permite buscas de estruturas complexas no treebank. Como resultado é possível visualizar as árvores de constituência e de dependência deste treebank.

Referências

Barreto, Florbela; Branco, Antóinio; Ferreira, Eduardo; Mendes, Amália; Nascimento, Maria Fernanda; Nunes, Filipe e Silva, João (2006), Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project, Proceedings of the 5th LREC, 2006. Genova, Italy.

Branco, António e Costa, Francisco (2008). A Computational Grammar for Deep Linguistic Processing of Portuguese: LX-Gram, version A.4.1 Relatório Técnico. Universidade de Lisboa. Departamento de Informática

Hellwig, Peter. (1986). Dependency Unification Grammar. In Proceedings of the 11th Conference on Computational Linguistics (Bonn, Germany, August 25 - 29, 1986). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 195-198.

Levy, Roger e Andrew, Galen (2006). Tregex and Tsurgeon: tools for querying and manipulation tree data structures. In Proceedings of The International Conference on Language Resources and Evaluation -LREC 2006. Disponibilizado em http://nlp.stanford.edu/pubs/levy_andrew_lrec2006.pdf.

Palmer, Martha; Gildea, Daniel e Kingsbury, Paul (2005). The Proposition Banks: An Annotated Corpus of Semantic Roles. Computational Linguistics vol. 31, No. 1, pp. 71-106.