Portuguese Analysis with Tree Adjoining Grammars

Karin Christine Kipper & Vera Lúcia Strube de Lima
kipper@brpucrsm.bitnet vera@brpucrsm.bitnet

PUCRS - Instituto de Informática
Av. Ipiranga 6681 prédio 30 bloco 4
90619-900 PORTO ALEGRE RS
BRASIL

Abstract

This article approaches syntactical analysis of Portuguese language based upon a formalism called Tree Adjoining Grammars (TAGs) [JOSHI 85]. It briefly describes the formalism and its main operations, outlines a Portuguese subset for analysis, and presents a parser developed according TAGs concepts in order to validate an application of the formalism for this language.

1. Introduction

This article describes an experiment approaching syntactical analysis of Portuguese based on Tree Adjoining Grammars (TAGs) [JOSHI 75]. It briefly presents the TAG formalism, placing it among other description formalisms used for natural language processing, and introduces a prototype which is being developed in order to validate application of this formalism to Portuguese language.

The present work concerns sentence analysis at syntactical level, which can be viewed as a process with two main functions for natural language processing : the identification of the input components through association of tree structures to sentences, and regularization of the identified structure in order to minimize the number of trees for each sentence [GRISHMAN 86].

Although Context-Free Grammars (CFG) have been the most studied ones in order to describe natural language, purely context-free grammars are not adequate for this description [RICH 91].
Context-Sensitive Grammars (CSG) are also used for description of natural languages, however they have not been proven to be a suitable formalism for stating most grammatical constraints [GRISHMAN 86].

Categorial Grammars (CG) seem to be a tendency for natural language description, including several related formalisms, all involved with the foundations of modern syntactic and semantics theories [STEEDMAN 93].

Among the formalisms related to Categorial Grammars we can mention Tree Adjoining Grammars (TAGs) [JOSHI 75] [JOSHI 85], Lexical Functional Grammar [BRESNAN 82], Dependency Grammar [HUDSON 82] and Generalized Phrase Structure Grammar [GAZDAR 85]. These grammatical formalisms and linguistic theories are based on unification and specification of constraints for definition of the possible structures to be unified.

This article is organized in four items. After a brief introduction, we present the Tree Adjoining Grammars formalism, describing its main components and operations, We comment our steps toward construction of a syntactical analyzer for Portuguese language and make some consideration about the prototype described.

2. Tree Adjoining Grammars

Tree Adjoining Grammars were first described by [JOSHI 75], as a tree based system, where the basic component is a set of elementary trees. Each tree represents a minimal linguistic structure and is a domain of locality. A TAG comprises two kinds of elementary trees:

- initial trees, which are complete structures, with pre-terminals on the leaves;
- auxiliary trees, which must have exactly one leaf node with the same syntactic category of the root node.

The elementary trees localize dependencies, like agreement, sub categorization, etc. and must have at least one terminal node.

Sentences generated from a language defined by a TAG can be derived by the composition of an initial tree and elementary trees, through two operations: substitution and adjunction.

Substitution, as showed in Fig 1, inserts an initial tree (or a tree derived from an initial tree) on the correspondent leaf node in the elementary tree.

Adjunction, as showed in Fig 2, inserts an auxiliary tree on the correspondent node in an elementary or derived tree.

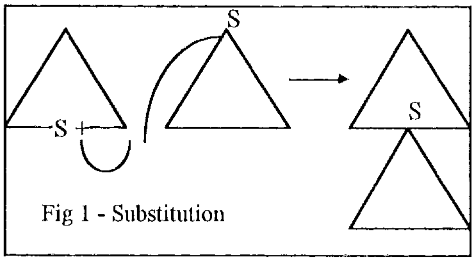


Fig 1 - Substitution

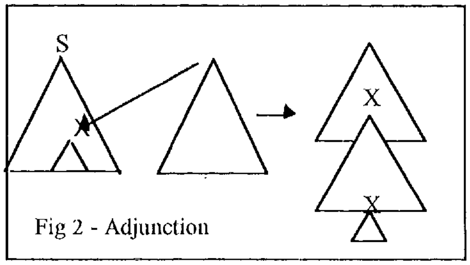


Fig 2 - Adjunction

The adjunction operation can be recursive, then an auxiliary tree can receive adjunction in itself. Adjunction allows an insertion of a complete structure on a node of another complete structure.

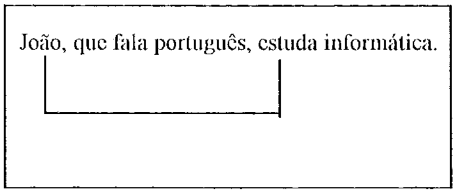
Adjunction makes TAGs a little more powerful then Context-Free Grammars (CFG), placing it in a class of grammars called Midly Context-Sensitive Grammars [JOSHI 85]. This operation preserves the dependencies among unbounded structures of the sentence.

3. Portuguese analysis with TAGs

Several research groups are working with Tree Adjoining Grammars. There are descriptions of grammars for French [ABEILLE 91], English [SCHABES 88], a study for German [RAMBOW 92], among other languages.

In order to analyze Portuguese language, there are many studies being developed, in Brazil and Portugal, which approach different formalisms. These researches focus punctual areas as lexical analysis [COURTIN 89], data-base queries using natural language [BIGOLIN 93], semantic analysis [FREITAS 93] [LUZ 93], etc.

In TAG formalism we can find aspects that help syntactic analysis of Portuguese, for example, the possibility to have unboundness dependencies, such as agreement, among nodes.



João, que fala português, estuda informática.

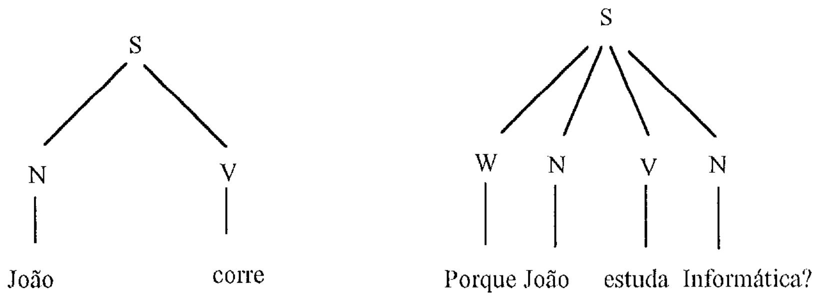
We are working on a grammar to describe Portuguese, and we are developing a syntactical analyzer for this grammar. One of the problems we faced was the absence of a description of the most common structures used for our language, something as "fundamental Portuguese", so we selected the subset to work with.

We decided by a large subset, which includes active and passive voice, relative and interrogative clauses, auxiliary and support verbs, and clitic pronouns.

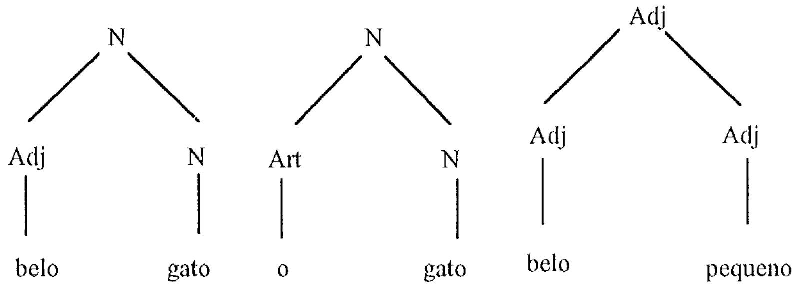
The syntactical categories included are verbs, nouns, pronouns, adjectives, adverbs, articles and prepositions. For each one of the categories there are syntactical traits associated like: concrete, abstract, number, gender, person, mode, voice, ...

The grammar is organized according to the formalism, using initial trees and auxiliary trees to describe surface structures of Portuguese language. These study was based on Portuguese normative grammars [ROCHA LIMA 92], and generative grammars [LOBATO 86].

Example of initial trees :



Example of auxiliary trees :



Its important to observe that each one of the nodes associated to a tree has traits used for unification, and can have dependency traits between unbounded nodes. These dependency traits are kept under an adjunction operation.

The first version of the syntactical analyzer, based upon TAGs, includes the acquisition of elementary trees, input of the sentence to be analyzed, construction of a solution tree (made by adjunction and substitution), and unification of the input sentence with the solution tree. Note that the analyzer must return all the derived trees for the given input sentence.

The elementary trees are supposed to contain information about the hierarchy of the nodes, type of that tree (relative, interrogative,...), operations that can be made on each node, and traits to be unified.
Syntactical analyzer input sentence comes from a morphological analyzer that splits this sentence in components such as words or expressions, associating them a set of traits.

Construction of the derived tree is made by adjunction and substitution operations over elementary trees. Unification compares traits of the input sentence with the traits described on TAG trees, producing the resulting trees.

Inclusion of semantic traits will allow us to upgrade this analyzer in a semantic-syntactic analyzer, anticipating evaluation of semantic traits to syntactical analysis, reducing the number of resulting trees.

4. Final remarks

In the scope of a project aiming to develop tools to treat Portuguese at morphological, syntactic and semantic levels, we started with morphological level, and we came to an implementation of a robust lexical-morphological analyzer through trie trees [STRUBE DE LIMA 93]. As a next step, we approached syntactical level looking for a formalism adequate to support Portuguese language. A large subset of this language was outlined, which should give rise to an experiment of implementation of algorithms and data structures for parsing Portuguese.

This seems to be the first study using Tree Adjoining Grammars for Portuguese language. Our contribution would state on description of a large subset of the language, construction of trees that represent syntactic structures for Portuguese, and development of a parser, according to the formalism.

We described around 300 inicial trees in order to cover the subset outlined, and developed a bottom-up LR parser working efficiently. We are now studying complementary data structures as a syntactical dictionary in order to improve the parser. This dictionary would be helpful to construct the solution tree, searching fastly the trees that can be used for a word. We are also adapting the output of the morphological analyzer in a model that fits the input of the syntactical analyzer developed.

Tree Adjoining Grammars formalism, to this moment, seems to present aspects that benefit treatment of Portuguese language in a robust way. Acquisition of new trees can be made easily, as well as describing semantic traits together with the syntactical ones.

5. Bibliography

[ABEILLE 91]

ABEILLE, Anne. "Une Grammaire Lexicalisée d'Arbres Adjoints pour le Français Application à l'analyse automatique". Thèse de Doctorat de linguistique. Université Paris 7, LADL, Janvier, 1991.

[BIGOLIN 93]

BIGOLIN, N. e CASTILHO, J. M. "Ferramenta de auxílio para a tradução de linguagens de especificação no desenvolvimento de sistemas de banco de dados". Simpósio Brasileiro de Banco de Dados, Campina Grande, 1993.

[BRESNAN 82]

BRESNAN, J., KAPLAN, R. "Lexical Functional Grammar: a formal system for grammatical representation". In: J.Bresnan (ed.), The Mental Representation of Grammatical Relations, MIT Press, 1982.

[COURTIN 89]

COURTIN, J. DUJARDIN, D., KOWARSKI, I, GENTHIAL, D., STRUBE DE LIMA, V.L. "Análise de textos escritos em português com PILAF: uma experiência e seus resultados". 18avas Jornadas de Informática e Investigación Operativa, Argentina, Agosto, 1989.

[FREITAS 93]

FREITAS, Sérgio, LOPES, José Gabriel. "Um sistema de representação do discurso utilizando DRT e a teoria do foco". X SBIA, Porto Alegre, 1993.

[GAZDAR 85]

GAZDAR, G.,KLEIN,E., PULLUM,G., SAG,I. "Generalized Phrase Structure Grammar". Harvard University Press, 1985.

[GRISHMAN 86]

GRISHMAN, R. "Computational Linguistics - An Introduction". Cambridge University Press, 1986.

[HUDSON 82]

HUDSON, Richard. "Word Grammar". Oxford: Blackwell, 1982.

[JOSHI 75]

JOSHI, A.K., LEVY, L. S., TAKAHASHI, M. "Tree Adjunct Grammars". Journal of the Computer and System Sciences, 10(1), 1975.

[JOSHI 85]

JOSHI, A. K. "Tree Adjoining Grammars : How much context-sensitivity is required to provide reasonable descriptions?". In: Natural Language Parsing, edited by D. Dowty, L. Kartunnen, A. Zwicky, Cambridge University Press, 1985.

[LOBATO 86]

LOBATO, L. "Sintaxe Gerativa do Português : da teoria padrão a regência e ligação". Belo Horizonte, Vigília, 1986.

[LUZ 93]

LUZ Filho, Saturnino de Brito. "Representação semântica de atitudes proposicionais através da teoria dos atos da fala". X SBIA, Porto Alegre, 1993.

[RAMBOW 92]

RAMBOW, Owen. "A Linguistic and Computational Analysis of the German Third Construction". 30th Annual Meeting COLING, July, 1992.

[RICH 91]

RICH, Elaine, KNIGHT, Kevin. "Inteligência Artificial". Mc Graw Hill, São Paulo, 1991.

[ROCHA LIMA 92]

ROCHA LIMA, C.H. "Gramática Normativa da Língua Portuguesa". Rio de Janeiro, José Olympio, 1992.

[SCHABES 88]

SCHABES, Yves, ABEILLE, Anne, JOSHI, Aravind. "Parsing Strategies with 'Lexicalized' Grammars: Applications to Tree Adjoining Grammars". COLING 88, Budapest, Hungary. August 1988.

[STEEDMAN 93]

STEEDMAN, Mark. "Categorial Grammar". In : Lingua 90. North-Holland, 1993.

[STRUBE DE LIMA 93]

STRUBE DE LIMA, V.L., KIPPER K.C. "Análise Morfológica de Textos Escritos em Português". Encontro de Processamento de Língua Portuguesa, Lisboa, 1993.