UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

ANDRÉA FEITOSA DOS SANTOS

MOVIMENTO DO VERBO E CATEGORIAS VAZIAS EM I E V EM UM FRAGMENTO DE GRAMÁTICA COMPUTACIONAL DO PORTUGUÊS

FORTALEZA

2009

ANDRÉA FEITOSA DOS SANTOS

MOVIMENTO DO VERBO E CATEGORIAS VAZIAS EM I E V EM UM FRAGMENTO DE GRAMÁTICA COMPUTACIONAL DO PORTUGUÊS

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Linguística, da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Mestre em Linguística.
Área de concentração: Descrição e Análise Linguística

Orientador: Prof. Dr. Leonel Figueiredo de Alencar

FORTALEZA

2009

ANDRÉA FEITOSA DOS SANTOS

MOVIMENTO DO VERBO E CATEGORIAS VAZIAS EM I E V EM UM FRAGMENTO DE GRAMÁTICA COMPUTACIONAL DO PORTUGUÊS

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Linguística, da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Mestre em Linguística.

Aprovada em 10/11/2009.

BANCA EXAMINADORA

_____

Professor Doutor Leonel Figueiredo de Alencar (Presidente)

Universidade Federal do Ceará - UFC

_____

Professor Doutor Antônio Luciano Pontes (1º. Examinador)

Universidade Estadual do Ceará - UECE

_____

Professora Doutora Rosemeire Selma Monteiro-Plantin (2ª. Examinadora)

Universidade Federal do Ceará - UFC

À minha irmã Maria Zélia,

Dedico

AGRADECIMENTOS

Ao Professor Leonel Figueiredo de Alencar, pois sem ele essa dissertação jamais existiria.

À CAPES, pelo apoio financeiro concedido com a manutenção da bolsa de estudos.

À Maria Elias Soares e Márcia Teixeira, por toda a contribuição dada ao andamento dessa pesquisa.

À professora Rosemeire Selma Monteiro-Plantin e ao professor Luciano Pontes, por aceitarem participar da banca examinadora desta dissertação.

Às professoras Ana Cristina Pelosi e Socorro Aragão, pelas maravilhosas aulas que proporcionaram à turma de mestrado de 2007.

Aos colegas da turma de mestrado 2007, pelas reflexões, conversas diálogos, críticas e bate-papos.

À Antônia e ao Eduardo, pelo ótimo serviço que prestam à coordenação do Programa de Pós-Graduação em Linguística.

Ao meu marido Juliano, por todo cuidado concedido a mim e a minha filha Helena com as coisas do nosso dia a dia, enquanto eu estive ocupada com esta dissertação e por toda ajuda dada à parte gráfica deste trabalho.

À minha irmã Zélia, por ter me dado as melhores coordenadas, desde a graduação até aqui.

À minha amiga Geórgia, pelo *Abstract*.

"Debulhar o trigo
Recolher cada bago do trigo
Forjar no trigo o milagre do pão
E se fartar de pão

Decepar a cana
Recolher a garapa da cana
Roubar da cana a doçura do mel
Se lambuzar de mel

Afagar a terra
Conhecer os desejos da terra
Cio da terra, a propícia estação
E fecundar o chão"

(O Cio da Terra, Milton Nascimento /
Chico Buarque)

RESUMO

Esse trabalho possui um recorte teórico-metodológico que se decompõe em dois domínios complementares: o Linguístico e o Computacional/Implementacional. Pelo seu cunho computacional, o escopo primeiro desse trabalho está diretamente ligado ao processamento de língua natural (PLN). Desse modo, implementa-se uma análise sintática automática (*parsing*) de expressões de língua portuguesa em programas da biblioteca em *Python* do NLTK, cujas análises são representadas em forma de configurações arbóreas que demonstram categorias vazias de sentenças finitas do português. Ainda pelo cunho computacional, esse trabalho elabora um fragmento de gramática, modelado para capturar traços específicos da estrutura linguística do português, com base no modelo formal de descrição linguística conhecido como Gramática Livre de Contexto (CFG) Baseada em Traços, com a finalidade de demonstrar como a biblioteca de programas do NLTK dá suporte à realização dos analisadores sintáticos na análise da estrutura de traços. Pelo seu cunho lingüístico, analisa-se, de acordo com a Teoria X-barra e o Programa Minimalista, frases nas variantes europeia e brasileira da língua portuguesa, obtidas de pesquisas em *corpora* eletrônicos disponíveis na *web*. E ainda nesse trabalho, descreve-se e discute-se a categoria IP (sintagma flexional) dentro da sua estrutura hierárquica de constituintes, de acordo com a hipótese da operação sintática de movimento visível e não visível dos elementos linguísticos, especificamente o movimento do verbo.

Palavras-chave: PLN. *Parsing*. *Python*. NLTK. Gramática Livre de Contexto. *Corpora* eletrônicos. Teoria X-barra. Programa Minimalista. Categoria IP. Operação sintática de movimento. Categorias Vazias.

ABSTRACT

This work has a theoretical and methodological framework that is divided into two complementary areas: the Language and Computational/Implementacional. For its computational stamp, the first scope of this work is directly linked to the processing of natural language (PNL). Thus, it implements an automatic syntactic analysis (parsing) of expressions of Portuguese in a computational program from the Python library of NLTK, whose tests are represented in tree configurations that show slash categories of finite sentences of Portuguese. Although the stamp computing, this research elaborates a grammar fragment, modeled to capture specific features of linguistic structure of Portuguese, based on the formal model of description linguistic known as Context Free Grammar (CFG) Based on Features, with the purpose of demonstrate how the library of NLTK programs supports the implementation of parsers for analyzing the feature structure. For its stamp of language, it analyzes, according to X-bar Theory and the Minimalist Program, sentences in European and Brazilian variants of Portuguese, obtained from surveys in electronic corpora available on the web. And this work describes and discusses the category IP (inflectional phrase) within the hierarchical structure of constituents, according to the hypothesis of syntactic operation of visible and invisible movement of elements of language, specifically the movement of the verb.

Keywords: PLN. Parsing. Python. NLTK. Context Free Grammar. Electronic *corpora*. Category IP. X-bar theory. Minimalist Program. Syntactic operation of movement. Slash Category.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EPP	Princípio de Projeção Estendida
IP	Sintagma Flexionado
VP	Sintagma Verbal
I	Flexão
V	Verbo
PLN	Processamento de Línguas Naturais
NLTK	*Natural Language Toolkit*
CFG	Gramática Livre de Contexto (do inglês *Context Free Grammar)*
PM	Programa Minimalista
PP	Princípios e Parâmetros
DP	Sintagma Determinante
PB	Português Brasileiro
FL	Faculdade da Linguagem
GU	Gramática Universal
Compl	Complemento
X''	X- duas barras
X'	X- barra
Spec	Especificador
N	Nome
P	Preposição
A	Adjetivo
PP	Sintagma Preposicionado
D	Determinante
AdvP	Sintagma adverbial
NP	Sintagma nominal
DS	Estrutura profunda (do inglês, *Deep Structure*)
SS	Estrutura de superfície (do inglês, *Surface Structure*)
LF	Forma lógica (do inglês, *Logic Form*)
PF	Forma fonética (do inglês, *Phonetic Form*)
SVO	Sujeito – verbo – objeto
ECP	Princípio de Categoria Vazia
NVAdvN	Nome – verbo – advérbio – nome
SVAdvO	Sujeito - verbo - advérbio – objeto
ModVP	Modificador de VP

SUMÁRIO

1 INTRODUÇÃO

A lingüística nem sempre é vista como uma disciplina científica que pode vir a contribuir com a ciência do modo como contribuem disciplinas de áreas como a tecnológica. Pensa-se freqüentemente em lingüística como uma disciplina das 'letras', que se comunica apenas com a literatura e com o ensino de língua portuguesa. Muito embora, pense-se que a lingüística pode estender sua comunicação a disciplinas como a Sociologia e Antropologia. Mas, o que podemos pensar de uma disciplina que traz atrelado ao seu nome o termo 'computacional'? Como as 'letras' podem colaborar com pesquisas tecnológicas, especialmente na área da Computação?

Há algumas décadas vem sendo desenvolvida, principalmente nos Estados Unidos, uma disciplina que aterrissou no campo da ciência da computação e vem colaborando significativamente com o desenvolvimento de programas e softwares de máquinas que simulam o comportamento humano. Podemos nos referir àquelas máquinas comumente utilizadas por empresas de telecomunicações como atendentes virtuais. Estes atendentes virtuais necessitam de um conjunto de informações diretamente ligado ao sistema fonético de uma língua, uma vez que para interagir com alguém na linha, a máquina precisa desempenhar as funções de 'ouvir' e 'falar'. Desse modo, cabe ao lingüista desenvolver um modo de implementar no programa que desempenha estas funções um processador de fala que seja capaz de compreender o que está sendo dito pelo usuário e que também seja capaz de interagir com ele através de perguntas ou sugerindo opções.

Outro exemplo prático de aplicação computacional da lingüística é largamente distribuído na internet na forma de tradutores automáticos de expressões lingüísticas. É possível encontrarmos tradutores na *web* que podem traduzir palavras, frases, textos e expressões idiomáticas. Para que um programa deste funcione é necessário que ele faça uso de conhecimentos sintáticos e semânticos na composição do vasto conjunto de informações lingüísticas que o constituem. O que podemos pensar é que se alguém está usando o programa para traduzir expressões do português para o inglês, este usuário espera que este programa seja 'inteligente' o bastante para traduzir, por exemplo, uma expressão idiomática como 'bateu as botas', não apenas no sentido literal da expressão, mas no sentido conotativo, ou seja, 'morreu'. Desse modo, cabe ao lingüista pensar em como atribuir ao programa a capacidade de fazer interpretações não apenas sintáticas, mas também semânticas das expressões lingüísticas solicitadas.

Com os exemplos acima observamos que para que sejam desenvolvidas máquinas 'pensantes' que desempenhem comportamentos típicos dos humanos, é imprescindível que um sistema lingüístico seja incorporado à máquina, pois é impossível conceber o ser humano sem pensar no aspecto lingüístico que o compõe. E para que os programas computacionais façam uso dos diversos tipos do conhecimento lingüístico é que certas pesquisas na área da lingüística computacional vêm sendo desenvolvidas.

Há pesquisas que tentam simular o modo como o conhecimento lingüístico é processado pela mente. Outras tentam implementar em máquinas pensantes os diversos tipos de conhecimento lingüístico do ser humano: fonético, morfológico, sintático, semântico e pragmático. Nossa pesquisa tenta contribuir com pesquisas de cunho linguístico-computacional, desenvolvendo um meio de fazer com que fragmentos de gramática se tornem objetos computacionalmente tratáveis.

Em nosso trabalho, especificamente, falaremos do modo como se pode implementar regras particulares da língua portuguesa em programas computacionais exclusivos para a realização de análises sintáticas automáticas. No caso dos analisadores sintáticos automáticos, o desafio é dar conta das especificidades de cada língua, principalmente no tocante a casos como os de concordância nominal, os de subcategorização verbal e categorias vazias.

De acordo com uma teoria da gramática, segundo a qual o conhecimento sintático que o falante tem de uma língua se organiza em forma de um sistema de regras que restringem o modo como os elementos lingüísticos se combinam para formar constituintes sintáticos, é possível se pensar na utilização de meios formais na descrição e na explicação das ocorrências dos fenômenos sintáticos. E ainda, com base na idéia de que a linguagem ocorre na mente de modo computacional, ou seja, de modo lógico e sistemático, centralizamos nosso objetivo em torno da demonstração de como uma mini-gramática, contendo um conjunto de regras sintáticas e regras lexicais pode ser implementada em um programa computacional de análise sintática, demonstrando, principalmente, o mecanismo desenvolvido pelo linguista para que o programa não analise sentenças agramaticais.

Com base nos níveis de representação do movimento do verbo em expressões finitas da língua portuguesa, acreditamos que em português brasileiro o movimento de subida do verbo pode ocorrer de forma não-visível. Para nós é a posição do advérbio entre o sujeito e

o verbo o indício de que o fenômeno de subida do verbo também ocorre de forma não-visível. Com base no Minimalismo, descrevemos, analisamos e principalmente, implementamos computacionalmente análises sintáticas de sintagmas do português como o sintagma flexionado, o sintagma determinante, o sintagma nominal, o sintagma verbal e o adverbial.

No caso específico dessa dissertação acreditamos que do ponto de vista puramente linguístico os assuntos tratados podem contribuir com pesquisas que possam preencher lacunas na área de descrições dos sintagmas da língua portuguesa. Silva (1996), com base no Princípio de Projeção Estendida (EPP), proposto por Alexiadou e Anagnostopolou (1998), aborda o sintagma flexionado (IP) em relação à natureza do expletivo que ocupa a posição de especificador do sintagma flexionado em sentenças inacusativas. Em nosso trabalho descrevemos IP em relação ao sintagma verbal (VP), visto que a categoria flexão (I), é o alvo do movimento do verbo (V).

Já do ponto de vista linguístico-computacional, o tema tratado pode colaborar ainda mais com o processamento de línguas naturais (PLN), tornando viável a implementação computacional em um número maior de linguagens de programação. Othero (2005) faz uma descrição sintática de expressões linguísticas do português visando à implementação computacional em linguagem *Prolog*. Entretanto, implementações computacionais para processamento de línguas naturais podem ser feitas, segundo Bird, Klein e Loper (2009), utilizando-se ao invés de *Prolog*, uma linguagem de programação mais simples, contudo poderosa, como é o caso de *Python*. Com base nisso, fazemos uma implementação computacional com o suporte da biblioteca de programas do NLTK (Natural Language Toolkit), utilizando para isso um formalismo que viabiliza o tratamento computacional das sentenças a serem analisadas, o *Context Free Grammar* (CFG).

Com base nessas orientações, nosso estudo tem como objetivos: 1. fazer uma análise de *inputs* do fragmento de gramática modelado em CFG, utilizando uma ferramenta computacional própria para o processamento de línguas naturais, especificamente um programa computacional de análise sintática (*parser*), capaz de analisar frases gramaticais da língua portuguesa, a partir da interação entre o léxico e as regras que constituem o nosso fragmento; 2. elaborar um fragmento de gramática, modelada de acordo com o formalismo CFG, capaz de capturar aspectos linguísticos como concordância nominal, acordo e projeções sintagmática e lexical em sentenças finitas do português; 3. fazer uma descrição sintática gerativo-transformacional do sintagma flexionado, com base na Teoria X-barra e no Programa

Minimalista (PM), a partir de uma discussão em torno das operações de movimento do verbo e das categorias vazias geradas em I e V.

Para alcançar e demonstrar nossos objetivos, este trabalho teórico-analítico-implementacional sobre o movimento de verbo dentro do sintagma flexionado do português está organizado em quatro capítulos, descritos nesta dissertação da seguinte forma:

No capítulo 2, apresentamos um apanhado teórico dos pressupostos que norteiam nossa pesquisa: a. expomos noções centrais da gramática gerativa e da sintaxe formal e à luz destas noções recorremos ao módulo da gramática gerativa Teoria X-barra para elucidar ideias como projeção de categorias, mecanismo de adjunção, movimento de constituintes dentro de expressões linguísticas, e diferenças entre categorias funcionais e categorias lexicais e no Programa Minimalista para explicar a diferença entre movimento visível e não visível de V; b. apresentamos os critérios adotados na utilização de modelos formais do funcionamento da linguagem e na implementação computacional de fragmentos de gramática elaborados com base nestes modelos.

No terceiro capítulo fazemos um levantamento das opções metodológicas adotadas em nosso trabalho. Nesse capítulo, apresentamos também a abordagem computacional da constituição dos *corpora*. Para esse fim, utilizamos uma ferramenta computacional chamada de "O Constructor", desenvolvida por Alencar (2002), que funciona como um tradutor automático com o fim de buscar na *web*, em *corpora* eletrônicos, expressões semelhantes à expressão linguística da busca.

No quarto capítulo, propomos uma revisão e uma discussão dos mecanismos envolvidos na estrutura do IP, a partir da descrição dos tipos de expressões linguísticas retirados de *corpora* tanto do português brasileiro como do português europeu. Será feita uma análise de sentenças simples retiradas dos nossos *corpora,* segundo critérios baseador nas condições de restrições de movimento de núcleos propostas por Raposo (1995) e no movimento de núcleo dentro de IP, segundo o Programa Minimalista em sua primeira versão.

Por fim, no último capítulo, mostramos a construção de um fragmento de gramática implementado computacionalmente capaz de analisar alguns tipos de expressões em língua portuguesa a partir da hipótese do movimento de subida do verbo de modo visível e da hipótese de o verbo permanecer também *in situ*.

2 SINTAXE GERATIVA: FORMALISMO E APLICAÇÃO COMPUTACIONAL

Nossos estudos se inserem em um contexto cujos postulados integram os fundamentos da teoria gerativa de investigação linguística. Acreditamos em língua, aquela adquirida naturalmente por nós, como um sistema modular de conhecimentos interiorizados na mente humana prontos para ser acessados. Conforme Raposo (1992. p. 28-30), essas regras atuam sobre o conhecimento linguístico de forma computacional resultando em representações mentais das formas linguísticas.

Conforme essa teoria, a gramática de uma língua é um sistema formado por princípios universais regidos por uma espécie de gramática internalizada e por parâmetros que vão sendo fixados durante o processo de aquisição da linguagem. O modelo que se baseia nesse sistema é chamado de **Princípios e Parâmetros** (doravante PP), porque sob esse ponto de vista, a língua é um sistema regido por regras (princípios rígidos) e um conjunto de valores fixados (parâmetros) obtidos a partir do meio linguístico, durante o processo de aquisição da linguagem pela criança.

Segundo Raposo (1992, p.52-55), as propriedades centrais da linguagem são determinadas por princípios e estruturas mentais de conteúdo especificamente linguístico. Com base nesse pressuposto, os princípios e os parâmetros de uma língua são motivos para se fazer pesquisas sobre as regras que regem as estruturas das línguas naturais. Desse modo, possíveis realizações gramaticais das línguas devem ser descritas e analisadas, até que se chegue às regras que a descrevam. De acordo com a arquitetura da linguagem no modelo de PP, uma investigação linguística deve estudar os aspectos da mente de um indivíduo relacionados com a compreensão que ele tem da linguagem e com o uso que faz dela.

Atualmente, novos rumos têm sido dados ao modelo PP com o intuito de diminuir a arquitetura desse modelo. Trata-se do modelo denominado **Programa Minimalista**, cujo objetivo é tentar minimizar e adequar o seu aparato técnico para uma melhor explicação dos fenômenos linguísticos (MODESTO, 2009, p.2).

A gramática gerativa constitui uma tentativa de formalização dos fatos linguísticos. O módulo da gramática gerativa que representa de modo formal os fenômenos sintáticos da língua chama-se de **Teoria X-barra.** Conforme esse modelo, uma expressão linguística constitui uma estrutura de constituintes, cujos núcleos são responsáveis pelo nível de projeção dos elementos linguísticos, que se dá pela combinação de cada núcleo com especificadores e complementadores.

Segundo Raposo (1992), esse tipo de tratamento só pode ser aplicado se o princípio da criatividade for considerado, pois para que uma gramática seja gerativa ela precisa traduzir o aspecto criativo da linguagem por meio de regras e processos explícitos, precisos e de aplicação automática.

Do ponto de vista gerativista, acredita-se que as propriedades de expressão de uma língua podem ser descritas em regras de uma **Gramática Livre de Contexto** (do inglês *Context Free Grammar*) que permitem construir um modelo da gramática mental para a implementação computacional em analisadores automáticos, visando a modelar as estruturas mentais do sistema linguístico e a explicar da forma como os princípios e as regras atuam na gramática mental.

Nossa discussão parte da relação entre a posição do advérbio em sentenças finitas do português e o fenômeno de movimento do verbo (V). Mais especificamente, o fenômeno de subida de V, na sintaxe visível, em direção à categoria funcional chamada de flexão (I, do inglês *inflection*). Com base nisso, demonstraremos como a categorias vazia pode ser gerada em V, como conseqüência de uma aplicação visível do movimento do verbo para I, ou como a categoria vazia em I pode ser decorrência da inexistência de representação fonética de I, extraída do léxico. Nossa demonstração será feita através da implementação computacional em programas próprios para o processamento automático de línguas naturais, codificados na linguagem de programação chamada *Python*.

Partimos dos princípios da Teoria X-barra para discutir o fenômeno de movimento de subida de V dentro do sintagma flexionado (IP), em sentenças finitas do português. Consideram-se sentenças finitas aquelas constituídas por um verbo flexionado, cuja flexão tem uma intrínseca relação de concordância de número e pessoa com o sintagma determinante (DP), que funciona como o sujeito da sentença, i.e. o DP na posição de especificador de IP. Faremos uma discussão em torno do movimento de V, da categoria vazia gerada e da posição dos atributos.

Como já foi dito, partimos da hipótese de que, dentro de IP, V faz um movimento de subida de encontro a I e gera uma categoria vazia. De acordo com essa hipótese, a posição do advérbio entre o DP sujeito e o verbo é um indício de que em português essa ordem dos elementos resulta da inexistência do movimento do verbo em sintaxe visível.

Com base nos níveis de representação do movimento de V acreditamos que em português brasileiro, de modo semelhante ao inglês, o movimento de subida do verbo pode ocorrer de forma não-visível. Para nós é a posição do advérbio entre o sujeito e o verbo o indício de que o fenômeno de subida de V também ocorre de forma não-visível.

O nosso argumento para sustentar essa hipótese de que a categoria vazia pode ser gerada em V ou em I, está relacionado com a retirada dos elementos fonéticos antes do movimento do verbo. Nós nos baseamos no contraste, feito por Raposo (1999, p.33), entre o português e o inglês, em relação à posição dos advérbios 'brutalmente' e '*brutally',* retirados dos exemplos[1] abaixo:

a) Eles agrediram brutalmente o prisioneiro

b) *Eles brutalmente agrediram o prisioneiro

c) They brutally hit the prisioner

d) *They hit brutally the prisioner

Em sua discussão, Raposo (1999, p.33) defende que (2) não é uma sentença gramatical do português, assim como (4) não o é em inglês. O asterisco utilizado no início da sentença é a forma utilizada pelo autor para demonstrar a agramaticalidade da expressão linguística. Para ele, a ordem dos elementos em (2) não é possível.

Segundo Chomsky (1957, p.13), a estrutura das sequências gramaticais de uma língua constitui a gramática de uma língua L. Esta gramática será um esquema que gera todas das sequências gramaticais de L e nenhuma das agramaticais.

Com base nas ideias de Chomsky (1957), assumimos a discussão feita em torno do que se pode considerar como 'gramatical' e 'agramatical'. Segundo Chomsky (1957, p.13), uma forma de testar a adequação de uma gramática proposta por L é determinar se as sequências que ela gera são ou não atualmente gramaticais, ou seja, aceitáveis por um falante nativo.

Desse modo distribuímos e discutimos o aparato teórico desse trabalho do seguinte modo: na seção 2.1 deste capítulo faremos um breve percurso para demonstrar os

[1] Exemplos retirados de Raposo (1999, p. 33)

desdobramentos da gramática gerativa desde *Syntactic Structures* até a teoria de Princípios e Parâmetros na abordagem do conhecimento linguístico de um indivíduo.

A seguir, nas seções 2.2 e 2.3 apresentamos um apanhado teórico dos pressupostos que norteiam nossa pesquisa, como a sintaxe formal e o módulo da gramática gerativa Teoria X-barra, para elucidar ideias como projeção de categorias, mecanismos de adjunção e movimento de constituintes dentro de expressões linguísticas, assim como a diferença entre categorias funcionais e categorias lexicais.

Nas duas últimas seções deste capítulo, demonstramos o embasamento teórico adotado na utilização de modelos formais do funcionamento da linguagem como as CFGs e as CFGs baseadas em traços e na implementação computacional de fragmentos de gramática elaborados com base nesses modelos.

2.1 De *Syntactic Structures* à Princípios e Parâmetros

Lingüistas gerativos defendem que a mente de cada indivíduo possui aspectos dedicados especificamente à linguagem. Esses aspectos compõem o que Raposo (1992, p.16) chama de **Faculdade da Linguagem** (doravante, FL). Segundo Raposo (1999, p.17), ao nascermos, a FL encontra-se em um estado inicial. Esse estado se desenvolve até chegar a um estado final. O estado final é em parte determinado pelo ambiente linguístico onde vivemos e em parte determinado por princípios internos uniformes para toda espécie humana.

Com base nesta FL muitos foram os desdobramentos da teoria gerativa desde sua origem até os dias atuais. Da ***Syntactic Structures*** até **Princípios e Parâmetros**, estas teorias foram passando a ser menos descritivistas, tiveram seu formato simplificado, impuseram, cada vez mais, maiores restrições ao formato das regras, e chegaram, por conseguinte, a um modelo apenas de princípios extremamente gerais, distribuídos pelos vários componentes de domínios da linguagem.

Em ***Syntactic Structures,*** passou-se a se supor que por trás da língua há uma capacidade dos falantes em produzir exatamente os enunciados que podem ser feitos. De acordo com Neto (2005, p.99), "a questão fundamental da teoria é a determinação das regras que regem o conhecimento compartilhado sobre os enunciados que podem e os que não podem ser produzidos e é este conhecimento que precisa ser descrito e explicado". Segundo Neto (2005, p.101) o modelo de análise de *Syntactic Structures* consiste de dois componentes sintáticos, um que forma expressões e outro que transforma expressões, e um componente morfofonêmico que atribui leituras fonológicas ao *output* do componente transformacional.

Na concepção de gramática da **Teoria *Standard***, a língua era o resultado da aplicação de regras de reescrita categorial e regras transformacionais. Sendo as primeiras as responsáveis pela derivação da estrutura profunda das frases. As regras transformacionais aplicavam-se sobre esta estrutura sucessivamente e geravam a estrutura de superfície.

Depois disso, em resposta a esse modelo demasiado descritivista e de regras muito flexíveis, surgiu a **Teoria *Standard* Estendida**. Uma reformulação do modelo anterior, que simplificou o seu formato através da redução das regras de reescrita e da formulação de princípios gerais da linguagem. Isto é, formulam-se menos regras particulares às línguas e se atribuem à **Gramática Universal** (doravante GU) [2] os princípios gerais.

No entanto, o número de regras ainda era grande e daí surgiu a necessidade de se restringir ainda mais os princípios, de modo a que fossem extremamente gerais. Desse modo, o modelo da Teoria *Standard* Estendida, constituído por regras e princípios, foi revisto e dele surgiu o modelo de PP.

Conforme Princípios e Parâmetros, a gramática de uma língua é um sistema formado por princípios universais regidos por uma espécie de gramática internalizada e por parâmetros que vão sendo fixados durante o processo de aquisição da linguagem. Nesse modelo a língua é um sistema regido por regras (princípios rígidos) e um conjunto de valores fixados (parâmetros) obtidos a partir do meio linguístico, durante o processo de aquisição da linguagem pela criança.

2.2 Sintaxe formal

Inicialmente, é necessário dizer que concebemos esse trabalho a partir da distinção feita entre sintaxe e gramática, proposta por Sag, Wasow, Bender (2003, p.1-9). Para eles, o termo 'gramática' deve ser empregado amplamente de forma que possa abranger todos os aspectos da estrutura da língua, no tocante aos componentes que formam o conhecimento linguístico de um indivíduo. O termo deve abranger o universo que compreende a semântica, a morfologia, a fonologia e a sintaxe, pois muitos fenômenos em línguas naturais envolvem mais que um desses componentes.

No entanto, é especificamente o termo sintaxe que se refere aos tipos de combinações de palavras dentro dos sintagmas e dos sintagmas dentro de frases. Portanto,

[2] Trata-se de termo próprio do Gerativismo. Para melhor compreensão desse trabalho sugerimos conhecimentos prévios de alguns aspectos dessa teoria.

estudar sintaxe pode ser um meio importante para a compreensão dos processos mentais e cognitivos envolvidos no uso da língua.

A língua, segundo Sag, Wasow, Bender (2003, p.9-16), é um corpo imensamente rico e sistemático de conhecimentos inconscientes que envolvem diversos tipos de conhecimentos e aplicações. Suas considerações sobre sintaxe formal se sustentam em três dos pilares centrais da abordagem chomskyana da sintaxe.

O primeiro deles é que qualquer hipótese levantada em discussões sobre estruturas linguísticas deve ser feita de modo suficientemente preciso para ser testada. Para isso, fundamenta-se no postulado de que a mente humana funciona de forma computacional no processamento da linguagem. Portanto, tentar processar computacionalmente uma língua natural, i.e. tentar reproduzir esse funcionamento através de máquinas e programas computacionais pode ser uma boa forma de simular os processos mentais envolvidos durante o uso da linguagem.

Qualquer lingüista que venha a concentrar seus estudos na área da sintaxe gerativa, na acepção chomskyana, deve aceitar como pilar primordial sua tese inatista de que as habilidades linguísticas fazem parte da base biológica do ser humano. Muitas complexidades da língua não precisam ser aprendidas porque muito do nosso conhecimento sobre ela é inato, pois só um tipo de conhecimento inato tornaria possível a aquisição da linguagem.

O terceiro pilar chomskyano, em que se sustenta o trabalho de Sag, Wasow, Bender (2003, p.9), é a delimitação precisa do campo de estudo da sintaxe. Acredita-se que o objeto de estudo da sintaxe deve ser o conhecimento inconsciente do falante de uma língua natural sobre a estrutura das sentenças de sua língua.

Segundo Sag, Wasow, Bender (2003, p. 14), que temos em mente, grosso modo, é algum tipo de aplicação computacional que envolve língua maturais. O processamento da linguagem integra diversos tipos de conhecimentos de forma muito rápida. E embora não se saiba exatamente como modelar esta integração, sabe-se que o conhecimento sintático exerce um papel crucial em tudo isto e impõe restrições no modo como as sentenças podem ou não ser construídas.

A sintaxe exerce particular importância nos modelos de processamento da linguagem e em relação a outros tipos de conhecimento. O conhecimento sintático é o domínio do conhecimento linguístico que pode ser caracterizado mais precisamente (SAG, WASOW, BENDER, 2003, p. 9-14). E segundo eles, pesquisas sintaticistas podem ajudar na construção de tecnologias para o processamento da língua natural.

Para a gerativista Klenk (2003, p.15), entende-se por sintaxe de uma dada língua a descrição das estruturas de expressão das orações dessa língua, das quais as palavras e os morfemas são elementos de base.

É possível, então, pensarmos em sintaxe como um mecanismo pelo qual as orações de uma língua podem ser produzidas e reconhecidas na sua estrutura.

2.3 Teoria X-Barra

Já sabemos que o conhecimento gramatical é um dos conhecimentos internalizados pertencentes à FL[3] nos humanos. Um ponto de vista central dos gerativistas é de que a gramática de um indivíduo se constitui de diversos tipos de conhecimento, tais como o conhecimento sintático, o conhecimento semântico e o conhecimento fonético, e que estes são aplicados de forma modular. Conforme esse modelo, a organização da gramática de uma língua é feita por módulos autônomos, cada um deles com uma organização interna extremamente simples, de princípios diferentes e que mantém uma rede de interações com os outros módulos (RAPOSO, 1992).

A partir dessa premissa, podemos dizer que a Teoria X-barra é um desses módulos. Ela é o módulo de conhecimento da gramática gerativa que permite a representação de constituintes sintáticos, conforme sua natureza, sua hierarquização e suas relações internas.

Como módulo sintático que é, a Teoria X-barra começa a representação dos constituintes das sentenças a partir da delimitação de seus núcleos. Um elemento sintático se agrupa a outro elemento sintático para formar um constituinte de nível hierárquico imediatamente superior ao que os inclui. Concordamos com Raposo (1992, p.65-67) ao dizer que a estrutura de constituintes de uma expressão linguística é constituída de um número finito de categorias gramaticais, elementos que vão desde os itens lexicais até a frase.

[3] Empregamos em nosso texto o termo Faculdade da linguagem na acepção de Raposo (1999).

Na perspectiva da Teoria X-barra, a frase é uma estrutura de constituintes que representa as relações sintáticas entre os núcleos funcionais e os núcleos lexicais que os compõem. Segundo Raposo (1992, p.66), uma frase se organiza em grupos de constituintes hierárquicos. Eles são construídos pela inclusão sucessiva de elementos de nível inferior em grupos maiores, começando pelos itens lexicais.

Como as frases se organizam em termos de estrutura de constituintes, sugere-se a existência de regras de reescrita categorial ou regras sintagmáticas que permitam engendrar essas estruturas. O conjunto dessas regras é capaz de gerar um conjunto potencialmente infinito de frases e também de representá-las estruturalmente.

Nós já sabemos que a sintaxe é apenas um dos módulos do conhecimento linguístico, mas que, no entanto, exerce grandes restrições na aplicação de regras em estruturas linguísticas. Como já foi dito, a sintaxe gerativa tem tentado não só descrever, mas também explicar os fenômenos presentes nas estruturas linguísticas das mais diversas línguas. Ela faz isso por meio de formalismos abstratos, com a finalidade de encontrar uma forma precisa capaz de dar conta do amplo conjunto de regras envolvidas no conhecimento linguístico que os falantes têm de sua língua. Como qualquer módulo da gramática gerativa, essa teoria deve ser capaz de captar a estrutura interna dos sintagmas de qualquer língua e também dar conta da variação nas diferentes línguas.

Já vimos que uma das formas encontradas pela sintaxe gerativa para tentar fazer esse tipo de descrição tem sido desenvolvida com o nome de Teoria X-barra. O módulo da gramática chamado de teoria do componente categorial. Nessa perspectiva a língua é vista como uma estrutura de constituintes, os quais se combinam entre si para gerar constituintes de níveis mais altos. Sendo umas das principais propostas da teoria a redução das regras sintagmáticas a um esquema universal extremamente simples e suscetível de parametrização.

Essa teoria procura construir hipóteses restritivas sobre a forma das regras de uma língua e sobre a forma das estruturas que resultam da ação dessas regras (RAPOSO, 1992, p.159). Segundo Mioto, Silva e Lopes (2005, p. 41), são essas regras que permitem a representação de um constituinte. Elas podem explicar a sua natureza e as relações estabelecidas dentro dele e ainda representar o modo como os constituintes se hierarquizam para formar uma sentença.

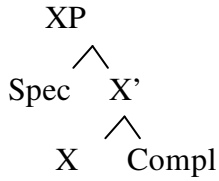
Baseamo-nos nos estudos de Raposo (1992) para demonstrar o percurso dos desdobramentos da Teoria X-barra desde a inicialmente proposta por Chomsky (1970) até a versão mais atualizada de Fukui (1986).

A tese central da teoria X-barra de Chomsky (1970) é que existem dois tipos de categorias, as sintagmáticas e as lexicais. As categorias sintagmáticas são projetadas com base em categorias lexicais. De acordo com o esquema X-barra, reconhecem-se três níveis hierárquicos, alcançados por duas projeções sucessivas das categorias lexicais (RAPOSO, 1992, p.168).

Segundo Chomsky (*apud* RAPOSO, 1992, p. 168-170), X' é resultado da relação entre uma categoria lexical com um **complemento**, reduzido ao símbolo de (Compl). E X'' é resultado da relação entre X' com um **especificador**, utilizado com a notação (Spec).

Ao final das duas projeções é possível se obter a configuração esquemática (12), na qual os especificadores, tal como os complementos, são elementos facultativos na estrutura das categorias sintagmáticas (RAPOSO, 1992, p.170).

(1)



As regras nesse esquema não são em si mesmas regras gramaticais, apenas representam a forma geral que as regras do componente categorial devem tomar nas gramáticas particulares das línguas humanas.

Percebe-se, então, devido às duas projeções de X, que é possível definir duas noções funcionais de núcleo, respectivamente o núcleo de X' e núcleo de X'' (Raposo, 1992, p.175). O núcleo de X' corresponde à categoria lexical X imediatamente dominada por X' e o núcleo de X" corresponde à categoria X' imediatamente dominada por X", sendo a categoria lexical X igualmente núcleo da projeção máxima X".

A propriedade da língua que faz com que o núcleo de um sintagma deva corresponder necessariamente ao mesmo núcleo da categoria lexical corresponde ao **Princípio da Endocentricidade**, pertencente à Gramática Universal. Isso explicaria o fato de que regras do tipo XP -> ... Y ... não sejam adquiridas.

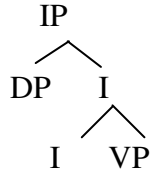
Vimos que na teoria X-barra de Chomsky o nível máximo de projeção de uma categoria é X''. Veremos agora como Fukui (1986) apresenta um novo esquema da teoria X-barra, também com duas projeções. Nesse novo desdobramento da Teoria X-barra o número máximo de barras permitido também é X'', contudo, há entre a Teoria X-barra de Chomsky e a teoria X-barra de Fukui uma diferença fundamental: a questão da **Uniformidade**.

A uniformidade está ligada ao número de barras que as categorias projetam. Para Raposo (1992, p. 175-176) na teoria de Chomsky[4], o número máximo permitido de barras é o mesmo para todas as categorias que projetam, mas na Teoria X-barra de Fukui a projeção máxima das categorias é variável. Por exemplo, se uma determinada categoria X não possuir um sistema de especificadores, não projetará o nível X'' e sua projeção máxima será X'. Na atual proposta, apenas as categorias funcionais projetam o nível X'', através da combinação da projeção X' com um especificador. As categorias lexicais, pelo contrário, não têm especificador, e, portanto projetam apenas o nível X'.

Portanto, para nosso trabalho levamos em conta as considerações de Fukui (1986, p.25) acerca das diferenças entre a projeção de categorias lexicais como N, V, P e A, mas especificamente V e outro tipo de projeção, a de categorias funcionais como, por exemplo, I. Conforme essa teoria, I representa a marca de concordância de pessoa e número de um verbo em sua forma finita.

Nesta nova análise, I é uma categoria de grau zero e sua projeção máxima é a categoria frásica IP. A primeira projeção de I contém I e o VP da oração como complemento. A projeção máxima de I, ou seja, I''[5] contém I' e o especificador DP. Assim como o VP fica reduzido ao estatuto de complemento, DP fica reduzido ao de especificador, conforme podemos ver na árvore a seguir:

(2)



Podemos representar esse esquema da seguinte forma:

[4] (Chomsky, 1970 *apud* Raposo, 1992)
[5] I'' corresponde à projeção de I', ou seja, corresponde à projeção máxima de I. Em nosso trabalho nos referimos à IP para falar de I''.

(3) IP -> DP I'

I' -> I VP

Conforme Mioto (2005), I encabeça o sintagma flexional IP e codifica certas propriedades gramaticais que definem uma sentença como finita ou infinitiva. Consideremos (4) [6]:

(4) a. ele chegará

b. *ele chegar

Nos exemplos acima podemos perceber que a marca de tempo do verbo faz com que apenas (4a) seja uma sentença do português. Portanto, há aí um indício de que a flexão verbal é o núcleo da sentença finita. Ao identificarmos I como a flexão verbal, assumimos então, que I só pode ser combinado com verbos, portanto o complemento de I só pode ser uma categoria de natureza verbal, como vermos a seguir, na seção 2.3.1

Vimos aqui brevemente que o tipo de descrição sintática X-barra de Fukui (1986) se aplica a qualquer constituinte lexical ou funcional. Sendo assim, I se associa a um complemento projetando I', que por sua vez se associa a um especificador e projeta IP. Trabalharemos, então, com as noções de categorias lexicais e categorias funcionais que encabeçam o constituinte para descrever e analisar sentenças da língua portuguesa de acordo com o esquema universal da Teoria X-barra em relação aos níveis de projeção de V e I.

2.3.1 Categorias funcionais e categorias lexicais

Como já sabemos, as frases nas línguas humanas são formadas por uma seqüência linear ordenada de itens lexicais. Os itens lexicais se combinam para formar grupos sintáticos hierarquicamente superiores. A organização da frase em grupos hierárquicos chama-se **estrutura de constituintes**. A estrutura de constituintes de uma expressão linguística se constitui de vários elementos que vão desde os itens lexicais até as frases e estes são classificados em um número finito de categorias gramaticais (RAPOSO, 1995, p.67). As categorias gramaticais podem ser do tipo **lexical** e do tipo **sintagmática.** Uma preposição (P), por exemplo, é um a categoria lexical, já o sintagma preposicional (PP) corresponde a uma categoria sintagmática.

[6] Exemplo retirado de Mioto (2005, p.58)

Para reconhecer uma categoria sintagmática é necessário identificar o seu núcleo e as relações estabelecidas a partir dele. Os núcleos, por sua vez, são categorias gramaticais que, entretanto podem ser de natureza lexical e de natureza funcional. Cada núcleo corresponde ao elemento central de uma categoria sintagmática. Por exemplo, P é uma categoria gramatical de natureza lexical, que é o núcleo da categoria sintagmática. De acordo com Fukui (1986), I é uma categoria funcional, núcleo da categoria sintagmática IP.

Portanto, categorias sintagmáticas são categorias superiores construídas com base em categorias lexicais ou funcionais. Segundo essa condição, a variável X é usada para representar qualquer núcleo de uma categoria XP qualquer.

As categorias lexicais são definidas pela combinação de dois traços distintivos fundamentais: +/- N(nominal) e +/-V (verbal). Essas características são apresentadas pelas categorias lexicais e só as categorias que apresentam esses traços selecionam argumentos (FUKUI, 2005:53).

Tomemos o exemplo (5), a seguir:

(5) As crianças narraram enfaticamente o episódio.

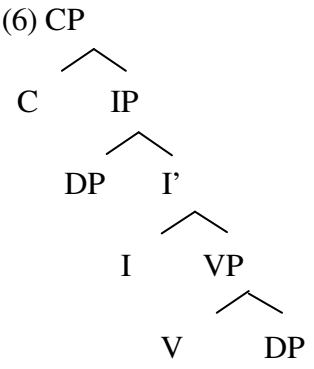
Um radical como /*narr*-/, que chamamos de V, estabelece o sentido lexical da palavra. Dele podemos derivar uma palavra como o verbo *narrar*. O núcleo *narrar* pode ser definido pelos traços [-N, +V]. Os núcleos lexicais têm a propriedade de **s-selecionar**[7] seus argumentos. Isto é, o complemento e o especificador s-selecionados pelos núcleos lexicais devem possuir propriedades semânticas compatíveis com as do núcleo. Assim, um verbo como *narrar* s-seleciona um sujeito e um objeto com propriedades semânticas relativas às necessidades do verbo.

O contrário acontece com as categorias funcionais. Estas não s-selecionam. Seu esquema dispõe de um complemento e uma posição de especificador, mas ao contrário dos núcleos lexicais, diz-se que esse tipo de núcleo apenas **c-seleciona**[8], ou seja, só precisa satisfazer a necessidade de selecionar a categoria do complemento que selecionará.

Na versão de Fukui, as categorias funcionais correspondem às categorias gramaticais C (Comp), I e D (Determinante). No caso, em particular, da categoria funcional I,

[7] S-selecionar deve ser lido como 'selecionar semanticamente'.
[8] C-selecionar é uma forma curta de se referir a uma categoria que apenas seleciona complemento.

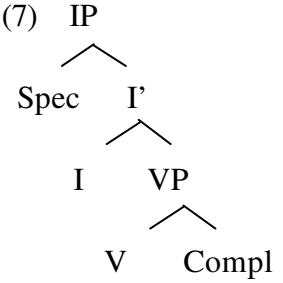
o papel reservado à projeção de I, ou seja, a posição de IP é o de complemento de C. Assim como o papel reservado à projeção da categoria lexical V, ou seja o VP, é o de complemento de I. A título de exemplo, confiramos a configuração em (6), abaixo:

(6)


Para Klenk (2003), características que determinam toda a sentença, como as marcas de tempo, modo e aspecto verbal são expressas através de afixos em verbos flexionados. No caso da língua portuguesa são introduzidos na forma do constituinte I como núcleo de uma sentença.

De acordo com a configuração acima, a categoria I é uma categoria funcional que projeta duas barras, enquanto V é uma categoria lexical que projeta apenas uma. No caso específico da configuração IP, a categoria I licencia um sujeito na posição de especificador de IP, ao passo que VP seleciona um objeto como complemento de V. Consideraremos que o complemento é um DP, assim como a abordagem dada à estrutura do VP neste trabalho não atribui especificador à V.

Como os núcleos funcionais são essencialmente gramaticais, podem aparecer em muitas línguas na forma de afixos. Vejamos então a flexão I encabeçando o sintagma IP. I é o elemento que define, entre outras, se uma sentença é finita ou infinitiva. Ele é o núcleo do constituinte IP e é representado esquematicamente do seguinte modo:

(7)


2.3.2 Mecanismo de adjunção

Uma categoria lexical pode possuir, além de um complemento, outro elemento que o modifique. Nesse caso, a modificação pode ser representada estruturalmente por meio de uma configuração de adjunção, como já foi visto anteriormente.

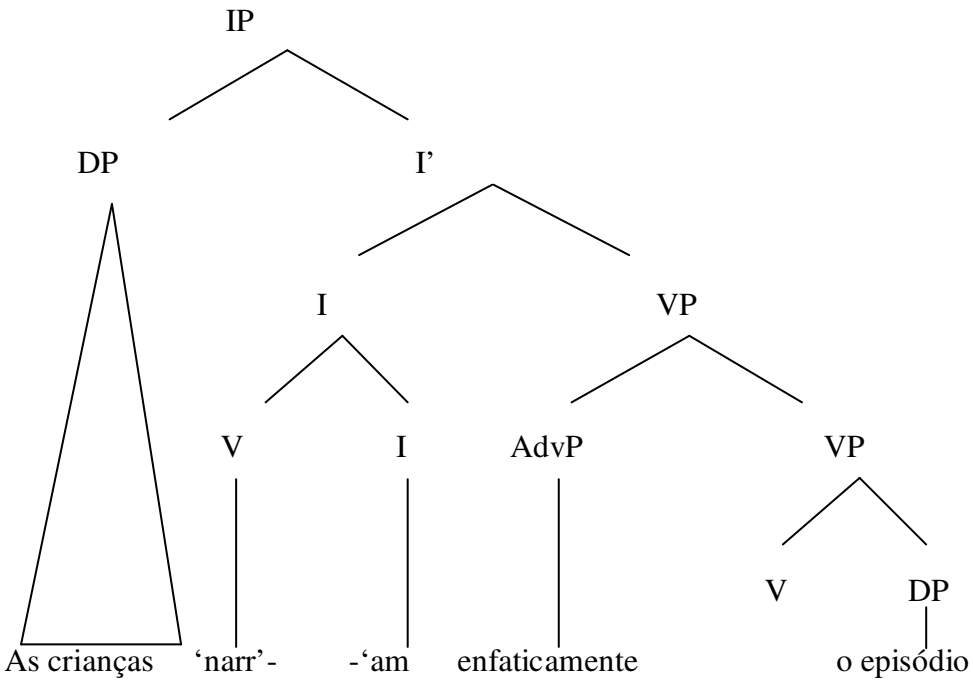
2.3.2.1 Adjuntos

No módulo da Teoria X-barra, os núcleos das categorias sintagmáticas não selecionam apenas complementadores e especificadores. Há numa sentença, ainda, constituintes que são licenciados sem serem complementos ou especificadores. São os chamados adjuntos.

Uma frase finita como (5), repetida em (8) abaixo, pode ser representada como na configuração esquemática em (9).

(8) As crianças narram enfaticamente o episódio

(9)



No caso de VPs[9], considera-se um adjunto aquele constituinte que existe além do(s) argumento(s) do verbo. Mioto, Silva e Lopes (2005, p.84) utilizam as noções de estar **incluído** e estar **contido** para diferenciar respectivamente os adjuntos dos argumentos.

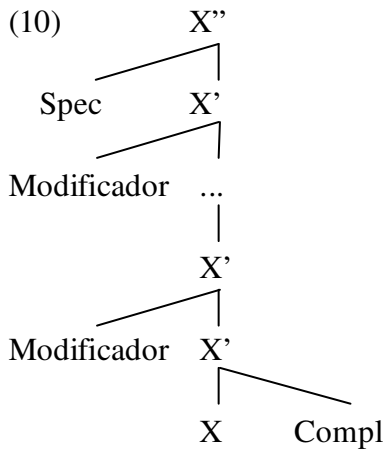
[9] De acordo com Mioto, Silva e Lopes (2005), um adjunto é um constituinte que está contido tanto em VP como em NP

Conforme essa afirmação, um adjunto é um constituinte que está contido em uma projeção máxima, dobrando o VP, a categoria da qual o constituinte é adjunto. O processo de representação é sempre o mesmo: dobrar a categoria da qual o constituinte é adjunto (MIOTO, SILVA e LOPES, 2005:86).

Pensando em termos de estruturas de constituintes, conforme o esquema X-barra, os adjuntos são elementos de uma expressão linguística, que exercem uma função de modificadores.

Como nós já sabemos uma categoria V só projeta uma categoria V' se estiver associada a um especificador. Em uma configuração de adjunção, a categoria V' não se expande para a projeção máxima, ao invés de se projetar uma categoria de duas barras, duplica-se uma projeção do tipo V'. Desse modo, o núcleo do sintagma pode ter que se expandir mais de uma vez como categoria V'.

(10)



No exemplo (10) acima podemos conferir o esquema X', em uma configuração de adjunção[10]. Conforme Klenk (2003, p. 83), um adjunto será representado como um Modificador. Adotamos essa denominação por motivos práticos e o utilizamos para nos referir à etiqueta do nó referente a um sintagma adverbial.

Para Raposo (1992, p. 200), numa configuração de adjunção a categoria adjunta é ao mesmo tempo irmã e filha da categoria à qual é adjunta. De acordo com ele, podemos atribuir à configuração de adjunção as seguintes propriedades: i. o adjunto e a categoria

[10] Esse esquema está exposto em Klenk (2003, p.83) adaptado para a língua alemã, na qual a configuração de adjunção ocorre à direita. Para nosso trabalho adaptamos a configuração para o português, na qual a esquerda do verbo.

modificada são irmãos; ii. o adjunto e a categoria modificada são dominados por um constituinte de categoria e número de barras idêntico ao da categoria modificada.

Com base no contraste feito entre inglês e português sobre posição de advérbios como *brutalmente* e '*brutally*', Raposo (1999, p.33) postula diferentes extensões para a regra de movimento do verbo, em português *brutalmente* ocupa a posição imediatamente posterior ao verbo, dentro do sintagma verbal e em inglês a posição imediatamente anterior. Veja os exemplos abaixo[11]:

(11) a. eles agrediram *brutalmente t* o prisioneiro

a'. *eles *brutalmente* agrediram o prisioneiro

b. they brutally hit the prisoner

b'. *they hit brutally *t* the prisoner

Note que *t* é uma representação do elemento movido. Segundo Raposo (1999), sentenças que apresentam advérbios como *brutalmente*, ordenados na frase antes do verbo são tidas como agramaticais.

Se supusermos que advérbios de um mesmo tipo são todos gerados na mesma posição e que não existe uma regra própria de movimento dos advérbios, os julgamentos inversos do inglês com respeito ao português fornecem uma prova de que em português, o exemplo (11a) o verbo não se move em sintaxe visível. Em nosso trabalho, ao invés de falarmos em movimento de V de modo não visível, achamos que é mais explicativo dizer, no contexto dessa pesquisa, que o verbo permanece *in situ* na sintaxe visível.

Em nossa pesquisa, ao contrário do que diz Raposo (1999) sobre a agramaticalidade de *a'* em português europeu, marcada como agramatical pelo asterisco, consideramos que *a*' é gramatical em português brasileiro.

2.3.2.2 Alguns critérios semânticos

Não é nosso objetivo nesse estudo fazermos uma incorporação da semântica aos nossos dados, mas alguns aspectos semânticos serão acrescidos, de forma breve, à nossa pesquisa no que diz respeito à abordagem dos tipos de advérbios.

[11] Estes exemplos foram retirados de Raposo (1995 p. 33).

Nosso estudo utiliza as noções adotadas por Ilari (1991, p.48) para caracterizar semanticamente alguns advérbios. Consideramos boas essas noções já que seu estudo sobre advérbios está diretamente relacionado à posição e ao movimento de verbos. Sua primeira colocação sobre posições adverbiais trata basicamente de duas classes: a dos **advérbios altos** e a dos **advérbios baixos**, tais como veremos a seguir.

Trazendo essas noções para nosso estudo, podemos dizer que os altos ocupam a posição inicial de IP e os baixos dispõem da posição inicial do VP. Como veremos abaixo, os advérbios que só aparecem à direita do verbo, os baixos, devem seguir obrigatoriamente o verbo. Em relação à classe dos advérbios altos, estes podem aparecer em uma posição mais alta que a do verbo.

Os advérbios baixos, que seguem obrigatoriamente o verbo, possuem características semânticas ligadas à noção de maneira, como *completamente*; de instrumento, como *manualmente*; de quantificação, com *muito* e *demais*; e advérbios orientados em direção ao verbo, como *corretamente* e *bem* (ILARI, 1991:50).

Observemos os exemplos (12) e (13) abaixo, retirados de Ilari (1991, p.50):

(12) O João perdeu completamente a cabeça

*O João completamente perdeu a cabeça

(13) O João tinha perdido completamente a cabeça

*O João tinha completamente perdido a cabeça

*O João completamente tinha perdido a cabeça

Conforme Ilari (1991, p.50), a impossibilidade de colocar esse advérbio entre o auxiliar e o particípio ou entre sujeito e auxiliar sugere que o português prefere gerar esses elementos na posição inicial do VP.

Em relação à classe dos advérbios altos, que podem aparecer em uma posição mais alta que a do verbo, segundo Ilari (1991, p.51), estes advérbios ocupam a posição entre sujeito e verbo. São os advérbios pragmáticos como *felizmente*; advérbios modais, como *provavelmente*; e advérbios orientados em direção ao sujeito, como *deliberadamente.*

Em relação às colocações acima, o que podemos dizer quanto ao advérbio dado nas frases (14) abaixo, em contradição ao que foi dito em Ilari (1991), quanto aos exemplos de (13)?

(14) a. Eles *brutalmente* agrediram os manifestantes

a'. Eles agrediram *brutalmente* os manifestantes

De acordo com os exemplos, somos levados a pensar que não é o advérbio o elemento móvel da frase. Do mesmo jeito que o movimento de V em direção a I não é visível no inglês, acreditamos que no português o fenômeno de subida do verbo como uma operação de movimento não visível justifica a conseqüência fonética acarretada à posição do advérbio nas frases do português.

2.3.2.3 Critérios morfossintáticos

Para Perini (2005, p. 118), o termo tradicional de adjunto adverbial corresponde, no nível da oração, à classe dos constituintes que ocupam funções "adverbiais" na oração. Mesmo assim, as funções ditas adverbiais são bastante diferentes entre si. Para ele, os advérbios constituem um grupo muito grande e heterogêneo de funções e ainda falta uma definição que delimite com alguma clareza essas entidades, por isso há certa dificuldade para sistematizá-los.

Com base no que foi exposto até aqui, entendemos que um sintagma adverbial é um lugar na oração ocupado pelo conjunto das funções de um advérbio. Os elementos que ocorrem com freqüência como sintagmas adverbiais são englobados pela gramática tradicional sob o rótulo "adjunto adverbial". Conforme Perini (2005), o rótulo adjunto adverbial abrange um conjunto bastante variado de funções. Segundo ele, é necessário saber como o constituinte se comporta para poder atribuir a ele o termo adjunto.

Perini (2005, p.118-120) classifica os adjuntos adverbiais (tal como classificados na gramática tradicional) em classes de funções do tipo: i. **atributo**; ii. a**djunto adverbial**; iii. **adjunto oracional**, iv. **adjunto circunstancial,** v**. negação verbal**. Para Perini (2005), os adjuntos são caracterizados pelas cinco funções distintas que desempenham.

A primeira função se caracteriza por ocupar uma posição relativamente fixa na oração e é chamado de adjunto adverbial (apesar da semelhança dos nomes, esse grupo de

advérbios compreende uma pequena parte dos termos tradicionalmente chamados de adjuntos adverbiais e constitui apenas um dos tipos de advérbio, na proposta de Perini (2005, p.118)).

Um exemplo desse tipo de advérbio pode ser observado nas sentenças abaixo:

(15) a. Emanuel ornamentou o escritório completamente.

a' *Completamente, Emanuel ornamentou o escritório.

Podemos perceber que ao mudar o termo *completamente* do seu lugar de origem, obtemos uma frase agramatical no português.

Outra classe de advérbios é chamada de adjunto oracional. Observando o adjunto *francamente* em (16), podemos pensar que ele não parece compor constituinte com nenhum outro elemento, sendo uma peça integrante da oração e não de um elemento específico. Isso tem sido considerado na literatura linguística moderna como um elemento anexo à oração (PERINI, 2005:86).

(16) Aquele moço, francamente, é um tolo.

Vejamos a diferença do adjunto oracional *francamente* dado acima do termo *frequentemente* no exemplo (17) [12].

(17) Jeremias reclama frequentemente.

Na sentença em (17), *frequentemente* compõe constituinte com outro elemento, o verbo *reclama*. Para Perini (2005, p.84), esse termo está contido no grupo dos adjuntos que funcionam como atributo do verbo, daí receber esse nome.

O atributo se caracteriza pela possibilidade de ser ocupar posições como a posição imediatamente anterior ao núcleo do predicado ou o início da oração.

(18) a. Jeremias frequentemente reclama.

a'. Frequentemente, Jeremias reclama.

A frase 'a' do exemplo (18) exprime a propriedade do atributo de ocorrer entre o sujeito e o núcleo do predicado.

[12] Os exemplos (16), (17) (18), (19) foram retirados de Perini (2005, p.85) para nos orientar na classificação dos advérbios que analisamos neste trabalho.

Outro tipo de advérbio é classificado por Perini (2005, p.88) como adjunto circunstancial. Esse adjunto não pode ocorrer na posição entre sujeito e núcleo de predicado.

Desse modo, uma frase como (19) é agramatical no português brasileiro.

(19) *Jeremias muito reclama.

Por fim, nos referiremos à negação verbal como um tipo de função sintática desempenhada, por exemplo, pela palavra *não*. Essa palavra é conhecida como um advérbio. Sua posição original é anteposta ao núcleo do predicado, logo após o sujeito. Desse modo, *não* sugere ser parte do predicado. Diferentemente do atributo, a negação verbal só pode ocorrer logo antes do núcleo do predicado (PERINI, 2005, p.85).

2.3.3 O movimento de constituintes

2.3.3.1 De acordo com o modelo de Princípios e Parâmetros

Segundo a teoria gerativa, a faculdade da gramática de um indivíduo em seu estado inicial representa aquilo que os gerativistas chamam de gramática universal. A partir daí, podemos pensar que, entre outros princípios pertencentes à GU, o movimento de constituintes é um deles.

No modelo PP cada descrição estrutural é um conjunto de quatro níveis de representação simbólica: **estrutura-D** (DP, do inglês, *Deep Structure*), **estrutura-S** (SS, do inglês, *Surface Structure*), **forma lógica** (LF, do inglês *Logic Form*) e **forma fonética** (PF, do inglês, Phonetic Form) [13]. Cada nível de representação capta propriedades diferentes das expressões linguísticas.

A estrutura profunda é uma interface entre a derivação sintática e o léxico. A forma fonética é gerada pelo componente fonológico, presente naquele sistema computacional. É ela que recolhe as propriedades fonéticas da expressão linguística. A forma lógica, por sua vez, recolhe as propriedades semânticas derivadas com base nas propriedades dos itens lexicais. Tanto forma fonética como forma lógica são níveis de representação que servem de interface com sistemas de pensamento que usam ou interpretam a linguagem. Por fim, a estrutura-S é o nível de representação da expressão. É onde estão representadas as

[13] Por conveniência, adotamos esses quatro termos de acordo com Raposo (1999). Esses termos se referem à níveis de representações linguísticas. A saber *estrutura profunda, estrutura de superfície, forma fonética e forma lógica.*

operações do sistema computacional que têm um reflexo fonético. É nesse ponto de derivação onde se aplicam vários princípios do modelo de princípios e parâmetros.

Conforme Mioto, Silva e Lopes (2005, p.249), o mecanismo de movimento é um fenômeno sintático que serve para deslocar sintagmas da posição em que foram gerados em estrutura-D para alocá-los em outras posições na sentença. Isso seria uma tentativa de explicar o fato de os seres humanos compreenderem estruturas sintáticas, nas quais certos sintagmas aparecem nas sentenças em posições diferentes daquelas em que são marcadas tematicamente. Veja o exemplo[14]:

(20) Quem que a Maria encontrou?

De acordo com (20) podemos compreender que a pergunta recai sobre o objeto do verbo *encontrar* apesar deste objeto não ocupar a posição originária do complemento do verbo. De acordo com a ordem SVO (sujeito-verbo-objeto), no caso de línguas românicas como a nossa, a posição argumental do português é após o verbo.

Há muitos tipos de movimento. O exemplo acima é considerado um movimento de sintagmas interrogativos. Esse movimento é considerado movimento de posição argumental, ou seja, **movimento A**. Além dele, há o movimento de posições não argumentais, chamado **movimento A-barra** e, ainda, o **movimento de núcleos**.

Dependendo do tipo de movimento e da configuração sintática, para que sejam licenciadas, é preciso que as regras de movimento sigam uma série de restrições de localidade. Outra necessidade para o licenciamento das regras de movimento é a sua articulação com a noção de regência. Segundo Mioto, Silva e Lopes (2005, p.250), essa articulação se dá de acordo com o tipo e a distribuição das categorias vazias que estão em questão.

No caso do movimento de um núcleo para uma posição de núcleo, certas exigências de localidade pesam sobre a estrutura. Para a satisfação dessas exigências há um tipo de restrição conhecida como **Restrição de Movimento de Núcleo**. De acordo com essa generalização, o alvo do movimento de um núcleo só pode ser outro núcleo que o c-comande[15]. Segundo Mioto, Silva e Lopes (2005, p.251), um exemplo desse tipo de

[14] Para uma compreensão mais ampla desse assunto, sugerimos a leitura do capítulo Mova α, do qual retiramos o exemplo discutido aqui.
[15] Definição de C-COMANDO, segundo Mioto (2005):

movimento ocorre com os verbos. Eles se deslocam de sua posição de base dentro do VP por razão de ordem morfológica.

Esses movimentos são operações conhecidas na teoria como **Mover α**. Segundo Raposo (1999), elas são aplicadas nas sentenças para, entre outras coisas, satisfazer propriedades morfológicas das expressões sintáticas que não poderiam ser satisfeitas de outro modo. Um exemplo de núcleo deslocado por Mover α é o do movimento do verbo para I. Isto é, V se desloca de sua posição para se completar morfologicamente em I.

Segundo Mioto, Silva e Lopes (2005, p.264), as operações de movimento aplicadas às sentenças devem satisfazer o **Princípio das Categorias Vazias** (ECP,) para explicar através da noção de regência as restrições que pesam sobre o movimento. Para Raposo (1992), é esse princípio que impõe as condições sobre a posição que um vestígio pode ocupar. O objetivo da ECP é caracterizar a posição legítima de um vestígio de Mover α.

Dizemos que o movimento de um constituinte de uma posição A para uma posição B deixa na posição originaria A uma cópia sem conteúdo fonético do constituinte movido, de acordo com a teoria dos vestígios (RAPOSO, 1992). Essa categoria vazia é designada de **vestígio**. O constituinte movido é, por sua vez, o antecedente do seu vestígio. A dependência que existe entre o constituinte movido e o seu vestígio é representada por índices idênticos. Usualmente, o vestígio é representado pela letra *t* (do inglês *trace*).

2.3.3.2 De acordo com o Programa Minimalista

"O programa minimalista constitui um conjunto de diretrizes metodológicas [...] na tentativa de minimizar, como o próprio nome diz, o aparato técnico de Princípios e Parâmetros" (MODESTO, 2009, p.2). Diferente de PP, no qual os níveis de representações linguísticos se dividem em quatro[16], o PM propõe apenas dois níveis de representações: a forma lógica e a forma fonética, i.e. apenas os níveis de representação internos, ou seja, os de interface.

Como o modelo minimalista reflete uma arquitetura da linguagem com bases empíricas, reduzida àquelas propriedades necessárias sem as quais não haveria o seu entendimento, "sua idéia-chave é remover do modelo o que não é estritamente necessário,

---

α c-comanda β se e somente se β é o irmão de α ou filho do irmão de α.

[16] Estrutura-D e estrutura-S como níveis de representação internos e Forma Fonética e Forma Lógica como níveis de representação externos.

quer do ponto de vista da inserção da linguagem na mente e dos seus mecanismos internos, quer do ponto de vista da parcimônia do modelo" (Raposo, 1999, p.23). Nesta concepção mínima do que pode ser a linguagem humana, então só a LF e a PF existem no programa minimalista.

De acordo com o PM, a faculdade da linguagem (FL) tem de associar a cada expressão gerada, o nível de representação que entra em contato com cada um dos sistemas de performances (RAPOSO, 1999, p.25-27). Estes sistemas constituem o sistema de performance articulatório-perceptual (A-P) e o sistema de performance conceitual-intencional (C-I), com os quais a faculdade da linguagem entra em contato.

O nível de representação forma fonética é a interface da faculdade da linguagem com o sistema articulatório-perceptual e o nível forma lógica é a interface da faculdade da linguagem com o sistema conceitual-intencional. Os sistemas de performances possuem estrutura própria e independente de FL. Eles impõem à FL que esta satisfaça as condições de legibilidade impostas pelo **Principio da Interpretação Plena.**

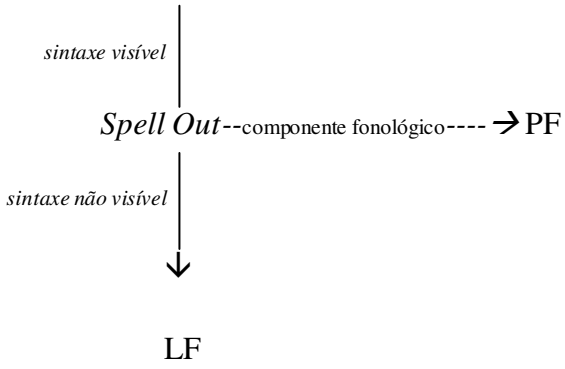
Segundo Raposo (1999, p. 27), tanto PF como LF não podem possuir elementos que não possam ser interpretados por eles, obedecendo ao requisito do Princípio da Interpretação Plena. A interface PF é exaustivamente constituída de traços fonéticos, estruturas silábicas e prosódia, por exemplo. Já em LF são interpretáveis, segundo Raposo (1999, p.28), estruturas sujeito-predicado, por exemplo.

Segundo Modesto (2009, p. 5), assume-se que há um ponto na derivação da estrutura sintática em que os traços fonéticos são retirados e mandados para PF e que os traços semânticos continuam em direção à LF.

Esta operação de mandar os traços fonéticos da estrutura para a PF é chamada de *Spell-out*. Diz-se que as operações efetuadas antes deste ponto são operações sintáticas visíveis, já que o resultado implica reflexos fonéticos. Ao contrário, as operações efetuadas após a aplicação de *Spell Out* são tidas como operações sintáticas não visíveis.

Uma leitura gráfica do modelo aqui descrito pode ser feita através do seguinte esquema:

(21) LÉXICO



No programa minimalista, as operações do tipo Mover, ou seja, os movimentos aplicados, podem ser visíveis ou não-visíveis, no sentido de alterarem ou não a forma fonética.

A existência de movimento, segundo Modesto (2009), é explicada através da postulação de que os traços sintáticos por não serem interpretáveis pelas interfaces LF e PF têm de ser eliminados na derivação sintática. A diferença entre traços fortes e fracos, conforme Modesto (2009, p.7), explica-se pela necessidade de cada tipo de ser checado na derivação. Os traços fortes precisam ser checados imediatamente, ao contrário da checagem dos fracos que pode ser adiada até após de *Spell Out*. Todos os traços precisam ser checados antes da LF, que só lê traços semânticos. Entretanto, só os traços fortes precisam ser checados antes de *Spell Out* (MODESTO, p.7).

2.4 Modelos formais do funcionamento da linguagem

A nossa concepção de língua é a de que esta é adquirida naturalmente por nós. Acreditamos que a língua é a manifestação de uma capacidade inata ao ser humano, biológica e mental, de articular diversos conhecimentos interiorizados na mente humana de forma modular que estão prontos para serem acessados através das interfaces mantidas nos sistemas que.

Nós pensamos que através da descrição de minigramáticas, torna-se mais claro perceber o aspecto formal das regras gramaticais de uma língua e então este aspecto formal é o que permite a sua adaptação aos sistemas de linguagem que compõem as máquinas. Como sabe-se, nas máquinas as linguagens implementadas são linguagens de programação e não línguas naturais.

Mais acima falamos da gramática de estrutura sintagmática e vimos que esse tipo de gramática combina constituintes sintáticos, de forma sistemática e hierárquica, até produzir uma estrutura de constituintes. De acordo com esse modo de descrever estruturas linguísticas, "dada uma sentença com sua estrutura de constituintes em forma de um diagrama arbóreo, deixando fora os nós mais baixos (as palavras), preserva-se uma estrutura de constituintes generalizada, à qual diversas outras frases se enquadram" (KLENK, 2003).

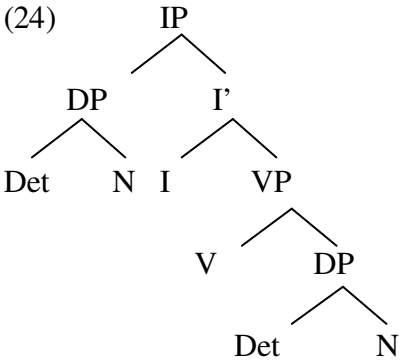
Dessa forma podemos dizer, então, que sentenças como (31) e (32), e muitas outras sentenças do português, possuem a mesma estrutura de constituintes.

(22) A estudante prepara um seminário.

(23) A mulher contempla a paisagem.

Vejamos em (24) a configuração arbórea dessas sentenças:

(24)



Para Klenk (2003, p.35), métodos como este de descrição sintática buscam encontrar um sistema que descreva todas as sentenças de uma língua com suas estruturas de constituintes. Nesta perspectiva, pensa-se na elaboração de um sistema de regras com base na estrutura de constituintes de uma língua. Esse sistema de regras, o qual consiste apenas de regras livres de contexto, é chamado de **Gramática Livre de Contexto**.

2.4.1 Gramática Livre de Contexto (CFG)

Considerável atenção vem sendo dada à idéia de que, com uma teoria de características sintáticas e princípios gerais apropriadamente projetados, as CFG podem servir como uma teoria da sintaxe de línguas naturais empiricamente adequada (SAG, WASOW, BENDER, 2003:36). Contudo, essa teoria pretende estar apta a esclarecer o enorme número de línguas existentes, os casos de ambigüidades estruturais e principalmente esclarecer a enorme riqueza da sintaxe de línguas naturais.

De acordo com Klenk (2003, p. 42), uma CFG descreve objetos linguísticos por meio de um sistema de regras, cujas relações são apresentadas através de árvores, por meio de nós que representam o grau de parentesco entre os constituintes de cada estrutura. Esses objetos são seqüências de elementos, válidos como elementos bases da sintaxe. Dependendo do começo da descrição, esses elementos podem ser palavras ou morfemas.

Por abranger categorias lexicais e categorias sintagmáticas, uma gramática livre de contexto deve consistir de certos componentes. Segundo Sag, Wasow, Bender (2003, p. 26), um destes componentes corresponde ao léxico da gramática. Ou seja, uma lista de palavras com suas respectivas categorias sintáticas. O outro, por sua vez, corresponde ao conjunto de regras feito a partir de categorias lexicais e/ou categorias funcionais.

Uma CFG possui um 'símbolo inicial' usualmente anotado de forma abreviada como 'S' (do inglês, *sentence*). Segundo Bird, Klein e Loper (2009) todas as árvores bem formadas devem ter este símbolo como sua etiqueta raiz. O lado esquerdo de cada regra especifica um tipo de sintagma. Do lado direito são dados os possíveis modelos para aquele tipo de sintagma (SAG, WASOW e BENDER, 2003, p.27). Portanto, essa gramática licenciará qualquer seqüência de palavras que possa ser derivada de 'S' por meio de aplicações seqüentes das regras da gramática.

Podemos ver a generalização feita por Sag, Wasow, Bender (2003, p.26) para representar a forma desse tipo de regras:

(25) A -> ... , onde A é uma categoria não lexical e '...' representa a expressão [...] formada. A seta é um símbolo que pode ser lido como 'consiste de'.

Para ilustrar como uma CFG trabalha, vamos utilizar a seguinte gramática:

(26)IP -> DP I’
DP -> Det N
I’ -> I VP
VP -> V DP
D -> ‘a’ | ‘um’
N -> ‘estudante’ | ‘seminário’
V -> ‘prepara’ |
I -> ‘prepara’

Conforme essa gramática, a sentença é encabeçada por IP, já que é a primeira categoria à esquerda da seta. A primeira regra permite a possibilidade de substituirmos IP por DP I’. O intuito é substituir todas as categorias sintagmáticas à direita da regra até que se chegue apenas a uma seqüência de categorias lexicais. Desse modo, numa substituição passo a passo, IP pode consistir de uma seqüência do tipo Det N I V Det N, a qual pode ser convertida para a seqüência desejada pela inserção apropriada de palavras no local de suas categorias lexicais.

Gramáticas livres de contexto podem gerar infinitas coleções de sentenças de uma língua através de simples regras. Elas também podem fornecer uma representação de certas ambigüidades. Contudo, esse formalismo, conhecido como CFG, não pode ser tido satisfatoriamente como uma tentativa de analisar a estrutura de uma língua. Do modo como foi apresentada, essa CFG pode hipergerar estruturas linguísticas. Hipergerar significa gerar estruturas linguísticas que não são bem-formadas ou agramaticais.

Desse modo, de volta àquele mecanismo de substituição passo a passo de categorias à direita da seta, visto mais acima, IP pode consistir de uma seqüência do tipo D N I V D N, e esta ao ser convertida em uma sentença pela inserção apropriada das palavras do léxico, gera uma sentença como:

(27) O menino prepara prepara um seminário

Isto ocorre porque ‘prepara’ está subcategorizado como I e V, nesse caso a gramática não foi elaborada prevendo uma categoria vazia e um elemento preenchedor que acomodasse o elemento movido.

Nesse modelo formal do funcionamento da linguagem, problemas específicos de determinadas línguas, como a questão da subcategorização ou da ambigüidade estrutural são resolvidos utilizando-se diagramas arbóreos distintos, cujos nódulos demonstram o tipo de relação entre os sintagmas. Tais ambigüidades surgem sempre que uma seqüência de palavras pode formar constituintes de mais de uma maneira. Portanto, CFG provê mecanismos claros para expressar tais ambigüidades.

Outro problema que pode surgir na derivação de sentenças de uma língua de acordo com o formalismo da CFG diz respeito à questão da concordância, no caso de línguas românicas como o português, segundo a qual, por exemplo, um sujeito de uma sentença tem de estar de acordo em número e pessoa com o verbo finito. Sozinha, uma regra do tipo DP -> D N pode hipergerar um sintagma do tipo 'a seminário' devido à ausência de traços como os de gênero e número.

Vejamos na próxima seção como alguns dos problemas da Gramática Livre de Contexto podem ser resolvidos através do aprimoramento dos sistemas categoriais.

2.4.2 Gramática Livre de Contexto Baseada em Traços

Como vimos acima, a CFG precisa de aprimoramentos para que não venha a hipergerar. Um dos aprimoramentos para eliminar expressões incorretas em relação, por exemplo, à concordância entre um determinante e um determinado substantivo em uma língua românica como o português, é colocar um apurado sistema de categorias e ao invés de uma única regra formular um conjunto de regras.

Já sabemos que as categorias sintagmáticas são nomeadas de acordo com a categoria lexical da qual é uma projeção e que esta categoria é parte obrigatória desse tipo de sintagma. Esse fenômeno das línguas naturais sugere a noção de nuclearidade. Segundo Sag, Wasow, Bender (2003, p.37), esse aspecto desenvolve um papel crucial em todas as línguas humanas.

Portanto, mais do que escolher pares como Det e DP, a arquitetura dessa teoria gramatical deve enriquecer a representação das categorias gramaticais e expressar diretamente o que um sintagma tem em comum com o seu núcleo. Essa necessidade de representar estruturas nucleares conduziu a teoria a uma dramática redução no número de regras gramaticais requeridas. Desse modo, mostraremos como o conjunto de regras dos exemplos

que mostraremos abaixo chegarão ao reduzido conjunto de regras do fragmento de gramática apresentado no capítulo 5.

Conforme Klenk (2003), um exemplo de descrição de concordância entre determinante e substantivo pode ser demonstrada do seguinte modo: um sintagma determinante do português pode ter uma estrutura interna do tipo DP -> D N.

Primeiro, sabemos que em nossa língua existem dois gêneros, um masculino e outro feminino. Segundo, sabemos que existem dois números, um singular e um plural. Desse modo, obtemos quatro classes de determinantes e quatro de substantivos, os quais serão indicados com novos símbolos categoriais. Essas regras fornecem a combinação desejada para a expansão dos sintagmas e os novos símbolos devem ser substituídos apenas por palavras das suas respectivas categorias.

Vejamos:

(28) DPms -> Detms Nms

DPfs -> Detfs Nfs

DPmp -> Detmp Nmp

DPfp -> Detfp Nfp

Repare-se que as categorias Det e N estão associadas aos símbolos ‘f’, ‘m’, ‘p’ e ‘s’. Isto quer dizer, por exemplo, que para cada Detfp (feminino, plural) há um Nfp, também feminino plural.

Esse mesmo método pode ser usado para descrever de forma livre de contexto, também, a concordância entre adjetivos e substantivos e ainda a concordância entre sujeito e verbo finito.

Muitos verbos e substantivos em português possuem a forma plural e a forma singular. É necessário, portanto que a pluralidade do núcleo substantivo de um sujeito DP, no presente do indicativo, esteja de acordo com o verbo. A estratégia então adotada é dividir as categorias gramaticais em categorias menores, distinguindo as formas plurais e singulares.

Um exemplo de regras de estrutura sintagmática é recolocado abaixo de forma mais específica e categorias lexicais como Det se distinguem pelo acréscimo de informações como ‘sg’ e ‘pl’, singular e plural.

(29) S -> DP VP

S -> DPsg VPsg

S -> DPpl VPpl

O cuidado em desenvolver gramáticas modeladas de acordo com CFG que não venham a hipergerar passa pela revisão não só das noções de concordância e nuclearidade, como também pelas noções de subcategorização e transitividade.

Em relação à transitividade verbal, já sabemos que há verbos transitivos, intransitivos e bitransitivos. Sendo assim, uma gramática livre de contexto precisa distinguir as subcategorias da categoria V, pois dessa forma, V não é insuficiente para uma descrição que leve em conta as três categorias: verbos transitivos, intransitivos, bitransitivos.

A descrição de uma língua natural através de uma gramática livre de contexto, mesmo quando é teoricamente possível, depara-se com muitas dificuldades práticas. Uma das dificuldades de ordem prática é a elaboração de um enorme conjunto de regras para, sozinhas, compreenderem todas as suas restrições de ocorrência.

As restrições de co-ocorrência que as palavras precisam trazer juntas a si não viabilizam o uso do formalismo padrão de gramática livre de contexto. Elas acabam criando uma enorme quantidade de redundâncias em CFG. Para Klenk (2003, p.56), a divisão dos componentes sintáticos da gramática em um léxico e uma otimizada gramática de estruturas sintagmáticas foi uma medida de aprimoramento para o contínuo desenvolvimento de gramáticas livres de contexto.

O comportamento sintático das palavras é descrito através de outras informações adicionais nas entradas lexicais. Elas são apresentadas por meio de especificações de traços. Um conjunto de especificações de traços para um elemento linguístico chama-se estrutura de traços. Para substantivos, adjetivos e determinantes, são especificados traços de *gênero e/ou número e/ou caso e/ou pessoa.* Essas especificações não necessariamente precisam ser inventariadas juntas na mesma entrada lexical.

Uma estrutura de traços é um meio de representar informações gramaticais formalmente. Ela consiste da especificação de um conjunto de características, cada uma das quais é formada em par com um valor particular (SAG, WASOW, BENDER 2003, p.50).

Por exemplo, nós podemos tratar a categoria de uma palavra como *seminário* em termos de estrutura de traços que especifiquem apenas sua parte do discurso, seu número e gênero. Conforme Sag, Wasow e Bender (2003, p.51), essa categoria precisa incluir especificações apropriadas para três características: a parte do discurso (POS - *Part of Speech*) é substantivo (por conveniência chamaremos de nome), seu número (NUM) é singular (sg) e seu gênero (GEN) é masculino (m).

No caso de especificação de traços de verbos transitivos diretos que subcategorizam um objeto direto, cada verbo incorpora no léxico a característica SUBCAT e um número como valor.

Um verbo como '*prepara'* requer apenas um DP como objeto direto. Vejamos, então, o exemplo (39).

(30) Léxico:
prepara V[SUBCAT 2]
Regra:
VP -> V[SUBCAT 2] DP

A aplicação linguística de estruturas de traços necessita da divisão das categorias de estruturas de traços. Isto é, cada elemento linguístico é de um tipo, ao qual apenas algumas características são apropriadas. De acordo com Sag, Wasow, Bender (2003, p.61), o uso de tipos de elementos linguísticos funciona como uma base para classificar as estruturas de traços e as suas restrições.

A primeira distinção feita está entre *palavras* e *sintagmas.* Conforme a teoria, as regras gramaticais especificam as propriedades dos sintagmas. E o léxico fornece uma teoria de palavras. Em uma árvore, os nós do tipo IP, DP, VP são todos sintagmas. Já os nós do tipo D e V são apenas palavras.

Podemos então representar a categoria DP assim:

(31) DP = [sintagma; núcleo determinante]

A representação da estrutura de traços de uma entrada lexical para um nome como *seminário* é a seguinte:

(32) seminário, [palavra; núcleo substantivo]

A primeira propriedade que *palavras* e *sintagmas* têm em comum é a categoria. Como a teoria afirma, toda expressão especifica valores para suas características, pode-se então dizer que essas características constituem o núcleo. Chama-se núcleo devido à dependência que o sintagma tem do núcleo filho (SAG, WASOW, BENDER, 2003, p.59-61). Ou seja, um DP é determinante porque possui um Det dentro dele e Det é o núcleo filho da estrutura DP. O valor do núcleo serve para indicar a parte do discurso da expressão.

2.5 Análise sintática computacional (*parsing)*

Há algum tempo, o interesse em processar línguas naturais automaticamente tem se tornado um desafio para lingüistas e cientistas da computação. No campo da inteligência artificial, é possível pensarmos em máquinas ou programas capazes de traduzirem textos automaticamente ou em reconhecedores de vozes na comunicação entre pessoa-máquina.

Para esses fins, no entanto, as estruturas sintáticas de expressões linguísticas precisam ser compreendidas pelas máquinas ou pelos programas e por isso, de alguma forma se faz necessário utilizar analisadores sintáticos automáticos no desenvolvimento de programas capazes de simular aspectos da racionalidade humana que são aplicados, entre outras coisas, ao processamento da linguagem natural (ALENCAR, 2006, p. 17).

Ao longo desse trabalho nos ocupamos com a gramática de estrutura sintagmática sob o aspecto da produção de sentenças e do reconhecimento de expressões linguísticas. Como já sabemos, ao produzirmos ou compreendermos expressões linguísticas nós, de alguma forma, fazemos uma análise das estruturas dessas expressões. O termo utilizado em inglês para se referir a uma análise sintática é *parsing*. Por isso, é comum nos referirmos aos analisadores sintáticos automáticos como *parsers.*

Conforme Vieira e Lima,

> [...] um analisador sintático trabalha em nível de frase ou sintagma e irá reconhecer uma seqüência de palavras como constituindo uma frase da língua ou não. Poderá também construir uma árvore de derivação, que explicita as relações entre as palavras que compõem a sentença (2001, p.62)

Entende-se por *parsing* um procedimento definido em passos elementares que vão sendo executados através da interação do léxico, que reúne o conjunto de itens lexicais da língua, e de uma gramática, que define as regras de combinação dos itens na formação das frases (VIEIRA e LIMA, 2001, p.52).

A construção de fragmentos de gramática a serem processados por analisadores sintáticos automáticos passa pela compreensão dos meios utilizados por esses analisadores para o processamento das estruturas de línguas naturais.

Uma das formas utilizadas em processamentos linguísticos automáticos são as interfaces gráficas interativas. Entre as diversas formas de processamento de línguas naturais, há programas sendo desenvolvidos para que sejam capazes de analisar estruturas linguísticas e descobrir seus significados. Portanto, um intrigante desafio para o processamento automático de língua natural é o uso formal de gramáticas para descrever a estrutura de um conjunto ilimitado de sentenças.

Segundo Bird, Klein e Loper (2009), o NLTK (do inglês *Natural Language Toolkit*), que começou a ser desenvolvido na Universidade da Pensilvânia, é uma entre outros kits para o desenvolvimento de gramáticas computacionais.

A partir daí, dezenas de contribuintes passaram a ajudar nesse processo e muitas universidades passaram a adotar essa ferramenta em disciplinas afins. De modo geral, pode-se dizer que NLTK é uma biblioteca implementada na linguagem de programação *Python*. O NLTK consiste em um conjunto de programas utilizados na construção de outros programas.

O NLTK é uma ferramenta fácil de usar e desenvolvida para proporcionar uma abordagem tanto teórica quanto prática da exploração de estruturas de dados. Conforme Bird, Klein e Loper (2009), o NLTK é uma ferramenta que deve garantir consistência às estruturas de dados e às interfaces na realização de tarefas a partir de uma estrutura uniforme. Para garantir expansão e modificação da ferramenta o NLTK deve ser modular. O NLTK possui estruturas de dados e implementações próprias e contém uma infra-estrutura básica para ser utilizada na construção de programas.

Essa ferramenta de processamento automático de língua natural também é organizada como uma coleção de componentes para tarefas específicas. Conforme Bird, Klein e Loper (2009), cada módulo é uma combinação das estruturas de dados para a representação de um tipo particular de informação.

Um dos módulos fundamentais do NLTK é utilizado para criar e manipular informações linguísticas estruturadas. Esse módulo inclui árvores - para representação do processamento da análise das expressões*;* estrutura de traços - para construir e unificar estruturas de valores; gramáticas livres de contexto; e *parser* - para criar árvores de análise de uma estrutura a partir de um *input* (BIRD, KLEIN e LOPER, 2009).

Os programas que constituem o NLTK e manipulam os dados linguísticos operam em linguagem *Python*. Uma linguagem bastante funcional no processamento de dados linguísticos. Conforme Alencar (2007, p.1), caracteriza-se por ser uma sofisticada e fácil linguagem de programação e facilitar a exploração interativa. Uma das coisas mais amigáveis de *Python* é que ela permite ao usuário digitar diretamente dentro de um interpretador interativo – um programa que processará os programas executáveis em *Python*.

O NLTK tem utilizado a forma de representar a estrutura das sentenças através de árvores sintáticas. Desse modo, os *parsers* analisam uma sentença e automaticamente constroem uma árvore sintática. Uma árvore é um conjunto de nós etiquetados e conectados, cada um deles estendido por um único caminho de um nó até a raiz. As árvores podem ser usadas para codificar uma estrutura hierárquica que expande uma seqüência de formas linguísticas. No NLTK as árvores são criadas a partir das etiquetas dos nós e da lista dos filhos dados (BIRD, KLEIN e LOPER, 2009).

Segundo Bird, Klein e Loper (2009), há aspectos sistemáticos do significado que são muito mais fáceis de capturar uma vez identificadas as estruturas das sentenças. Alguns dilemas gramaticais como os casos de ambigüidade são alvos constantes de investigação e modelagem segundo esse método computacional e formal de descrição e análise gramaticais. Modelos bem-formados ou mal-formados de seqüência de palavras podem ser captados de acordo com a estrutura sintagmática e as relações de parentesco nela contidas.

Na literatura, vários tipos de *parsers* já foram propostos. Eles diferem entre si por meio da estratégia e do método de *parsing* empregados. Dentre os métodos de *parsing,* há o método que trabalha de modo descendente recursivo. Nesse caso, o *parser* realiza a análise de

um dado *input* aplicando a gramática de cima para baixo, começando pela categoria mais alta expandindo esse símbolo conforme as especificações à direita até se chegar a um elemento terminal (uma palavra). O *parser* processa esse elemento, o compara com o *input* e, no caso de se verificar uma correspondência entre eles, passa ao elemento seguinte (ALENCAR, 2007, p.17).

Por exemplo, uma regra como IP -> DP Ibar permite que o analisador automático substitua IP por um DP e um Ibar. Cada uma dessas categorias é substituída por outras categorias usando as categorias que estão à direita da seta nas regras de DP e de Ibar. Desse modo, a gramática é interpretada como uma especificação de como derivar o nível mais alto da sentença, o nível IP, em níveis mais baixos. Observe que utilizamos a notação Ibar para nos referirmos ao nível de projeção I'. Como veremos no capítulo 5, também o número de barras de uma projeção sintagmática pode ser modelada nas regras da gramática.

A cada estágio, o *parser* consulta a gramática para encontrar as regras de produção que podem ser usadas para alargar a árvore, até que se chegue a uma regra lexical. Daí, a palavra é comparada com o *input*. Após uma análise completa, o *parser* retrocede e procura por mais análises. Durante esse processo, o *parser* é forçado a escolher entre diversas possibilidades de produção.

Já sabemos que línguas naturais possuem uma enorme variedade de construções gramaticais que são difíceis de serem descritas com métodos simples de descrição. Fragmentos de gramática elaborados a partir da CFG são limitados em relação a aspectos das línguas como concordância e regência. Daí a necessidade de aumentar com as estruturas de traços as gramáticas livres de contexto.

Vimos que CFG baseada em traços deixam de lado aquelas noções de etiquetas atômicas como DP e V, por exemplo, e as decompõem em estruturas, cujas características podem receber uma variedade de valores. Não se pode gerar estrutura apenas com regras de reescrita, pois estas apresentam o problema da hipergeração. Para que isso não ocorra é necessário acrescentar traços morfossintáticos aos itens que compõem o léxico.

Conforme Alencar (2007, p.22) a capacidade do fragmento de gramática baseado em traços de analisar como gramaticais sintagmas nominais do tipo de '*aquele dentista', 'vocês alunas'* e '*estas meninas'* e ao mesmo tempo de rejeitar construções do tipo '*aquele dentistas', 'você alunos'* e *'estas meninos'* é transcrita nas seguintes regras:

(33) Fragmento de CFG enriquecida de traços

```
DP[num=?n, gend=?g, pers=?p] -> Det[num=?n, gend=?g, pers=?p] N[num=?n
gend=?g]
Det [num='sg', gend='m', pers= 3] -> 'estas' | 'vocês' | 'aquele'
N[num='pl', gend='f'] -> 'alunas' | 'meninas'
N [num='sg'] -> 'dentista'
```

Transcrição retirada de Alencar (2007, p. 22)

Podemos ler os traços da seguinte forma: 'num', 'gend' e 'pers' são termos empregados para designar atributos como os traços de número, gênero e pessoa. O símbolo '?' seguido por um elemento como 'n', 'g' ou 'p' corresponde a uma variável desses atributos.

Percebemos, com isso, que fragmentos de gramáticas modeladas de acordo com CFG baseada em traços, por possuírem vários tipos de informação sobre as entidades gramaticais, podem ser construídos e manipulados por programas do NLTK.

3 HIPÓTESES E PROCEDIMENTOS METODOLÓGICOS

Neste capítulo descreveremos o percurso feito nesta pesquisa desde o levantamento das questões e das hipóteses, até a metodologia adotada e o modo de análise.

Na seção 3.1 apresentamos os nossos questionamentos e as nossas hipóteses com base em um modelo formal de descrição gramatical, baseado nos pressupostos da teoria gerativa e visando a aplicação computacional. A seguir, na seção 3.2 demonstramos o método adotado na coleta feita para a constituição do nosso *corpus* e fazemos o levantamento do número de frases, palavras e gênero textuais de cada *corpus* eletrônico consultado. E na seção 3.3 fazemos uma demonstração do modo como procedemos na análise dos resultados.

3.1 Questões e hipóteses

Em nossa pesquisa partimos do pressuposto teórico de que a estrutura sintática de qualquer língua natural é formada por sintagmas ou categorias, constituídos de forma hierarquizada, que mantém entre si relações de parentesco. Com base nisso, nos fixamos no módulo da gramática gerativa denominado Teoria X-barra para buscar esclarecimentos sobre questões ligadas aos fenômenos de movimento e adjunção, aplicados por meio de operações sintáticas no processo derivacional da estrutura de sentenças encabeçadas por um sintagma IP, cujo especificador e complemento são respectivamente um DP sujeito e um DP objeto.

No nosso trabalho consideramos o pressuposto de que durante o processo derivacional de uma sentença são aplicadas algumas operações sintáticas de movimento, responsáveis pela cópia e pelo apagamento de itens lexicais ou constituintes sintáticos. Segundo Modesto (2009:6), de acordo com o Programa Minimalista, as operações podem ocorrer de modo visível, no componente aberto, ou de modo não visível, no componente coberto. Com base nessas ideias, utilizamos as noções de operações de movimento para pesquisar e elucidar o fenômeno de subida do verbo dentro de IP em frases finitas do português brasileiro. Por se tratar de uma pesquisa de cunho computacional, assumimos o pressuposto de que toda língua natural pode ser descrita de forma matemática ou lógica (DAVID, 2007, p.42) e pode ser implementada computacionalmente para fins diversos.

Portanto, com base nesses pressupostos, levantamos as seguintes questões:

a) Em português brasileiro, a aplicação de operações de movimento do verbo numa frase finita, ocorre apenas antes de *Spell-Out*, havendo assim apenas movimento

visível do verbo, ou seja, com conseqüência fonética? E em que medida podemos analisar sentenças finitas do PB, nas quais o movimento do verbo lexical ocorre de modo não visível?

b) Em que medida um fragmento de gramática do português, modelado de acordo com uma CFG baseada em traços, pode ter uma estrutura de traços capaz de representar características particulares da língua portuguesa como a concordância nominal e a concordância verbal, o fenômeno da subcategorização verbal, a projeção de barras e as categorias vazias?

c) Em que medida um analisador gramatical de expressões da língua portuguesa, modelado de acordo com CFG baseada em traços, pode ser processado computacionalmente por programas específicos para o processamento de línguas naturais?

E propomos as seguintes hipóteses:

a) Acreditamos que o português brasileiro aplica as operações de movimento tanto antes quanto após *Spell-Out,* permitindo que o movimento do verbo lexical também ocorra sem trazer conseqüências fonéticas ao processo derivacional da sentença. Numa categoria IP na qual o movimento de V ocorre de modo não-visível, como pensamos acontecer numa sentença como '*eles brutalmente agrediram o prisioneiro',* a categoria vazia é gerada em I em decorrência de V permanecer *in situ* na sintaxe visível.

b) Na medida em que um fragmento de gramática é modelado de acordo com CFG baseada em traços, ele se torna capaz de representar características próprias da língua portuguesa, se for enriquecido pela especificação de uma estrutura de traços, atribuídos a valores diretamente inseridos no léxico. No caso da relação temática estabelecida entre DP e V, seus traços correspondem aos valores de gênero e número, no caso da subcategorização de V, seu traço é atribuído ao valor de transitivo ou intransitivo e no caso da relação estrutural da projeção dos sintagmas, seus traços são atribuídos aos valores de uma ou duas barras.

c) Na medida em que um analisador gramatical de expressões da língua portuguesa é modelado de acordo com CFG baseada em traços, ele pode ser processado computacionalmente por programas específicos para o processamento de línguas naturais, desenvolvidos na linguagem de programação *Python*, disponíveis através da biblioteca de programas em forma de módulos.

3.2 Metodologia empregada

Inicialmente, tecemos nossa hipótese de que em português brasileiro o verbo dentro de IP pode permanecer *in situ,* em sintaxe visível, de modo que uma frase como *eles brutalmente espancaram os prisioneiros*[17] pode ser considerada uma frase gramatical, em oposição a Raposo (1999, p.33).

Para a nossa pesquisa optamos por fazer busca por dados em *corpora* eletrônicos. Essa opção se deu pela facilidade em acessar na rede, livremente e sem custos, o conjunto de recursos para a engenharia da linguagem em português, desenvolvidos no âmbito da Linguateca. Desse modo, através do projeto AC/DC tivemos acesso aos *corpora* a partir de um único local.

A Linguateca é um centro de recursos para processamento automático da língua portuguesa que mantém um portal constantemente atualizado e distribui livremente seus recursos através de internet. Esses recursos são distribuídos através de projetos desenvolvidos com o objetivo de disponibilizar *corpora* para lingüistas ou não, que pretendam fazer pesquisas em *corpora* eletrônicos do português.

Contudo, como os *corpora* disponíveis na Linguateca estão codificados em linguagem de programação, um lingüista leigo em programação muitas vezes precisa utilizar ferramentas especializadas em traduzir expressões linguísticas do português para esse tipo de linguagem. Daí a necessidade de se utilizar interfaces amigáveis que facilitem a consulta aos *corpora*.

Em nossa pesquisa utilizamos uma ferramenta computacional interativa chamada de "O Constructor"[18]. Esta interface nos auxilia na construção de comandos específicos destinados à obtenção de expressões linguísticas presentes nos *corpora* da Linguateca. Ele auxilia na construção das expressões pelo pesquisador e as interpreta para a linguagem de programação na qual os dados estão codificados.

Na página do sítio onde estão disponíveis os *corpora* eletrônicos, temos a opção de fazer um pedido de concordância em contexto de expressões linguísticas. No caso, os nossos pedidos em contexto constituíam-se por itens encadeados nas seguintes ordens:

[17] Exemplo retirado de Raposo (1999, p.33).
[18] Ferramenta computacional desenvolvida por Alencar (2002)

a) N-Adv-V-N
b) N-V-Adv-N

Então essas duas ordens são interpretadas pela interface *O Constructor* como expressões do tipo:

a) **[pos="N.*"&pos!="NUM.*"] [ ] {0,0} [word=".+mente"&pos="ADV.*"] [ ] {0,0} [pos="V.*"] within 1 s;** e
b) **[pos="N.*"&pos!="NUM.*"] [ ] {0,0} [word=".+mente"&pos="ADV.*"] [ ] {0,0} [pos="V.*"] within 1 s;**.

Inicialmente, começamos a nossa coleta de dados utilizando o *corpus* do PB chamado NILC/São Carlos (batizado de NILC por ter sido desenvolvido pelo Núcleo Interinstitucional de Linguística Computacional da Universidade de São Carlos), dessa coleta inicial uma ocorrência já sustenta a nossa hipótese, contrariando Raposo (1999).

Veja abaixo o modo apresentado pelo programa após a busca. No quadro consta uma lista com a o pedido de concordância em contexto, o nome do *corpus* consultado e o número de ocorrências do tipo de expressão linguística solicitada.

Resultados da procura - Thu Nov 27 14:09:17 WET 2008

Pedido de uma concordância em contexto
Corpus: NILC/São Carlos v. 8.0
4185 ocorrências.
Concordância
Procura: **[pos="N.*"&pos!="NUM.*"] [ ] {0,0} [word=".+mente"&pos="ADV.*"] [ ] {0,0} [pos="V.*"] within 1 s;**.

par=Esporte-94b-des-2: Ao preencher e assinar as fichas, os **interessados automaticamente autorizam** as investigações .

Quadro 1 - Resultado da busca no *corpus* NILC/São Carlos do PB retirado da Linguateca

Com base nos resultados do PB, nosso próximo passo foi recorrer a um *corpus* do português europeu chamado DiaCLAV (uma abreviação de 'Diário de Coimbra, Leiria, Aveiro e Viseu') com a finalidade de encontrar amostras de frases correspondentes à sequência dos itens da busca no *corpus* NILC/São Carlos, a fim de confrontá-las com as amostras do português brasileiro.

Veja abaixo os dados:

Resultados da procura - Thu Dec 18 17:35:15 WET 2008
Pedido de uma concordância em contexto
Corpus: DiaCLAV v. 3.0
769 ocorrências.

Concordância

Procura: **[pos="N.*"&pos!="NUM.*"] [ ] {0,0} [word=".+mente"&pos="ADV.*"] [ ] {0,0} [pos="V.*"] within 1 s;**.

---

par=DA-N0608-1: Os **pais facilmente abriram** a porta, vendo que se tratava do filho, pois de acordo com fontes próximas do casal, estes nunca abriam a porta a desconhecidos.

Quadro 2: resultado da busca no *corpus* DiaCLAV do PE retirado da Linguateca

No quadro acima, vê-se que nesse corpus do PE também já há um exemplo que contraria Raposo (1999). Após a consulta a esse dois *corpora* e com essas duas amostras obtidas, nosso próximo passo foi fazer novos levantamentos de dados em outros *corpora* da Linguateca, com o intuito de obter mais amostras e tornar os números finais mais expressivos. A partir do número total de ocorrências, procedemos desconsiderando todas as frases que não apresentassem um verbo transitivo direto ou que apresentassem advérbios diferentes daqueles que consideramos como atributos, conforme foi discutido no capítulo 2 dessa dissertação.

Na tabela abaixo podemos ver a descrição dos *corpora* consultados para essa pesquisa, de acordo com o tamanho (em unidades, palavras e frases), o gênero do seu conteúdo e a variante. O sítio da Linguateca disponibiliza todos os *corpora* que constituem o projeto AC/DC (Acesso a *corpora*/Disponibilização de *corpora*).

Tabela 1 – Breve descrição do tamanho em unidades, frases, palavras e a variante linguística em que consistem os *corpora*. Informações obtidas no site da Linguateca.

Corpora	**Tamanho (unidades)**	**Tamanho (palavras)**	**Tamanho (frases)**	**Variante (s)**	**Breve descrição**
AmostRA-NILC	124.836	98.505	4.965	BR	AmostRA-NILC
ANCIB	1.575.659	1.172.782	75.681	BR	Correio eletrônico correspondente ao tráfego na lista ANCIB
Avante!	7.768.261	6.503.189	204.414	PT	Semanário político Avante!, 1997-2002
Clássicos LP/Porto Editora	1.922.433	1.304.284	74.690	PT	Clássicos da literatura portuguesa, séc. XV e XIX
CONDIVport	7.117.746	5.570.215	328.214	PT BR	Jornais desportivos e revistas de moda e saúde
CoNE	925.228	685.231	31.561	PT BR	Mensagens de correio eletrônico não-endereçadas

Corpora	Tamanho (unidades)	Tamanho (palavras)	Tamanho (frases)	Variante (s)	Breve descrição
DiaCLAV	7.683.593	6.651.554	213.305	PT	Diário de Coimbra, Diário de Leiria, Diário de Aveiro, Viseu Diário
ECI-EBR	914.750	724.006	44.381	BR	Texto do corpo Borba-Ramsey,
Natura/Minho	2.156.187	1.749.083	68.910	PT	Jornal regional Diário do Minho, antes da revisão
Natura/Público	7.369.349	6.274.542	225.752	PT	Jornal PÚBLICO, dois parágrafos por notícia, 1991-1994
NILC/São Carlos	42.157.263	32.459.483	1.963.795	BR	Texto do corpo NILC, contendo majoritariamente texto jornalístico, mas também cartas comerciais e textos didáticos
Vercial	8.968.057	8.376.956	383.805	PT	Clássicos da literatura portuguesa, século XIX
Total	447.198.624	362.188.147	16.105.168	PT BR	todos os *corpora*

Como podemos ver na tabela acima, o tamanho total dos *corpora*, constitui um número bastante grande de frases.

A nossa intenção de fazer busca por dois tipos diferentes de sequências, justifica-se pela necessidade de compararmos os números das expressões linguísticas do PB e do PE.

3.3 Descrição da análise

A partir do número das ocorrências obtidas em cada *corpus*, eliminamos o que consideramos estruturas insatisfatórias, como por exemplo, uma sentença que apresentasse um verbo no particípio, e assim chegamos ao número de amostras correspondentes à uma expressão como *eles agrediram brutalmente o prisioneiro*, como mostraremos nas tabelas a seguir.

Do total de *corpora* apresentado na tabela 1, apenas em alguns encontramos amostras para nossa pesquisa.

Abaixo veremos duas tabelas organizadas em duas colunas. A primeira coluna informa o nome do *corpus* e a segunda o número de amostras obtidas da busca. Nas tabelas, demonstramos uma comparação dos resultados do PB e do PE em relação ao número de amostras na sequência N V Adv N.

Tabela 2 – número de amostras do PB por número de ocorrências obtidas da sequência N V Adv N

***Corpora* do PB**	**Número de amostras**
ANCIB v.3.4	2
ECI/EBR v. 5.0	2
NILC/São Carlos v. 8.0	31
Total:	35

Tabela 3 - número de dados do PE por número de ocorrências obtidas da sequência N V Adv N

***Corpora* do PE**	**Número de amostras**
Avante! V. 1.3	7
Clássicos da Literatura Portuguesa/Porto Editora	6
DiaCLAV v. 3.0	16
Natura/Minho v. 4.1	1
Natura/Público v. 5.0	10
Total:	40

Na tabela 2, referente ao PB vemos um total de 35 amostras satisfatórias à nossa pesquisa, para um número total de 4587 ocorrências em contexto, nos três *corpora*. Desconsideramos aqui os *corpora* que não apresentaram amostras. Já em PE, de acordo com a tabela 3, 5 *corpora* um total de 40 amostras, de um total de 8019 ocorrências totais obtidas nos *corpora* citados. Do mesmo modo desconsideramos os *corpora* que não apresentaram amostras.

As próximas duas tabelas comparam os resultados do PB e do PE em relação ao número de ocorrências de expressões lingüísticas na ordem e N Adv V N, tal qual uma

expressão como *eles brutalmente agrediram o prisioneiro*, demonstrando exemplos que podem contrariar Raposo (1999), segundo o qual trata-se de uma frase agramatical.

Tabela 4 - número de dados do PB por número de ocorrências obtidas da sequência N Adv V N

Corpus do PB	**Número de amostras**
ANCIB v.3.4	1
NILC/São Carlos v. 8.0	1
Total:	2

Tabela 5 - número de dados do PE por número de ocorrências obtidas na sequência N Adv V N

Corpus do PE	**Número de amostras**
DiaCLAV v. 3.0	1
Natura/Minho v. 4.1	1
Total:	2

Conforme vemos, as tabelas 4 e 5 apresentam ambas 3 amostras. Vê-se que a sequência N Adv V N aparece em um número bastante reduzido, tanto em PB como em PE, se comparado às amostras da sequência N V Adv N, vistas nas tabelas 2 e 3.

A partir das amostras das tabelas acima, concluímos então que os números referentes à sequência NAdvVN, tanto do PB como do PE não são significativos o suficiente para expressar uma contestação à afirmação de Raposo (1995) de que a sequência N Adv V N não se trata de um paradigma distribucional da língua portuguesa. Nesse caso sugerimos uma consulta ampliada à *corpora* da língua portuguesa, a fim de elucidar questões pertinentes ao

fenômeno de movimento de verbo, principalmente em relação à força dos traços na sintaxe visível e não visível da derivação.

4 DESCRIÇÃO SINTATICA GERATIVO-TRANSFORMACIONAL DE SENTENÇAS FINITAS DA LINGUA PORTUGUESA

O capítulo 4 consiste de uma seção, dividida em duas subseções. Na subseção 4.1.1, utilizaremos as condições de restrições de movimento, com base em Raposo (1992, p.228), para exemplificar o movimento de V. Utilizaremos estas condições de restrições à luz do PM sobre as regras de movimento do verbo e mostraremos brevemente como este autor analisa, de modo paramétrico, o movimento do verbo no português de Portugal e em inglês.

A seguir, na subseção 4.2, consideramos como modelo para nossa análise o esquema X-barra de Fukui (1986) de que a projeção de uma expressão linguística como, por exemplo, *Pedro visitou Maria,* se dá a partir de I e não de V. Ainda nesta seção faremos análises de três tipos de frases do português. Descreveremos, na análise 1, três frases do tipo SVO (sujeito-verbo-objeto), como ponto de partida para as duas análises seguintes. Na análise 2 e na análise 3, descrevemos seis expressões da língua portuguesa encabeçadas pelo sintagma IP, nas sequências NVAdvN e NAdvVN, obtidas com base nas amostras do nosso conjunto de *corpora*. Por fim, ao final de cada análise apresentamos o sistema de regras que constitui o fragmento de gramática livre de contexto (CFG) elaborado em nosso trabalho, o qual traz em sua estrutura regras sintagmáticas e um léxico que comporta itens lexicais encontrados nas expressões linguísticas analisadas nesta pesquisa. A partir desta CFG construímos um analisador automático e o aplicamos a um *parser,* em linguagem de programação *Python*, distribuído pelo NLTK, como será visto no capítulo 5 dessa dissertação.

4.1 Sentenças finitas analisadas de acordo com Fukui (1986)

Nesta seção, primeiramente, utilizaremos as condições de restrições de movimento, com base em Raposo (1992, p.228), para exemplificar o movimento de V. Utilizaremos estas condições de restrições à luz do PM e mostraremos brevemente como este autor analisa, de modo paramétrico, o movimento do verbo no português de Portugal e em inglês.

4.1.1 Condições de restrições ao movimento de núcleos

Uma das primeiras formulações sobre condições de economia em relação à operação Mover foi proposta por Chomsky (1986b)[19], através da idéia de que a operação

[19] Chomsky, N. *Knowledge of language: its nature, origins and use*. New York, Praeger, 1986b.

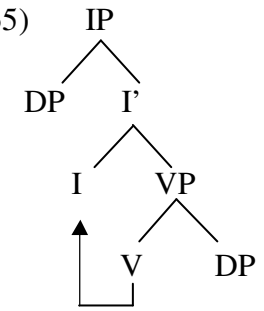
Mover só se aplica para satisfazer propriedades morfológicas das expressões sintáticas (*apud* Raposo, 1999, p. 30). No caso específico do sintagma IP, podemos dizer então que V se completa morfologicamente em I através do movimento de subida que realiza.

Conforme Raposo (1992, p 228), com base nos resultados dos estudos de Travis (1984), uma das propriedades essenciais dos movimentos de núcleos é a **Restrição Sobre Movimentos Nucleares**. Essencialmente, esta restrição propõe que o alvo do movimento de um núcleo é sempre outro núcleo, assim como o núcleo alvo do movimento é aquele imediatamente superior ao núcleo movido.

Na árvore (35) podemos ver configuração arbórea da frase abaixo:

(34) A criança utiliza o computador

(35)



Conforme se vê na configuração acima, I é núcleo e filho de I' e V constitui o núcleo e o filho de VP. Ambos são núcleos e ambos mantém uma relação de parentesco com I'. As restrições impostas ao sentido do movimento faz com que o alvo do movimento de V seja I. Podemos dizer então que apenas V pode se mover em direção a I para se complementar morfologicamente com suas marcas flexionais de I.

4.1.2 O movimento de núcleo dentro de IP no Programa Minimalista

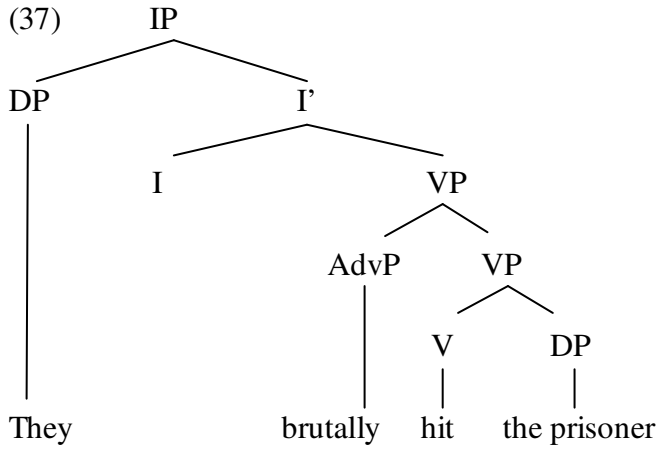
Consideramos a idéia de que as interfaces LF e PF não possuem elementos estranhos aos seus respectivos sistemas de performances. Assim, de acordo com o Princípio da Interpretação Plena, elementos como os traços de concordância, não são elementos legítimos da forma lógica, exaustivamente constituída por entidades semânticas, organizadas no modo exigido pelo sistema C-I (RAPOSO, 1999, p.27-28).

Já vimos que o fenômeno de subida de V para a categoria funcional I é efetuado por uma operação Mover. Observemos agora a seguinte sentença do inglês, retirada de Raposo (1999, p.33)

(36) They brutally hit the prisoner

Vejamos a árvore abaixo:

(37)



Para Raposo (1999, p.33), a ordem dos elementos da sentença em inglês difere do português. Ele considera agramatical em português uma sentença como a que apresentamos abaixo, cujos elementos estão dispostos em uma ordem que chamamos Adv V DP, do modo como ocorre em inglês.

Vejamos o exemplo retirado de Raposo(1999) tido como agramatical :

(38) *Eles brutalmente agrediram os manifestantes

Contudo, em nossas buscas em *corpora* eletrônicos, encontramos exemplos como a sentença (39) retirada do *corpus* NILC/São Carlos referente ao português de Brasil e a sentença (40), retirada do *corpus* DiaCLAV, referente ao português de Portugal. Em ambas um atributo com as mesmas características semânticas e funcionais do advérbio da sentença em inglês vista acima, constitui uma frase na ordem Adv V DP, exatamente como em inglês.

(39) Os interessados automaticamente autorizam as investigações

(40) O discurso unicamente particulariza a informação

Conforme as orientações teóricas do programa minimalista, segundo as colocações de Raposo (1999, p. 27) a respeito da arquitetura geral do modelo, a posição do atributo não é decorrente do sentido do movimento do verbo, pois como já sabemos, há condições que são impostas ao movimento de núcleos.

Com base nas discussões aqui levantadas, somos levados a atribuir o que chamamos de sequência Adv V DP, das sentenças acima, ao tipo de operação de movimento aplicada no processo derivacional da estrutura. Isto é, quando dizemos que o parâmetro Adv V DP também é possível em português, queremos dizer que as operações de movimento ocorreram de modo não visível. Nessa pesquisa, para fins explicativos, analisamos algumas sentenças da sequência Adv V DP dizendo que não houve movimento do verbo.

Os exemplos (39) e (40) corroboram com a nossa proposta de que assim como em inglês, também em português o atributo modificador de VP pode preceder o verbo. Contrariamente à posição de Raposo (1999) de que em português, essas sentenças são agramaticais.

Levando em conta as colocações acima e considerando que a projeção de uma expressão linguística como, por exemplo, *os familiares ajudam seus parentes,* se dá a partir de I e não de V, utilizaremos o esquema X-barra de Fukui (1986) como modelo para as análises que faremos a seguir.

Referir-se a sentenças finitas significa falar em sentenças cuja estrutura sintagmática é encabeçada por um sintagma flexionado que subcategoriza um complemento. Usando termos mais tradicionais pode-se dizer, então, que IP encabeça uma sentença que possui um I que c-seleciona um VP e um V que s-seleciona um DP objeto.

Nas análises que veremos a seguir, apresentamos a categoria IP e sua representação arbórea, referente a três tipos diferentes de sentenças. Dizemos aqui que cada tipo de sentença tem uma ordem específica de elementos. Cada análise corresponde a um tipo de ordem diferente de frase: i. SVO; ii. SVAdvO; iii. SAdvVO.

ANÁLISE 1

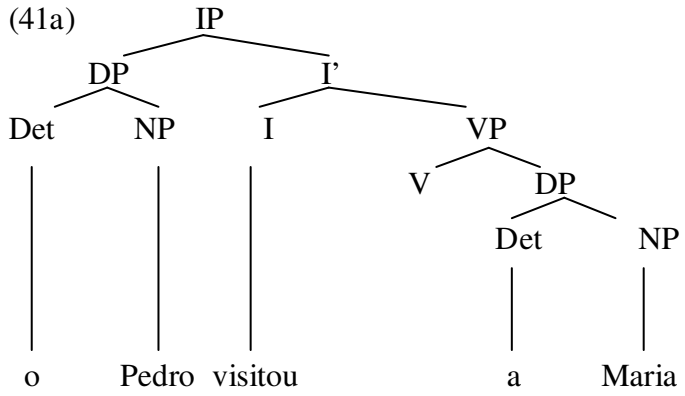
Nessa análise consideramos apenas uma expressão linguística encabeçada por IP que consiste de DP na posição de especificador de IP e de DP como argumento interno. Observemos o exemplo a seguir:

O Pedro visitou a Maria

Queremos deixar claro para o leitor que essa sentença foi aleatoriamente escolhida para exemplificar uma sequência do tipo SVO. Acreditamos que o leitor não achará improvável que uma frase como essa já tenha sido expressa ou de que possa vir a ser.

De acordo com a análise feita pelo nosso analisador ao ser implementado computacionalmente, a frase acima demonstra ser passível de duas analises, conforme veremos nas árvores abaixo.

(41a)



A partir dos nós da arvore acima nós podemos extrair as seguintes regras:

(41b) IP -> DP I'

DP -> Det NP

I' -> I VP

VP -> V DP

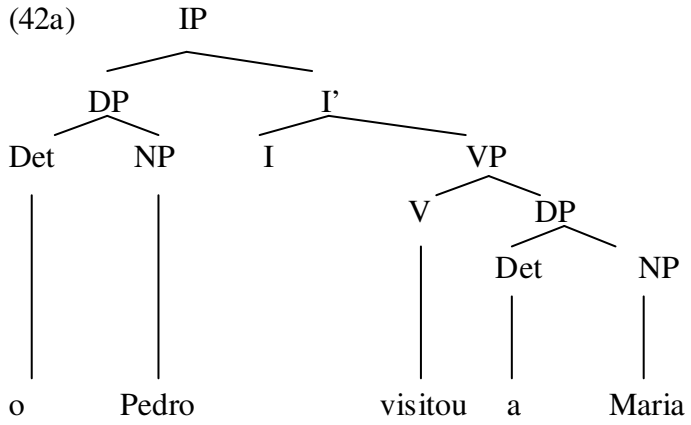
NP -> 'Pedro' | 'Maria'

Det -> 'o' | 'a'

I -> 'visitou'

V ->

(42a)



E extraímos as seguintes regras da árvore acima:

(42b) IP -> DP I'
DP -> Det NP
I' -> I VP
VP -> V DP
Det -> 'o' | 'a'
NP -> 'Pedro' | 'Maria'
I ->
V-> 'visitou'

Desse modo, nós podemos dizer que no caso da árvore (50a) houve movimento visível do verbo, provavelmente pela necessidade de checagem de traços antes da retirada dos elementos da Forma Fonética, gerando em V a categoria vazia. Em (50b), pelo contrário, dizemos que não houve movimento visível do verbo, que não sobe na sintaxe visível para juntar-se a I, deixando essa categoria foneticamente vazia.

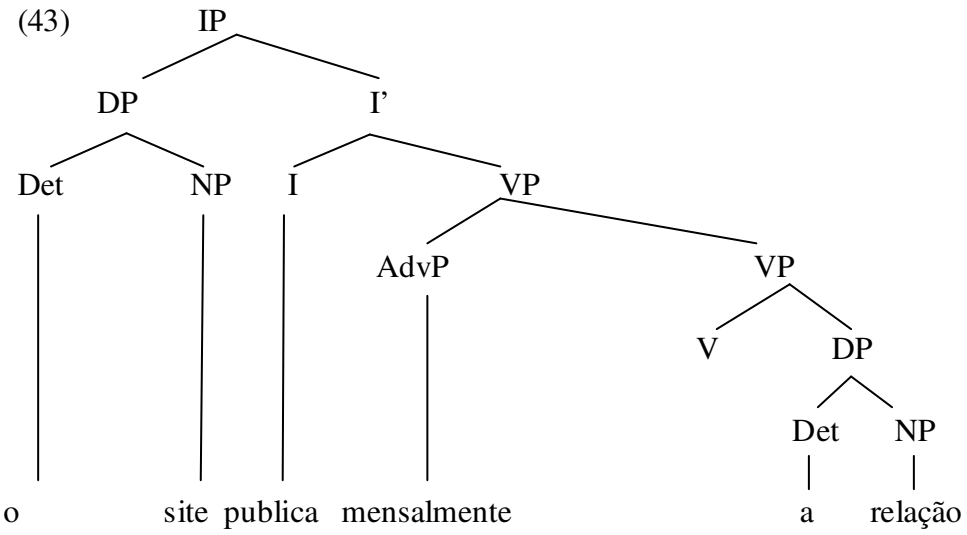
ANÁLISE 2

A partir das sentenças a seguir veremos como um IP pode sofrer uma modificação na seqüência SVO apresentada acima, pelo mecanismo de adjunção de um atributo à posição direita de V. Através desse modificador, a posição da análise 1, SVO, passa a constituir a ordem SVAdvO, como veremos nas frases abaixo.

O site publica mensalmente a relação
Esses serviçais tratavam invariavelmente a clientela.
O governo puniu cruelmente os rebeldes

As frases acima foram obtidas dos *corpora* ANCIB, ECI-EBR e NILC/São Carlos, respectivamente. Nestas frases o advérbio ocorre à direita do verbo. Nesse caso dizemos que o movimento de V para se juntar ao seu morfema flexional I foi feito na sintaxe visível.

O esquema (43) é a representação arbórea das frases acima.



Para essa frase, podemos fazer uma análise dizendo que ao pronunciarmos uma sentença com o advérbio ocupando a posição posterior ao verbo, este verbo faz um movimento de subida que é anterior a *Spell Out*, portanto um movimento visível.

Observe que apresentamos I numa forma amalgamada de I com V, conforme elucidação à pagina 28.

Desse esquema arbóreo obtemos as seguintes regras:

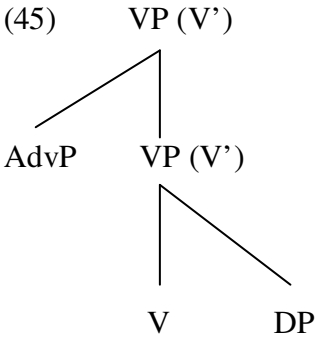
(44) IP -> DP I’
DP -> Det NP
I’ -> I VP
VP -> AdvP VP | V DP
NP -> ‘relação’ | ‘site’
Det -> ‘a’ | ‘o’
V ->
AdvP -> ‘mensalmente’
I -> ‘publica’

Como não podemos representar o movimento que supomos haver no processo de derivação, apenas demonstramos o seu resultado Note que nas regras acima V não possui nenhuma entrada lexical. Esse modo de fazer a regra é uma forma de representar a categoria vazia.

ANÁLISE 3

Consideramos agora o terceiro caso, no qual o atributo antecede V na ordem dos elementos da sentença. Eles constituem uma frase com os elementos na ordem SAdvVO. O modificador do VP ocupa a posição esquerda desta categoria. Vejamos uma configuração de adjunção deste tipo:

(45)



Na configuração acima, o VP imediatamente superior a V é um V’ e este V’ projeta um outro sintagma V’ pela adjunção de AdvP.

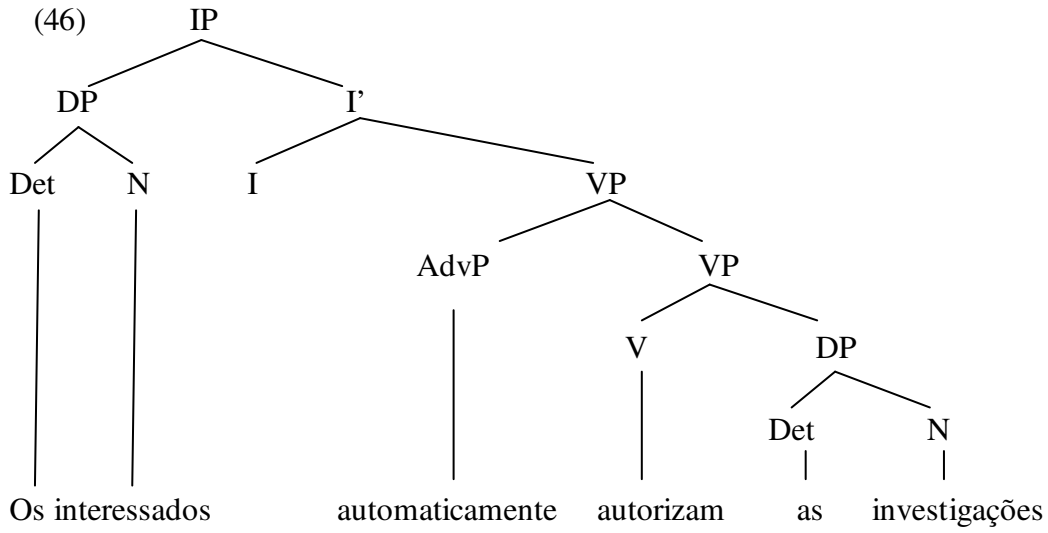
Os pais facilmente abriram a porta
Os interessados automaticamente autorizam as investigações
Os procuradores persistentemente transgrediram as regras

As frases acima foram retiradas dos *corpora* DiaCLAV, NILC/São Carlos e Natura/Minho, respectivamente Ao observá-las, podemos ver os atributos *facilmente, automaticamente* e *persistentemente* ocupando a posição à esquerda do verbo. Para esta análise, recorremos às regras transformacionais de movimento para tentar explicar o que levaria o modificador de VP a ser pronunciado ora entre o verbo e o seu objeto, ora entre o sujeito e o verbo.

Nesta análise levamos em conta as noções de sistema de projeção de categorias lexicais e funcionais. Segundo Fukui (1986, p. 2), essas se diferenciam entre si por possuírem, a primeira uma estrutura conceitual-lexical e a segunda uma função de ligar duas unidades sintáticas, através de algum tipo de 'regência' ou 'concordância'.

Diante dessa questão da posição do atributo nas frases acima, corroborada pelos resultados que constituem nosso *corpus*, atendemos ao questionamento primeiro do nosso trabalho sobre a gramaticalidade da ordem SAdvVO em PB e em PE.

Vejamos a representação arbórea das frases acima:



Primeiramente consideramos que palavras como *abriram*, *autorizam* e *transgrediram* são constituídas de duas partes: a primeira corresponde a V, categoria lexical identificada nos verbos acima como os radicais '*abr*'-, '*autoriz*'-, '*transgred*-; a segunda parte

corresponde ao que chamamos de I, categoria funcional com marcas próprias de sua categoria como tempo, número, pessoa, identificadas como -‘*am*’.

Podemos perceber que V, sendo uma categoria lexical, possui uma noção conceitual, também chamada de noção interpretativa. Esta precisa ser adicionada das informações funcionais próprias de I. São estas informações funcionais que ligam a noção conceitual de V a uma unidade sintática como as marcas de concordância de I.

Na análise 3 podemos observar o atributo na posição entre V e o DP sujeito, pelo motivo de V permanecer *in situ*. Ao contrário das frases da análise 2, as quais nos mostram o atributo ocupando o lugar de modificador de VP, na posição entre V e seu DP objeto.

Em nossa análise consideramos que a posição do atributo entre o DP sujeito e V é decorrente do fato de não haver operação de movimento do verbo em sintaxe visível. Assumimos em nossa análise, portanto que em sentenças do tipo visto acima não há movimento de V para I, o que leva o processo derivacional a gerar em I a categoria vazia.

E do esquema arbóreo apresentado em (46), retiramos dos seus nós as regras abaixo.

(47) IP -> DP I’
DP -> Det NP
I’ -> I VP
VP -> AdvP VP | V DP
NP -> ‘interessados’ | ‘investigações’
AdvP -> ‘automaticamente’
Det -> ‘os’ | ‘as’
V -> ‘autorizam’
I ->

Nesta seção nós vimos a análise de três diferentes estruturas da língua portuguesa. Essas análises foram feitas com a ajuda do nosso analisador gramatical, como apresentaremos no próximo capítulo. A partir de configuração arbórea esboçada em cada análise, fomos montando passo a passo o conjunto de especificações das regras da nossa gramática.

Da elaboração das regras de cada tipo de árvore acima partiremos enfim para a elaboração final do nosso fragmento de gramática, enriquecendo-o com uma estrutura de traços que o torne capaz de analisar frases de uma lista previamente estabelecidas, contendo valores específicos para a unificação dos elementos linguísticos.

5 ANALISADOR SINTÁTICO AUTOMÁTICO (*PARSER*) DO PORTUGUÊS

O presente capítulo divide-se em duas seções. Na primeira seção é feita uma demonstração do modo como estruturas linguísticas modeladas de acordo com CFG pode ser expandida para a acomodação da estrutura de traços.

Na seção 5.2, faremos uma demonstração de como um analisador gramatical, elaborado no formalismo CFG baseada em traços, é capaz de analisar, de modo automático, estruturas linguísticas e representar em árvores as relações de parentesco estabelecidas entre os sintagmas, pela unificação dos elementos da sentenças.

5.1 Fragmento de gramática computacional do português

Segundo Bird, Klein e Loper (2009, p. 50), os grupos de traços e valores, conhecidos como estrutura de traços, farão com que os analisadores automáticos combinem apenas elementos que possuem valores compatíveis entre si. Isso faz com que os *parsers* não analisem sentenças agramaticais ao processar sentenças. Mostraremos como a estrutura de um fragmento de gramática, modelado segundo CFG, pode ser expandido para acomodar categorias vazias na estrutura de traços.

Com base nas frases analisadas no capítulo anterior desenvolvemos um fragmento de gramática da língua portuguesa. A partir dos exemplos abaixo, podemos ter um fragmento de gramática modelado para representar categorias vazias através de regras que demonstrem que para cada elemento vazio da frase, chamado por Bird, Klein e Loper (2009) de **lacuna** (em inglês, *gap*), deve haver um elemento que funcione como um **preenchedor** (em inglês, *filler*) dessa lacuna.

(48) O site publica mensalmente a relação

(49) Esses serviçais tratavam invariavelmente a clientela.

(50) O governo puniu cruelmente os rebeldes

Nas frases acima, considera-se que o verbo se movimenta de modo visível até I, gerando em V uma categoria vazia. Portanto, na construção do nosso analisador com base na estrutura de traços da CFG, refinamos o sistema de regras das produções lexicais e gramaticais através do enriquecimento das regras com traços que podem capturar os traços do preenchedor e os da lacuna.

Como é possível observar, no fragmento em (51) abaixo, temos o símbolo '/?x' demonstrando a ausência de um elemento movido, no nosso caso, o verbo.

(51) IP -> DP Ibar
DP -> Det NP
NP -> N
Ibar -> V VP/?x
VP/?x -> AdvP VP/?x | V/?x DP
AdvP -> Adv
Adv -> 'mensalmente' | 'invariavelmente' | 'cruelmente'
V/V ->
V -> 'publica' | 'tratavam' | 'puniu'
Det -> 'o' | 'a' | 'esses' | 'os'
N -> 'site' | 'relação' | 'serviçais' | 'clientela' | 'governo' | 'rebeldes'

De acordo com o suporte do NLTK para *parsers* de CFG, deve haver, contudo, um momento da produção gramatical do nosso *parser*, no qual o verbo precisa se acomodar. Note que, para fins de implementação computacional, reescrevemos a regra Ibar -> V VP/?x com o núcleo V ao invés de I, em violação ao princípio de endocentricidade, conforme discussão da página 23 do capítulo 2. Fizemos isto para demonstrar que V é a representação do resultado da concatenação do verbo aos seus morfemas flexionais, I. Desse modo podemos entender que I continua presente na expansão do sintagma.

As frases acima foram utilizadas para demonstrar como modelar categorias vazias. Agora, utilizamos as frases abaixo para pensar no conjunto de regras que dê conta não só do tipo de frase SVAdvO, mas também SAdvVO. Essa frases foram retiradas do conjunto de frases que constituem nosso *corpora*. As regras compreendem a sentenças do tipo SAdvVO e SVAdvO, conforme a análise 2 e a análise 3, feitas na seção 2 do capítulo anterior.

Desse modo, com o conjunto de regras que nós veremos, nosso fragmento também será capaz de analisar as frases abaixo, que, de acordo com a nossa hipótese de movimento visível e não visível do verbo, não apresentam movimento visível de V.

(52) Os pais facilmente abriram a porta

(53) Os procuradores persistentemente transgrediram as regras

(54) O discurso unicamente particulariza a informação

(55) Os interessados automaticamente autorizam as investigações

(56) O preso frequentemente vomita sangue

Para analisar as sentenças acima, precisamos incluir no fragmento (51), outros tipos de regras. Conforme veremos em (57), o fragmento pode conter diferentes expansões de uma dada categoria, separadas por barras verticais, representando cada uma, nós específicos das suas respectivas árvores.

(57) IP -> DP Ibar

DP -> Det NP

Ibar -> V VP/?x | I VP

VP/?x -> AdvP VP/?x | V/?x NP | V/?x DP

VP -> AdvP VP | V DP | V NP

NP -> N

AdvP -> Adv

Adv -> ‘mensalmente’ | ‘invariavelmente’ | ‘cruelmente’ | ‘facilmente’ | ‘persistentemente’ | ‘unicamente’ | ‘automaticamente’ | ‘frequentemente’

V/V ->

V -> ‘publica’ | ‘tratavam’ | ‘puniu’ | ‘abriram’ | ‘transgrediram’ | ‘particulariza’ | ‘autorizam’ | ‘vomita’

I ->

Det -> ‘o’ | ‘a’ | ‘esses’ | ‘os’ | ‘as

N -> ‘site’ | ‘relação’ | ‘serviçais’ | ‘clientela’ | ‘governo’ | ‘rebeldes’ | ‘pais’ | ‘porta’ | ‘procuradores’ | ‘regras’ | discurso’ | ‘informação’ | ‘interessados’ | ‘investigações’ | ‘preso’ | ‘sangue’

Até agora modelamos nosso fragmento para que o analisador automático analise frases do tipo SVAdvO e SAdvVO. Porém, como partimos de uma frase base do tipo SVO,

consideramos necessário modelar nosso fragmento com regras que também sejam capazes de analisar esse tipo de estrutura.

(58) O Pedro visitou a Maria

De acordo com as frases acima, incluímos mais uma vez ao nosso fragmento as regras obtidas da frase acima, como se pode ver na análise 1 do capítulo anterior.

(59) IP -> DP Ibar

DP -> Det NP

Ibar -> V VP/?x | I VP

VP/?x -> AdvP VP/?x | V/?x NP | V/?x DP

VP -> AdvP VP | V DP | V NP

NP -> N

AdvP -> Adv

Adv -> ‘mensalmente’ | ‘invariavelmente’ | ‘cruelmente’ | ‘facilmente’ | ‘persistentemente’ | ‘unicamente’ | ‘automaticamente’ | ‘frequentemente’

V/V ->

V -> ‘publica’ | ‘tratavam’ | ‘puniu’ | ‘abriram’ | ‘transgrediram’ | ‘particulariza’ | ‘autorizam’ | ‘vomita’ | ‘visitou’

I ->

Det -> ‘o’ | ‘a’ | ‘esses’ | ‘os’ | ‘as

N -> ‘site’ | ‘relação’ | ‘serviçais’ | ‘clientela’ | ‘governo’ | ‘rebeldes’ | ‘pais’ | ‘porta’ | ‘procuradores’ | ‘regras’ | discurso’ | ‘informação’ | ‘interessados’ | ‘investigações’ | ‘preso’ | ‘sangue’ | ‘Pedro’ | ‘Maria’

Vimos até aqui como modelar regras de sentenças nas ordens SVO, SVAdvO e SAdvVO. Estas mesmas regras também foram modeladas para analisar sentenças que possuem categorias vazias.

Partiremos, então, para a modelação do nosso fragmento com o enriquecimento da sua estrutura de traços, além da modelação das categorias vazias. Esse enriquecimento dará conta de fenômenos sintáticos como concordância nominal e verbal, subcategorização, além

de receber valores que atribuem barras aos itens lexicais e sintagmáticos das estruturas linguísticas em questão.

Nós começaremos pelo fenômeno da concordância sintática. Como já foi visto, gramáticas livres de contexto não são satisfatórias para dar conta de relações como a relação de concordância entre Det e N, ou entre DP sujeito e I'. Vejamos a seguir:

(60) os psiquiatras preencheram os prontuários

(61) *o psiquiatra preencheram os prontuários

Assim como Det e N concordam em número e gênero nos exemplos acima, no exemplo (61) o sujeito e o verbo não concordam em número, o que a torna uma sentença agramatical. Os exemplos acima mostram que os verbos têm pelo menos duas formas de flexão: uma para pessoa e outra para número. Portanto, nossa gramática precisa conter regras, cujas marcas de concordância tenham seus traços de pessoa e número atribuídos a valores como feminino ou masculino, singular ou plural.

Para que uma gramática não hipergere coisas do tipo (61), enriquecem-se as regras de produção gramatical e as regras de produção lexical com certos traços e certos valores próprios das propriedades de cada item lexical ou gramatical.

Da mesma forma, nossa gramática baseada em traços foi enriquecida com traços e valores correspondentes a tempo e acordo e subacategorização verbal, além do número de barras referente à projeção dos itens lexicais. Em nosso fragmento, atribuímos às regras traços como tempo verbal 'presente' e 'passado', acordo verbal de 'número' e 'pessoa', verbo transitivo e número de barras.

Desse modo, veremos em (62) abaixo, o que acontece quando nós codificamos as restrições de unificação dos traços às regras do fragmento visto em (59).

.

(62) I[BAR=2, AGR=?a] -> D[BAR=1, AGR=?a] I[BAR=1, AGR=?a]

D[BAR=1, AGR=?a] -> D[BAR=0, AGR=?a] N[BAR=1, AGR=?a]

N[BAR=1, AGR=?a] -> N[BAR=0, AGR=?a]

I[BAR=1, AGR=?a, TENSE=?t] -> I[BAR=0] V[BAR=1, TENSE=?t, AGR=?a] | V[BAR=0, AGR=?a, SUBCAT=?s, TENSE=?t] V[BAR=1]/?x

V[BAR=1]/?x -> Adv[BAR=1] V[BAR=1]/?x | V[BAR=0]/?x D[BAR=1] | V[BAR=0]/?x N[BAR=1]

V[BAR=1, TENSE=?t, AGR=?a] -> Adv[BAR=1] V[BAR=1, AGR=?a, TENSE=?t] | V[BAR=0, TENSE=?t, AGR=?a, SUBCAT=?s] D[BAR=1] | V[BAR=0, TENSE=?t, AGR=?a, SUBCAT=?s] N[BAR=1]

Adv[BAR=1] -> Adv[BAR=0]

Adv[BAR=0] -> ‘mensalmente’ | ‘invariavelmente’ | ‘cruelmente’ | ‘facilmente’ | ‘persistentemente’ | ‘unicamente’ | ‘automaticamente’ | ‘frequentemente’

I[BAR=0] ->

V[BAR=0]/V ->

V[BAR=0, SUBCAT=trans, TENSE=past, AGR=[NUM=sg, PER=3]] -> 'visitou' | ‘puniu’

V[BAR=0, SUBCAT=trans, TENSE=pres, AGR=[NUM=sg, PER=3]] -> 'particulariza' | 'vomita' | 'publica'

V[BAR=0, SUBCAT=trans, TENSE=past, AGR=[NUM=pl, PER=3]] -> 'abriram' | 'transgrediram'

V[BAR=0, SUBCAT=trans, TENSE=pres, AGR=[NUM=pl, PER=3]] -> 'autorizam' | ‘tratavam’

N[BAR=0, AGR=[NUM=sg, GND=fem]] -> 'Maria' | 'porta' | 'informação' | 'relação' | ‘clientela’

D[BAR=0, AGR=[NUM=sg, GND=fem]] -> 'a'

N[BAR=0, AGR=[NUM=pl, GND=fem]] -> 'regras' | 'investigações'

D[BAR=0, AGR=[NUM=pl, GND=fem]] -> 'as'

N[BAR=0, AGR=[NUM=sg, GND=masc]] -> 'Pedro' | 'preso' | 'sangue' | 'site' | 'discurso' | ‘governo’

D[BAR=0, AGR=[NUM=sg, GND=masc]] -> 'o'

N[BAR=0, AGR=[NUM=pl, GND=masc]] -> 'pais' | 'procuradores' | 'interessados' | ‘serviçais’ | ‘rebeldes’

D[BAR=0, AGR=[NUM=pl, GND=masc]] -> 'os' | ‘esses’

Em (63) listamos as frases utilizadas para a modelação das regras do nosso fragmento de gramática.

(63) O site publica mensalmente a relação
Esses serviçais tratavam invariavelmente a clientela.
O governo puniu cruelmente os rebeldes
Os pais facilmente abriram a porta
Os procuradores persistentemente transgrediram as regras
O discurso unicamente particulariza a informação
Os interessados automaticamente autorizam as investigações
O preso frequentemente vomita sangue
O Pedro visitou a Maria

Observe que estamos utilizando o símbolo *?a* como uma variável sobre os valores de concordância (AGR, do inglês *Agreement*), subcategorização (SUBCAT) e tempo (TENSE, do inglês *Tense*). No exemplo da segunda produção gramatical, D[BAR=1, AGR=?a] -> D[BAR=0, AGR=?a] N[BAR=1, AGR=?a], seja qual for o valor que D pegue para o traço AGR, N deve pegar o mesmo valor, por que há entre eles uma relação de concordância.

Desse modo, se D contém as mascas de NUM=sg e GND=fem, também N possuirá esses traços. Assim como utilizamos a notação AGR (abreviatura do inglês *Agreement*) para nos referirmos à concordância, também usamos as abreviações do inglês para os termos número, gênero e pessoa (NUM, GND e PER).

Assim concluímos o enriquecimento do nosso fragmento de gramática e partimos então para a análise automática das frases.

5.2 *Parsing* automático de um fragmento computacional

De acordo com a perspectiva metodológica do programa gerativista, é parte fundamental da pesquisa do lingüista de um programa como esse criar sistemas computacionais que sirvam de modelo para a descrição do conhecimento linguístico dos falantes/ouvintes. Portanto, nosso objetivo nessa seção é demonstrar computacionalmente a realização de análises automáticas feitas pelo nosso analisador gramatical, o qual representa

alem dos traços vistos, as categorias vazias resultantes da permanência do verbo *in situ* ou do movimento deste para I.

Já que as estruturas de traços, segundo Bird, Klein e Loper (2009), são uma estrutura de dados gerais para representar informações de qualquer tipo, nós iremos olhar para ela de um ponto de vista mais formal e ilustrar o suporte para estrutura de traços oferecido pelo NLTK na construção de fragmentos de CFG. Desse modo, podemos analisar sentenças de uma lista de frases a partir de *input* fornecido com base no léxico das regras da gramática.

Conforme Alencar (2009), "para analisar frases conforme a gramática fornecida por nós, utilizamos uma interface amigável para os *parsers* de CFG da biblioteca em *Python* do NLTK 2.0, conhecida como Donatus Versão 2009".

Para utilizarmos os módulos do NLTK, a partir do *Windows,* é necessário fazer download e instalar *Python* e NLTK no computador, a partir do sítio do NLTK[20] na internet.

Em um computador operado pelo *Windows,* utilizamos uma interface chamada IDLE, na qual se abre a interface Donatus2009 e executa-se o comando F5 para que ele importe do NLTK os módulos necessários de *parsers* da biblioteca do NLTK.

Dentro do IDLE e com o Donatus2009 executado, então, solicitamos do Donatus que este exerça determinadas funções através dos comandos que constam em sua lista. A partir de comandos dados pelo usuário, o Donatus é capaz de: i.iniciar a gramática do usuário que o utiliza; ii. analisar frases; iii.exibir o comprimento da lista de árvores, i.e. a quantidade de análises; iv. mostrar as análises em formato arbóreo numa outra janela; v. criar uma lista de frases; vi. analisar todas as construções da lista armazenada na variável frases; entre outros comandos.

A fim de ilustração do que dissemos acima, apresentamos abaixo os comandos dados pelo usuário durante a utilização do Donatus2009.

(64) inicia("nome da gramática do usuário")

(65) arvores=analisa_frase("frase da lista do usuário")

(66) len(arvores)

[20] http://www.nltk.org/

(67) mostra(arvores)

(68) frases=abre("nome do arquivo da lista de frases")

(69) analisa(frases)

A seguir mostramos, passo a passo, o modo como trabalhamos com o Donatus para utilizarmos *parsers* de CFG da biblioteca em *Python* do NLTK 2.0. O primeiro comando dado ao Donatus é inicia("gramaticafinal.fcfg"). Esse é o nome da nossa gramática. Assim ele constrói o parser '*nltk.parse.featurechart.FeatureChartParser*' para a nossa gramática, que está localizada em C:\Documents and Settings\ Administrador\ nltk_2011\ gramaticas\gramaticafinal.fcfg.

Após a construção do *parser*, damos ao Donatus o comando frases=abre("listadefrases.txt"). Com esse comando o programa irá analisar todas as frases que estão em no arquivo com o nome 'listadefrases'. Ele nos dirá através do comando analisa(frases) quantas análises cada frase possui.

Vejamos abaixo o recorte dos passos dados dentro do programa para obter as análises de cada frase da lista, até que o programa abra uma janela com as representações arbóreas de cada uma.

Python 2.6.2 (r262:71605, Apr 14 2009, 22:40:02) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.

****
Personal firewall software may warn about the connection IDLE
makes to its subprocess using this computer's internal loopback
interface. This connection is not visible on any external
interface and no data is sent to or received from the Internet.
****

IDLE 2.6.2
>>> ========== RESTART ==========
>>>
Olá! Sou o Donatus Versão 2009: uma interface amigável para os *parsers* de CFG da biblioteca em *Python* do NLTK 2.0, descritos no livro:

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with *Python*: analyzing text with the Natural Language Toolkit. Sebastopol, CA: Oâ€™Reilly, 2009. 502p. Disponível em: <http://www.nltk.org/book> Acesso em: 25 abr. 2009.

```
Posso analisar frases conforme gramáticas fornecidas pelo usuário. Para informações
sobre Python e o NLTK em português, consulte:

ALENCAR, Leonel Figueiredo de. Fundamentos de Python para a linguística
computacional. Manuscrito. Fortaleza: Universidade Federal do Ceará, 2007.

Para obter ajuda, execute o comando ajuda()

Diretório de trabalho default do Donatus:C:\Documents and Settings\ Administrador.
MICROSOF-960202\ nltk_2011\ gramaticas

Para maior comodidade, salve suas gramáticas nesse diretório.

Diretório atual da shell de Python:
C:\Documents and Settings\Administrador.MICROSOF-960202\nltk_2011\gramaticas
>>> inicia("gramaticafinal.fcfg")
Construção do parser <class 'nltk.parse.featurechart.FeatureChartParser'> para a
gramática em C:\Documents and Settings\Administrador.MICROSOF 960202\
nltk_2011\ gramaticas\ gramaticafinal.fcfg.
>>> frases=abre("listadefrases.txt")
>>> analisa(frases)
o site publica mensalmente a relação: 1 análise(s)
esses serviçais tratavam invariavelmente a clientela:1 análise(s)
o governo puniu cruelmente os rebeldes: 1 análise(s)
os pais facilmente abriram a porta: 1 análise(s)
os procuradores persistentemente transgrediram as regras:  1 análise(s)
o discurso unicamente particulariza a informação:   1 análise(s)
os interessados automaticamente autorizam as investigações: 1 análise(s)
o preso frequentemente vomita sangue: 1 análise(s)
o Pedro visitou a Maria: 2 análise(s)
>>> mostra(8)
Veja árvore em janela.
Feche janela para voltar.
```

Note que os comandos são dados ao programa sempre que este apresenta o símbolo '>>>' no canto esquerda da tela. De acordo com os dados acima, podemos ver que o programa mostra que última frase possue duas análises. Através do comando mostra(8) obtivemos do programa uma janela com as representações em árvores das duas análises.

Abaixo podemos ver as duas configurações arbóreas da frase referente ao que chamamos ao longo deste trabalho de SVO: '*o Pedro visitou a Maria*'.

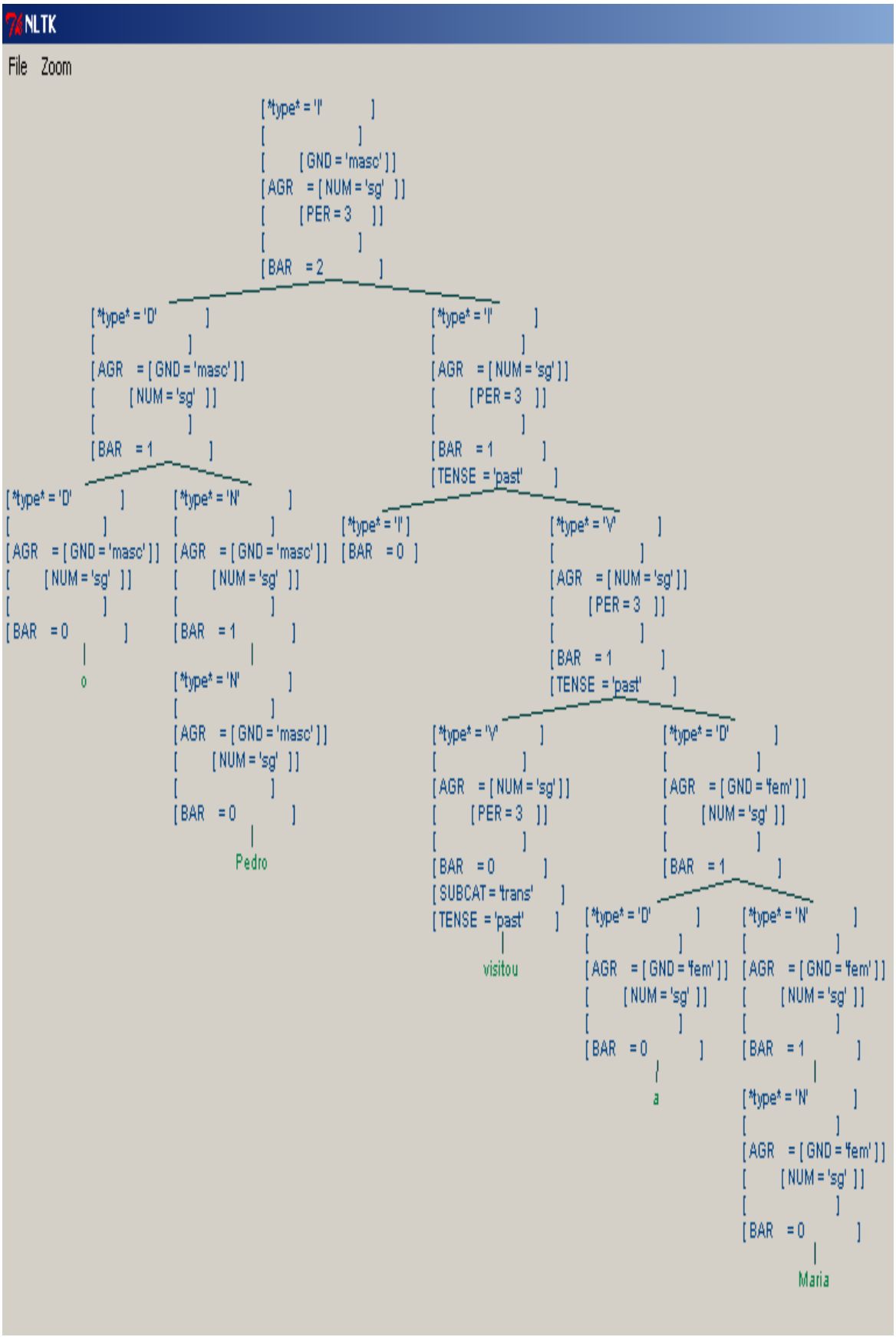


Figura 1 – Janela do NLTK com a primeira representação arbórea da frase ‘o Pedro visitou a Maria’, da nossa lista de frases.

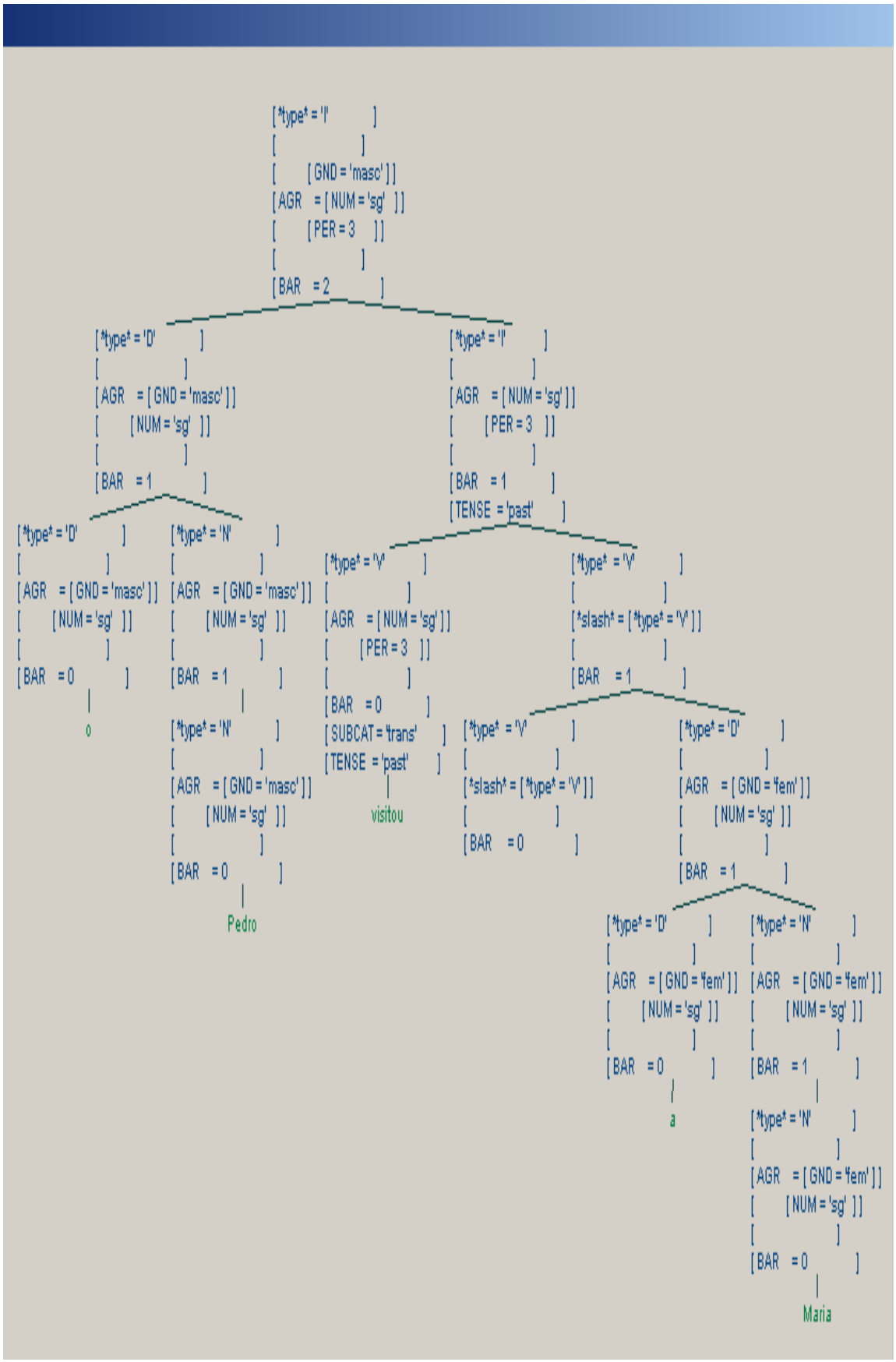


Figura 2 – Janela do NLTK com a segunda representação arbórea da frase ‘o Pedro visitou a Maria’, da nossa lista de frases.

Após a análise de uma sentença do tipo SVO, faremos abaixo uma demonstração de como o programa representa em árvore a análise a frase '*os procuradores automaticamente transgrediram as regras*', uma sentença do tipo SAdvVO.

Do mesmo modo como fizemos anteriormente, solicitamos do programa a configuração arbórea da frase acima, com o comando >>> mostra (6). O programa abre uma janela com a árvore da sétima frase da lista do arquivo analisado, conforme figura abaixo.

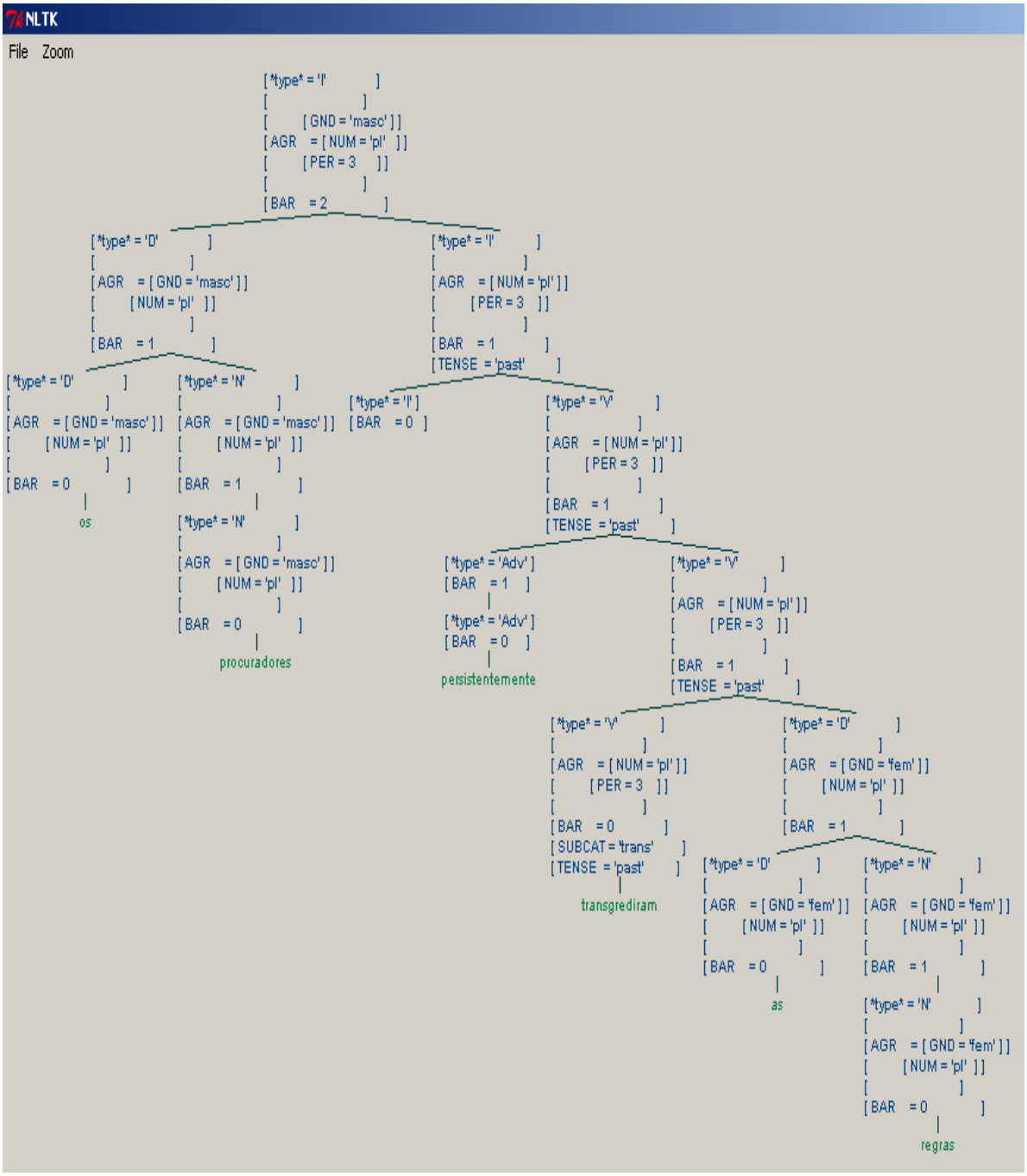


Figura 3 – Janela do NLTK com a representação arbórea da frase 'os procuradores automaticamente transgrediram as regras'

Agora analisaremos a sentença '*o site publica mensalmente a relação*'. Trata-se de uma sentença que chamamos de SVAdvO. Ao utilizarmos o programa, seguindo os comandos descritos mais acima, obtivemos do programa a seguinte análise.

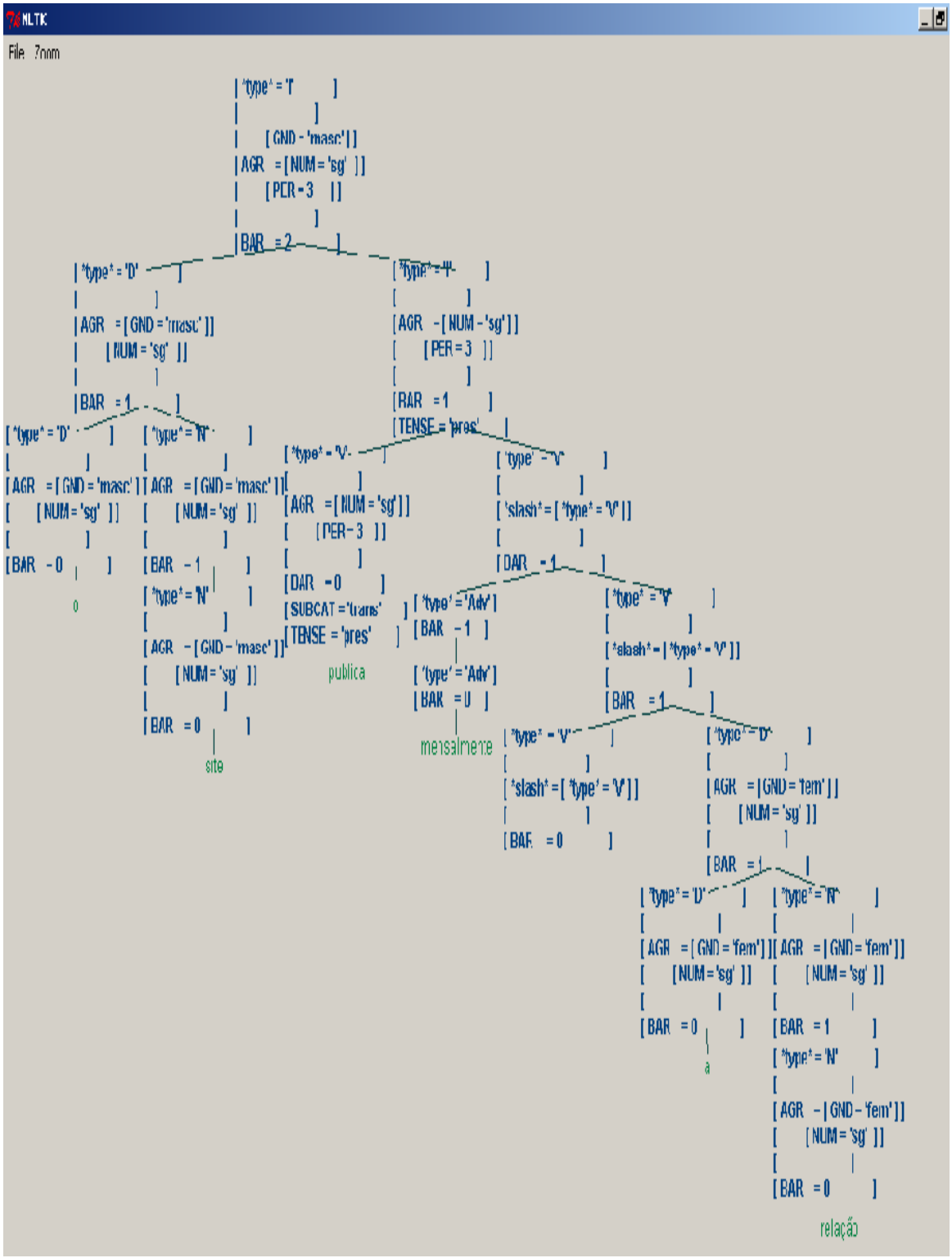


Figura 4 – Janela do NLTK com a representação arbórea da frase 'o site publica mensalmente a relação'

Chegamos ao fim do capítulo 5 desse trabalho. Nele nós vimos como elaborar regras de um fragmento de gramática da língua portuguesa capazes de capturar traços linguísticos envolvidos no processo derivacional da estrutura.

Como objetivo principal desse trabalho move-se em torno do processamento computacional de estruturas linguísticas, no âmbito do processamento sintático, utilizamos as janelas do NLTK como demonstração do produto de cada análise realizada pelo nosso analisador gramatical.

CONSIDERAÇÕES FINAIS

Chegamos ao fim desse trabalho. Ao longo de sua trajetória nos debruçamos sobre o modelo formal próprio da teoria gerativa para estabelecer o aparato teórico, no âmbito da sintaxe, que fundamentaria a nossa pesquisa. De forma breve, descrevemos as versões gerativas que se desenvolveram de *Syntactic Strucutres* até a Teoria de Princípios e Parâmetros, em uma das suas mais recentes abordagens, o Programa Minimalista.

De acordo com a Teoria X-barra de Fukui (1986), situamos o nosso trabalho dentro de uma descrição formal de aspectos linguísticos para demonstrar as relações estabelecidas na estrutura hierárquica dos constituintes. O esquema X-barra é o módulo da gramática gerativa onde estão previstos fenômenos como os de projeção sintagmática e de adjunção. Discutimos também o fenômeno de movimento de constituintes, de acordo com a hipótese de movimento dos verbos em sintaxe visível e sintaxe não visível e a questão dos níveis de representação das estruturas linguísticas em Forma Lógica e Forma Fonética, à luz da primeira versão do Programa Minimalista de Chomsky (1995), apresentada por Raposo (1999) e Modesto (2009).

Nosso trabalho partiu da hipótese de que o fenômeno de movimento do verbo em língua portuguesa pode ocorrer em sintaxe não visível. Dessa hipótese partimos para uma descrição de sentenças finitas com a finalidade de estabelecer uma relação entre o movimento do verbo no processo derivacional da estrutura e a posição dos constituintes da sentença. Com base nessas noções utilizamos como recorte para nossa pesquisa, sentenças finitas da língua portuguesa que apresentassem em sua estrutura advérbios, que nessa pesquisa foram classificados de acordo com os critérios de Perini(2005) e Ilari(1991).

Constituiu-se uma lista de frases que apresenta um número de expressões linguísticas, que embora pequeno, pode refutar a tese de Raposo (1995) de que em língua portuguesa o fenômeno de movimento do verbo ocorre de modo apenas visível. Dessa forma, esse trabalho suscita novas pesquisas que visem discutir e ampliar os dados apresentados nessa dissertação. Consideramos essa pesquisa importante pelas contribuições descritivistas que parecem, no entanto, contrariar Raposo (1999) e Ilari (1991) no que diz respeito à posição dos advérbios de modo.

A partir desse tipo de sentenças recorremos ao formalismo CFG Baseada em Traços para modelar um fragmento de gramática e explicar como as regras desse fragmento

são capazes de analisar apenas sentenças gramaticais da língua portuguesa. Para constituir a lista de frases a serem analisadas buscamos expressões linguísticas em *corpora* eletrônicos do português brasileiro e do português europeu, disponíveis na internet, em um sítio do Ministério da Tecnologia de Portugal, o Linguateca - centro de recursos para o processamento computacional da língua portuguesa.

Por se tratar de uma pesquisa de cunho computacional, essa pesquisa passou da descrição e análise linguística para o processamento computacional das especificações da nossa gramática, por meio de um analisador gramatical implementado em linguagem *Python*, através de programas da biblioteca do NLTK específicos para o processamento de línguas naturais. O analisador é capaz de processar sintaticamente *inputs* lexicais modelados em regras capazes de unificar os traços linguísticos de cada categoria gramatical, seja lexical ou sintagmática.

A análise sintática computacional apresentada nesse trabalho constitui o ponto alto da nossa pesquisa. Já que o gerativismo é uma teoria linguística de caráter mentalista, o processamento automático de expressões de línguas naturais representa uma possibilidade de testar e verificar hipóteses acerca de como os fenômenos linguísticos são processados pela mente humana.

Contudo não queremos deixar transparecer que a implementação computacional de analisadores sintáticos no formalismo CFG baseada em traços seja capaz de abranger o grande número de especificações que cada língua possui. Um bom exemplo do que não é possível fazer com o fragmento de gramática apresentado aqui pode ser exposto aqui em relação ao sintagma adverbial. Em nosso trabalho trabalhamos com essa categoria como um elemento da sequência que pode ser pronunciada em duas posições distintas da sentença. Mas se nós ao invés de fazermos uso de um advérbio como '*persistentemente',* quiséssemos utilizar *'com persistência'* para nos expressar? Daria conta a nossa gramática desse fenômeno sintático-semântico? Com estas questões queremos demonstrar nossa consciência sobre as coisas que não são contempladas em nosso pesquisa e que podem vir a ser temas de pesquisas futuras.

No entanto, mesmo com suas limitações, não podemos deixar de reconhecer que esse tipo de investigação linguística ligada à aplicação computacional representa uma

tendência das últimas décadas no modo de especular e tratar os fenômenos linguísticos, sobretudo em importantes centros de pesquisas linguísticas do Brasil e do mundo.

Portanto pensar em investigações linguísticas visando à implementação computacional é uma forma de estabelecer um elo entre ciências tecnológicas e ciências humanas, propício a elucidação de fatos das línguas naturais e do funcionamento da linguagem na mente humana, e principalmente propício ao desenvolvimento de tecnologias, especialmente tecnologias da informação. Chomsky (1957) já faz uma comparação entre certos sistemas formais por exemplo da matemática com os sistemas das línguas naturais em sua forma escrita ou falada.

No caso específico do nosso trabalho, o nosso intuito é demonstrar como certos aspectos específicos de cada língua, no tocante ao conhecimento sintático, podem ser desenvolvidos para facilitar a sua implementação em programas computacionais e também para facilitar que os programas tenham mais condições de analisar somente expressões gramaticais, sem que reconheçam frases agramaticais ou sem sentido. Porém, este é apenas um pequeno aspecto do que pode ser feito na lingüística computacional, pois como já sabemos o que temos em mente é um complexo e intrincado sistema lingüístico que permite uma atuação interacional entre os conhecimentos fonéticos, sintáticos, semânticos e pragmáticos.

Pesquisas desse tipo podem tentar reproduzir em computadores as hipóteses levantadas acerca, sobretudo dos meios trilhados pelo diversos tipos de conhecimento lingüístico durante o processo de produção e de compreensão de enunciados. Por isso, acreditamos que uma boa sugestão de pesquisa linguístico-computacional seria a observação da interface mantida na análise sintática de uma expressão linguística com o seu nível representacional. Assim, tendo como base Raposo (1999) e suas considerações acerca do sistema de performance conceitual-intencional, um objeto de pesquisa como o que sugerimos pode constituir uma forma intrigante de tentar compreender os processos de enunciação.

Por isso, as pesquisas lingüísticas desenvolvidas para contribuir com a computação no sentido de enriquecer as capacidades das máquinas de atuarem de modo semelhante ao ser humano ainda é um desafio não só para a Inteligência Artificial como também para a Lingüística, pois os estudos na área de Inteligência Artificial não podem deixar de incluir nas máquinas conhecimentos lingüísticos específicos dos humanos e por outro lado a Lingüística precisa passar a refletir no modo como especificar os vários conhecimentos

lingüísticos específicos da espécie humana de modo que se tornem computacionalmente tratáveis.

REFERÊNCIAS

ALENCAR, Leonel Figueiredo de (2002). **O *Constructor*.** Disponível em: http://www.geocities.com/briefaustausch/indexc.htm Acesso em: 18 dez.2008.

_____, Leonel Figueiredo de. Linguagem e Inteligência Artificial. In: MATTES, Marlene (Org.). **Linguagem. As expressões do múltiplo.** Fortaleza: Premius, 2006a.

_____, Leonel Figueiredo de. **Teoria da gramática: uma abordagem computacional.** Manuscrito não publicado. Fortaleza: Universidade Federal do Ceará, 2006b.

_____, Leonel Figueiredo de. Der *Constructor* – ein interaktives Werkzeug für Recherchen in portugiesischen Korpora auf dem WWW. In: PUSCH, Claus D.; RAIBLE, Wolfgang (Orgs.). **Romanistische Korpuslinguistik. Korpora und gesprochene Sprache.** Tübingen: Narr, 2002, p. 147-154.

_____, Leonel Figueiredo de. **Fundamentos de *Python* para a linguística computacional.** Manuscrito. Fortaleza: Universidade Federal do Ceará, 2007.

_____, Leonel Figueiredo de. Sintaxe formal de línguas não configuracionais num ambiente computacional: o caso do latim. **Revista Virtual de Estudos da Linguagem** – ReVEL. V. 6, n. 10, março de 2008. ISSN 1678-8931 [www.revel.inf.br]

ALEXIADOU, Artemis; Elena ANAGNASTOPOLOU. Parametrizing Agr: word order, V-movement and EPP-checking. **Natural Language and Linguistic Theory,** n.6, p.491-539, 1998.

BELLETTI, A. **Generalized verb movement.** Turin: Rosenberg e Sellier, 1990.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward (2009). **Natural Language Processing with *Python***. Disponível em: http://www.nltk.org/book Acesso em: 01.04.2009.

CHOMSKY, Noam. **Syntactic Structures. 2ª. ed. Berlin: Mouton de Gruyter, 2002.** Disponível em: http://d-nb.info/965414604 Acesso em: 29 dez.2009.

_____, Noam. Remarks on Nominalization. In: Jacobs, R. A e P. Rosenbaum (orgs.). **Readings in English Transformational Grammar.** Massachusetts: Ginn and Company, 1970.

Chomsky, Noam. **Knowledge of language: its nature, origins and use**. New York, Praeger, 1986b.

_____, Noam. **The Minimalist Program.** Cambridge, Massachusetts: MIT Press, 1995.

DAVID, K. A. **Sintaxe das expressões nominais no português do Brasil: uma abordagem computacional.** 2007. 117 f. Dissertação (Mestrado em Linguística) – Centro de Humanidades, Universidade Federal do Ceará, Fortaleza, 2007.

FUKUI, Naoki. **A theory of category prejection and its applications.** 1986. 281 f. Tese (Doutorado em Linguística) – Massachusetts Institute of Technology, Massachusetts, 1986.

ILARI, Rodolfo. Considerações sobre a Posição dos Advérbios. In: CASTILHO, Ataliba Teixeira de. **Gramática do português falado: a ordem.** vol. 1. 2ª ed. São Paulo: Editora da UNICAMP/FAPESP, 1991.

KLENK, Ursula. **Generative Syntaxe**. Tübingen: Gunter Narr Verlag, 2003.

LANGENDOEN, D. Terence. Linguistic theory. In: BECHTEL, William; GRAHAM, George (Orgs.). **A companion to cognitive science**. Oxford: Blackwell, 1999. p. 235-244.

MIOTO, Carlos; SILVA, Maria Cristina Figueiredo; LOPES, Ruth. **Novo Manual de Sintaxe**. Florianopolis: Insular, 2005.

KAPLAN, Ronald M. Syntax. In: MITKOV, Ruslan. **The Oxford Handbook of Computational Linguistics.** Oxford: OUP, 2004.

MODESTO, Marcello. O Programa Minimalista em sua Primeira Versão. In: ALENCAR, Leonel Figueiredo de; OTHERO, Gabriel de Ávila; PAGANI, Luiz Arthur (Orgs.). **Abordagens computacionais da teoria da gramática.** 2009. Em preparo.

MOURA, Maria Denilda. Variação em Sintaxe. In: Denilda Moura; Jair Farias. (Org.). **Reflexões sobre a sintaxe do português**. 1ª ed. Maceió: EDUFAL, 2005, v. 1, p. 47-71.

NETO, José Borges. O empreendimento gerativo. In: MUSSALIM, Fernanda; BENTES, Anna Cristina (orgs.). **Introdução à Linguística: fundamentos epistemológicos.** vol. 3. 2ª ed. São Paulo: Cortez, 2005.

OTHERO, Gabriel de Ávila. **Teoria X-barra: descrição do português e aplicação computacional**. São Paulo: Contexto, 2006.

PAGANI, Luiz Arthur. Analisador gramatical em Prolog para gramáticas de estrutura sintagmática. In: **Revista Virtual de Estudos em Linguagem – ReVEL**. Ano. 2, n. 3, [www.revelhp.cjb.net], 2004.

PERINI, Mário A. **Gramática descritiva do português**. São Paulo: Ática, 2005.

RAPOSO, Eduardo. **Teoria da gramática. A faculdade da linguagem**. Lisboa: Editorial Caminho, 1992.

_____, Da Teoria dos Princípios e Parâmetros ao Programa Minimalista: algumas ideias-chave. Apresentação da versão portuguesa de CHOMSKY 1995. Lisboa: Caminho, 1999. p. 15-37.

SAG, Ivan A.; WASOW, Thomas; BENDER, Emily. **Syntactic Theory: a formal introduction.** 2ª ed. Stanford: CSLI Publications, 2003.

SILVA, Claúdia Roberta Tavares. A natureza dos expletivos em construções inacusativas: uma análise não unificada em língua PROP-DROP e não PROP-DROP. **GELNE:** Revista do Grupo de Estudos Linguísticos do Nordeste, João Pessoa, v.7, n.1, p. 89-102, mês. 2005.

TRAVIS, L. **Parameters and Effects of Word-Order Variation.** Tese (Pós-Doutorado em Linguística) – Massachusetts Institute of Technology, Massachusetts, 1984.

VIEIRA, Renata; LIMA, Vera Lúcia Strube de. Linguística Computacional: princípios e aplicações. In: FERREIRA, Carlos Eduardo (Ed.). **As tecnologias da informação e a questão social.** v.2, 2001. p. 47-86.

ANEXO A – Amostras do Português Brasileiro, na sequência N Adv V N, obtidas dos NILC/São Carlos v. 8.0 e *corpora* ANCIB v. 3.4, respectivamente.

Ao preencher e assinar as fichas, **os interessados automaticamente autorizam as investigações.**

Apesar de seu poder de convencimento e de sua promessa de verdade, **o discurso unicamente particulariza a informação**.

ANEXO B – Amostras do Português Europeu, na sequência N Adv V N, obtidas de buscas no *corpus* DiaCLAV v. 3.0 e no *corpus* Natura/Minho v. 4.1, respectivamente.

Os pais facilmente abriram a porta, vendo que se tratava do filho, pois de acordo com fontes próximas do casal, estes nunca abriam a porta a desconhecidos.

Os procuradores persistentemente transgrediram, mudaram e modificaram as regras com o fim de perseguir um objectivo político, insistiu Lockhart.