

# **Analisador gramatical em Prolog para gramáticas de estrutura sintagmática**

Luiz Arthur Pagani (UFPR)

arthur@ufpr.br

## ***Análise, analisador e gramática***

Um analisador gramatical (que em inglês é chamado de *parser*) pode ser inicialmente caracterizado como um mecanismo que aplica um conjunto de especificações que determinam uma língua a uma expressão lingüística para deduzir alguma informação sobre esta última. Em outras palavras, as expressões lingüísticas são associadas a análises gramaticais (*parsing*, em inglês) estabelecidas pela gramática daquela língua, e o analisador gramatical é justamente o que permite essa associação entre a expressão e sua respectiva análise através de uma gramática.

Em termos de processamento, as expressões lingüísticas são os dados a serem processados, a análise gramatical é o produto desse processamento, a gramática é a especificação abstrata sobre o que deve ser processado e o analisador gramatical é o algoritmo que orienta a execução do processamento. Nesse sentido, é costume associar a gramática à parte declarativa do processamento, enquanto a parte procedimental é associada ao analisador gramatical.

Da forma como foi definida aqui, a análise gramatical pode ser comparada à geração apenas invertendo-se suas duas primeiras partes: para a geração, os dados a serem processados são as análises gramaticais, enquanto o produto resultante do processamento é a expressão lingüística. Dessa forma, a diferença entre análise e geração é concebida apenas como a mudança de sentido do processamento: a análise é o processamento da expressão lingüística para sua respectiva análise gramatical, e a geração é o processamento desta para aquela. No entanto, em alguns casos o processamento parece não permitir essa bidirecionalidade, o que caracterizaria a geração e a análise como dois processos completamente diferentes; nessa hipótese, a análise continua sendo concebida como o processo que leva de expressões lingüísticas a análises gramaticais, enquanto a geração

passa a ser concebida como o processo que passa de representações abstratas (algumas vezes associadas a entidades mentais) para expressões lingüísticas. Mesmo que seja uma questão importante, ela não será discutida aqui, porque nos limitaremos exclusivamente a problemas relativos à análise gramatical.

Concebido como processador, um analisador gramatical pode ser definido em relação ao tipo de produto a que se chega no final do processamento. Em sua versão mais simples, esse produto pode ser a mera identificação das expressões de uma determinada língua. Ou seja, dada uma expressão lingüística e a gramática dessa língua, o resultado do processamento nos informa se aquela expressão faz parte ou não daquela língua. Quando o processador apenas distingue as expressões que fazem parte de uma língua das que não fazem, ele recebe o nome de “reconhecedor”.

Além dos meros reconhecedores, os analisadores gramaticais propriamente ditos são aqueles que não só executam o reconhecimento, mas que ainda nos informam sobre outros aspectos da estrutura lingüística que, como não aparecem explicitamente no simples encadeamento seqüencial das unidades lingüísticas, precisam ser deduzidos. Assim, o produto dos analisadores vai depender do aspecto gramatical a ser tratado; em sua forma mais recorrente, os analisadores deduzem a organização hierárquica em constituintes das expressões lingüísticas, mas eles podem ser concebidos de diversas maneiras: como relacionadores entre tipos de sentenças (entre passivas e suas respectivas ativas, por exemplo), como elaboradores de representação semântica etc. Algumas vezes, os analisadores são chamados de “transdutores”, porque seu produto é a própria entrada transformada de alguma maneira.

Mas, como dissemos, as versões mais difundidas de analisadores gramaticais são aquelas em que se reconhece a organização hierárquica dos constituintes lingüísticos. Uma das questões menos controversas na recente tradição lingüística é o fato de que as expressões lingüísticas não são formadas apenas pela seqüenciação de unidades elementares, mas que estas unidades se organizam antes em blocos maiores, que se combinam entre si para ir formando unidades lingüísticas de ordem cada vez maiores. A essa organização, os lingüistas chamam de “estrutura de constituintes”.

## ***Estrutura de constituintes***

A determinação da estrutura de constituintes de uma certa expressão lingüística respeita alguns critérios empíricos. Para justificar a postulação de uma determinada estrutura subjacente, o lingüista deve identificar alguma operação que seja executada intuitivamente pelos falantes da língua estudada; essa operação deve incidir sistematicamente sobre alguma parte das expressões lingüísticas em que ela se manifesta, de forma que o lingüista pode inferir que os constituintes são as partes recorrentemente afetadas em todas as suas manifestações.

Um dos critérios mais empregados para a postulação de estruturas subjacentes é o da coesão dos constituintes em relação à sua mobilidade. Segundo esse critério, um falante do português poderia descrever uma certa situação através de qualquer uma das sentenças de (1) a (4) (os exemplos são adaptados de Graffi 1994: 77, bem como alguns dos critérios que serão apresentados aqui, que podem ser encontrados nas pp. 76-86).

- (1) João voltou para sua cidade natal no ano passado.
- (2) No ano passado, João voltou para sua cidade natal.
- (3) João, no ano passado, voltou para sua cidade natal.
- (4) João voltou no ano passado para sua cidade natal.

Independentemente do seu significado, qualquer falante parece relacionar intuitivamente essas quatro sentenças através de uma operação quase mecânica que desloca a expressão “no ano passado” dentro da sentença. Com isso, esse tipo de teste mostra que “no ano passado” é um constituinte, já que ele pode ser sistematicamente movido como um bloco coeso.

Além disso, se acrescentarmos a observação de que nenhuma das sentenças de (5) a (7) é bem formada (ou seja, que os falantes do português não as produzam em situações normais,<sup>1</sup> o que é representado pela anteposição do asterisco), podemos deduzir ainda que

---

<sup>1</sup> Pelo termo “situações normais”, estamos querendo nos remeter à distinção já clássica na Lingüística entre competência e desempenho, que pode ser encontrada em quase todos os manuais de introdução à Lingüística, principalmente os que têm origem na gramática gerativa, já que a distinção se deve

“para a sua cidade natal” é um constituinte, já que a expressão apresenta uma certa coesão, de forma a não permitir a inserção de outro constituinte dentro dela. Esse comportamento define outro critério para a determinação dos constituintes, conhecido como “teste da inseribilidade”.

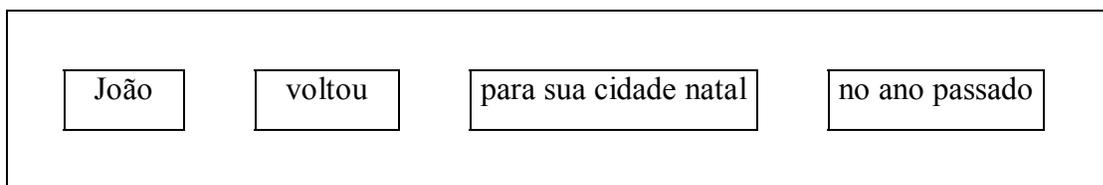
(5) \* João voltou para no ano passado sua cidade natal.

(6) \* João voltou para sua no ano passado cidade natal.

(7) \* João voltou para sua cidade no ano passado natal.

A partir desses testes, os lingüistas podem deduzir uma estruturação para além do simples encadeamento seqüencial, que em seus primeiros momentos foi representado por esquemas como o de (8), no qual a coesão dos constituintes é representada pelos retângulos em torno das expressões que os compõem.

(8)



Um outro teste, o da enunciabilidade em isolamento, nos ajudaria ainda a perceber que “voltou para sua cidade no ano passado” também é um constituinte, já que (10) é uma resposta plausível para (9).

(9) O que aconteceu com o João?

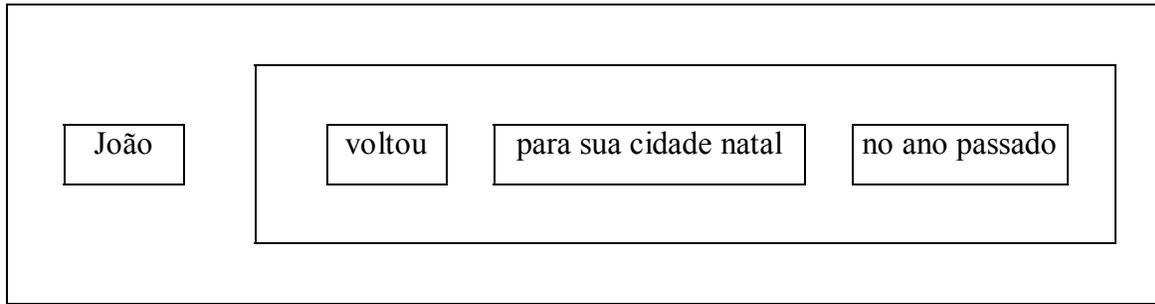
(10) Voltou para sua cidade natal no ano passado.

Dessa maneira, a estrutura de constituintes de “João voltou para sua cidade natal no ano passado” parece ser melhor representada por (11).

---

fundamentalmente a Chomsky (1965). Essa distinção tem por objetivo excluir da ontologia das teorias lingüísticas determinados acidentes de ordem psico-física, já que o objeto de estudo da Lingüística (a linguagem) é definido como uma capacidade (uma potência) e não uma mera manifestação física (um ato). Isso não exclui necessariamente a Lingüística da lista das ciências empíricas.

(11)



Uma das vantagens dessa estrutura de constituintes é que ela parece oferecer uma explicação mais sucinta para o fenômeno da passivização. Para uma sentença como “João ama Maria”, caso a passivização incida sobre toda a sentença, obtemos a passiva “Maria é amada por João”; mas a passivização também pode ser concebida como uma operação que incide apenas sobre o constituinte “ama Maria”, resultando na passiva “Maria é amada”. Por outro lado, devido a esse comportamento, a passivização também pode ser encarada como um critério para a determinação da estrutura de constituintes (perdendo, portanto, o seu caráter explicativo).

Um outro fenômeno menos mencionado na tradição gramatical, mas que também pode ser explicado pela estrutura de constituintes é a clivagem. Uma sentença clivada tem o formato “foi... que...” (ou “é... que...”), no qual a primeira lacuna é ocupada por um pedaço de uma sentença e a segunda pelo que sobrou da sentença sem aquele pedaço. Para uma sentença como (1), poderíamos construir as seguintes sentenças clivadas.

(12) Foi João quem voltou para sua cidade natal no ano passado.

(13) Foi para sua cidade natal que João voltou no ano passado.

(14) Foi no ano passado que João voltou para sua cidade natal.

Em termos de estrutura de constituintes, podemos dizer mais precisamente que a clivagem é uma operação que incide sobre uma sentença, separando um de seus constituintes para ocupar a posição entre “foi” e “que”, ficando o resto da sentença (sem esse constituinte) depois de “que”. Da mesma forma, a clivagem pode deixar de ser concebida como explicação e passar a ser também considerada um critério para a constituência.

Finalmente, ainda em relação ao poder explicativo da estrutura de constituintes, também é possível apresentar uma concepção mais clara da chamada ambigüidade estrutural, que é a ambigüidade proveniente de combinações diferentes dos constituintes envolvidos na expressão. Dessa forma, uma famosa sentença ambígua como “o homem viu a mulher com o telescópio”, que pode ser parafraseada ou como ‘o homem usou o telescópio para ver a mulher’, ou como ‘o homem viu a mulher que estava portando o telescópio’, pode ser entendida da seguinte maneira: a segunda interpretação só é possível caso analisemos “a mulher” e “com o telescópio” compondo um único constituinte, já na primeira interpretação é preciso que “a mulher” e “com o telescópio” façam parte de constituintes diferentes. Se pensarmos na passivização e da clivagem como testes, chegamos sistematicamente a dois resultados: 1) “a mulher com o telescópio foi vista pelo homem” e “foi a mulher com o telescópio que o homem viu” (para o caso de “a mulher com o telescópio” ser um único constituinte, o que vai corresponder à segunda interpretação), e 2) “a mulher foi vista com o telescópio pelo homem” e “foi a mulher que o homem viu com o telescópio” (na quais “a mulher” e “com o telescópio” pertencem a constituintes diferentes, o que vai caracterizar a primeira interpretação).

Porém, se os critérios listados até agora nos permitem decidir que partes de uma expressão mais complexa formam constituintes, e quais não, eles não nos dizem muito sobre as categorias a que os constituintes pertencem.

### ***Categoria dos constituintes***

Apresentaremos aqui apenas dois desses critérios que, além da estrutura de constituintes, também nos indicam os seus tipos:<sup>2</sup> o da coordenabilidade e o da intersubstituibilidade.

---

<sup>2</sup> É interessante observar que os testes servem para distinguir o conjunto de expressões que pertencem a um determinado tipo, ou seja, o conjunto das expressões que constituem as categorias gramaticais; os rótulos usados para designar essas categorias (abreviados como N, V, A, P) são distintos delas, apesar de serem frequentemente equiparados no discurso de alguns linguistas: os rótulos são apenas símbolos escolhidos arbitrariamente para nos referirmos às categorias.

As conjunções coordenadas são definidas na gramática tradicional como partículas cuja função é a de agrupar duas sentenças em uma sentença mais complexa. Assim, usando a conjunção “e” como exemplo, e sendo que “João fuma” e “Maria voltou para sua cidade natal no ano passado” são sentenças, elas podem ser coordenadas através daquela conjunção, formando a sentença composta “João fuma e Maria voltou para sua cidade natal no ano passado”.

No entanto, não é preciso muito esforço para perceber que a conjunção “e” não reúne apenas sentenças, mas pode combinar outros tipos de constituintes, formando um constituinte mais complexo daquele mesmo tipo. Por exemplo, a sentença “João fuma e voltou para sua cidade natal no ano passado” combina dois sintagmas verbais (SV)<sup>3</sup> – “fuma” e “voltou para sua cidade natal no ano passado” – resultando num sintagma verbal mais complexo “fuma e voltou para sua cidade natal no ano passado”. Algumas vezes, essa coordenação é explicada como se fosse formada de duas sentenças, “João fuma” e “João voltou para sua cidade natal no ano passado”, nas quais a identidade do sujeito permite a sua omissão na segunda sentença; assim, “João fuma e voltou para sua cidade natal no ano passado” pressupõe um estado anterior, no qual a sentença seria “João fuma e João voltou para sua cidade natal no ano passado”.<sup>4</sup>

Esse tipo de explicação ainda poderia ser adaptado para dar conta de exemplos de coordenação de sujeito (“João e Maria fumam”, que seria a combinação de “João fuma” e “Maria fuma”), de objeto direto (“João fuma cigarro e charuto”, construída a partir de “João fuma cigarro” e “João fuma charuto”), e de verbos (como em “João preparou e comeu o bolo”, que conteria “João preparou o bolo” e “João comeu o bolo”), ou mesmo de exemplos mais complicados com coordenação de mais de um constituinte (como em “João e Maria prepararam e comeram o bolo e a torta”).

---

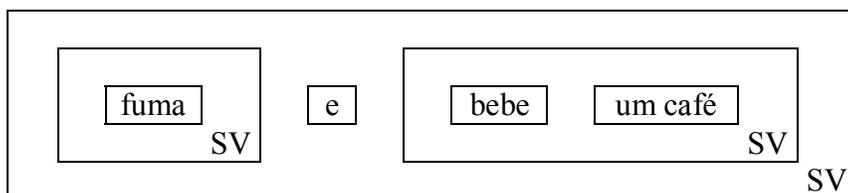
<sup>3</sup> O sintagma verbal corresponde mais ou menos ao que a gramática tradicional chamava de predicado.

<sup>4</sup> Esse estado anterior não é pressuposto ontologicamente (ou seja, não é preciso que as sentenças “João fuma” e “João voltou para sua cidade natal no ano passado” tenham sido anteriormente atestadas na língua); elas são pressupostas apenas logicamente: basta que a competência dos falantes da língua permita o seu reconhecimento.

Como a coordenação pode ainda ser aplicada a quaisquer outras categorias, como por exemplo, as preposições (“João viajou de e para sua cidade”) e os advérbios (“João fuma intensa e apaixonadamente”),<sup>5</sup> parece ser mais simples concebê-la como uma operação que reúne duas expressões da uma única categoria, formando outra expressão daquela mesma categoria.

Esse tipo de comportamento permite não apenas a dedução da estrutura de constituintes, mas também a de que constituintes são ou não do mesmo tipo. É dessa forma que se infere que o verbo “fuma”, usado intransitivamente, precisa ser do mesmo tipo que “bebe um café”, porque essas coisas podem ser reunidas na sentença “João fuma e bebe um café” (esse exemplo é adaptado de Chierchia 1997: 99). Como não gostaríamos de afirmar que “bebe um café” é um verbo intransitivo (como “fuma”), uma alternativa é a de dizer que “bebe um café” pertence a uma categoria complexa a que “fuma” também pertence, mesmo contendo apenas uma palavra; essa categoria tem sido chamada de sintagma verbal. Assim, a estrutura de constituintes mais adequada para “fuma e bebe um café” parece ser aquela representada em (15).

(15)



Essa mesma comparação entre “fuma” (no seu uso intransitivo) e “bebe um café” pode ser feita através da intersubstituibilidade. Analisando as sentenças “João fuma” e “João bebe um café”, podemos observar que, de uma para outra, há apenas a troca entre

---

<sup>5</sup> A coordenação parece reunir também, em alguns casos, coisas que não se parecem muito com constituintes, como “João foi para e voltou de sua cidade natal”, na qual “foi para” e “voltou de” não são constituintes, da forma como eles estão sendo definidos aqui. Uma alternativa interessante para manter a idéia de que a coordenação reúne dois constituintes de mesmo tipo, resultando num outro constituinte mais complexo desse mesmo tipo, foi largamente explorada pela gramática categorial, como podemos ver em Wood 1993: 114-121.

“fuma” e “bebe um café”, o que nos permite deduzir que essas duas coisas que podem ser substituídas uma pela outra devem pertencer ao mesmo tipo. O raciocínio é o mesmo: como não queremos dizer que esse tipo seja o dos verbos intransitivos, já que “bebe um café” não se parece com um verbo intransitivo, postulamos que “fuma”, além de pertencer aos verbos intransitivos, também pertence, por si só, à categoria dos sintagmas verbais, como “bebe um café”.

Porém, talvez por sua pouca praticidade, os diagramas em retângulos não foram muito usados pelos lingüistas. Duas outras representações mais difundidas são os diagramas em árvore e os colchetes rotulados.

### ***Colchetes rotulados e diagramas em árvore***

A notação de colchetes rotulados consiste em representar os constituintes dentro de um par de colchetes aos quais se subscreve o rótulo da respectiva categoria. Assim, uma representação para a estrutura de constituintes de “João voltou para sua cidade natal no ano passado” seria  $[_S[_{SN}João_{SN}] [_{SV}voltou [_{SP}para sua cidade natal_{SP}] [_{SP}no ano passado_{SP}]_{SV}]_S]$ ;<sup>6</sup> nesta representação, podemos constatar que a sentença é composta de um sintagma nominal e de um sintagma verbal ( $[_S[_{SN}\dots_{SN}] [_{SV}\dots_{SV}]_S]$ ) e que o sintagma verbal, por sua vez, é construído através de “voltou” anteposto a dois sintagmas preposicionados ( $[_{SV}voltou [_{SP}\dots_{SP}] [_{SP}\dots_{SP}]_{SV}]$ ).

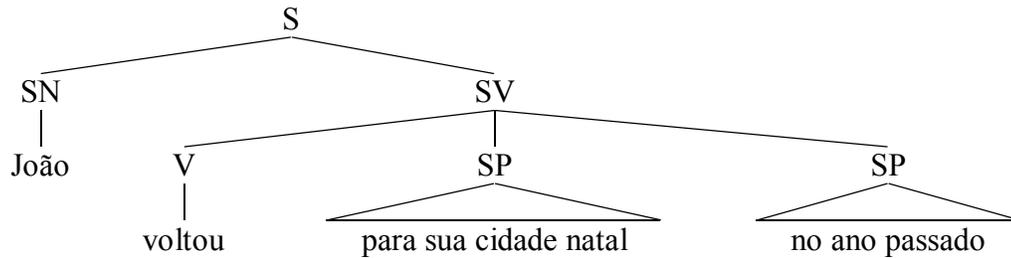
A mesma informação ainda pode ser apresentada num diagrama em árvore, no qual a estrutura de constituintes é representada através de ramificação. O mesmo exemplo corresponderia à árvore em (16).<sup>7</sup>

---

<sup>6</sup> Algumas vezes, quando a expressão não é muito complexa e sua estrutura pode ser facilmente apreendida, costuma-se omitir um dos rótulos categoriais subscritos; assim, a mesma representação poderia ser apresentada como  $[[João_{SN}] [voltou_v [para sua cidade natal_{SP}] [no ano passado_{SP}]_{SV}]_S]$  ou como  $[_S[_{SN}João] [_{SV}voltou [_{SP}para sua cidade natal] [_{SP}no ano passado]]]$ .

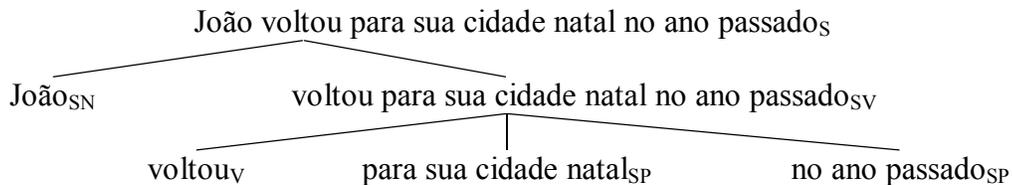
<sup>7</sup> Na árvore em (16), os triângulos são usados para abreviar o reconhecimento da existência de estruturas lingüísticas que, por algum motivo, não precisam ser exaustivamente representadas. No nosso caso, como a estrutura interna dos sintagmas preposicionados não são indispensáveis para a apresentação do problema discutido no texto, ela é omitida.

(16)



Nesse diagrama em árvore, a estrutura de constituintes é representada da seguinte maneira: “voltou” é um verbo, “para sua cidade natal” e “no ano passado” são sintagmas preposicionados, e juntos esses três constituintes formam o sintagma verbal “voltou para sua cidade natal no ano passado”; como “João” é um sintagma nominal, antepondo-se este ao sintagma verbal anterior, obtemos a sentença “João voltou para sua cidade natal no ano passado”. Nesse sentido, o diagrama em árvore, da forma como ele foi apresentado em (16), é uma abreviação que seria mais exhaustivamente apresentada como (17).

(17)



Mas como é possível perceber, essa representação exhaustiva nos obriga a repetir algumas expressões até três vezes. Dessa forma, se nos lembrarmos que, na árvore em (16), os nós nos quais aparecem apenas rótulos categoriais são, na verdade, abreviações para a concatenação das expressões que são dominadas por esses nós, então essa notação é mais prática do que a de (17).

O que foi dito até aqui se relaciona mais com o resultado da análise gramatical (o seu produto), do que com o que o licencia. Para representar o tipo de coisa que licencia o produto da análise gramatical, os linguistas costumam empregar a gramática de estrutura sintagmática (em inglês, *phrase structure grammar*, normalmente abreviado como PSG).

## **Gramática de estrutura sintagmática**

Uma PSG é uma forma de especificar quais são as expressões bem formadas de uma língua, especificando também a que estruturas ela pode estar relacionada, sem estabelecer como as expressões são construídas ou como suas estruturas são encontradas. Esse tipo de gramática é constituído por dois tipos de regras: as regras sintagmáticas e as regras de inserção lexical.

As regras sintagmáticas descrevem como as categorias sintagmáticas, ou seja, as categorias compostas, são formadas através da combinação de outras categorias sintagmáticas ou lexicais. As regras lexicais são responsáveis por introduzir as categorias lexicais, que são as categorias correspondentes às unidades básicas.

Na gramática em (18), que permite analisar a sentença “João ama Maria”, a diferença entre regras sintagmáticas e lexicais pode ser claramente identificada pelo fato de que as últimas definem conjuntos, que são os conjuntos dos itens lexicais relativos a cada uma das categorias lexicais (regras (18)(c) e (18)(d)); as regras sintagmáticas ((18)(a) e (18)(b)) introduzem outras categorias sintagmáticas.

(18)

(a)  $S \rightarrow N SV$

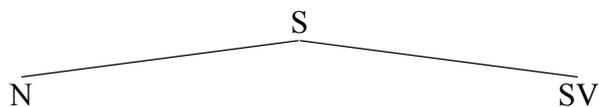
(b)  $SV \rightarrow V N$

(c)  $N \rightarrow \{\text{João, Maria}\}$

(d)  $V \rightarrow \{\text{ama}\}$

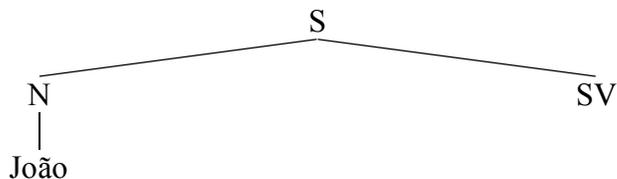
Por mais simples que a gramática em (18) possa parecer, ela já é suficiente para começarmos a apresentar a distinção entre a gramática e determinadas características que o analisador gramatical irá apresentar. Da forma como a gramática está concebida, por exemplo, poderíamos pensar em duas ordens possíveis para a construção da estrutura de “João ama Maria”. O primeiro passo em ambas as possibilidades seria o mesmo: aplicando a regra (18)(a), construímos a parte da árvore em (19).

(19)



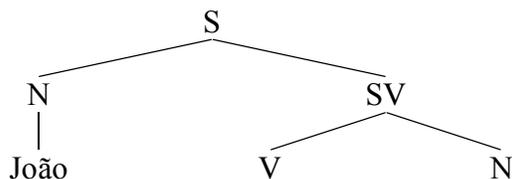
Na primeira possibilidade, porém, poderíamos decidir aplicar, na seqüência, a regra (18)(c), inserindo na árvore o nome “João”, como está em (20).

(20)



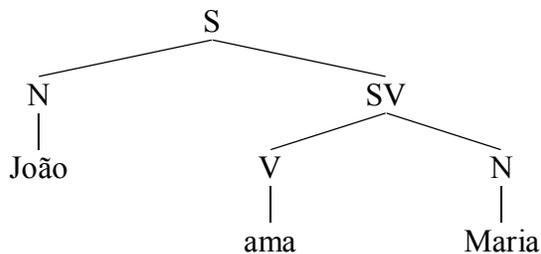
A seguir, aplicaríamos a regra (18)(b), o que resultaria na árvore em (21).

(21)



Finalmente, depois de aplicarmos as duas regras lexicais em (18)(d) e (18)(c), chegaríamos à árvore em (22).

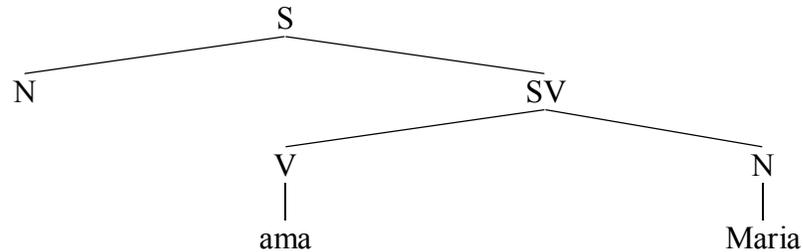
(22)



Numa segunda possibilidade, a partir da mesma árvore em (19), poderíamos decidir construir o SV antes de inserirmos o nome. Assim, aplicando as mesmas regras na mesma

seqüência em que elas foram aplicadas para construir as árvores em (21) e (22), chegaríamos à árvore em (23).

(23)

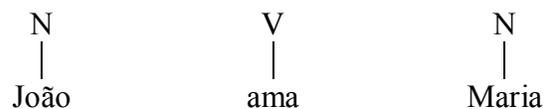


Agora sim, depois de aplicarmos a regra de inserção nominal de (18)(c), chegamos à mesma árvore de (22).

Dessa forma, como é possível perceber, podemos chegar de duas maneiras diferentes à mesma análise estrutura para a sentença “João ama Maria”, dependendo da escolha da ordem de aplicação das regras. No entanto, essa ordem de aplicação não é de forma alguma estabelecida pela gramática; o ordenamento da aplicação das regras deve ser explicitamente determinado pelo analisador gramatical.

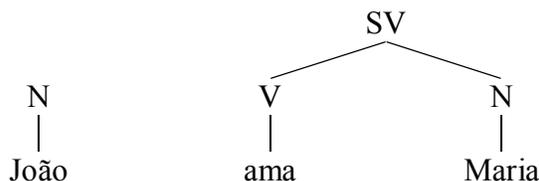
Além disso, as duas alternativas anteriores foram construídas descendentemente, ou seja, iniciando pelo topo e seguindo para baixo na árvore (por isso, esse estilo de analisador é chamado em inglês de *top-down*). Uma outra alternativa seria a de iniciar a análise de baixo, procedendo ascendentemente na árvore (o que em inglês é chamado de *bottom-up*). Dessa maneira, poderíamos chegar à mesma representação de (22), começando pelas inserções lexicais, obtendo então o diagrama em (24) (ainda não é possível dizer que este seja um diagrama em árvore porque não há nenhuma ramificação que o caracterize; na construção descendente já tínhamos uma árvore desde a primeira operação).

(24)



Como ainda não é possível aplicar a regra (18)(a), porque ainda não dispomos de um sintagma verbal, podemos aplicar a regra (18)(b) ao verbo e ao segundo nome, obtendo a árvore em (25).

(25)



Agora, finalmente, a regra (18)(a) pode ser aplicada, o que nos levaria a uma árvore exatamente como a de (22).<sup>8</sup>

### **Gramática de cláusula definida**

Tudo o que foi discutido acima sobre a estrutura de constituintes pode ser muito facilmente replicado em Prolog através de um recurso conhecido como gramática de cláusula definida (em inglês, *definite clause grammar*, sempre abreviado como DCG). Esse recurso foi apresentado pela primeira vez por Pereira & Warren 1980 e foi amplamente difundido por Pereira & Shieber 1987.

Através desse recurso, a PSG de (18) pode ser imediatamente transformada em um analisador gramatical através do programa em (26).

(26)

a.  $s \rightarrow n, sv.$

---

<sup>8</sup> Tudo o que foi dito sobre as árvores pode ser dito exatamente da mesma maneira para os colchetes rotulados. Assim, a árvore em (19) corresponderia a  $[[\dots_N] [\dots_{SV}]_S]$ . Na primeira possibilidade descendente, a árvore de (20) corresponderia a  $[[João_N][\dots_{SV}]_S]$  e a de (21) a  $[[João_N] [[\dots_V][\dots_N]_{SV}]_S]$ . Na segunda possibilidade descendente, a árvore de (23) corresponderia a  $[[\dots_N] [[ama_V] [Maria_N]_{SV}]_S]$ . E a árvore em (22), comum a ambas as possibilidades, corresponderia a  $[[João_N] [[ama_V] [Maria_N]]_S]$ . Na alternativa ascendente, a árvore de (24) corresponderia a  $[[João_N] [ama_V] [Maria_N]]$ , enquanto que a de (25) seria  $[[João_N] [[ama_V] [Maria_N]_{SV}]]$  (observe que a única diferença aqui é a ausência da informação de que esse constituinte é uma sentença).

- b.  $sv \rightarrow v, n.$
- c.  $n \rightarrow [jo\tilde{a}o]; [maria].$
- d.  $v \rightarrow [ama].$

Quando esse programa é carregado num interpretador de Prolog,<sup>9</sup> passamos a dispor de um mecanismo que não apenas gera a sentença “João ama Maria” (ou, mais precisamente, a sua representação em Prolog, que é “[joão, ama, maria]” – ou seja, uma lista de átomos que representam as palavras), mas que ainda funciona como um reconhecedor dessa mesma sentença.

Para operar o gerador, basta dirigir ao interpretador carregado com o programa em (26) a diretiva “s(X, []).”, que serão seqüencialmente apresentadas as seguintes respostas: “[joão, ama, joão]”, “[joão, ama, maria]”, “[maria, ama, joão]” e “[maria, ama, maria]”, que são as quatro sentenças (representadas em Prolog) que o programa consegue gerar. A solicitação do reconhecimento da sentença “João ama Maria” é feita através da diretiva “s([joão, ama, maria], []).”, para a qual obtém-se uma resposta afirmativa, o que indica que o interpretador reconheceu “João ama Maria” como sendo uma sentença aceita por aquela gramática.

Antes de começar a comentar o funcionamento de uma DCG como analisador gramatical, é preciso observar que o interpretador Prolog não opera diretamente as regras listadas em (26); quando um programa escrito na notação da DCG é carregado, suas regras são automaticamente traduzidas para um formato conhecido como diferença de listas. Assim, quando um interpretador de Prolog lê a gramática de (26), ele a transforma em (27).<sup>10</sup>

---

<sup>9</sup> Nem todo interpretador de Prolog dispõe de uma DCG; nesse caso, pode-se empregar diretamente a diferença de listas através das quais a DCG é processada em Prolog. No entanto, a maioria dos interpretadores já dispõe desse recurso.

<sup>10</sup> O que ocorre com o conjunto de itens lexicais de um mesmo tipo, quando definidos numa única cláusula disjuntiva, depende da solução escolhida na implementação de cada interpretador. De qualquer forma, como não se costuma usar essa definição disjuntiva, já estamos começando a empregar a definição mais corrente, em que cada item lexical é determinado independentemente um do outro.

(27)

- a.  $s(L0, L) :- n(L0, L1), sv(L1, L).$
- b.  $sv(L0, L) :- v(L0, L1), n(L1, L).$
- c.  $n([jo\tilde{a}o|L], L).$
- d.  $n([maria|L], L).$
- e.  $v([ama|L], L).$

É isso que exige que as diretivas mencionadas para a geração e para o reconhecimento tenham dois argumentos, ainda que, na gramática de (26), os predicados empregados não apresentem explicitamente nenhum argumento.

Dessa forma, o funcionamento do reconhecedor pode ser explicado da seguinte maneira. Ao se questionar o interpretador com “s([joão, ama, maria], []).”, essa cláusula unifica com a parte esquerda da regra em (27)a, de forma que a sua satisfação passa a depender da satisfação de duas outras cláusulas: “n([joão, ama, maria], L1)” e “sv(L1, []).”. Quando a primeira dessas novas cláusulas é unificada com a regra (27)c, o resultado é que a variável ‘L1’ é instanciada com o que resta da lista “[joão, ama, maria]”, excluído o seu primeiro elemento; ou seja, ‘L1’ passa a valer “[ama, maria]” (daí essa técnica se chamar diferença de listas). Com isso, a segunda nova cláusula passa a ser “sv([ama, Maria], []).”; unificada com a parte esquerda da regra (27)b, o interpretador passa a depender da satisfação de duas outras novas cláusulas: “v([ama, maria], L1)” e “n(L1, []).”. A primeira delas unifica com a regra (27)e, resultando na instanciação da variável ‘L1’ com a lista “[maria]”; assim, a segunda cláusula passa a ser “n([maria], []).”, que unifica com a regra (27)d. Com essa última unificação, todas as cláusulas são satisfeitas, de forma que o interpretador Prolog responde afirmativamente à solicitação de “s([joão, ama, maria], []).”.

### ***Analizador gramatical em DCG***

Para transformar o reconhecedor em (26) num analisador gramatical, basta acrescentar um argumento em cada um dos predicados usados, de forma a construir nesse argumento uma representação da estrutura gramatical. Normalmente, as árvores costumam ser representadas em Prolog através de predicados e argumentos; nessa representação, uma

sentença formada por um sintagma nominal e um sintagma verbal apareceria como “s(n(..), sv(...))” (na qual as reticências indicam locais que ainda precisam ser preenchidos). Assim, a estrutura de constituintes de uma sentença como “João ama Maria” acabará sendo representada por “s(n(João), sv(v(ama), n(Maria)))”. O programa listado em (28) constrói esse tipo de representação.<sup>11</sup>

(28)

- a. s(s(N,SV)) --> n(N), sv(SV).
- b. sv(sv(V,N)) --> v(V), n(N).
- c. n(n(joão)) --> [joão].
- d. n(n(maria)) --> [maria].
- e. v(v(ama)) --> [ama].

O efeito de se incluir argumentos nos predicados de uma gramática em DCG é que, quando ela é lida pelo interpretador Prolog, esses argumentos são incluídos antes dos dois argumentos em que a diferença de lista é processada.

Assim, para solicitarmos a análise gramatical de “João ama Maria”, usa-se a diretiva “s(EC, [joão, ama, maria], []).”. Essa cláusula unifica com o lado esquerdo da regra (28)a, instanciando a variável ‘EC’ (abreviatura para estrutura de constituintes) com “s(N, SV)” e as duas novas cláusulas a serem satisfeitas são “n(N, [joão, ama, maria], L1)” e “sv(SV, L1, []).”. A primeira delas unifica com o lado esquerdo da regra (28)c, instanciando a variável ‘N’ com “n(joão)” e a variável ‘L1’ com “[ama, maria]”; devido novamente à unificação, essa instanciação faz com que a variável ‘EC’ passe a ser instanciada com “s(n(joão), SV)”. Com essas novas unificações, a segunda cláusula passa a ser “sv(SV, [ama, maria], []).”,

---

<sup>11</sup> Uma alternativa mais simples, e que se pareceria mais com os colchetes, ainda que sem os rólulos, seria a de montar para uma sentença como “João ama Maria” uma lista do seguinte tipo: “[joão], [[ama], [maria]]”. Para isso, seria preciso alterar as regras (28)a e (28)b para, respectivamente “s([N,SV]) --> n(N), sv(SV).” e “sv([V,N]); as regras lexicais também precisariam ser adaptadas: “n([joão]) --> [joão].”, “n([maria]) --> [maria].” e v([ama]) --> [ama].”. Mais para a frente, adotaremos essa representação mais simples.

que unifica com o lado esquerdo da regra (28)b, de forma que a variável ‘SV’ seja instanciada com “sv(V, N)”, e duas outras novas cláusulas precisam ser satisfeitas: “v(V, [ama, maria], L1)” e “n(N, L1, [])”. Unificando com a regra (28)e, a variável ‘V’ é instanciada com “v(ama)” e a variável ‘L1’ com “[maria]”, o que exige a atualização da variável ‘SV’ como “sv(v(ama), N)” e a segunda cláusula como “n(N, [maria], [])”. Esta segunda cláusula unifica com a regra (28)d, causando a instanciação da variável ‘N’ com “n(maria)”, o que por sua vez reinstancia a variável ‘SV’ como “sv(v(ama), n(maria))”. Com todas as cláusulas finalmente satisfeitas, a variável ‘EC’ é definitivamente instanciada como “s(n(joão), sv(v(ama), n(maria)))”; assim, antes de responder afirmativamente a diretiva, o interpretador ainda apresenta a informação de que isso acontece com quando “EC = s(n(joão), sv(v(ama), n(maria)))”.

### ***Uma análise gramatical mais consistente***

As gramáticas apresentadas até agora tinham um interesse mais didático do que propriamente empírico: elas serviam mais para mostrar como os recursos funcionavam, e tinham pouco interesse descritivo ou explicativo em relação à língua. A sua excessiva simplicidade era proposital. Começemos a ver uma gramática empiricamente mais interessante, e através da qual podemos descrever e explicar alguns fenômenos lingüísticos relacionados à estrutura de constituintes do português brasileiro; outra aplicação interessante da gramática que vamos apresentar é que ela também serve para deixar claro exatamente o que a gente não consegue nem descrever nem explicar – o que já é um passo na direção de se encontrar uma solução melhor.

Uma gramática mais plausível reconheceria, por exemplo, que as expressões que se combinam com o verbo também podem ser compostas por um constituinte complexo, e não apenas por um único nome; além disso, parece que a coesão com que esses complementos se relacionam com o verbo também não é a mesma (o que é normalmente chamado de assimetria entre sujeito e objeto, que reconhece no objeto um complemento mais coeso ao verbo do que o sujeito). O próprio constituinte verbal, por sua vez, pode ser composto não apenas por um objeto direto: um segundo tipo de verbo não exige complemento nenhum, um terceiro tipo exige apenas um complemento preposicionado e quarto tipo exige ambos os complementos. Esse tipo de constatação resulta numa gramática como a de (29).

(29)

- (a)  $S \rightarrow SN SV$
- (b)  $SN \rightarrow N$
- (c)  $SN \rightarrow Det N$
- (d)  $SV \rightarrow V$
- (e)  $SV \rightarrow V SN$
- (f)  $SV \rightarrow V SP$
- (g)  $SV \rightarrow V SN SP$
- (h)  $SP \rightarrow P SN$
- (i)  $Det \rightarrow \{o, a, os, as, \dots\}$
- (j)  $N \rightarrow \{Jo\~{a}o, Maria, homem, menino, \dots\}$
- (k)  $V \rightarrow \{corre, ama, gosta, coloca, \dots\}$
- (l)  $P \rightarrow \{de, em, para, \dots\}$

Essa gramática de (29) é facilmente transformada em um analisador através de uma DCG, como pode ser constatado em (30).<sup>12</sup>

(30)

- (a)  $s([SN, SV]) \rightarrow sn(SN), sv(SV).$
- (b)  $sn(N) \rightarrow n(N).$
- (c)  $sn([Det, N]) \rightarrow det(Det), n(N).$
- (d)  $sv(V) \rightarrow v(V).$
- (e)  $sv([V, SN]) \rightarrow v(V), sn(SN).$
- (f)  $sv([V, SP]) \rightarrow v(V), sp(SP).$
- (g)  $sv([V, SN, SP]) \rightarrow v(V), sn(SN), sp(SP).$

---

<sup>12</sup> O resultado da análise de constituintes será apresentado em colchetes não rotulados.

- (h)  $sp([P, SN]) \rightarrow p(P), sn(SN)$ .
- (i)  $det(Det) \rightarrow [Det], \{det(Det)\}$ .
- (j)  $det(o)$ .
- (k)  $det(a)$ .
- (l)  $det(os)$ .
- (m)  $det(as)$ .
- (n)  $n(N) \rightarrow [N], \{n(N)\}$ .
- (o)  $n(jo\tilde{a}o)$ .
- (p)  $n(maria)$ .
- (q)  $n(homem)$ .
- (r)  $n(meninas)$ .
- (s)  $v(V) \rightarrow [V], \{v(V)\}$ .
- (t)  $v(corre)$ .
- (u)  $v(ama)$ .
- (v)  $v(gosta)$ .
- (w)  $v(coloca)$ .
- (x)  $p(P) \rightarrow [P], \{p(P)\}$ .
- (y)  $p(de)$ .
- (z)  $p(em)$ .
- (aa)  $p(para)$ .

Nessa nova vers\~ao, al\~em de incluir as observa\~oes sobre a estrutura de constituintes, foram inseridos tamb\~em alguns outros recursos da notac\~ao da DCG. As regras sintagm\~aticas n\~ao foram alteradas, mas as regras de inserc\~ao lexical deixaram de mencionar

explicitamente o item lexical e passam a ter a seguinte forma “ $x(X) \rightarrow [X], \{x(X)\}.$ ”;<sup>13</sup> nesse formato, a função das chaves é a de interromper a tradução em diferença de listas. Assim, uma cláusula como esta seria traduzida para o interpretador como “ $x(X, [X|P], P) :- x([X|P], P), x(X).$ ”; dessa maneira, os itens lexicais podem ser definidos apenas como “ $x(X)$ ”, simplificando um pouco as entradas lexicais.

Com essa DCG em (30), é possível analisar uma sentença como “o homem gosta de Maria”, através da diretiva “ $s(EC, [o, homem, gosta, de, maria], [])$ ”, e seu resultado será “ $EC = [[o, homem], [gosta, [de, maria]]]$ ”. No entanto, apesar das análises adequadas, essa DCG chega a resultados inadequados, permitindo análise para expressões nas quais a concordância (tanto nominal quanto verbal) e a transitividade dos verbos é desconsiderada; por exemplo, para uma expressão como “a meninas gosta João”, o resultado da análise seria “[ $[a, meninas], [gosta, João]$ ”, apesar de que nem o determinante nem o verbo estejam concordando com “meninas”, e de que a regência de “gostar” não esteja sendo respeitada. O problema decorre do fato de que, da forma como está, a gramática não tem controle nem sobre as características de concordância nominal e verbal, nem sobre a regência verbal (na verdade, essas características nem chegam a ser mencionadas na DCG de (30)).

O tipo de solução permitida pela DCG a distingue radicalmente das PSG.<sup>14</sup> Uma maneira clássica de forçar a concordância nominal consiste na especificação de um

---

<sup>13</sup> Isso não é uma cláusula de Prolog; estamos apenas tomando a liberdade de usar uma espécie de metavariable, para indicar a estrutura das cláusulas de inserção lexical. O que se está querendo representar é que a metavariable com letra minúscula deve ser preenchida com uma categoria lexical e a metavariable com letra maiúscula representa uma variável do Prolog (ligada por três ocorrências).

<sup>14</sup> Ao contrário das PSGs, as DCGs podem ser usadas para descrever línguas do tipo ‘ $a^n b^n c^n$ ’, o que demonstra a diferença no seu poder gerativo. Uma gramática para gerar expressões ‘ $a^n b^n c^n$ ’ seria:

$anbnncn(N) \rightarrow a(N), b(N), c(N).$

$a(0) \rightarrow [].$

$a(N) \rightarrow [a], \{M \text{ is } N - 1\}, a(M).$

$b(0) \rightarrow [].$

$b(N) \rightarrow [b], \{M \text{ is } N - 1\}, b(M).$

argumento no qual a concordância seja definida. Assim, as regras para o sintagma nominal e suas respectivas entradas lexicais podem ser redefinidas como em (31).<sup>15</sup> (A concordância verbal será apresentada logo a seguir, junto com as modificações para lidar com a transitividade dos verbos, já que ambas são operações que afetam o sintagma verbal.)

(31)

- (a)  $sn(\text{Conc}, N) \rightarrow n(\text{Conc}, N)$ .
- (b)  $sn(\text{Conc}, [\text{Det}, N]) \rightarrow \text{det}(\text{Conc}, \text{Det}), n(\text{Conc}, N)$ .
- (c)  $\text{det}(\text{Conc}, \text{Det}) \rightarrow [\text{Det}], \{\text{det}(\text{Conc}, \text{Det})\}$ .
- (d)  $\text{det}([\text{masc}, \text{sing}], o)$ .
- (e)  $\text{det}([\text{fem}, \text{sing}], a)$ .
- (f)  $\text{det}([\text{masc}, \text{plur}], os)$ .
- (g)  $\text{det}([\text{fem}, \text{plur}], as)$ .
- (h)  $n(\text{Conc}, N) \rightarrow [N], \{n(\text{Conc}, N)\}$ .
- (i)  $n([\text{masc}, \text{sing}], joao)$ .
- (j)  $n([\text{fem}, \text{sing}], maria)$ .
- (k)  $n([\text{masc}, \text{sing}], homem)$ .
- (l)  $n([\text{fem}, \text{plur}], meninas)$ .

---

$c(0) \rightarrow []$ .

$c(N) \rightarrow [c], \{M \text{ is } N - 1\}, c(M)$ .

No entanto, essa DCG não é capaz de analisar as expressões dessa mesma língua; mas essa questão não será explorada aqui.

<sup>15</sup> O grau não faz parte destas características porque ele não sofre concordância. Isso inclusive tem sido usado como argumento para não considerar o grau sequer como flexão; nessa perspectiva, a flexão seria a mudança morfológica devida à concordância.

Na DCG em (31),<sup>16</sup> a concordância nominal é estabelecida pela ligação da variável ‘Conc’ na regra (31)(b): tanto o determinante quanto o nome precisam apresentar as mesmas informações de pessoa e número (codificadas numa lista de dois elementos, como se fosse um par ordenado) para serem integrados num sintagma nominal. (Em relação apenas à concordância nominal, a ocorrência da variável ‘Conc’ como argumento de ‘sn’ seria desnecessária; no entanto, como ela terá um papel na concordância verbal, ela já aparece aqui.)

Para lidar com a transitividade dos verbos, é costume empregar um recurso um pouco diferente, incluindo nos símbolos relativos aos verbos índices correspondentes às categorias pertinentes (Pereira & Shieber 1987: 73, por exemplo). Assim, a regra do sintagma verbal para os verbos intransitivos é escrita como “SV  $\rightarrow$  V<sub>i</sub>”; as outras três regras ficam respectivamente como “SV  $\rightarrow$  V<sub>td</sub> SN” (para os transitivos diretos), “SV  $\rightarrow$  V<sub>ti</sub> SP” (para os transitivos indiretos) e “SV  $\rightarrow$  V<sub>tdi</sub> SN SP” (para os transitivos diretos e indiretos; ou bitransitivos). Mas a transitividade também pode ser especificada como uma característica sintática determinada como argumento dos predicados sintagmáticos e lexicais que definem os verbos (como em Pereira & Shieber 1987: 86).

(32)

- (a)  $s([SN, SV]) \rightarrow sn([\_ , Num], SN), sv(Num, SV)$ .
- (b)  $sv(Num, V) \rightarrow v(Num, i, V)$ .
- (c)  $sv(Num, [V, SN]) \rightarrow v(Num, td, V), sn(\_ , SN)$ .
- (d)  $sv(Num, [V, SP]) \rightarrow v(Num, ti, V), sp(SP)$ .
- (e)  $sv(Num, [V, SN, SP]) \rightarrow v(Num, tdi, V), sn(\_ , SN), sp(SP)$ .

---

<sup>16</sup> É importante observar que os próximos programas não constituem gramáticas por si sós. Eles são emendas que devem substituir as partes apropriadas da gramática completa em (30). Uma maneira de se fazer isso é apagando e inserindo em (30) os respectivos predicados; outra maneira é consultar (através do predicado pré-definido ‘consult/1’) o programa em (30) e depois ir reconsultando os novos programas (através do predicado pré-definido ‘reconsult/1’; cujo efeito é obtido no SWI-Prolog colocando-se o nome do arquivo a ser reconsultado entre colchetes).

- (f)  $v(\text{Num}, \text{Val}, \text{V}) \rightarrow [\text{V}], \{v(\text{Num}, \text{Val}, \text{V})\}$ .
- (g)  $v(\text{sing}, i, \text{corre})$ .
- (h)  $v(\text{plur}, i, \text{correm})$ .
- (i)  $v(\text{sing}, \text{td}, \text{ama})$ .
- (j)  $v(\text{plur}, \text{td}, \text{amam})$ .
- (k)  $v(\text{sing}, \text{ti}, \text{gosta})$ .
- (l)  $v(\text{plur}, \text{ti}, \text{gostam})$ .
- (m)  $v(\text{sing}, \text{tdi}, \text{coloca})$ .
- (n)  $v(\text{plur}, \text{tdi}, \text{colocam})$ .

Na DCG em (32), acima, define-se para cada entrada lexical relativa aos verbos seu número e sua transitividade. A regra de inserção lexical transfere essas informações sobre a valência e a concordância para as regras sintagmáticas, que controlam cada uma dessas características. A concordância verbal é estabelecida na regra (32)(a), através da ligação da variável ‘Num’, que exige que o sintagma nominal tenha o mesmo número que o sintagma verbal.<sup>17</sup> Já a transitividade é controlada da seguinte maneira: dependendo do valor que a variável ‘Val’ (valência) assume na regra de inserção lexical (32)(f), a regra sintagmática apropriada é selecionada; caso o valor de ‘Val’ seja “i” (intransitivo), a única unificação possível será com a regra (32)(b), que garante um sintagma verbal constituído apenas por um verbo; caso o valor de ‘Val’ seja “tdi” (transitivo direto e indireto), a unificação só poderá acontecer com a regra (32)(e), que exige que o sintagma verbal também contenha um sintagma nominal e um sintagma preposicionado, além do verbo.

---

<sup>17</sup> Como essa gramática ainda não dispõe de pronomes, a concordância verbal ainda não precisa lidar com a concordância de pessoa; de qualquer forma, essa concordância seria facilmente tratada, exigindo-se que o domínio de ‘Num’ fosse uma lista com as informações de concordância de forma que as informações de número e de pessoa pudessem ser controladas (para melhorar seu aspecto mnemônico, a variável poderia passar a se chamar ‘ConcVerb’, por exemplo).

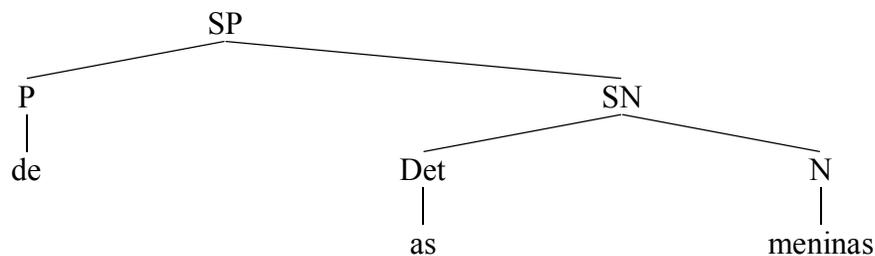
### **Alternativas incomuns**

Mesmo com todas as alterações propostas, a gramática apresentada aqui ainda está longe de ser empiricamente completa, ou mesmo minimamente aceitável. Além da falta de várias categorias lexicais (como a dos pronomes, já mencionada antes; e de adjetivos, advérbios e conjunções, que ainda nem tinham sido mencionados), muitas questões sintáticas também ficaram de fora (como a questão da estruturação sistemática de todas as categorias a partir de um núcleo, que originou a chamada teoria X-barras na gramática gerativa, conforme Miotto, Silva & Lopes 1999: 45-81).

### **Preposições flexionadas e regência**

Em todas as gramáticas apresentadas até agora, só apareceram preposições não flexionadas; no entanto, além da preposição “de”, ainda dispomos de outras quatro versões dessa mesma preposição: “do”, “da”, “dos” e “das”. Normalmente, se costuma explicar essas últimas quatro formas da preposição como se elas fosse o resultado da amalgamação da preposição “de” com uma das quatro versões do artigo definido; assim, adotando a hipótese de que o determinante esteja incluído no sintagma nominal, uma preposição “das” resultaria de alguma operação que afeta a estrutura em (33).

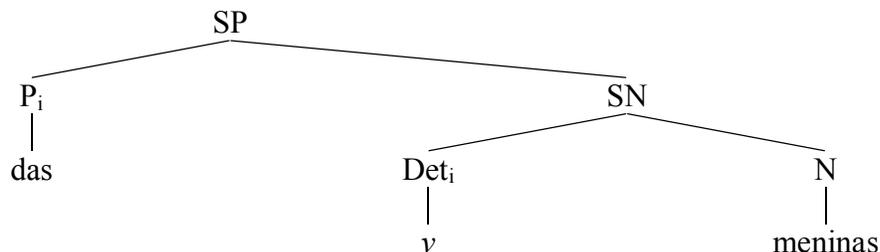
(33)



Alguma coisa (como o alçamento do determinante “as” para o núcleo do sintagma preposicionado, se juntando com a preposição “de”) precisa acontecer para que a preposição e o determinante fiquem próximos o suficiente para que possa acontecer uma fusão entre eles. Num certo sentido, esta é a antiga idéia de que “de+as” resulta em “das”, que já aparecia na gramática tradicional. Depois dessas operações de deslocamento e de amalgamação, a estrutura resultante seria (34), na qual “v” representa a posição esvaziada

pelo deslocamento do determinante e o índice “i” subscrito a P e a Det representam a ligação entre as posições de origem e destino do movimento.<sup>18</sup>

(34)



Uma alternativa para esse tipo de solução é a de supor que ao invés de uma única preposição, o léxico já disponha das cinco versões dela. Essas preposições se dividiriam em dois grupos: o das preposições flexionáveis e o das preposições “inflexionáveis”. Aquelas subcategorizariam<sup>19</sup> não sintagmas nominais inteiros, mas apenas uma parte desses sintamas; estas subcategorizariam sintagmas nominais integrais. Além de economizar recursos mais abstratos, como o deslocamento e a categoria vazia, essa explicação ainda tem a vantagem de justificar melhor a concepção de que a estrutura dos constituintes é estratificada; nessa concepção, generalizada pela chamada teoria X-barrá, um sintagma nominal como “as meninas” teria a seguinte estrutura: “[<sub>SN</sub>[<sub>Det</sub>as]][<sub>N</sub>:[<sub>N</sub>meninas]]”.

---

<sup>18</sup> Normalmente, os índices subscritos são colocados juntos às palavras; mas como não são apenas as palavras que podem se deslocar, e sim qualquer tipo de constituinte (e mesmo quando não há deslocamento, a ligação pode acontecer anaforicamente também entre constituintes), prefiro subscrever os índices às categorias pertinentes. Numa sentença como “de quem o João gosta?”, por exemplo, o deslocamento afeta um sintagma preposicionado inteiro, de forma que o índice não pode indicar apenas o pronome “quem”, pelo menos não enquanto operação sintática.

<sup>19</sup> A noção de subcategorização se parece de uma certa forma com o que a gramática tradicional chama de “regência”; assim, uma categoria subcategoriza outra quando a primeira exige um subtipo da segunda. Nesse sentido, os verbos transitivos diretos subcategorizam o sintagma nominal como tipo de complemento do verbo.

Essas observações podem ser incluídas na elaboração de uma DCG como (35).<sup>20</sup>

(35)

- (a)  $\text{sn}(\text{Conc}, \text{NBarra}) \text{-->} \text{n\_barra}(\text{Conc}, \text{NBarra})$ .
- (b)  $\text{sn}(\text{Conc}, [\text{Det}, \text{NBarra}]) \text{-->} \text{det}(\text{Conc}, \text{Det}), \text{n\_barra}(\text{Conc}, \text{NBarra})$ .
- (c)  $\text{n\_barra}(\text{Conc}, \text{N}) \text{-->} \text{n}(\text{Conc}, \text{N})$ .
- (d)  $\text{sp}([\text{P}, \text{SN}]) \text{-->} \text{p}(\text{inf}, \text{P}), \text{sn}(\_, \text{SN})$ .
- (e)  $\text{sp}([\text{P}, \text{NBarra}]) \text{-->} \text{p}(\text{Conc}, \text{P}), \text{n\_barra}(\text{Conc}, \text{NBarra})$ .
- (f)  $\text{p}(\text{Conc}, \text{P}) \text{-->} [\text{P}], \{\text{p}(\text{Conc}, \text{P})\}$ .
- (g)  $\text{p}(\text{inf}, \text{com})$ .
- (h)  $\text{p}([\_, \_], \text{de})$ .
- (i)  $\text{p}([\text{masc}, \text{sing}], \text{do})$ .
- (j)  $\text{p}([\text{fem}, \text{sing}], \text{da})$ .
- (k)  $\text{p}([\text{masc}, \text{plur}], \text{dos})$ .
- (l)  $\text{p}([\text{fem}, \text{plur}], \text{das})$ .
- (m)  $\text{p}([\_, \_], \text{em})$ .
- (n)  $\text{p}([\text{masc}, \text{sing}], \text{no})$ .
- (o)  $\text{p}([\text{fem}, \text{sing}], \text{na})$ .
- (p)  $\text{p}([\text{masc}, \text{plur}], \text{nos})$ .
- (q)  $\text{p}([\text{fem}, \text{plur}], \text{nas})$ .
- (r)  $\text{p}(\text{inf}, \text{para})$ .

---

<sup>20</sup> Nessa versão, escolhemos não representar a estrutura dos constituintes simples; ou seja, apesar da estratificação, a estrutura de “as meninas” continuará sendo “[as, meninas]”. Isso fica ainda mais evidente nos sintagmas preposicionados: “das meninas” é analisado como “[das, meninas]”, enquanto “com as meninas” vai corresponder à estrutura “[com, [as, meninas]]”. No entanto, caso fosse necessário, não seria difícil fazer as modificações para se chegar a estruturas como “[das, [meninas]]” ou “[com, [as, [meninas]]]”.

- (s) p([\_,\_], por).
- (t) p([masc,sing], pelo).
- (u) p([fem, sing], pela).
- (v) p([masc, plur], pelos).
- (w)p([fem, plur], pelas).

Nessa DCG, os sintagmas nominais são construídos em dois passos: um nome gera um sintagma de tipo N' ("n\_barra"), quando combinado com seu complemento (no nosso caso, essa possibilidade não foi expressa; normalmente, os complementos em sintagmas nominais são os adjetivos e os sintagmas preposicionados), e esse sintagma N' constitui o sintagma nominal, ou por si só ou quando vem junto com um determinante. No entanto, o mais interessante é mesmo o tratamento do sintagma preposicionado: com apenas duas regras e três tipo de preposições, obtemos apenas as expressões bem formadas.

Preposições como “com” e “para”, marcadas com “inf” no primeiro argumento dos predicados que as definem, são aceitas em expressões tanto com os determinantes (em “com as meninas”, por exemplo), como sem eles (em “como meninas”). Já preposições como “de” e “em”, cuja subcategorização é indicada por “[\_,\_]”, que não se flexionam como as duas primeiras, ao contrário daquelas, não aceitam os determinantes (\* “de as meninas”), mas aparecem em expressões sem os determinantes (“de meninas”). Finalmente, as preposições nas quais as marcas da flexão nominal aparecem explicitamente têm o mesmo comportamento das preposições como “de” e “em” (“das meninas”, em oposição a \* “das as meninas”), e são incompatíveis com os contextos sintáticos das preposições como “com” e “para”.

Esse tipo de solução para a distribuição das preposições sugere que existem dois grandes subconjuntos de preposições: as que são completamente inflexionáveis e as que aceitam flexão. Esse segundo subconjunto, por sua vez, se subdivide em outros dois

subconjuntos: aquelas marcadas explicitamente com a flexão nominal e aquelas que não são marcadas, mas que, por isso mesmo, são compatíveis com qualquer concordância.<sup>21</sup>

Um outro fenômeno que afeta as preposições e que as gramáticas apresentadas até aqui não dão conta é a chamada regência verbal: alguns verbos exigem determinadas preposições e não outras. O exemplo mais evidente é o do verbo “gostar” que só aceita a preposição “de”; além desse, um outro verbo que já apareceu nas nossas gramáticas também impõe uma regência específica: o verbo “colocar” exige que o seu sintagma preposicionado contenha a preposição “em” ou alguma de suas flexões.

A adaptação que precisa ser feita para podermos representar esse fenômeno numa DCG pode ser vista em (36).

(36)

- (a)  $sv(Num, [V, SP]) \rightarrow v(Num, ti(P), V), sp(P, SP)$ .
- (b)  $sv(Num, [V, SN, SP]) \rightarrow v(Num, tdi(P), V), sn(\_, SN), sp(P, SP)$ .
- (c)  $sp(Lex, [P, SN]) \rightarrow p(Lex, inf, P), sn(\_, SN)$ .
- (d)  $sp(Lex, [P, NBarra]) \rightarrow p(Lex, Conc, P), n\_barra(Conc, NBarra)$ .
- (e)  $p(Lex, Conc, P) \rightarrow [P], \{p(Lex, Conc, P)\}$ .
- (f)  $p(com, inf, com)$ .
- (g)  $p(de, [\_, \_], de)$ .
- (h)  $p(de, [masc, sing], do)$ .
- (i)  $p(de, [fem, sing], da)$ .
- (j)  $p(de, [masc, plur], dos)$ .

---

<sup>21</sup> A análise dessa questão sobre a concordância das preposições é bem mais complicada do que está sendo sugerido aqui. Como na gramática gerativa a concordância é uma característica sintática que parece se estabelecer entre o núcleo e o especificador, seria preciso inferir disso que o sintagma preposicionado fosse uma outra categoria que, além do determinante, pode constituir o especificador do sintagma nominal. No entanto, qualquer decisão sobre essa questão excede os objetivos propostos no presente texto.

- (k) p(de, [fem, plur], das).
- (l) p(em, [\_,\_], em).
- (m)p(em, [masc, sing], no).
- (n) p(em, [fem, sing], na).
- (o) p(em, [masc, plur], nos).
- (p) p(em, [fem, plur], nas).
- (q) p(para, inf, para).
- (r) p(por, [\_,\_], por).
- (s) p(por, [masc,sing], pelo).
- (t) p(por, [fem, sing], pela).
- (u) p(por, [masc, plur], pelos).
- (v) p(por, [fem, plur], pelas).
- (w)v(sing, ti(de), gosta).
- (x) v(plur, ti(de), gostam).
- (y) v(sing, tdi(em), coloca).
- (z) v(plur, tdi(em), colocam).

A regência é controlada transformando a informação sobre a transitividade dos verbos (que aparece nas regras (36)(a) e (36)(b), e nas informações lexicais em (36)(w)-(z)) num predicado de um argumento, o que permite que se exprima nesse argumento a preposição exigida. Essa exigência é transferida para o sintagma preposicionado através do primeiro argumento dos predicados ‘sp/2’ (regras (36)(c) e (36)(d)), ‘p/3’ (regras (36)(f)-(v)) e ‘p/5’ (regra (36)(e)), criado nessa versão especialmente para isso. Nas regras em que as informações lexicais das preposições são definidas, foi preciso acrescentar uma informação sobre o tipo da preposição, já que algumas são formas flexionadas (instâncias)

de uma determinada preposição (tipo); assim, “nas” é uma forma flexionada de “em” e deve ser reconhecida como preposição válida para a regência do verbo “colocar”.<sup>22</sup>

Assim, com a DCG em (36), chegamos a análises para as sentenças como “João gosta da Maria” e “as meninas colocam os livros em estantes” (respectivamente “[joão, [gosta, [da, maria]]]” e “[[as, meninas], [colocam, [os, livros], [em, estantes]]]”), mas não para sentenças como “João gosta para Maria” ou “as meninas colocam os livros com as estantes”.

### **Teoria X-barra**

De volta à observação sobre a propriedade endocêntrica dos constituintes (ou seja, “as propriedades do núcleo são preservadas em cada projeção”, segundo Miotto, Silva & Lopes 1999: 53), de acordo com a teoria X-barra, todos os constituintes são projetados a partir de um núcleo e dispõem de dois níveis de estratificação: sua projeção intermediária e a projeção máxima; “na projeção intermediária, o núcleo pode estar relacionado com complementos (Compl) e, na projeção máxima, pode estar relacionado com, no máximo, um especificador (Spec)” (Miotto, Silva & Lopes 1999: 50).<sup>23</sup>

Para núcleos com um único complemento,<sup>24</sup> poderíamos especificar estas disposições sobre a teoria X-barra através da DCG em (37).

---

<sup>22</sup> Essa informação foi inserida de uma forma redundante, porque não se está dando aqui atenção ao léxico.

<sup>23</sup> Na teoria X-barra, há ainda a adjunção, que é definida da seguinte maneira “um adjunto, por sua vez, é um constituinte que está apenas contido na projeção máxima de um núcleo” (Miotto, Silva & Lopes 1999: 69), mas não incluído; essa explicação pressupõe as definições de inclusão e continência (Miotto, Silva & Lopes 1999: 68). Como existem outras definições para a adjunção, e para não abordarmos essa controvérsia, a adjunção não será tratada aqui.

<sup>24</sup> Para núcleos com mais de um complemento, seria preciso decidir se os complementos se combinam todos de uma única vez com o núcleo, ou se eles se combinam um a um; nessa segunda alternativa, as definições dos níveis precisariam inclusive ser revista. Apesar de adotarem a segunda alternativa (Miotto, Silva & Lopes 1999: 76, por exemplo), os autores não redefinem o complemento, de forma que se cria uma assimetria entre o primeiro complemento (o que se combina diretamente com o núcleo) e os outros: o primeiro é contido e incluído na projeção intermediária, enquanto que os outros (de forma similar aos

(37)

- (a) `proj_max(Cat) --> proj_int(Cat, []), !.`
- (b) `proj_max(Cat) --> proj_max(Arg), proj_int(Cat,[Arg]).`
- (c) `proj_int(Cat, []) --> nucleo(Cat,[]), !.`
- (d) `proj_int(Cat, [Arg]) --> nucleo(Cat, [Arg]), !.`
- (e) `proj_int(Cat,Resto) --> nucleo(Cat, [Prim|Resto]), proj_max(Prim).`
- (f) `nucleo(Cat, Args) --> [Pal], {palavra (Pal,Cat, Args)}.`
- (g) `palavra (joao, n, []).`
- (h) `palavra (maria, n, []).`
- (i) `palavra(corre, v, [n]).`
- (j) `palavra(ama, v, [n, n]).`

Nas estradas lexicais, definidas pelo predicado “palavra”, além de sua especificação substancial (o que foi feito aqui através de sua representação ortográfica), estão definidas a categoria e os padrões argumentais dos itens lexicais que servirão de núcleo na nossa pequena gramática. Nas regras (37)(c)-(e), as projeções intermediárias são definidas da seguinte maneira: 1) caso o núcleo não aceite nem sequer especificador (“[]” no segundo argumento do predicado “núcleo”), ele sozinho constitui uma projeção intermediária; 2) caso o núcleo não exija complemento (especificado na variável ‘Arg’), ele constitui novamente sozinho a projeção intermediária; e 3) quando o núcleo exigir um complemento (determinado na variável ‘Prim’), ele só vai constituir uma projeção intermediária quando se combinar com uma projeção máxima do mesmo tipo especificado na variável ‘Prim’.<sup>25</sup> O mesmo mecanismo vai ser responsável pela construção da projeção máxima.

---

adjuntos) é apenas contido nela; no entanto, os autores não dizem explicitamente o que essa diferença entre os complementos significa. Mais uma vez para evitar uma controvérsia como essa, a questão não será tratada aqui.

<sup>25</sup> Essa abordagem da estrutura argumental concebida como pilha, que vai sendo consumida à medida que as concatenações adequadas acontecem, é inspirada na HPSG (Pollard & Sag 1994).

Com a DCG de (37), podemos constatar que “Maria corre” e que “João ama Maria” são sentenças dessa gramática; e ainda que as análises gramaticais não sejam explicitamente construídas pelo programa, ele trabalha com a perspectiva de que “Maria” e “João” são nomes que não precisam nem de complemento para constituir uma projeção intermediária, nem de um especificador para constituir uma projeção máxima. O verbo “corre” determina apenas a possibilidade de ter uma projeção máxima de tipo nominal como especificador (definida pela lista “[n]”), mas não permite nenhum complemento. O verbo “ama” seleciona primeiro como complemento uma projeção máxima nominal (estipulado pelo primeiro elemento na lista “[n, n]”, que é consumido e não é herdado pela projeção intermediária resultante); depois, essa projeção intermediária vai exigir uma projeção máxima também nominal como seu especificador para finalmente constituir uma projeção máxima verbal.

### **Conclusão**

É claro que todos os exemplos apresentados nesse texto ainda são rudimentarmente simples, mas eles já começam a demonstrar como é possível implementar certos aspectos dos fenômenos lingüísticos ainda com uma DCG, mesmo aqueles que estão relacionados à gramática gerativa, para os quais se espera uma gramática com poder gerativo maior do que apenas o das gramáticas independentes de contexto.

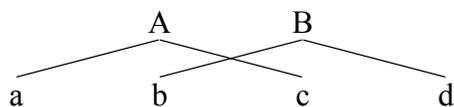
No entanto, o objetivo deste texto era o de apresentar algumas alternativas para lidar com esses fenômenos com recursos mais simples do que os postulados pela gramática gerativa. Além de serem mais simples de se implementar computacionalmente, o que facilita a automatização da análise gramatical, esse tipo de investigação ajuda a entendermos melhor a necessidade ou não de recursos mais poderosos, como os que a gramática gerativa costuma exigir. À medida que for impossível implementar um analisador gramatical com esses recursos mais simples para algum fenômeno lingüístico, fica empiricamente comprovada a necessidade do recurso mais robusto.

Como dissemos, o objetivo era apenas o de apresentar as alternativas, e não de tirar as conclusões sobre a necessidade desses recursos robustos para lidar com o português. Mesmo porque, esse tipo de conclusão exige que outros tipos de fenômenos sejam abordados, e esse texto foi apenas uma introdução a esse tipo de questão.

Na apresentação que fizemos aqui, ainda ficaram de fora muitas categorias, como os adjetivos, os advérbios, os pronomes, as conjunções etc. Além disso, tanto essas categorias que nem foram mencionadas no texto quanto às que foram um pouco mais exploradas, ainda apresentam subdivisões internas, como pudemos ver para os verbos e para as preposições; mas mesmo para estas duas, existem outras subcategorizações que podem ser pertinentes a uma gramática. Nada foi dito, por exemplo, dos verbos inacusativos ou ergativos; os adjetivos também podem apresentar uma subdivisão bastante complexa que ainda está para ser descoberta em relação à língua portuguesa, principalmente em sua manifestação brasileira. Em resumo, resta ainda muito o que fazer.

De qualquer forma, não seria honesto que o leitor terminasse essa leitura com a falsa impressão de que alguma PSG, por mais sofisticada que ela acabe ficando, possa dar conta de todos os fenômenos estruturais lingüísticos. Na verdade, o resultado dos últimos debates parece indicar que existem certos fenômenos lingüísticos que jamais poderiam ser representados através de uma PSG; o argumento empírico do texto em que Shieber (1985) mostra isso era do tipo ‘dependência cruzada’, em que dois constituintes diferentes apresentam subconstituintes intercalados, como no diagrama em (38), no qual dois ramos se cruzam. Só que nenhuma PSG permite a construção de um diagrama como esse.

(38)



Em português, esse tipo de configuração parece ocorrer em construções com o advérbio “respectivamente”, como em “João e Maria são, respectivamente, médico e advogada”. Nessa sentença, “João e Maria” seria um constituinte (o sujeito, para usar o termo da gramática tradicional), enquanto “médico e advogada” seria outro constituinte (o predicativo do sujeito, ainda nos termos da gramática tradicional). Até aí não há nada de errado, o problema é que a primeira parte do sujeito concorda com a primeira parte do predicativo e a segunda parte do sujeito concorda com a segunda parte do predicativo, caracterizando a dependência cruzada.

Assim, a conclusão final parece ser a de que nem todos os fenômenos lingüísticos podem ser representados por uma PSG. De qualquer forma, como uma grande parte desses fenômenos é independente de contexto, uma grande parte de uma determinada língua pode ser estudada através desse recurso. Além disso, como todo bom debate, este também ainda não terminou; desde os primeiros ataques (já nos primeiros textos de Chomsky, mas principalmente em Postal 1964), a PSG sempre encontrou seus defensores, como a tentativa da GPSG (Gazdar et al. 1985). E se o argumento de Shieber parece ser crucial para uma análise exclusivamente lingüística, ele não é tão forte para uma abordagem computacional, já que os recursos computacionais exigidos por uma gramática gerativa são muito maiores do que os de um PSG. Ou seja, ainda que uma PSG não seja uma ferramenta completa para a representação de todos os fenômenos lingüísticos, ela ainda funciona como uma boa primeira abordagem para entendê-los, mesmo que seja para percebermos os próprios limites de nossa ferramenta.

### **Referências bibliográficas**

Chierchia, Gennaro. 1997. **Semantica**. Milano: Il Mulino.

Chomsky, Noam. 1965. **Aspects of the Theory of Syntax**. Cambridge, Massachusetts: The MIT Press.

Gazdar, Gerald; Klein, Ewan; Pullum, Geoffrey K. & Sag, Ivan. 1985. **Generalized Phrase Structure Grammar**. Oxford: Basil Blackwell.

Graffi, Giorgio. 1994. **Sintassi**. Milano: Il Mulino.

Mioto, Carlos; Silva, Maria Cristina Figueiredo & Lopes, Ruth Elisabeth Vasconcellos. 1999. **Manual de Sintaxe**. Florianópolis: Insular.

Pereira, Fernando & Shieber, Stuart. 1987. **Prolog and Natural-Language Analysis**. Stanford: CSLI.

Pereira, Fernando & Warren, David. 1980. Definite clause grammars for language analysis – a survey of the formalism and a comparison with augmented transition networks. **Artificial Intelligence**, 13: 231-278.

Pollard, Carl & Sag, Ivan. 1994. **Head-Driven Phrase Structure Grammar**. Chicago: University of Chicago Press.

Postal, Paul M. 1964. Limitations of phrase structure grammars. *In* Fodor, Jerry A. & Katz, Jerrold J. (eds.); **The Structure of Language - Readings in the Philosophy of Language**. Englewood Cliffs, New Jersey: Prentice-Hall, pp. 137-151.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. **Linguistics and Philosophy**, 8(3): 333-343.

Wood, Mary McGee. 1993. **Categorial Grammars**. London: Routledge.

## **Apêndice**

Como apresentamos inicialmente uma DCG que foi sendo remendada à medida que as questões eram abordadas, apresentamos abaixo a versão definitiva da DCG para lidar com todas as questões apresentadas no texto. Além disso, nesta versão foi incluído um predicado para facilitar a inicialização da solicitação da análise. Esse predicado ‘s/1’ precisa apenas da expressão a ser analisada (em formato de lista), invocando a análise com a quantidade de argumentos apropriada (‘s/3’), de forma que a digitação da solicitação da análise fica mais simples. Além disso, antes de apresentar o resultado da análise em constituintes, esse predicado de inicialização indica explicitamente que vai fazer isso escrevendo “Estrutura de constituintes:” imediatamente antes do resultado.

```
% Inicialização do analisador
s(S) :- s(EC, S, []),
        write('Estrutura de constituintes: '),
        write(EC), nl.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Regras sintagmáticas %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% sentença
s([SN, SV]) --> sn([_, Num], SN), sv(Num, SV).

% sintagmas nominais
sn(Conc, NBarra) --> n_barra(Conc, NBarra).
sn(Conc, [Det, NBarra]) --> det(Conc, Det), n_barra(Conc, NBarra).

n_barra(Conc, N) --> n(Conc, N).
```

```

% sintagmas verbais
sv(Num, V) --> v(Num, i, V).
sv(Num, [V, SN]) --> v(Num, td, V), sn(_, SN).
sv(Num, [V, SP]) --> v(Num, ti(P), V), sp(P, SP).
sv(Num, [V, SN, SP]) --> v(Num, tdi(P), V), sn(_, SN), sp(P, SP).

% sintagma preposicionado
sp(Lex, [P, SN]) --> p(Lex, inf, P), sn(_, SN).
sp(Lex, [P, NBarra]) --> p(Lex, Conc, P), n_barra(Conc, NBarra).

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Regras lexicais %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

det(Conc, Det) --> [Det], {det(Conc, Det)}.
n(Conc, N) --> [N], {n(Conc, N)}.
v(Num, Val, V) --> [V], {v(Num, Val, V)}.
p(Lex, Conc, P) --> [P], {p(Lex, Conc, P)}.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Léxico %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Determinantes
det([masc, sing], o).
det([fem, sing], a).
det([masc, plur], os).
det([fem, plur], as).
det([masc, sing], um).
det([fem, sing], uma).
det([masc, plur], uns).
det([fem, plur], umas).

% Nomes
n([fem, sing], estante).
n([fem, plur], estantes).
n([masc, sing], homem).
n([masc, plur], homens).
n([masc, sing], João).
n([masc, sing], livro).
n([masc, plur], livros).
n([fem, sing], maria).
n([masc, sing], menino).
n([fem, sing], menina).
n([masc, plur], meninos).
n([fem, plur], meninas).
n([fem, sing], mulher).
n([fem, plur], mulheres).

% Verbos
v(sing, i, corre).

```

```
v(plur, i, correm).
v(sing, td, ama).
v(plur, td, amam).
v(sing, ti(de), gosta).
v(plur, ti(de), gostam).
v(sing, tdi(em), coloca).
v(plur, tdi(em), colocam).

% Preposições
p(com, inf, com).
p(de, [_,_], de).
p(de, [masc,sing], do).
p(de, [fem, sing], da).
p(de, [masc, plur], dos).
p(de, [fem, plur], das).
p(em, [_,_], em).
p(em, [masc, sing], no).
p(em, [fem, sing], na).
p(em, [masc, plur], nos).
p(em, [fem, plur], nas).
p(para, inf, para).
p(por, [_,_], por).
p(por, [masc,sing], pelo).
p(por, [fem, sing], pela).
p(por, [masc, plur], pelos).
p(por, [fem, plur], pelas).
```