

# Gramática e Processamento da Linguagem Natural: Fundamentos

António Branco  
Francisco Costa

DI-FCUL

TR-2008-16

September 2008

Departamento de Informática  
Faculdade de Ciências da Universidade de Lisboa  
Campo Grande, 1749-016 Lisboa  
Portugal

Technical reports are available at <http://www.di.fc.ul.pt/tech-reports>. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.



# Gramática e Processamento da Linguagem Natural: Fundamentos

António Branco

Francisco Costa



# Resumo

O presente documento tem o objectivo didáctico de apresentar uma breve introdução à área do processamento de linguagem natural que pode ser utilizada por estudantes ou investigadores interessados. Esta introdução é elaborada do ponto de vista do processamento linguístico profundo com base na gramática das línguas naturais. Adopta como enquadramento metodológico a Head-Driven Phrase Structure Grammar (HPSG).

O presente documento serve também de introdução metodológica ao documento TR-2008-17, que descreve a concretização de um fragmento da gramática para a língua portuguesa de acordo com os princípios aqui apresentados.

Embora possam ser usados autonomamente, estes dois documentos devem ser vistos como constituindo duas partes de um mesmo texto.

O presente documento foi escrito em português por forma a tornar os seus temas disponíveis pela primeira vez para um público de falantes desta língua — outros textos em língua inglesa já existem com objectivos aproximados. O documento TR-2008-17, por sua vez, foi escrito em inglês por forma a que os resultados aí apresentados possam ser utilizados por toda a comunidade científica.

**Palavras-chave:** processamento de linguagem natural, linguística computacional, processamento linguístico profundo, gramática computacional, HPSG, gramática universal, estruturas de traços, unificação, léxico, subcategorização, constituência, funções gramaticais, concordância, dependências de longa-distância, configuracionalidade, representação e composição do significado, ambiguidade e sub-especificação.



# Conteúdo

<b>Introdução</b>	<b>7</b>
<b>1 Arquitectura</b>	<b>9</b>
1.1 Introdução . . . . .	9
1.2 Assunções Fundacionais . . . . .	9
1.3 Bases Empíricas . . . . .	11
1.4 Estruturas de Modelação . . . . .	11
1.5 Meios Descritivos . . . . .	13
1.6 Dispositivos Computacionais . . . . .	17
1.7 Leituras Subsequentes . . . . .	20
<b>2 Sintaxe</b>	<b>23</b>
2.1 Introdução . . . . .	23
2.2 Tipos e Estruturas de Traços . . . . .	23
2.3 Léxico e Subcategorização . . . . .	26
2.4 Constituintes e Funções Gramaticais . . . . .	28
2.5 Concordância . . . . .	35
2.6 Dependências de Longa Distância . . . . .	39
2.7 Configuracionalidade . . . . .	45
2.8 Leituras Subsequentes . . . . .	45
<b>3 Semântica</b>	<b>47</b>
3.1 Introdução . . . . .	47
3.2 Representação do Significado . . . . .	48
3.3 Composição do Significado . . . . .	49
3.4 Ambiguidade e Sub-especificação . . . . .	54
3.5 Representações Semânticas em HPSG . . . . .	58
3.6 Outros Exemplos . . . . .	63
3.7 Leituras Subsequentes . . . . .	68
<b>Bibliografia</b>	<b>71</b>





# Introdução

Como acontece em qualquer tentativa de definição, quando se procura por uma definição do que é o processamento computacional da linguagem natural, encontrar-se-á um amplo leque de diferentes propostas.

Num dos extremos do espectro encontrar-se-á a indicação de que a investigação desenvolvida na área do processamento de linguagem natural tem por objectivo último tornar possível interagir com dispositivos e agentes artificiais através da utilização simples e directa da linguagem natural, tal como esta interacção tem lugar entre seres humanos.

No outro extremo do espectro, dando enfoque a subtarefas de mais baixo nível e a aplicações tecnológicas de viabilidade mais imediata, o processamento de linguagem natural será visto como tendo por objectivo a explicitação ou extracção de informação linguística a partir de materiais textuais por forma a apoiar a execução de uma ampla gama de tarefas e sistemas, tais como categorização, sumarização, indexação ou recuperação de textos, extracção de informação, resposta a perguntas, geração de relatórios, sistemas de diálogo, memórias de tradução, tradução automática, etc.

Qualquer que seja porém a perspectiva em que se encare o processamento da linguagem natural, um dos aspectos centrais desta área consiste em lidar com o conhecimento linguístico, modelando as suas regularidades e sobretudo o modo como forma e significado linguísticos se encontram associados de forma sistemática.

Do ponto vista computacional, esta associação é explorada procurando extrair-se automaticamente uma representação do significado a partir de uma forma linguística superficial, ou então gerando-se a partir de uma representação do significado as formas linguísticas superficiais que o podem expressar. Estas duas tarefas são agrupadas de baixo do que é designado por processamento linguístico profundo, para o qual a HPSG proporciona um dos enquadramentos mais avançados e abrangentes.

“HPSG” é o acrónimo da expressão inglesa *Head-Driven Phrase Structure Grammar*, que designa um enquadramento metodológico para o estudo da linguagem natural humana e em particular para o estudo científico daquilo que, na sequência de uma tradição milenar, se designa por gramática.

O presente conjunto de notas didácticas procura oferecer uma breve introdução à área do processamento de linguagem natural do ponto de vista da gramática e do processamento linguístico profundo, adoptando por enquadramento a HPSG.

Esta publicação é composta por três capítulos. No primeiro capítulo, é apresentada

a arquitectura básica do enquadramento metodológico adoptado.

Nos dois capítulos seguintes é progressivamente descrito o conteúdo de uma pequena gramática de um fragmento do Português. O Capítulo 2 ocupa-se essencialmente de aspectos relativos à sintaxe. O Capítulo 3, por sua vez, ocupa-se da semântica. A gramática é aí ampliada com dispositivos que permitem representações do significado para expressões arbitrariamente complexas.

A disponibilização do presente documento na série Technical Reports do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa visa divulgar uma versão pré-final de um texto que pode vir a encontrar edição comercial, pelo que todos os direitos se mantêm reservados aos seus autores.

# Capítulo 1

## Arquitectura

### 1.1 Introdução

No presente capítulo são apresentados os primitivos conceptuais, as assunções fundacionais, as bases empíricas e os dispositivos descritivos e computacionais básicos da gramática e do processamento da linguagem natural.

Na Secção 1.2 descrevem-se os objectivos principais do processamento linguístico profundo no quadro da HPSG: caracterização explícita e precisa do conhecimento linguístico, capacidade preditiva, neutralidade procedimental, processamento incremental, tratabilidade computacional, distinção entre princípios gramaticais universais e generalizações específicas a cada língua, e capacidade de integração da informação linguística com outras fontes de informação.

A Secção 1.3 discorre sobre a obtenção de dados empíricos que alimentam a indução de generalizações gramaticais a ser incluídas na teoria.

Nas Secções 1.4 e 1.5 apresenta-se o formalismo utilizado para modelar os fenómenos linguísticos: estruturas de traços e hierarquias de tipos.

Na Secção 1.6 fala-se das operações computacionais de maior relevo assumidas pela gramática.

A terminar, a Secção 1.7 apresenta referências para algumas obras introdutórias, assim como outras fontes bibliográficas de interesse.

### 1.2 Assunções Fundacionais

Na HPSG é assumido que a faculdade da linguagem é uma capacidade cognitiva específica, com autonomia considerável em relação a outras capacidades cognitivas dos sujeitos humanos. É assumido também que os estudos gramaticais têm por objectivo a descoberta do conhecimento linguístico subjacente a essa capacidade cognitiva. A gramática é por conseguinte entendida como uma teoria explícita acerca desse conhecimento, que assegura uma caracterização precisa do mesmo.

Enquanto teoria, a gramática apresenta capacidade preditiva: através de uma caracterização finita, representa o conhecimento potencial que um sujeito detém acerca de um número eventualmente ilimitado de putativas entidades linguísticas. Esse conhecimento abrange quer entidades linguísticas pertencentes à linguagem específica de que o sujeito é um utilizador competente (e.g. Português, Chinês, Polaco, etc), quer para-entidades linguísticas não pertencentes a essa linguagem.

Não obstante a sua especificidade, a faculdade da linguagem é uma de entre as diversas faculdades cognitivas. Nesta medida, é concebida no quadro das assunções fundacionais da Ciência Cognitiva, em que os processos cognitivos são entendidos como tendo por base o processamento de informação cujo melhor modelo até à data é justificado pelo Resultado Fundamental da Teoria da Computação. Deste modo, a gramática de uma linguagem  $L$  é uma teoria compatível com um algoritmo de *parsing* que, para cada putativa entidade linguística, permite decidir se essa entidade pertence a  $L$ .

Tratando-se de uma teoria do conhecimento linguístico, a gramática oferece a melhor compatibilidade com a modelação do comportamento linguístico e em particular com o processamento de informação linguística. Nesta medida, a gramática é compatível com um algoritmo de *parsing* se não eficiente em termos gerais do ponto de vista da complexidade computacional, pelo menos tratável dentro das condições típicas da execução do *parser* mental natural.

Por outro lado, para ser compatível com diferentes regimes de processamento linguístico, e.g. compreensão, produção ou tradução, etc, a descrição do conhecimento gramatical é feita com neutralidade procedimental através da utilização de um dispositivo descritivo declarativo.

Adicionalmente, a gramática é compatível com modelos de processamento parcial de entidades linguísticas e com a articulação flexível entre os seus sub-módulos (i.e. fonologia, morfologia, sintaxe, semântica, pragmática). Nesta medida, o dispositivo descritivo que usa permite também a integração flexível das contribuições funcionais dos diferentes sub-módulos de forma incremental.

Em face de aspectos comuns que se abstraem das gramáticas de diferentes línguas naturais, neste enquadramento epistemológico é assumido que existe um núcleo de conhecimento gramatical comum às diferentes linguagens. Esse conhecimento é designado por Gramática Universal. Cada gramática de um idioma individual integra a descrição desse conhecimento universal comum e apresenta uma sua especificação e extensão particular, a qual constitui o conhecimento específico desse idioma particular. Deste modo, a gramática é compatível com uma leque de modelos de aquisição individual da linguagem, incluindo os que advogam alguma forma de inatismo.

Quando instanciado numa dada entidade linguística particular, o conhecimento gramatical associa a forma dessa entidade (viz. marcas simbólicas sonoras) e o seu respectivo significado—ou os seus respectivos significados, no caso de haver ambiguidade para a qual o contexto não é suficiente para possibilitar a sua resolução por retenção apenas do significado intencionado. Por conseguinte, a par da descrição precisa do conhecimento gramatical, a gramática garante a descrição precisa da informação

linguística associada a cada entidade linguística em que esse conhecimento se instancia. Desta maneira, permite associar a representação da forma linguística e a do respectivo significado, no que se designa de representação gramatical.

As representações gramaticais são compatíveis com a integração da representação de informação especificamente linguística com a representação do conhecimento genérico, não-linguístico, do mundo. Permitem também a integração da representação do significado com modelos de processamento inferencial. Nesta medida, a gramática é compatível com o conhecimento não linguístico e com a dimensão regular do raciocínio.

### 1.3 Bases Empíricas

A base empírica para a construção de uma gramática de uma linguagem  $L$  é a base habitual do trabalho de investigação sobre gramática, a qual vem na linha de uma tradição milenar. Os dados empíricos são recolhidos através de acesso introspectivo a juízos acerca da pertença de putativas entidades linguísticas a  $L$ , também referidos como juízos de gramaticalidade ou de aceitabilidade.

A assunção aqui é a de que o investigador que executa esses juízos é um falante plenamente competente de  $L$  e a de que, em larga medida e na extensão relevante para a construção da gramática, a sua competência gramatical é idêntica à dos restantes falantes de  $L$ . A validade e produtividade desta metodologia é assegurada pela dimensão colectiva do trabalho científico, a qual neste caso se traduz na avaliação crítica dos dados empíricos recolhidos por parte de outros investigadores, também falantes de  $L$ , que confirmam se os dados utilizados coincidem com os juízos obtidos por si relativamente às mesmas entidades.

Para melhor aceder a esses juízos e aclarar as suas implicações, estes podem ser executados sobre pares de entidades linguísticas. Com o objectivo de ajudar a isolar determinadas propriedades linguísticas específicas, é comum procurar-se que as entidades assim colocadas em contraste difiram entre si de forma mínima e apenas em resultado de diferenças relativas a essas propriedades sob investigação.

Outra fonte de dados empíricos encontra-se nos repositórios de produções linguísticas conhecidos como *corpora*, e cuja utilização contribui para assegurar que a gramática possui uma cobertura com máxima abrangência. As entidades linguísticas que se encontram nos *corpora* ajudam a sugerir dados empíricos que, sem o recurso à análise destes repositórios, poderiam escapar a um escrutínio baseado apenas na consideração de entidades linguísticas obtidas por manipulação mental com vista a se construir contrastes mínimos.

### 1.4 Estruturas de Modelação

Sendo uma teoria sobre um domínio empírico, a gramática e as entidades desse domínio são mediados por uma estrutura matemática que serve de modelo deste último. Esta

estrutura e as suas partes estão em correspondência com as entidades observáveis por forma a que estas se encontrem representadas por entidades da estrutura simbólica que as modela. Neste arranjo ontológico tripartido—observáveis, modelo e teoria—, a gramática discorre acerca das entidades linguísticas discorrendo acerca das suas representações, sendo assim interpretada no modelo simbólico do seu domínio empírico.

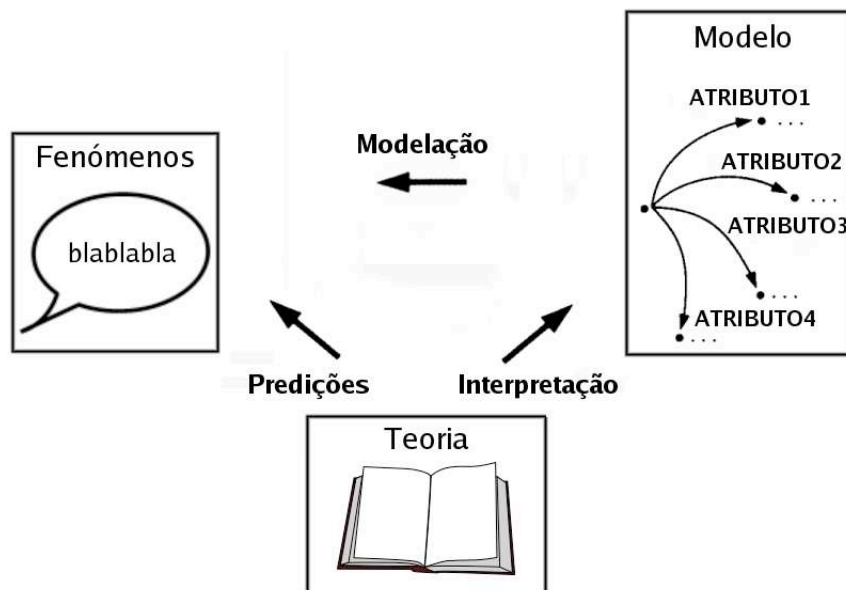


Figura 1.1: Diagrama esquemático das relações entre observáveis, modelo e teoria. Exemplo adaptado de (Pollard e Sag, 1994, p. 9).

Os observáveis relevantes do domínio a modelar são os tipos de expressões da linguagem natural assim como os das suas subexpressões. Um tipo de expressão, entre inúmeros outros, será por exemplo aquele que se representa por “mesa”, o qual se obtém por abstracção a partir das diferenças acidentais presentes nos enunciados particulares de “mesa”.

Para representar os observáveis no modelo, são usadas hierarquias de tipos, as quais são ordens parciais, e estruturas de traços (*feature structures*), que são entidades definidas na teoria dos grafos.

Uma hierarquia de tipos é um grafo acíclico em que os nós são etiquetados com etiquetas de tipos e em que um nó  $s_2$  onde chega um arco que sai de um nó  $s_1$  é um subtipo de  $s_1$ , e por isso mais específico que este último.

Uma estrutura de traços é um grafo em que os nós são etiquetados com etiquetas de tipos e os arcos com etiquetas de atributos. Por conseguinte, de uma dada estrutura de traços não atômica de tipo  $s$ —i.e. um grafo cujo nó de topo está etiquetado com  $s$ —, diz-se que tem atributos—aqueles que etiquetam os arcos que saem do nó  $s$ . Por sua vez, de um atributo  $a$ , diz-se que tem um valor, que é uma outra estrutura de traços (atômica ou não) para que o arco etiquetado com  $a$  aponta.

Dado que são modelos totais de entidades linguísticas, as estruturas de traços têm de ser (i) totalmente bem-tipadas e (ii) tipo-resolvidas. Em termos informais, isto impõe que: (i) para cada nó  $s$  no grafo, cada arco que devia sair de  $s$  por ser apropriado para caracterizar  $s$  está de facto presente; (ii) cada nó é etiquetado com o tipo  $s$  que é maximamente específico na hierarquia de tipos que  $s$  integra.

Uma propriedade crucial das estruturas de traços é que dois caminhos distintos no grafo podem levar a um mesmo nó. Nesta medida, a partilha de estrutura é possível ou seja, é possível dois caminhos no grafo terem uma mesma estrutura como seu valor comum.

Neste enquadramento, a teoria é uma especificação a ser interpretada no modelo do domínio constituído por estruturas de traços. Essa especificação é um conjunto de restrições que definem uma caracterização intensional finita de um conjunto eventualmente infinito de entidades desse modelo. Nesta medida, essas restrições estabelecem a predição de quais entidades de entre as que estão disponíveis no modelo (e através da sua mediação, que entidades de entre as que se encontram no domínio empírico) são entidades que pertencem à linguagem natural para a qual a teoria em questão é uma gramática.

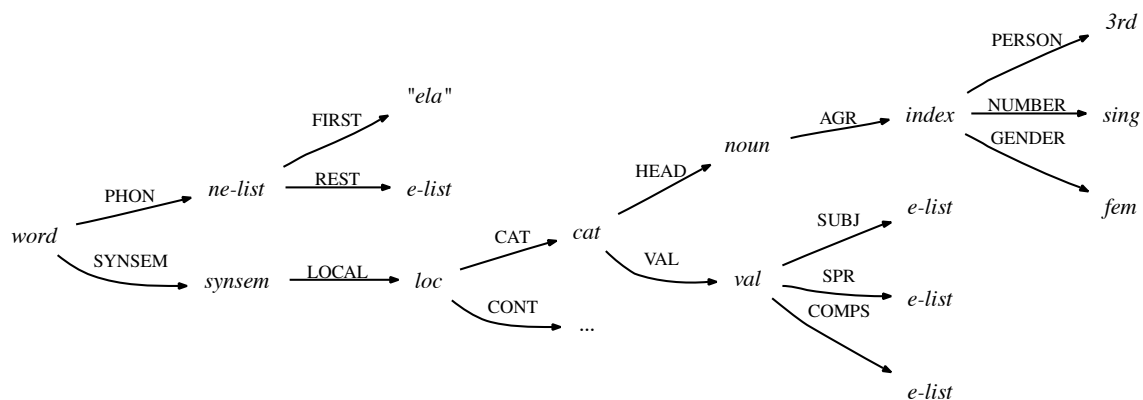


Figura 1.2: Exemplo de um grafo que representa uma entidade linguística no modelo, no caso a palavra “ela” (o valor da maior parte das etiquetas é apresentado no Capítulo 2). Exemplo adaptado a partir de (Pollard e Sag, 1994, p. 17).

## 1.5 Meios Descritivos

De um ponto de vista algébrico, uma gramática é uma especificação, que comporta uma assinatura e uma apresentação.

A assinatura define que tipos de dados estão disponíveis para representar entidades linguísticas. Inclui uma hierarquia de tipos e uma definição de adequação.

A hierarquia de tipos é uma ordem parcial de tipos em que os tipos possíveis para entidades linguísticas e seus subtipos são estabelecidos.

A definição de adequação, também designada por declaração de traços, estabelece quais são as características apropriadas para cada tipo na hierarquia. Consiste em associar a cada tipo restrições que definem quais as estruturas de traços apropriadas para esse tipo de entidade. A hierarquia de tipos é uma árvore taxonómica em que cada tipo herda as restrições de adequação dos seus supertipos.

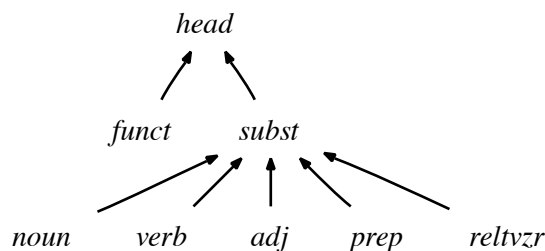


Figura 1.3: Exemplo de uma hierarquia de tipos, no caso relativa à hierarquia dos subtipos de *núcleo* (*head*), que inclui no nível imediatamente inferior *núcleo funcional* (*funct*) e *núcleo substantivo* (*subst*), o qual tem como subtipos *preposição* (*prep*), *relativizador* (*reltvzr*), *nome* (*noun*), *adjectivo* (*adj*) e *verbo* (*verb*).

A apresentação é um conjunto de restrições implicacionais que são interpretadas (como sendo verdadeiras ou falsas) no modelo. São entendidas como descrições que circunscrevem o conjunto das estruturas de dados que, obedecendo ao formato dos tipos de dados estabelecido na assinatura, a teoria prediz serem entidades da linguagem natural em questão.

As restrições que integram a gramática são expressas numa linguagem de descrição específica cujas expressões são matrizes de atributos-valores (MAV). Numa apresentação abreviada da sua sintaxe, pode dizer-se que as MAVs não atómicas consistem em matrizes de duas colunas, em que na primeira coluna se encontram os atributos e na segunda coluna os valores correspondentes. Os valores dos atributos são MAVs. As matrizes recebem um subscripto ou superscrito à esquerda que indica o seu tipo. Uma MAV atómica consiste apenas num tipo atómico.

$$\left[ \begin{array}{l} \textit{tipo} \\ \text{ATRIBUTO}_1 \quad \textit{mav}_1 \\ \dots \quad \dots \\ \text{ATRIBUTO}_n \quad \textit{mav}_n \end{array} \right]$$

Figura 1.4: Exemplo esquemático de uma matriz de atributos-valores (MAV) ou de traços.

Existe uma simbologia reservada para indicar partilha de estrutura. Quando cada um de dois atributos tem por valor uma etiqueta formada por uma cercadura em que se



inscreve um número, e esses dois números são idênticos, isso indica que esses atributos têm o mesmo valor, ou seja que os seus valores são uma e a mesma instância. Essas etiquetas podem ser seguidas por MAVs, caso em que se está a expressar qual o valor a que a etiqueta corresponde.

As descrições de conjuntos são feitas da forma esperada, através do uso de chavetas em torno dos seus elementos, separados por vírgulas, e as descrições de listas são feitas através do uso similar de parênteses angulares.

Cabe salientar que as MAVs expressam restrições sobre as entidades (totais) do universo do modelo, ou seja são descrições (parciais) das entidades para cuja caracterização contribuem.

Dependendo da natureza da informação linguística que capturam, as restrições constantes da apresentação costumam ser designadas por princípios gramaticais ou entradas lexicais.

A gramática de uma dada linguagem natural inclui um conjunto de princípios que se assume, por generalização indutiva, serem parte da gramática de qualquer linguagem natural. Estes princípios são designados como princípios da Gramática Universal.

$$\begin{array}{c} \text{Princípio do Traço do Núcleo} \\ \left[ \begin{array}{l} \textit{headed-phrase} \\ \text{SYNSEM|LOCAL|CAT|HEAD } \boxed{1} \\ \text{HEAD-DTR } \left[ \begin{array}{l} \textit{sign} \\ \text{SYNSEM|LOCAL|CAT|HEAD } \boxed{1} \end{array} \right] \end{array} \right] \end{array}$$

Figura 1.5: Exemplo de um princípio gramatical, no caso o Princípio do Traço do Núcleo (*Head Feature Principle*), que impõe que se se trata de um sintagma com núcleo (*headed-phrase*), então os valores do atributo HEAD do constituinte que é o núcleo e os do sintagma são idênticos. Os valores do atributo HEAD são do tipo *head*, definidos na hierarquia da Figura 1.3. Duas etiquetas idênticas denotam identidade entre instâncias. Neste caso, a identidade entre o atributo HEAD do nó mãe e o atributo HEAD do núcleo (a HEAD-DTR) é representada pelas duas etiquetas  $\boxed{1}$ .

Juntamente com estes princípios, cada gramática para um dado idioma particular inclui princípios que valem apenas para esse idioma ou para um conjunto ou família de línguas a que esse idioma pertence. Todos estes princípios integram a gramática como uma conjunção de restrições, que qualquer entidade linguística tem de satisfazer.

A par do conjunto dos princípios gramaticais, as restrições que correspondem a entradas lexicais formam o conjunto dos princípios satisfazíveis disjuntivamente.

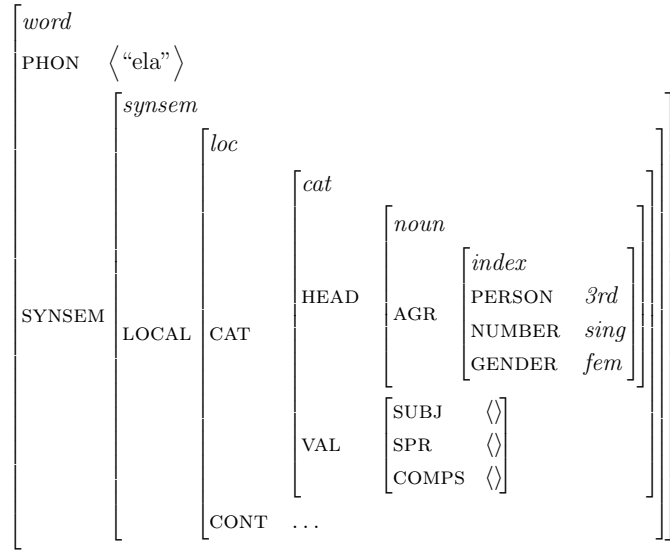


Figura 1.6: Exemplo da descrição de uma entrada lexical, no caso da palavra “ela”.

Outras restrições disjuntivas são também incluídas na gramática e são conhecidas como esquemas (para relações de dominância e de precedência).

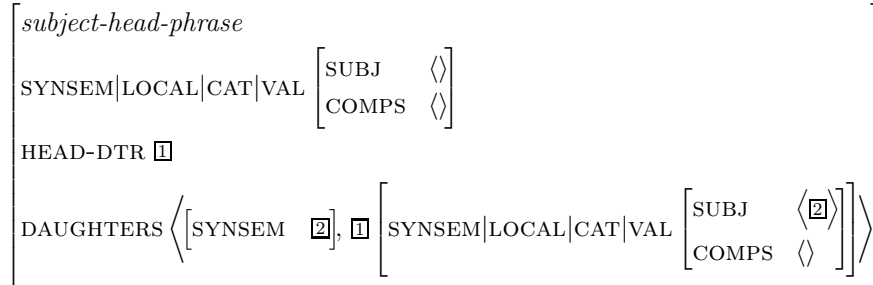


Figura 1.7: Exemplo de um Esquema, no caso o Esquema Núcleo-Sujeito (*subject-head-phrase*), que impõe entre outras coisas que os atributos SUBJ e COMPS se encontram vazios, o que corresponde à circunstância de que um sintagma com o sujeito realizado não dispõe de capacidade de subcategorizar qualquer outro argumento.

Finalmente, um outro conjunto ainda de restrições, conhecidas como regras lexicais, inclui restrições com propriedades formais distintas, usadas para captar generalizações acerca de entradas lexicais. Uma regra lexical é um par de “meta-descrições” visto relacionar entidades da especificação (descrições) e não entidades do modelo (estruturas de traços). Ao se aplicar uma regra lexical, a restrição que é mencionada no seu lado esquerdo é usada para circunscrever as estruturas de traços que a satisfazem, às quais supletivamente se aplicam as restrições mencionadas no lado direito da regra.

$$\left[ \begin{array}{l} \text{PHON } \boxed{1} \\ \text{SYNSEM } \boxed{2} \left[ \text{LOCAL|CAT|HEAD|AGR|NUMBER } \textit{sing} \right] \\ \text{DAUGHTERS } \left\langle \begin{array}{l} \textit{noun-lexeme} \\ \text{PHON } \boxed{1} \\ \text{SYNSEM } \boxed{2} \end{array} \right\rangle \end{array} \right]$$

Figura 1.8: Exemplo de uma Regra Lexical, no caso a que produz a forma singular dos nomes. A unificação dos atributos PHON modela o facto de que a forma singular de um nome é idêntica ao seu lema.

Na Figura 1.9 encontra-se uma sinopse das diferentes classes de restrições que integram a gramática de uma linguagem  $L$ .

$\text{GU} = P_1 \wedge \dots \wedge P_n$ $\text{Linguagem } L = \text{GU} \wedge P_{n+1} \wedge \dots \wedge P_{n+m} \wedge (S_1 \vee \dots \vee S_p \vee L_1 \vee \dots \vee L_q)$	
<i>Em que:</i>	
$P_1, \dots, P_n$	são princípios da Gramática Universal (GU)
$P_{n+1}, \dots, P_{n+m}$	são princípios específicos de $L$
$S_1, \dots, S_p$	são esquemas agrupados em princípios específicos de $L$
$L_1, \dots, L_q$	são signos lexicais de $L$ (básicos ou o resultado de regras lexicais)

Figura 1.9: Sinopse das diferentes classes de restrições implicacionais que integram uma gramática de uma linguagem natural  $L$ .

Um exemplo acessível de uma gramática HPSG para um fragmento de uma língua natural, no caso o Inglês, encontra-se nas 9 páginas do Apêndice de (Pollard e Sag, 1994).

## 1.6 Dispositivos Computacionais

O conhecimento linguístico tal como é descrito na gramática constitui uma teoria com capacidade preditiva. Essa capacidade é tornada operacional no desempenho de duas tarefas computacionais chave: (i) síntese (ou geração): dada uma representação semântica, obter o conjunto de frases cuja representação gramatical inclui essa representação semântica, ou seja, a lista de frases com significado idêntico ao que é expresso pela representação semântica de partida; (ii) análise: dada uma expressão, verificar se essa expressão pertence à linguagem em questão, caracterizada pela gramática, e em caso afirmativo obter o conjunto das suas representações gramaticais.

Diversas linguagens de programação, também conhecidas por formalismos de implementação, têm sido desenvolvidas de tal forma que uma gramática pode ser escrita

como um programa numa dessas linguagens, o qual é executável num computador. Estas linguagens de programação oferecem meios descritivos que se aproximam dos dispositivos descritivos definidos na HPSG, ao mesmo tempo que asseguram que a execução dos seus programas permitem o desempenho das tarefas acima referidas.

Dois procedimentos computacionais desempenham um papel nuclear na execução desses formalismos de implementação, o mecanismo de unificação e o de *parsing*.

O mecanismo de *parsing* permite decidir se uma expressão pertence ao conjunto das expressões definidas intensionalmente por uma gramática como formando uma linguagem.

Em termos muito genéricos, e no quadro dos formalismos de implementação, o mecanismo de unificação permite a construção de representações através da combinação de representações parciais compatíveis e da eventual instanciação de variáveis nessas representações.

$$\left[ \begin{array}{l} \textit{noun-lexeme} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD|AGR } \square \\ \text{VAL|SPR } \langle \left[ \text{LOCAL|CAT|HEAD|AGR } \square \right] \rangle \end{array} \right] \end{array} \right]$$

Figura 1.10: Exemplo de uma operação de unificação, no caso a que impõe que os traços de género e número de um nome (representados sob o atributo AGR) devem ser idênticos aos traços de género e número do seu especificador (no traço SPR).

No seguimento da tarefa de análise, alguns formalismos de implementação permitem a execução automática de uma terceira tarefa, dita de resolução de ambiguidade. Essa tarefa consiste em indicar, de entre as representações gramaticais possíveis para uma dada expressão de partida, aquela que mais provavelmente é veiculada por essa expressão. Para esta finalidade, são usados procedimentos automáticos de classificação baseados em parâmetros estocásticos obtidos a partir de *corpora*, conhecidos como bancos de árvores (*treebanks*), em que cada frase foi previamente anotada com a representação gramatical mais favorecida pelo seu contexto.

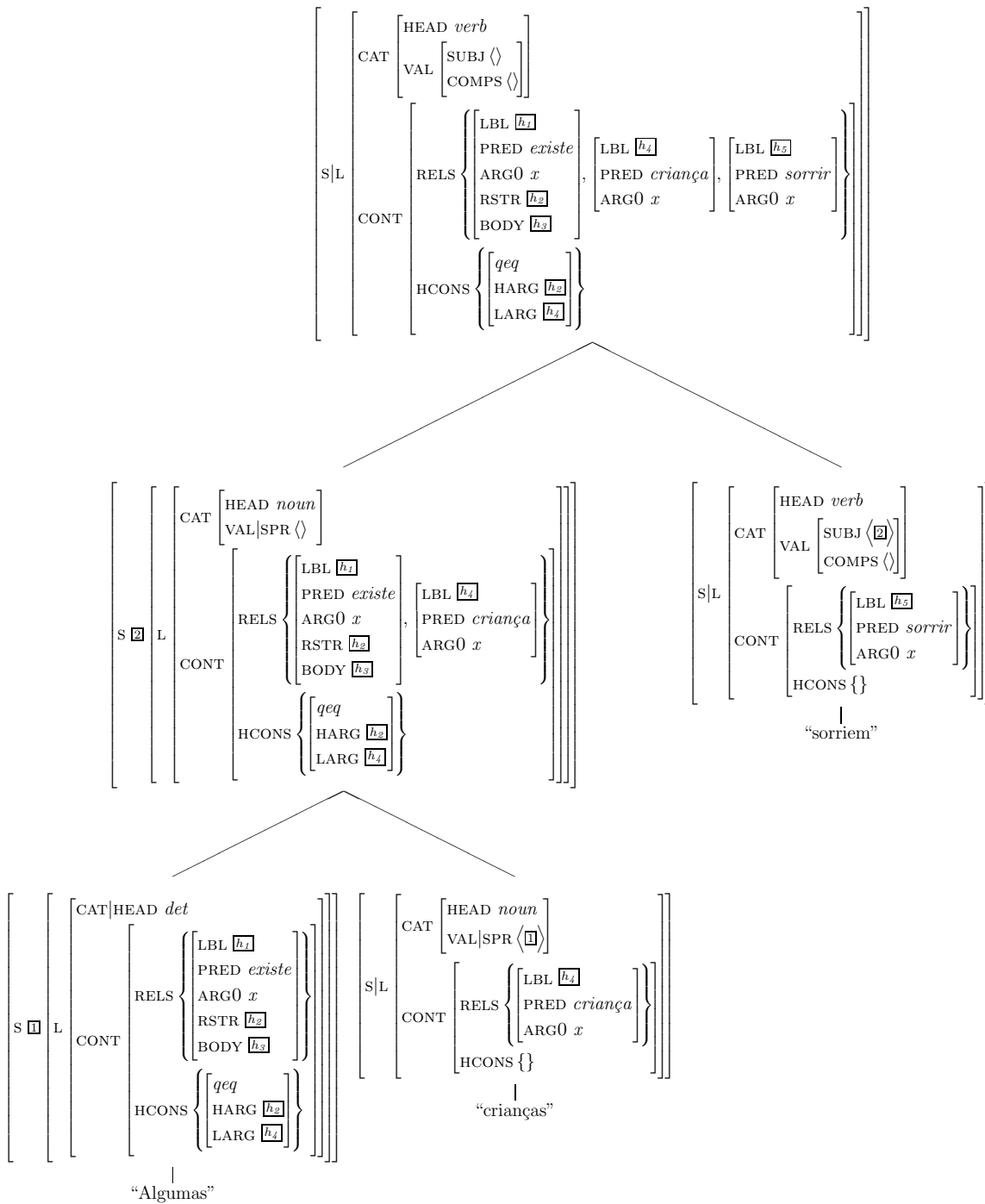


Figura 1.11: Exemplo de uma representação gramatical abreviada, no caso a que é relativa à frase “Algumas crianças sorriem”. O atributo SYNSEM está abreviado como S, e LOCAL como L. Este exemplo ilustra a forma como a subcategorização é codificada na HPSG, através dos atributos dentro de VAL: SUBJ e SPR (vd. Capítulo 2). Também ilustra a representação semântica atribuída aos vários constituintes, dentro do atributo CONT: RELS é uma coleção de relações e HCONS restringe as possíveis relações de âmbito entre elas (vd. Capítulo 3).

## 1.7 Leituras Subsequentes

Para uma introdução ao estudo científico da gramática, a partir de um nível básico e sem pré-requisitos, (Kim e Sells, 2008) é um excelente manual pedagógico que progressivamente expõe conceitos linguísticos chave, exercita técnicas nucleares de investigação e argumentação e usa meios formais de modelação e descrição. Para o leitor que procura uma exposição mais alongada e que não se inicie a partir de um nível tão básico, (Sag *et al.*, 2003) é um ponto de partida alternativo.

O enquadramento epistemológico da HPSG encontra-se definido nos dois volumes fundacionais da autoria de Pollard e Sag (1987, 1994), os quais constituem leituras complementares e de aprofundamento em relação a estas duas obras de introdução referidas acima.

Os leitores de um nível avançado têm como fonte principal de documentação o repositório bibliográfico *online* da HPSG, com acesso livre à maior parte das publicações aí listadas, que se encontra em (HPSG, Página Web).

As assunções fundacionais da HPSG têm a sua base nos pressupostos epistemológicos que na segunda metade do século XX estiveram na origem e amadureceram sob a forma de Ciência Cognitiva, sendo que de entre estes, têm especial relevo os pressupostos que emergiram no âmbito dos estudos sobre a chamada gramática formal. Os manuais de Friedenberg e Silverman (2006) e de Fromkin (2000) são introduções conceituadas a cada uma destas áreas e oferecem um amplo leque de referências a outras obras de maior profundidade e especialidade.

A obra de referência para os meios descritivos explorados na HPSG é a de Carpenter (1992). Para uma panorâmica sobre questões relacionadas com os fundamentos computacionais do processamento de linguagem natural, incluindo os que se aplicam no quadro da HPSG, veja-se (Branco, 2005a) e as referências para que remete. Uma introdução à utilização do mecanismo de unificação em gramáticas encontra-se em (Shieber, 1986).

O desenvolvimento de dispositivos computacionais para gramáticas da família da HPSG foi alvo de diversas iniciativas durante a última década do século XX, documentadas em (Uszkoreit *et al.*, 1996). Dessas contribuições, destacam-se os sistemas de implementação TRALE (TRALE, Página Web) e LKB (LKB, Página Web), por o seu desenvolvimento continuar a ser mantido.

De entre estes sistemas, cabe assinalar o LKB (*Linguistic Knowledge Builder*) por um conjunto de factores de que a seguir se destacam alguns dos mais importantes. Trata-se de *software* cujo acesso é gratuito, encontra-se apropriadamente documentado, oferece uma interface gráfica amigável em relação ao utilizador e apresenta níveis de eficiência de desempenho que permitem a sua utilização em aplicações computacionais comerciais.

Existe uma obra de introdução acessível ao desenvolvimento de gramáticas computacionais através da utilização do LKB (Copestake, 2002), com pré-requisitos mínimos relativamente às competências dos leitores.

O LKB é utilizado pela maior comunidade de investigadores em engenharia de

gramáticas para o processamento linguístico profundo, reunidos no consórcio Delph-In (DELPH-IN, Página Web).

LXGram é uma gramática computacional para o Português desenvolvida sobre o LKB, no âmbito deste consórcio, pela Universidade de Lisboa. Encontra-se disponível em (LXGram, Página Web).





# Capítulo 2

## Sintaxe

### 2.1 Introdução

Este capítulo apresenta uma introdução à representação sintáctica em HPSG. Para tanto, são apresentadas nos seus aspectos essenciais as análises gramaticais para alguns fenómenos sintácticos básicos.

Cabe lembrar que na HPSG, a representação gramatical é uma descrição integrada das diversas dimensões linguísticas. A apresentação separada da componente de representação sintáctica faz-se apenas por conveniência de exposição.

Na Secção 2.2, são apresentados os traços e tipos mais importantes para a organização da informação necessária para modelar os fenómenos sintácticos abrangidos neste capítulo.

De seguida, é apresentada na Secção 2.3 a forma de codificar a informação categorial e de subcategorização nos itens lexicais, e na Secção 2.4 são apresentadas as regras sintácticas principais.

As secções seguintes debruçam-se sobre o tratamento da concordância, Secção 2.5, e das dependências de longa distância, Secção 2.6.

Inclui-se ainda na Secção 2.7 uma breve referência ao tratamento da ordem de palavras. O presente capítulo termina com a Secção 2.8, em que é feita uma apresentação da literatura HPSG produzida até ao momento sobre a língua portuguesa.

### 2.2 Tipos e Estruturas de Traços

As entidades linguísticas encontram-se organizadas num sistema de tipos, pelo que cada objecto linguístico é de um determinado tipo. Em particular, qualquer constituinte sintáctico—seja uma palavra ou um sintagma—é uma instância do tipo *sign*.

Cada tipo está associado a um conjunto de traços, os quais são pares formados por atributos e seus valores. Os atributos definidos para o tipo *sign* são PHON (PHONOLOGY)<sup>1</sup> e SYNSEM (SYNTAX-SEMANTICS). O valor do atributo PHON guarda

---

<sup>1</sup>Em implementações computacionais, utiliza-se o atributo ORTH (por vezes chamado STEM), uma

a representação ortográfica do constituinte a que diz respeito, enquanto o valor de SYNSEM contém informação categorial e semântica.

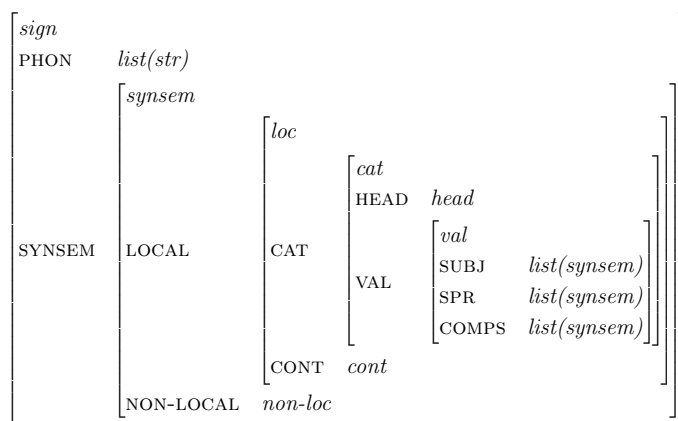
Os valores destes atributos são também objectos linguísticos, tendo por isso um tipo associado. O atributo SYNSEM toma como valores instâncias do tipo *synsem*, para as quais dois atributos são apropriados: LOCAL e NON-LOCAL. O atributo NON-LOCAL contém informação necessária para acomodar as dependências de longa distância (vd. Secção 2.6). O atributo LOCAL, por sua vez, guarda informação acerca da sintaxe e acerca da semântica, que se encontra nos valores dos atributos CAT (CATEGORY) e CONT (CONTENT), respectivamente.<sup>2</sup>

A informação semântica é discutida no Capítulo 3. Quanto à informação sintáctica, sob CAT encontram-se os traços HEAD, que codifica informação sobre a classe de palavras, e VAL (VALENCE), que guarda informação sobre a valência nos sub-traços SUBJ (SUBJECT), SPR (SPECIFIER) e COMPS (COMPLEMENTS).

Em suma, utilizando o formato MAV, as definições de adequação para os tipos *sign*, *synsem*, *loc*, *cat* e *val* são portanto as seguintes:

$$\left[ \begin{array}{l} \textit{sign} \\ \text{PHON} \quad \textit{list(str)} \\ \text{SYNSEM} \quad \textit{synsem} \end{array} \right] \left[ \begin{array}{l} \textit{synsem} \\ \text{LOCAL} \quad \textit{loc} \\ \text{NON-LOCAL} \quad \textit{non-loc} \end{array} \right] \left[ \begin{array}{l} \textit{loc} \\ \text{CAT} \quad \textit{cat} \\ \text{CONT} \quad \textit{cont} \end{array} \right] \left[ \begin{array}{l} \textit{cat} \\ \text{HEAD} \quad \textit{head} \\ \text{VAL} \quad \textit{val} \end{array} \right] \left[ \begin{array}{l} \textit{val} \\ \text{SUBJ} \quad \textit{list(synsem)} \\ \text{SPR} \quad \textit{list(synsem)} \\ \text{COMPS} \quad \textit{list(synsem)} \end{array} \right]$$

A representação gramatical de um objecto do tipo *sign* em termos de estrutura de traços toma então o seguinte aspecto:



Há mais atributos sob os traços CONT e NON-LOCAL, que são apropriados para os tipos *cont* e *non-loc*, que se encontram omitidos aqui. Estes são apresentados no Capítulo 3e na Secção 2.6, respectivamente.

vez que o que interessa é processar texto. Na literatura é comum ver-se o atributo PHON, mas por simplificação muitas vezes o valor representado é uma representação ortográfica. É este o uso que adoptamos aqui.

<sup>2</sup>Adicionalmente, no traço LOCAL encontra-se também um atributo CTXT (CONTEXT) para representar informação pragmática, que não utilizaremos nestes capítulos.

O valor do atributo PHON é uma lista de representações fonológicas (ou ortográficas).

O valor dos atributos SUBJ, SPR e COMPS é uma lista de objectos do tipo *synsem*, onde se representa informação de subcategorização para o sujeito, o especificador e os complementos. Isto significa que a informação sobre o que pode ser subcategorizado acertadamente não inclui informação fonológica, uma vez que, embora PHON seja um traço de *sign*, não é um traço apropriado para o tipo *synsem*.

Os tipos pertencem a uma hierarquia de herança que se organiza debaixo da relação “é-um”. Por exemplo, qualquer sintagma (tipo *phrase*) e qualquer palavra (tipo *word*) são também um signo (tipo *sign*) e portanto *sign* é um supertipo de *phrase* e *word*, sendo por isso os tipos *phrase* e *word* subtipos de *sign*.

A Figura 2.1 apresenta uma hierarquia para estes três tipos agora referidos e ainda para os que serão apresentados nas secções seguintes. Esta hierarquia servirá de base para a exposição do material que se segue neste capítulo.<sup>3</sup>

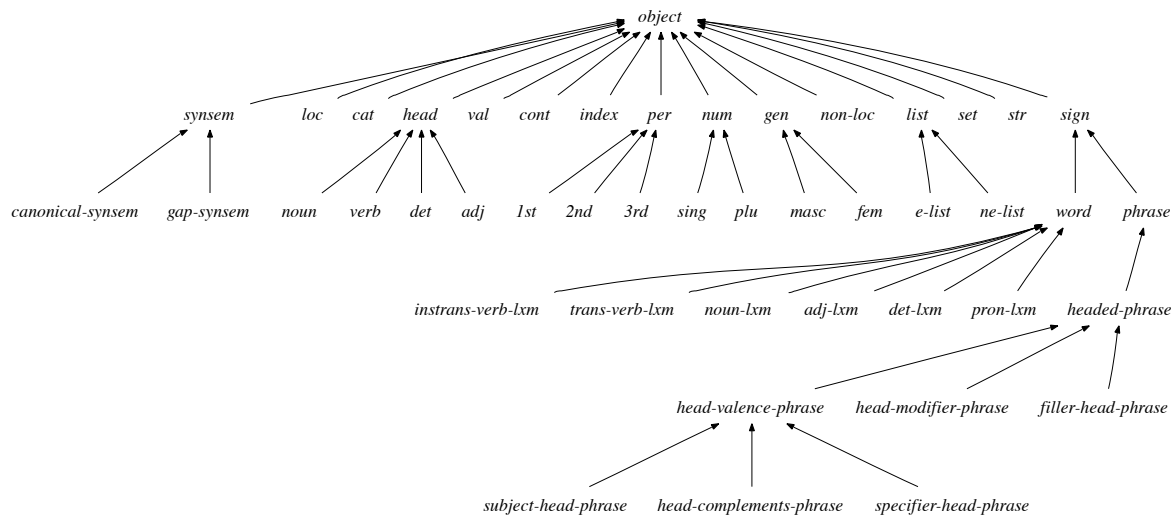


Figura 2.1: Hierarquia de tipos

Um tipo herda as restrições definidas para todos os seus supertipos. Assim, todas as instâncias de *word* e *phrase*, por exemplo, têm um traço PHON e um traço SYNSEM uma vez que estes dois traços são apropriados para o seu supertipo *sign*.

Juntamente com a definição de adequação, a hierarquia de tipos ajuda a definir os valores que são possíveis para os vários traços. Por exemplo, por os tipos *noun*, *verb*, *det* (*determiner*) e *adj* (*adjective*) serem subtipos de *head*, eles constituem valores possíveis para o atributo HEAD, o qual na definição de adequação respectiva é declarado como tomando valores do tipo *head*. Em contraponto, por exemplo, o tipo *verb* não pode ser um valor válido para o atributo CAT pois não é um subtipo de *cat*.

Há atributos que tomam listas como valores. Estas podem ser vazias (tipo *e-list* (*empty-list*)) ou conter elementos (tipo *ne-list* (*non-empty-list*)). No último caso, o

<sup>3</sup>Cabe notar que esta hierarquia se encontra simplificada relativamente a uma hierarquia mais completa para uma gramática que abranja mais fenómenos linguísticos que os abordados aqui.

primeiro elemento da lista é representado no atributo FIRST, tomando o atributo REST como valor a lista que contém os restantes elementos. As listas não vazias têm portanto a seguinte definição de adequação (em que  $\sigma$  é o tipo dos elementos da lista):

$$\left[ \begin{array}{l} ne-list(\sigma) \\ \text{FIRST } \sigma \\ \text{REST } list(\sigma) \end{array} \right]$$

Por exemplo, uma lista com “a” e “b” (com “a” à frente na lista) é uma lista  $ne-list(str)$  cujos elementos são do tipo  $str$  (*string*), representada na seguinte MAV:

$$\left[ \begin{array}{l} ne-list(str) \\ \text{FIRST } \text{“a”} \\ \text{REST } \left[ \begin{array}{l} ne-list(str) \\ \text{FIRST } \text{“b”} \\ \text{REST } e-list(str) \end{array} \right] \end{array} \right]$$

A notação  $\langle \text{“a”}, \text{“b”} \rangle$  é uma simplificação notacional equivalente. A lista vazia ( $e-list$ ) também pode ser representada por  $\langle \rangle$ .

## 2.3 Léxico e Subcategorização

Para exemplificação de um tipo lexical, em baixo encontra-se a estrutura de traços associada a um outro dos tipos da hierarquia, o *intrans-verb-lxm* (*intransitive-verb-lexeme*). Este tipo contém as restrições apropriadas para todos os verbos intransitivos.<sup>4</sup>

$$\left[ \begin{array}{l} intrans-verb-lxm \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD } verb \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \text{LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD } noun \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \rangle \\ \text{SPR } \langle \rangle \\ \text{COMPS } \langle \rangle \end{array} \right] \rangle \\ \langle \rangle \\ \langle \rangle \end{array} \right] \rangle \\ \langle \rangle \end{array} \right] \end{array} \right] \end{array} \right]$$

Esta MAV indica que este objecto é da categoria verbo (valor *verb* para o traço HEAD) e selecciona como sujeito um elemento com o valor *noun* para o seu atributo HEAD.<sup>5</sup> Para além disso, o sujeito tem de estar saturado (visto ter listas vazias para os traços de valência) isto é, tem de ser um SN completo.

<sup>4</sup>A notação ATRIBUTO1|ATRIBUTO2  $x$  é equivalente a ATRIBUTO1 [ATRIBUTO2  $x$ ]. É usual omitir a referência aos tipos quando estas são recuperáveis pelo contexto. Também é usual omitir os traços que são herdados de um supertipo se não estão sujeitos a restrições adicionais (é o caso de PHON neste exemplo).

<sup>5</sup>Tanto os nomes como os pronomes pessoais estão definidos no léxico com o valor *noun* para este atributo.

A entrada lexical para um verbo intransitivo como “correr” resulta de se associar este item lexical a esse tipo e de se preencher o valor do atributo PHON com a seguinte lista singular de *strings* < “correr” >:

$$\left[ \begin{array}{l} \textit{intrans-verb-lxm} \\ \text{PHON} \langle \text{“correr”} \rangle \end{array} \right]$$

Como é ilustrado neste exemplo, a informação de subcategorização associada a um item lexical encontra-se codificada no valor do traço VAL.

O traço SUBJ recolhe informação sobre o sujeito. Por sua vez, os atributos SPR e COMPS contêm informação sobre o especificador e os complementos de um determinado item, respectivamente.

Por exemplo, os nomes contáveis encontram-se no léxico com o requisito de se terem de combinar com um determinante, o que resultará na seguinte estrutura de traços para o seu tipo lexical:

$$\left[ \begin{array}{l} \textit{noun-lxm} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \textit{noun} \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \langle \rangle \\ \text{SPR} \langle \langle \text{LOCAL|CAT|HEAD} \textit{det} \rangle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \end{array} \right] \end{array} \right]$$

A informação sobre complementos é dada de forma análoga. Os verbos transitivos seleccionam um SN sujeito e um SN complemento, e o seu tipo lexical é o seguinte:

$$\left[ \begin{array}{l} \textit{trans-verb-lxm} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \textit{verb} \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \langle \langle \text{LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \textit{noun} \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \langle \rangle \\ \text{SPR} \langle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \end{array} \right] \rangle \rangle \\ \text{SPR} \langle \rangle \\ \text{COMPS} \langle \langle \text{LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \textit{noun} \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \langle \rangle \\ \text{SPR} \langle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \end{array} \right] \rangle \rangle \end{array} \right] \end{array} \right] \end{array} \right]$$

Os itens que não seleccionam dependentes têm os atributos dentro do valor de VAL todos com listas vazias. Para pronomes pessoais, por exemplo, o seu tipo lexical consiste na seguinte estrutura de traços:

$$\left[ \begin{array}{l} \textit{pron-lxm} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \textit{noun} \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \langle \rangle \\ \text{SPR} \langle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \end{array} \right] \end{array} \right]$$

## 2.4 Constituição e Funções Gramaticais

Qualquer construção sintáctica é um objecto linguístico que é um subtipo de *phrase*. Este tipo *phrase* herda os traços do seu supertipo *sign* e como tal possui os atributos PHON e SYNSEM. Adicionalmente, possui o atributo DAUGHTERS, que representa os constituintes sintácticos imediatos da construção em causa. Estes constituintes imediatos podem ser palavras ou outros sintagmas, pelo que o valor deste atributo DAUGHTERS é uma lista de objectos do tipo *sign*.

Na definição de adequação de *phrase*, o valor de PHON do nó mãe é a concatenação do valor de PHON dos nós filhos. Sendo a concatenação de listas representada pelo símbolo  $\oplus$ ,<sup>6</sup> a estrutura de traços de *phrase* vem:

$$\left[ \begin{array}{l} \textit{phrase} \\ \text{PHON} \quad \boxed{1} \oplus \dots \oplus \boxed{n} \\ \text{DAUGHTERS} \quad \textit{ne-list} \left\langle \left[ \begin{array}{l} \textit{sign} \\ \text{PHON} \quad \boxed{1} \end{array} \right], \dots, \left[ \begin{array}{l} \textit{sign} \\ \text{PHON} \quad \boxed{n} \end{array} \right] \right\rangle \end{array} \right]$$

A ordem dos elementos na lista de DAUGHTERS corresponde à ordem de palavras (e.g. o primeiro elemento corresponde ao filho mais à esquerda, etc.).

Cabe notar que etiquetas  $\boxed{i}$  idênticas denotam identidade entre valores. Neste caso, por exemplo, a identidade entre o primeiro operando da operação de concatenação que constrói o valor de PHON do nó mãe e o valor de PHON do filho mais à esquerda é representada pelas duas etiquetas  $\boxed{1}$ .

Um dos subtipos de *phrase* é o *headed-phrase* (vd. Figura 2.1). Na definição deste último, é introduzido o atributo HEAD-DTR, cujo valor corresponde à caracterização do núcleo do constituinte com esse tipo. O valor desse atributo HEAD-DTR é identificado com um dos elementos da lista em DAUGHTERS, consoante a ordem de palavras em construções específicas (e.g. com o primeiro elemento dessa lista se se tratar de uma construção em que o núcleo ocorre no início, à esquerda portanto).

Um dos princípios sintácticos nucleares é o Princípio do Traço do Núcleo (*Head Feature Principle*), que capta o facto de os sintagmas e os seus núcleos partilharem características categoriais. Este princípio estabelece que em qualquer constituinte que possui núcleo, i.e. que seja do tipo *headed-phrase*, o valor do atributo HEAD é idêntico ao valor do atributo com esse mesmo nome no nó filho que corresponde ao núcleo. Por exemplo, numa construção para realizar ou projectar os complementos à direita de um núcleo, o valor deste atributo é idêntico ao do filho mais à esquerda.

<sup>6</sup>Assumimos a definição habitual para a concatenação de listas de elementos de tipo  $\sigma$ :

- $e\text{-list}(\sigma) \oplus \boxed{1} \textit{list}(\sigma) = \boxed{1}$
- $\left[ \begin{array}{l} \textit{ne-list}(\sigma) \\ \text{FIRST} \quad \boxed{1} \sigma \\ \text{REST} \quad \boxed{2} \textit{list}(\sigma) \end{array} \right] \oplus \boxed{3} \textit{list}(\sigma) = \left[ \begin{array}{l} \textit{ne-list}(\sigma) \\ \text{FIRST} \quad \boxed{1} \\ \text{LAST} \quad \boxed{2} \oplus \boxed{3} \end{array} \right]$

Este princípio corresponde à seguinte restrição no tipo *headed-phrase*, que é super-tipo de todas as construções em que um dos filhos é um núcleo:

$$\begin{array}{c} \text{Princípio do Traço do Núcleo} \\ \left[ \begin{array}{l} \textit{headed-phrase} \\ \text{SYNSEM|LOCAL|CAT|HEAD } \boxed{1} \\ \text{HEAD-DTR } \left[ \begin{array}{l} \textit{sign} \\ \text{SYNSEM|LOCAL|CAT|HEAD } \boxed{1} \end{array} \right] \end{array} \right] \end{array}$$

A identidade entre os dois atributos HEAD é representada pelas duas etiquetas  $\boxed{1}$ .

Todas as construções que possuem um núcleo herdam de *headed-phrase*. Como tal, também herdam esta restrição.

Desta forma, como veremos de seguida, todo o aparato necessário para codificar regularidades relativas à estrutura sintáctica de constituição consiste na definição apropriada da hierarquia de subtipos de *phrase* e nas definições de adequação para cada um deles.

## Sujeitos

A construção em que o sujeito surge projectado à esquerda do SV é de tipo *subject-head-phrase*, que é um subtipo de *headed-phrase* com as seguintes restrições adicionais:

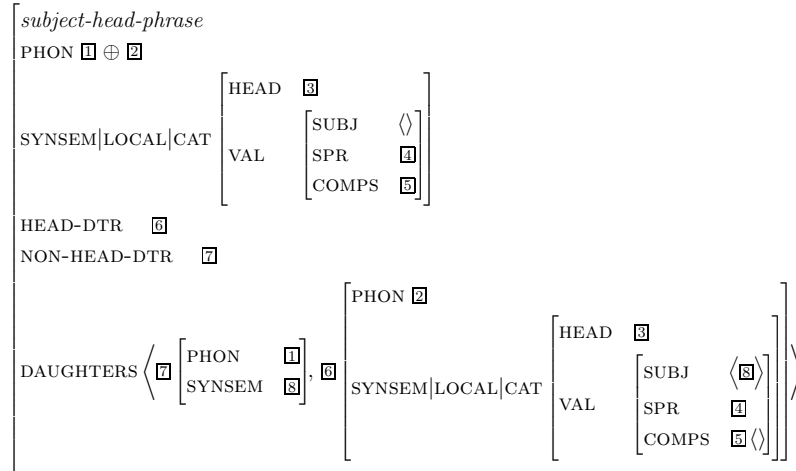
$$\left[ \begin{array}{l} \textit{subject-head-phrase} \\ \text{SYNSEM|LOCAL|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \rangle \\ \text{SPR } \boxed{1} \\ \text{COMPS } \boxed{2} \end{array} \right] \\ \text{HEAD-DTR } \boxed{3} \\ \text{NON-HEAD-DTR } \boxed{4} \\ \text{DAUGHTERS } \left\langle \boxed{4} \left[ \text{SYNSEM } \boxed{5} \right], \boxed{3} \left[ \text{SYNSEM|LOCAL|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \boxed{5} \rangle \\ \text{SPR } \boxed{1} \\ \text{COMPS } \boxed{2} \langle \rangle \end{array} \right] \right] \right\rangle \end{array} \right]$$

Esta definição de adequação estabelece que o núcleo da construção é o nó filho da direita uma vez que através de  $\boxed{3}$  o atributo HEAD-DTR está unificado com o último elemento de DAUGHTERS. O SYNSEM do nó filho da esquerda (o Sujeito) é identificado através de  $\boxed{5}$  com o elemento que está no atributo SUBJ do núcleo. Desta forma as propriedades do objecto linguístico que instanciar o Sujeito têm de ser compatíveis com as propriedades impostas pelo núcleo relativamente a essa posição subcategorizada por si. O nó mãe tem a lista vazia como valor do atributo SUBJ e os restantes traços de valência são idênticos aos do seu núcleo.<sup>7</sup>

As regras sintácticas utilizadas numa construção são os subtipos de *phrase* que não têm subtipos. Nos exemplos vistos até agora, isto só se aplica ao tipo *subject-head-phrase*.

<sup>7</sup> A restrição sobre o atributo COMPS ser adicionalmente a lista vazia é explicada na Subsecção **Complementos**.

Este é um subtipo de *headed-phrase* e de *phrase* pelo que herda as restrições impostas nestes últimos. Deste modo, todas as restrições sobre *subject-head-phrase*, incluindo as herdadas dos seus supertipos, são as seguintes:<sup>8</sup>



Tendo sido apresentadas as estruturas de traços lexicais para verbos intransitivos e para pronomes assim como a estrutura de traços para as construções intransitivas, apresenta-se na Figura 2.2 o resultado da sua unificação e da instanciação com as entradas lexicais para o exemplo “Eles correm”.

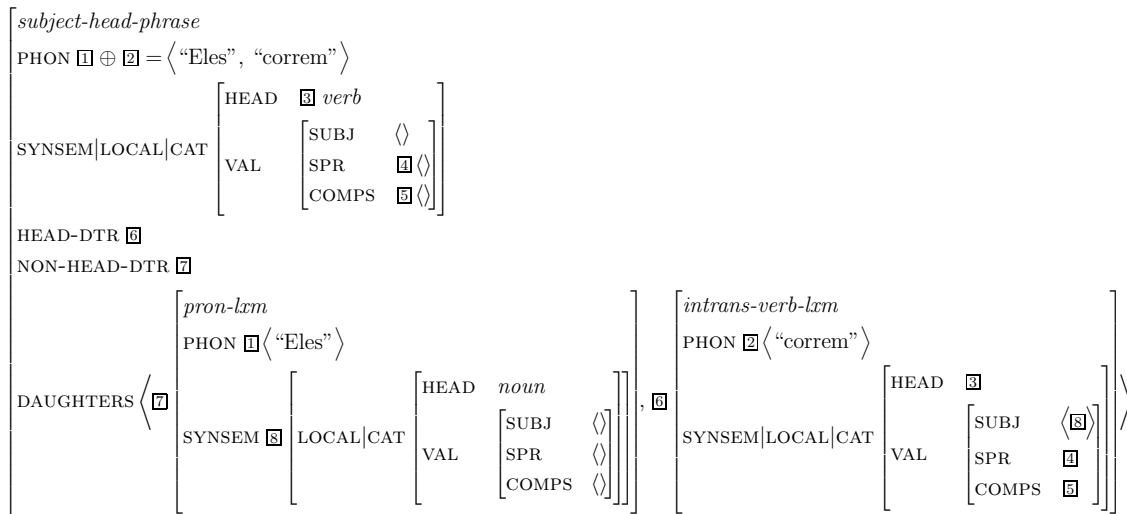


Figura 2.2: Análise para a frase “Eles correm”: representação em MAV

As relações de dominância imediata e de precedência linear encontram-se codificadas no valor do atributo DAUGHTERS. Pode dar-se a estas relações uma representação alternativa em termos diagramáticos, em que a representação da Figura 2.2 recebe a representação equivalente em árvore na Figura 2.3.

<sup>8</sup> Por agora, são omitidas aqui as restrições definidas no tipo *head-valence-phrase*, que também é um supertipo de *subject-head-phrase* e que são apresentadas mais à frente na Secção 2.6.



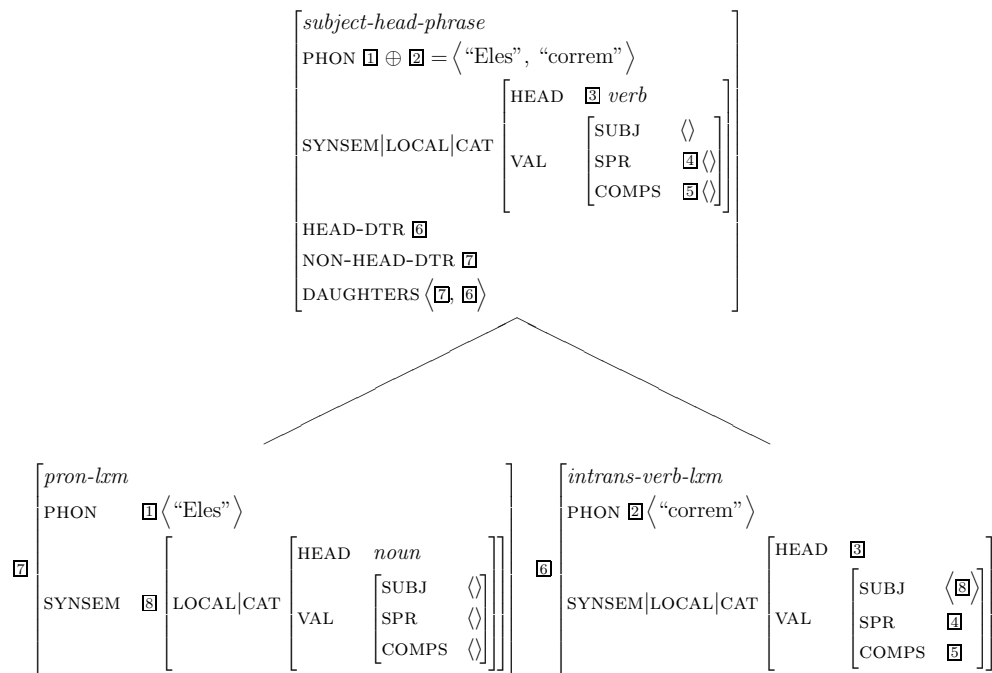
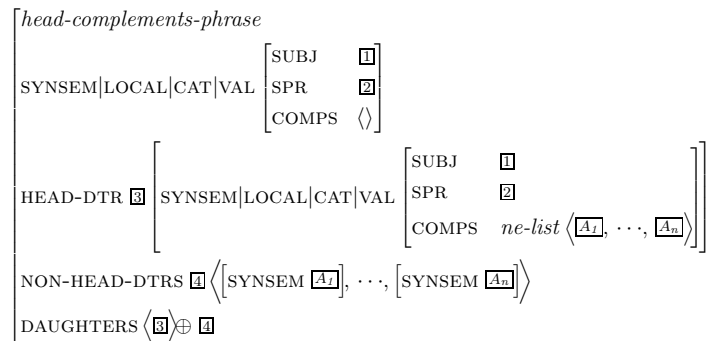


Figura 2.3: Análise para a frase “Eles correm”: representação mista em MAV e “árvore” de constituintes

## Complementos

As expressões em que os complementos do núcleo se encontram projectados são caracterizadas por um outro subtipo de *headed-phrase*, o tipo *head-complements-phrase*. Este último possui as seguintes restrições adicionais:



Esta regra projecta todos os complementos de um núcleo. Daí que o nó mãe apresente a lista vazia para o valor do atributo COMPS. O núcleo é o filho mais à esquerda desta construção. Os restantes filhos têm o seu SYNSEM unificado com os elementos do COMPS do núcleo, pela mesma ordem em que lá ocorrem.

As restantes listas de valência, nomeadamente SUBJ e SPR, são idênticas às listas de valência do núcleo.

A restrição nesta regra segundo a qual o COMPS do núcleo é uma lista não vazia (*ne-list*) impede que a regra se aplique sobre si própria, produzindo um ciclo infinito.

Cabe assinalar que a restrição no *subject-head-phrase* segundo a qual o COMPS do núcleo dessa construção é vazio garante que os complementos são projectados numa posição sintáctica inferior àquela em que o sujeito ocorre. Isto é, a construção *subject-head-phrase* pode aplicar-se tomando como HEAD-DTR uma construção do tipo *head-complements-phrase*, mas a HEAD-DTR de uma construção *head-complements-phrase* não pode ser uma instância de *subject-head-phrase*.

Note-se que a regra *head-complements-phrase*, tal como foi apresentada aqui, não permite a não realização de complementos opcionais, nem contempla as diferentes possibilidades de realização dos complementos que diferem quanto à ordem por que são realizados. Estas questões não são abordadas neste capítulo introdutório.

## Especificadores

Como decorre do que foi apresentado anteriormente, os núcleos seleccionam o seu especificador através do atributo SPR. Por sua vez, os especificadores também têm visibilidade sobre o núcleo, nomeadamente através de um outro traço, o traço SPEC, que se encontra sob HEAD. As restrições que os especificadores impõem sobre o seu núcleo, através deste traço, são sobretudo de natureza semântica e são apresentadas no Capítulo 3. Neste capítulo, vamos considerar que os determinantes se encontram na posição de especificador<sup>9</sup> pelo que se assume que o atributo SPEC é apropriado para o tipo *det*, o qual recebe a seguinte definição:

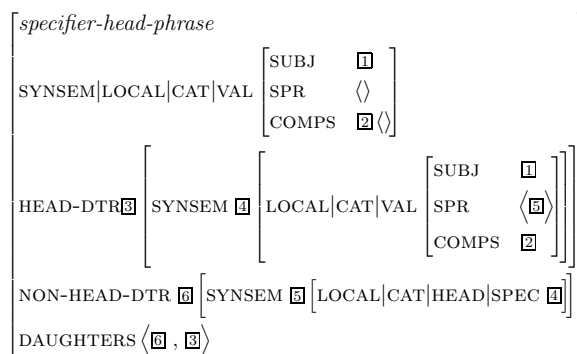
$$\left[ \begin{array}{l} det \\ SPEC \quad synsem \end{array} \right]$$

As restrições sintácticas nas entradas lexicais dos determinantes são as seguintes:

$$\left[ \begin{array}{l} det-lam \\ \\ SYNSEM|LOCAL|CAT \end{array} \left[ \begin{array}{l} HEAD \quad det \\ VAL \quad \left[ \begin{array}{l} SUBJ \quad \langle \rangle \\ SPR \quad \langle \rangle \\ COMPS \quad \langle \rangle \end{array} \right] \end{array} \right] \right]$$

Por sua vez, a construção em que os especificadores se encontram projectados é do tipo *specifier-head-phrase* e tem a definição:

<sup>9</sup> Esta é a análise adoptada em (Pollard e Sag, 1994). Numa análise alternativa, a chamada hipótese DP (*Determiner Phrase*), os determinantes são o núcleo e o nó irmão é seleccionado como seu complemento.



Os atributos SUBJ e COMPS do núcleo são percolados para o nó mãe. Esta construção também requer que o traço COMPS do núcleo tenha como valor a lista vazia, obrigando eventuais complementos do nome a serem projectados numa posição sintáctica mais baixa do que aquela em que os determinantes se realizam.

Para ilustrar a unificação das três regras sintácticas discutidas até ao momento, nomeadamente *subject-head-phrase*, *head-complements-phrase* e *specifier-head-phrase*, a Figura 2.4 apresenta a representação sintáctica para a frase exemplo “Eles viram os livros”.

## Modificadores

As estruturas de modificação vão ser apresentadas tomando os adjectivos como pretexto. Note-se porém que os dispositivos gramaticais responsáveis por dar conta das construções com modificadores são independentes da categoria sintáctica do modificador e do núcleo presentes, e que as restrições relevantes sobre os modificadores de outras categorias são semelhantes às que serão apresentadas para os adjectivos.

Os modificadores possuem um atributo MOD (MODIFIED). Este atributo encontra-se sob o traço HEAD, tem um objecto de tipo *synsem* como seu valor, e descreve as propriedades sintácticas do núcleo que uma dada categoria pode modificar.

Para os adjectivos, o atributo MOD é introduzido no tipo *adj*, o qual é o valor do atributo HEAD dos adjectivos, e constitui um subtipo de *head*:

$$\left[ \begin{array}{l} \text{adj} \\ \text{MOD} \quad \text{synsem} \end{array} \right]$$

Uma vez que os adjectivos modificam projecções nominais que não são SNs completos, as restrições para as entradas lexicais dos adjectivos sem complementos são as seguintes:<sup>10</sup>

<sup>10</sup> Por simplificação, assumimos que os adjectivos não têm sujeito. Isto pode ter de ser alterado para se abranger adjectivos em posição predicativa.

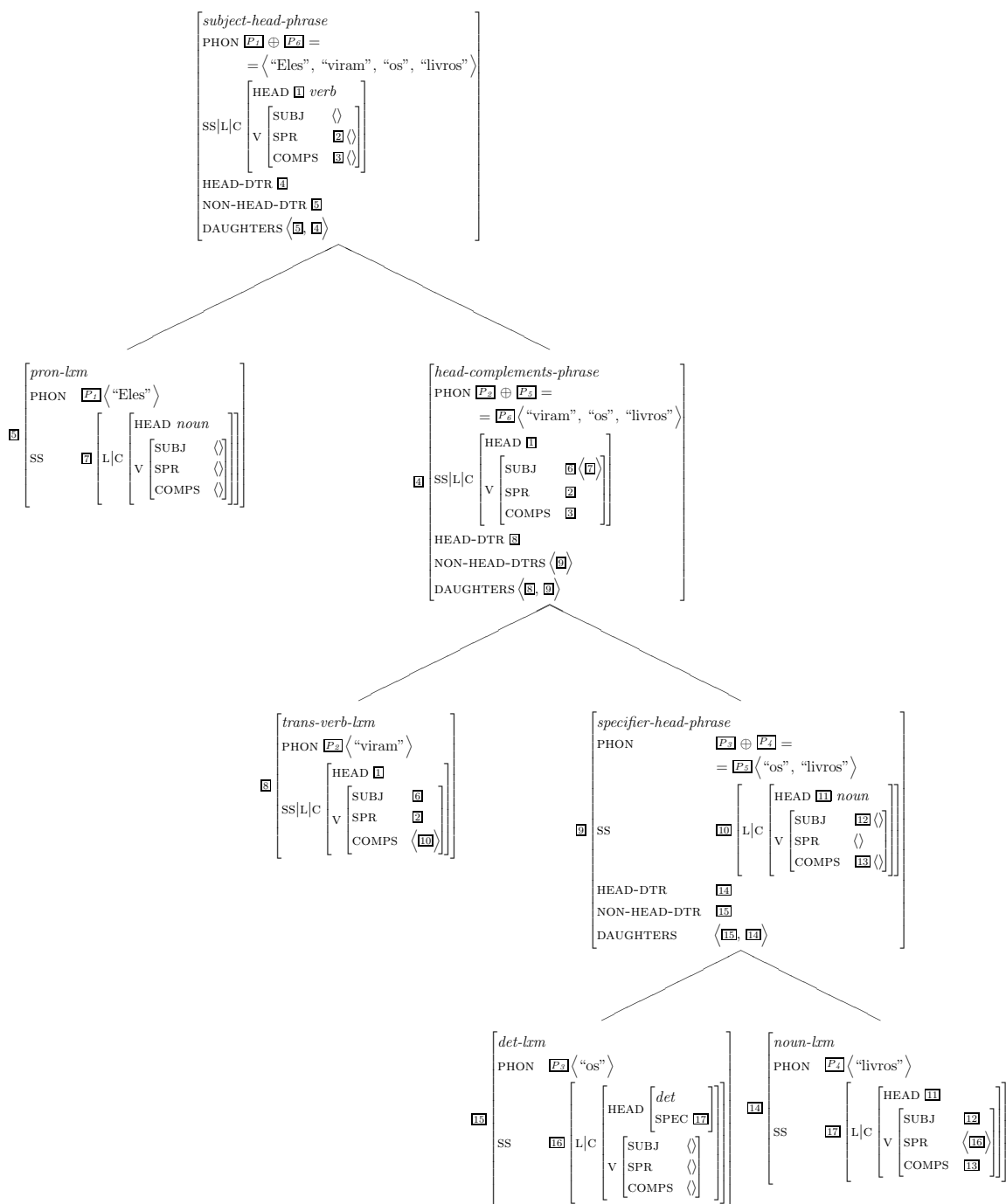


Figura 2.4: Análise sintáctica da frase "Eles viram os livros". Para economia de espaço, SS abrevia SYNSEM, L abrevia LOCAL, C abrevia CAT e V abrevia VAL.

$$\left[ \begin{array}{l} \text{adj-}l_{xm} \\ \text{SYNSEM|LOCAL|CAT} \end{array} \left[ \begin{array}{l} \text{HEAD} \left[ \begin{array}{l} \text{adj} \\ \text{MOD|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \quad \textit{noun} \\ \text{VAL|SPR} \quad \textit{ne-list} \end{array} \right] \end{array} \right] \\ \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \quad \langle \rangle \\ \text{SPR} \quad \langle \rangle \\ \text{COMPS} \quad \langle \rangle \end{array} \right] \end{array} \right] \right]$$

Os adjetivos seleccionam constituintes nominais com a posição de especificador por projectar, visto o valor do SPR do constituinte seleccionado ser uma lista não vazia (*ne-list*). Desta forma, não podem por exemplo modificar pronomes pessoais, que vêm no léxico com a lista vazia para o atributo SPR.

As construções de modificação são do tipo *head-modifier-phrase*, um subtipo de *headed-phrase*. A par dos traços e restrições que herda deste último, o tipo *head-modifier-phrase* integra restrições adicionais, ilustradas na seguinte definição de adequação para uma construção que permite projectar um adjunto à direita do seu núcleo:

$$\left[ \begin{array}{l} \textit{head-modifier-phrase} \\ \text{SYNSEM|LOCAL|CAT|VAL} \quad \boxed{1} \\ \text{HEAD-DTR} \quad \boxed{2} \left[ \text{SYNSEM} \quad \boxed{3} \left[ \text{LOCAL|CAT|VAL} \quad \boxed{1} \right] \right] \\ \text{NON-HEAD-DTR} \quad \boxed{4} \left[ \text{SYNSEM|LOCAL|CAT|HEAD|MOD} \quad \boxed{3} \right] \\ \text{DAUGHTERS} \quad \langle \boxed{2}, \boxed{4} \rangle \end{array} \right]$$

Apresentamos a análise para o SN “o carro azul pequeno” na Figura 2.5. Como se pode ver pelo exemplo, este tipo *head-modifier-phrase* permite recursão.

## 2.5 Concordância

O estabelecimento de concordância morfo-sintáctica resulta da unificação entre atributos que representam a informação relevante.

Essa informação (de pessoa, género e número) é colocada nos atributos PERSON, NUMBER,x GENDER, os quais são agrupados no tipo *index*. Este tipo é o valor do atributo AGR (AGREEMENT), que se encontra sob o atributo HEAD.

$$\left[ \begin{array}{l} \textit{noun} \\ \text{AGR} \quad \textit{index} \end{array} \right] \quad \left[ \begin{array}{l} \textit{index} \\ \text{PERSON} \quad \textit{per} \\ \text{NUMBER} \quad \textit{num} \\ \text{GENDER} \quad \textit{gen} \end{array} \right]$$

Na mini-gramática que está a ser apresentada ao longo deste capítulo só precisamos do traço AGR para nomes e pronomes, por isso definimos o traço AGR como apropriado para *noun*.

Os subtipos dos três tipos *per*, *num* e *gen* apresentados na hierarquia na Figura 2.1 são os valores possíveis para estes atributos.

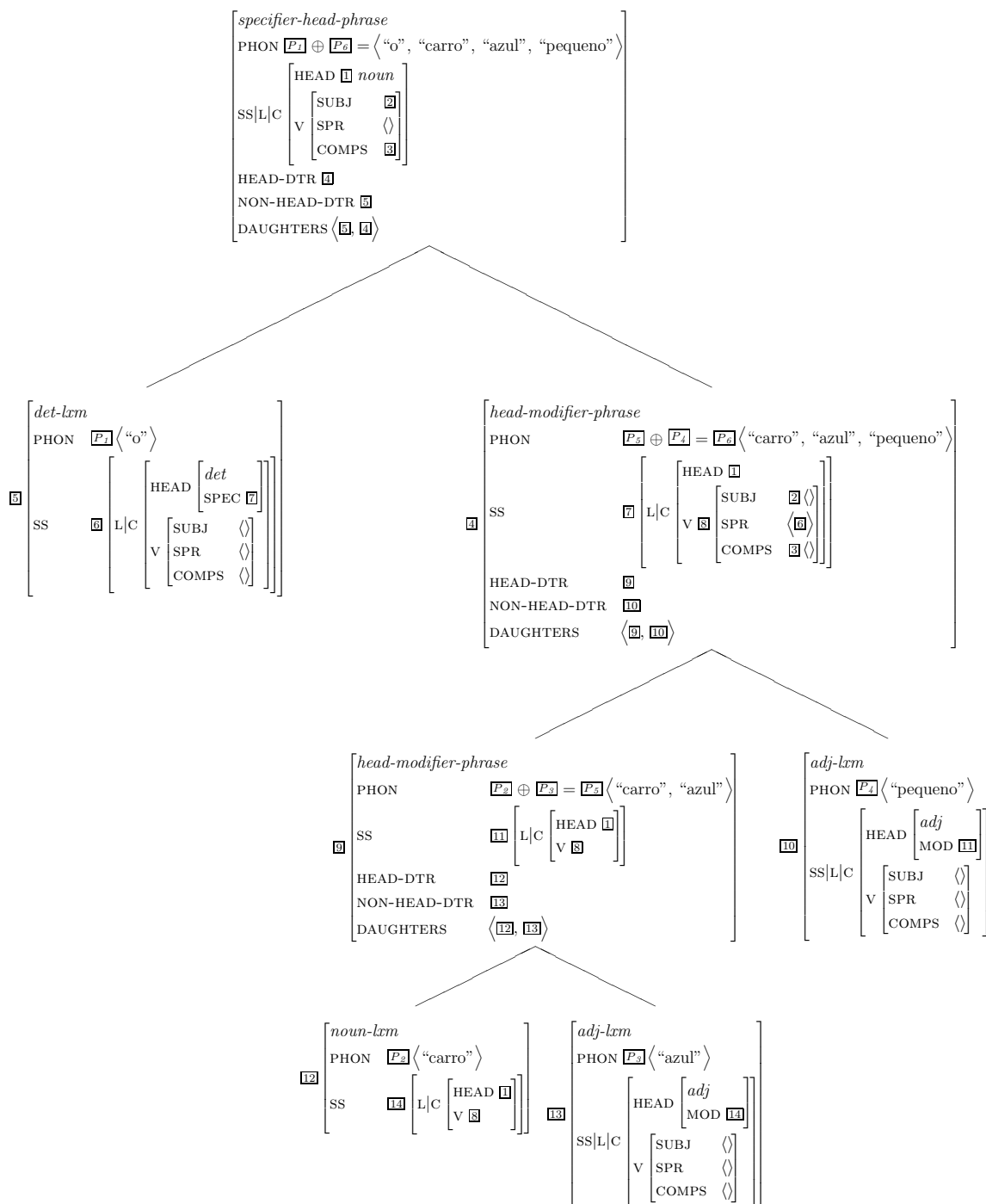


Figura 2.5: Análise para o SN “o carro azul pequeno”. Por economia de espaço, SS abrevia SYNSEM, L abrevia LOCAL, C abrevia CAT e V abrevia VAL.

A informação sobre o valor destes três atributos vem codificada nas entradas lexicais. Por exemplo, as estruturas de traços associadas aos nomes contêm informação de género. Eis um exemplo de entrada lexical, para o nome “lápiz”:

$$\left[ \begin{array}{l} \textit{noun-lxm} \\ \text{PHON} \langle \text{“lápiz”} \rangle \\ \text{SYNSEM|LOCAL|CAT|HEAD} \left[ \begin{array}{l} \textit{noun} \\ \text{AGR} \left[ \begin{array}{l} \text{PERSON} \quad \textit{3rd} \\ \text{NUMBER} \quad \textit{num} \\ \text{GENDER} \quad \textit{masc} \end{array} \right] \end{array} \right] \end{array} \right]$$

Este exemplo tem o valor do traço NUMBER sub-especificado, uma vez que o nome em questão não apresenta morfologia diferente em virtude de diferente número gramatical. De forma análoga, os nomes comuns de dois, como “pianista”, têm o traço GENDER sub-especificado, apresentando este o valor mais geral que lhe é apropriado, *gen*.

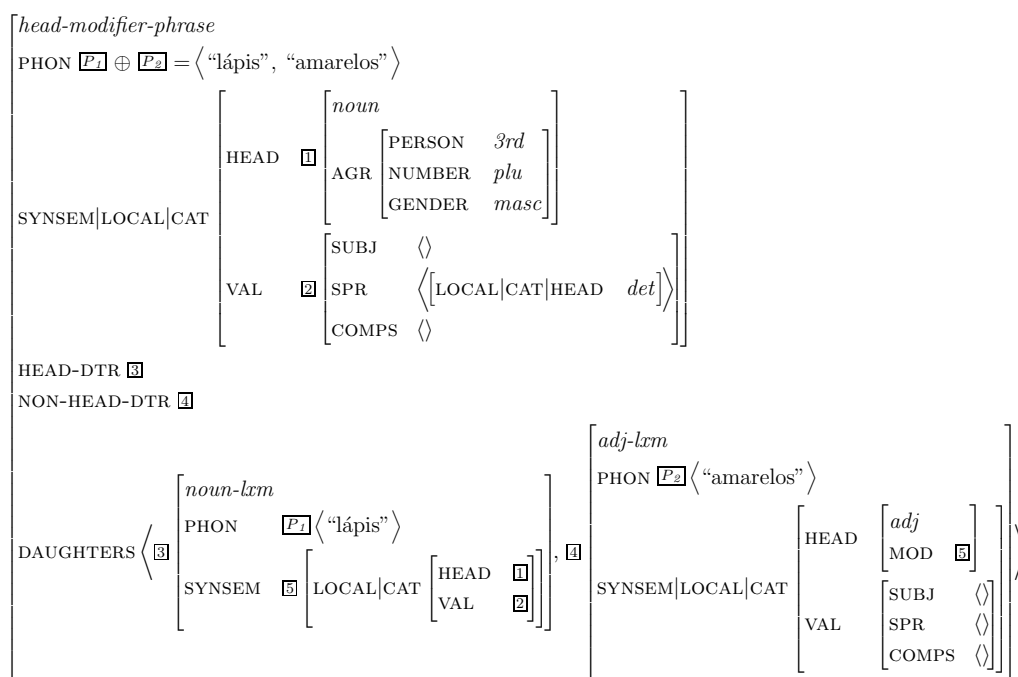
Para simplificar a exposição, neste capítulo de introdução não se apresentam as regras lexicais para a flexão pelo que se assume um léxico com todas as formas flexionadas.

Através do seu traço MOD, os adjectivos restringem os atributos apropriados sob o traço AGR da categoria que modificam. Os determinantes, por sua vez, restringem o AGR do seu nó irmão através do atributo SPEC. Tal é ilustrado, respectivamente, com os exemplos do adjectivo “amarelos” e do determinante “bastantes”:

$$\left[ \begin{array}{l} \textit{adj-lxm} \\ \text{PHON} \langle \text{“amarelos”} \rangle \\ \text{SYNSEM|LOCAL|CAT|HEAD} \left[ \begin{array}{l} \textit{adj} \\ \text{MOD|LOCAL|CAT|HEAD} \left[ \begin{array}{l} \textit{noun} \\ \text{AGR} \left[ \begin{array}{l} \text{NUMBER} \quad \textit{plu} \\ \text{GENDER} \quad \textit{masc} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$$

$$\left[ \begin{array}{l} \textit{det-lxm} \\ \text{PHON} \langle \text{“bastantes”} \rangle \\ \text{SYNSEM|LOCAL|CAT|HEAD} \left[ \begin{array}{l} \textit{det} \\ \text{SPEC|LOCAL|CAT|HEAD|AGR|NUMBER} \quad \textit{plu} \end{array} \right] \end{array} \right]$$

Para ilustrar uma estrutura em que é estabelecida uma relação de concordância, no exemplo a seguir apresenta-se a representação sintáctica para o sintagma “lápiz amarelos”:



Uma vez que o *synsem* do nome é unificado com o valor do traço MOD do adjetivo e o atributo HEAD do nome é unificado com o mesmo traço do nó mãe, este último vai ter um AGR que é a unificação das duas estruturas seguintes:

$$\left[ \begin{array}{l}
 \text{PERSON } 3rd \\
 \text{NUMBER } num \\
 \text{GENDER } masc
 \end{array} \right] \sqcap \left[ \begin{array}{l}
 \text{PERSON } per \\
 \text{NUMBER } plu \\
 \text{GENDER } masc
 \end{array} \right] = \left[ \begin{array}{l}
 \text{PERSON } 3rd \\
 \text{NUMBER } plu \\
 \text{GENDER } masc
 \end{array} \right]$$

A primeira corresponde ao AGR do nome, com o valor de número sub-especificado e portanto com o valor mais geral que lhe é apropriado, *num*; a segunda corresponde ao AGR seleccionado pelo adjetivo, com o valor de pessoa sub-especificado.

A unificação opera de forma recursiva e todos os subtraços são também unificados— assim, a unificação de *synsems* implica a unificação do atributo LOCAL|CAT|HEAD|AGR. Para além disso, a unificação de dois tipos é o tipo mais específico dos dois: por exemplo, no caso do valor de número, o resultado da unificação de *num* e *plu* é *plu* porque *plu* é um subtipo de *num*.

Cabe notar que o sintagma resultante não se poderia combinar com a forma do artigo definido “as”, uma vez que esta requer como nó irmão um constituinte com o valor *fem* para o traço NUMBER: *fem* e *masc* são incompatíveis, porque nenhum é um subtipo do outro nem têm nenhum subtipo comum na hierarquia de tipos assumida. Este sintagma pode contudo combinar-se com a forma “bastantes”, apresentada acima, pois os valores de pessoa, número e género são unificáveis.

A concordância sujeito-verbo, por sua vez, é modelada através de restrições nas formas verbais relativas ao traço AGR do sujeito. Eis um exemplo para uma forma de terceira pessoa singular:



$$\left[ \begin{array}{l} trans\text{-}verb\text{-}lcm \\ PHON \langle \text{“viu”} \rangle \\ SYNSEM|LOCAL|CAT|VAL|SUBJ \left\langle \left[ LOCAL|CAT|HEAD|AGR \left[ \begin{array}{l} PERSON \quad 3rd \\ NUMBER \quad sing \end{array} \right] \right] \right\rangle \end{array} \right]$$

## 2.6 Dependências de Longa Distância

Em alguns tipos de construções, como as orações interrogativas ou relativas, por exemplo, um constituinte pode não ocorrer na posição canônica em que ocorreria numa construção declarativa análoga não marcada. Apesar de se conformar às restantes dependências sintáticas e semânticas a que se conformaria caso ocorresse nessa posição canônica, esse constituinte pode ocorrer numa posição irrestritamente distante.

Para ilustrar a análise deste tipo de dependências ditas de longa-distância, a seguir apresenta-se a análise de uma construção de topicalização, seguindo a abordagem proposta em (Bouma *et al.*, 2001). A frase que vai servir de exemplo de trabalho será “Isso, eles viram”.

A ideia subjacente para se entender os pormenores da análise da topicalização apresentados a seguir é a de que, relativamente a um dado predicador, um dos seus argumentos não ocorre localmente na posição canônica: em vez disso, a informação quanto às restrições de subcategorização desse argumento é posta a circular pela árvore de constituição até encontrar um nó em que um dos seus nós filhos se conforme a essas restrições. Esse nó filho é o argumento topicalizado.

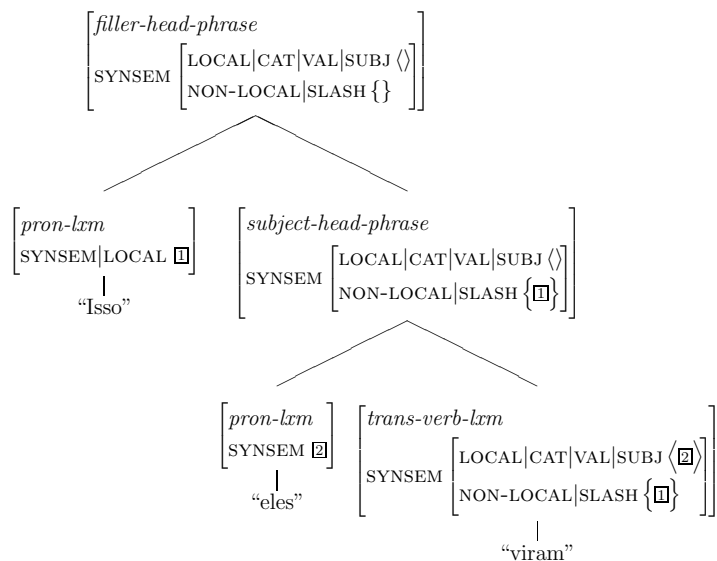
Esta abordagem envolve assim três componentes principais: (i) um tipo de construção em que o nó filho da esquerda é um constituinte topicalizado; (ii) um dispositivo de percolação da informação de subcategorização acerca do argumento topicalizado; (iii) um mecanismo que, relativamente a um predicador, remove a informação sobre o argumento topicalizado da sua lista de subcategorização e a coloca a circular ao longo da árvore de constituição.

Estes três componentes têm em comum integrarem um traço que guarda e percola a informação sobre o constituinte subcategorizado. Esse traço é SLASH, que se encontra sob o atributo NON-LOCAL e que toma como valores conjuntos de objectos do tipo *loc*:<sup>11</sup>

$$\left[ \begin{array}{l} non\text{-}loc \\ SLASH \quad set(loc) \end{array} \right]$$

A seguinte árvore ilustra o papel do atributo SLASH na percolação da informação de subcategorização entre o predicador e a periferia esquerda da frase, onde se encontra o constituinte topicalizado:

<sup>11</sup> Na análise de outras construções com dependências de longa distância, faz-se uso de mais dois outros traços sob NON-LOCAL, os traços REL e QUE, para dependências de longa distância que envolvem elementos-QU. Por limitação de espaço, estes traços não são apresentados neste capítulo introdutório.



Esta análise é melhor apreendida com a descrição de cada um dos três componentes principais das dependências de longa distância mencionados acima.

## Tópico

A construção responsável por projectar um constituinte na periferia esquerda da frase é do tipo *filler-head-phrase*, com a seguinte definição:<sup>12</sup>

$$\left[ \begin{array}{l} \textit{filler-head-phrase} \\ \text{SYNSEM} \left[ \begin{array}{l} \text{LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \quad \textit{verb} \\ \text{VAL} \quad \boxed{1} \left[ \begin{array}{l} \text{SUBJ} \langle \rangle \\ \text{SPR} \langle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \end{array} \right] \\ \text{NON-LOCAL|SLASH} \boxed{2} \end{array} \right] \\ \text{HEAD-DTR} \quad \boxed{3} \left[ \text{SYNSEM} \left[ \begin{array}{l} \text{LOCAL|CAT|VAL} \boxed{1} \\ \text{NON-LOCAL|SLASH} \boxed{2} \cup \{ \boxed{4} \} \end{array} \right] \right] \\ \text{NON-HEAD-DTR} \quad \boxed{5} \left[ \text{SYNSEM|LOCAL} \boxed{4} \right] \\ \text{DAUGHTERS} \quad \langle \boxed{5}, \boxed{3} \rangle \end{array} \right] \quad \text{tal que } \boxed{2} \cap \{ \boxed{4} \} = \{ \}$$

Note-se que o constituinte da direita, com a etiqueta  $\boxed{3}$ , tem um SLASH não vazio, fazendo com que este sintagma seja bem formado quando algures em nós inferiores existe um predador sem um dos seus argumentos a ocorrer localmente e que, por isso, tenha posto a circular informação através de SLASH. O constituinte da esquerda, por sua vez, tem o valor de LOCAL unificado com um dos elementos do SLASH do constituinte direito.

<sup>12</sup> Para efeitos da presente exposição introdutória, trata de uma simplificação relativamente ao tipo *filler-head-phrase* definido em (Bouma *et al.*, 2001).

Em virtude de o argumento subcategorizado a longa distância ter já ocorrido, a informação acerca desse elemento não é percolada para nós superiores, pelo que na construção *filler-head-phrase* o SLASH do nó mãe é um conjunto sem essa informação.

## Percolação

A segunda componente da análise das dependências de longa distância consiste na percolação do valor de SLASH.

As restrições que unificam os valor de SLASH no nó mãe de uma construção e nos nós filhos estão associadas ao tipo *head-valence-phrase*:

$$\left[ \begin{array}{l} \textit{head-valence-phrase} \\ \text{SYNSEM|NON-LOCAL|SLASH } \square \\ \text{HEAD-DTR|SYNSEM|NON-LOCAL|SLASH } \square \end{array} \right]$$

Este é um supertipo de várias das construções *subject-head-phrase*, *head-complements-phrase* e *specifier-head-phrase* já apresentadas (vd. Figura 2.1).

## Lista de Subcategorização

Resta considerar o terceiro componente da análise de dependências de longa distância, relativo à definição do tipo do predicador, e como nesta definição o valor de SLASH é devidamente instanciado.

O tipo *synsem* tem dois subtipos, nomeadamente *canonical-synsem* e *gap-synsem*.

O valor de SYNSEM para os constituintes, sejam palavras ou sintagmas, é do tipo *canonical-synsem*. Por conseguinte, o tipo *sign* tem a seguinte restrição adicional, a qual recebe a designação de Princípio de Canonicidade (*Principle of Canonicity*):

$$\left[ \begin{array}{l} \textit{sign} \\ \text{SYNSEM } \textit{canonical-synsem} \end{array} \right]$$

O tipo *canonical-synsem* é também o subtipo de *synsem* dos elementos que se encontram nas listas de subcategorização de um predicador e que, em relação a esse predicador, ocorrem localmente.

O tipo *gap-synsem*, por sua vez, é o o subtipo de *synsem* dos elementos que se encontram nas listas de subcategorização de um predicador e que ocorrem a longa distância deste.

É de notar que nenhuma regra sintáctica é capaz de projectar um elemento de SUBJ ou de COMPS que seja do tipo *gap-synsem*, uma vez que estes elementos são unificados com o SYNSEM de outro constituinte, o qual em virtude do Princípio de Canonicidade tem de ter sempre o tipo *canonical-synsem*.

Para além desta extensão em termos dos subtipos de *synsem*, os traços em que a informação de subcategorização de um predicador é codificada vão também ser alargados. É utilizado um traço adicional, o traço DEPS (DEPENDENTS), que tal como os

traços VAL|SUBJ e VAL|COMPS, também se encontra sob CAT. O seu valor é a concatenação das listas de valência SUBJ e COMPS. O valor deste atributo DEPS só é relevante para palavras pelo que o valor de DEPS é definido no tipo *word* através das seguintes restrições adicionais, as quais recebem a designação de Princípio de Realização do Argumento (*Argument Realization Principle*):

**Princípio de Realização do Argumento**

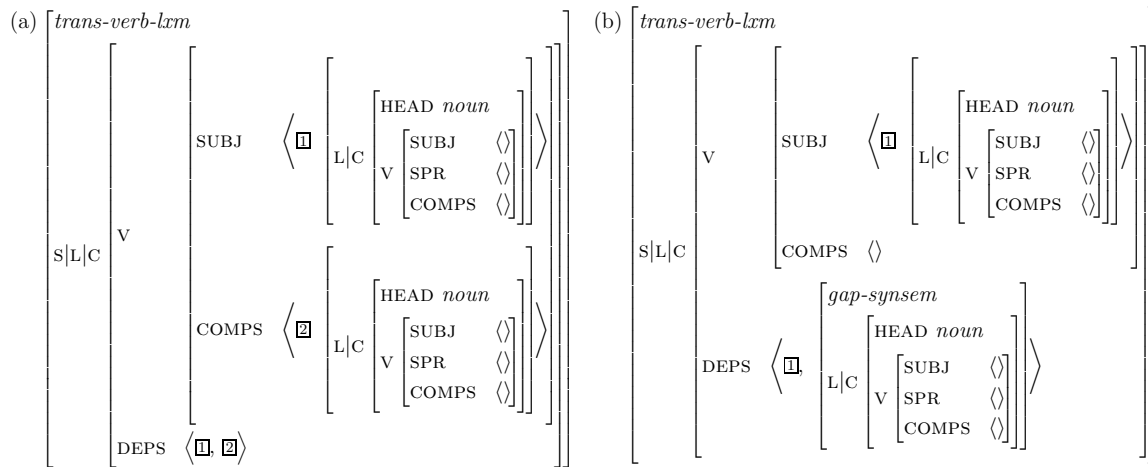
$$\left[ \begin{array}{l} \textit{word} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{VAL} \left[ \begin{array}{l} \text{SUBJ} \quad \boxed{1} \\ \text{COMPS} \quad \boxed{2} \ominus \textit{list}(\textit{gap-synsem}) \end{array} \right] \\ \text{DEPS} \quad \boxed{1} \oplus \boxed{2} \end{array} \right] \end{array} \right]$$

Sendo o seu valor uma lista de elementos de tipo *synsem*, e podendo por isso conter elementos do tipo *gap-synsem*, o valor do traço DEPS inclui a concatenação dos elementos de SUBJ com os elementos de COMPS que não são do tipo *gap-synsem*.

Para tirar partido deste alargamento, e em vista da análise das dependências de longa distância, os verbos transitivos recebem a seguinte definição:

$$\left[ \begin{array}{l} \textit{trans-verb-lxm} \\ \text{SYNSEM|LOCAL|CAT} \left[ \begin{array}{l} \text{HEAD} \quad \textit{verb} \\ \text{VAL} \quad \left[ \begin{array}{l} \text{SUBJ} \quad \langle \textit{synsem} \rangle \\ \text{COMPS} \quad \langle \rangle \end{array} \right] \\ \text{DEPS} \quad \left\langle \left[ \begin{array}{l} \textit{synsem} \\ \text{L|C} \left[ \begin{array}{l} \text{HEAD} \quad \textit{noun} \\ \text{V} \left[ \begin{array}{l} \text{SUBJ} \quad \langle \rangle \\ \text{SPR} \quad \langle \rangle \\ \text{COMPS} \quad \langle \rangle \end{array} \right] \end{array} \right] \right], \left[ \begin{array}{l} \textit{synsem} \\ \text{L|C} \left[ \begin{array}{l} \text{HEAD} \quad \textit{noun} \\ \text{V} \left[ \begin{array}{l} \text{SUBJ} \quad \langle \rangle \\ \text{SPR} \quad \langle \rangle \\ \text{COMPS} \quad \langle \rangle \end{array} \right] \end{array} \right] \right] \right\rangle \end{array} \right] \end{array} \right]$$

Nesta definição não se especifica o subtipo de *synsem* dos SNs seleccionados pelo verbo. Consequentemente, estas restrições impostas na entrada lexical de um verbo transitivo podem ser satisfeitas por duas estruturas distintas, consoante o segundo elemento de DEPS, relativo ao Objecto Directo, acontece ser resolvido para um objecto do subtipo *canonical-synsem*, em (a), ou do subtipo *gap-synsem*, em (b):



Finalmente, há que estabelecer que quando um elemento de DEPS é do tipo *gap-synsem*, como em (b), a lista no SLASH do verbo tem como elemento o LOCAL desse *gap-synsem*. Isto é obtido através da seguinte definição para o tipo *gap-synsem*:

$$\left[ \begin{array}{l} \textit{gap-synsem} \\ \text{LOCAL } \boxed{1} \\ \text{NON-LOCAL|SLASH } \{ \boxed{1} \} \end{array} \right]$$

Esta definição faz com que no nosso exemplo (b) o valor de SLASH do SN complemento contenha como membro o LOCAL desse SN.

Para além desta definição de *gap-synsem*, há que impor uma restrição adicional sobre o tipo *word*, chamada Amálgama SLASH (SLASH *Amalgamation*):<sup>13</sup>

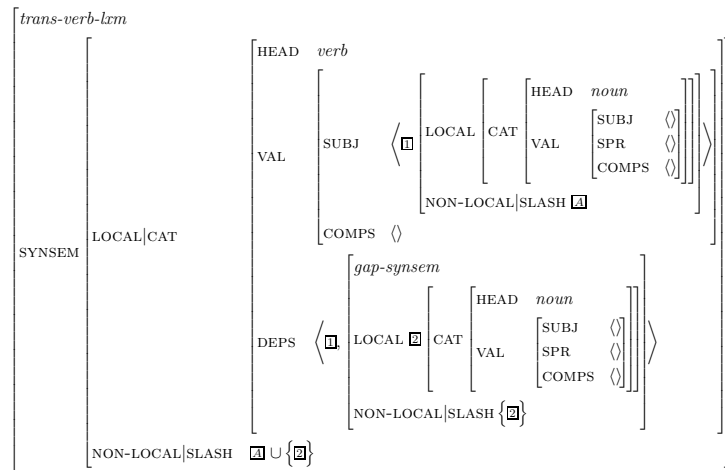
$$\begin{array}{c} \textbf{Amálgama SLASH} \\ \left[ \begin{array}{l} \textit{word} \\ \text{SYNSEM} \left[ \begin{array}{l} \text{LOCAL|CAT|DEPS} \langle [ \text{NON-LOCAL|SLASH } \boxed{S_1} ], \dots, [ \text{NON-LOCAL|SLASH } \boxed{S_n} ] \rangle \\ \text{NON-LOCAL|SLASH } \boxed{S_1} \cup \dots \cup \boxed{S_n} \end{array} \right] \end{array} \right] \\ n \geq 0 \end{array}$$

Esta restrição estabelece que o SLASH de uma palavra é a união dos traços SLASH de todos os seus dependentes sintáticos.

Desta forma, o LOCAL do complemento de tipo *gap-synsem* de um verbo é um membro do SLASH desse verbo.

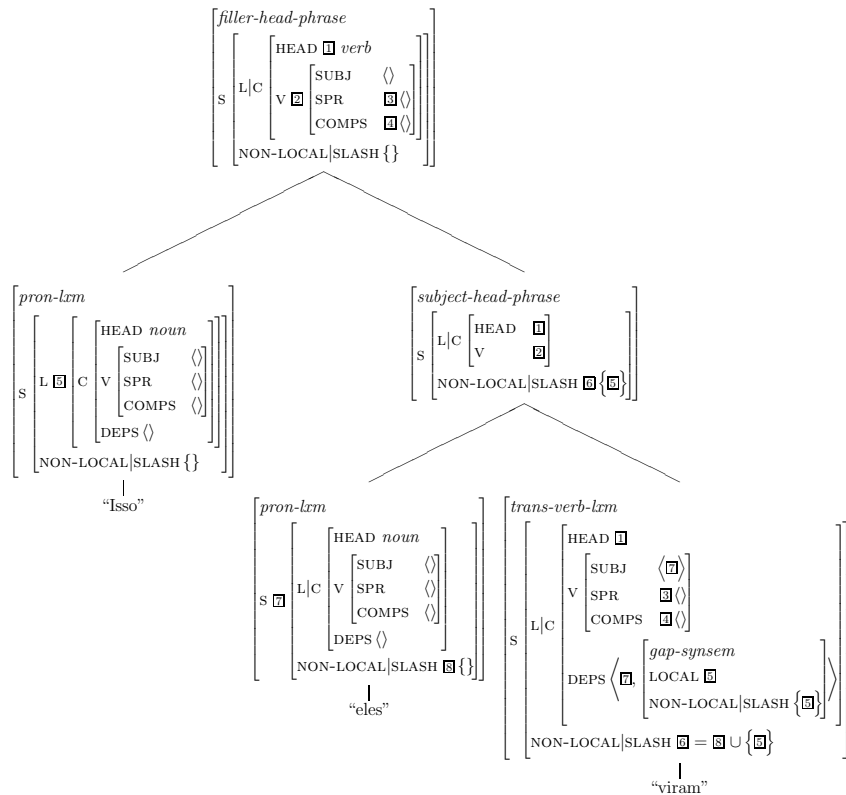
Com estas restrições acrescentadas à gramática, o caso (b) contém a seguinte informação:

<sup>13</sup>A definição que apresentamos da Amálgama SLASH é uma simplificação relativamente à apresentada em (Bouma *et al.*, 2001), uma vez que neste capítulo só estamos interessados em abranger topicalizações, dentro das dependências de longa distância.



É de notar que o valor de SLASH com a etiqueta  $\boxed{A}$  toma o conjunto vazio quando o verbo é combinado com o Sujeito pela construção *subject-head-phrase*.

Após a apresentação dos dispositivos acima, é possível retomar o exemplo de trabalho, “Isso, eles viram”, apresentando a sua análise de forma mais completa no que toca aos dispositivos chave envolvidos nas dependências de longa distância:



## 2.7 Configuracionalidade

Para efeitos da exposição neste capítulo, estabeleceu-se que a ordem superficial de palavras, representada na ordem dos elementos de PHON, é replicada pela ordem dos elementos de DAUGHTERS, através da seguinte definição do tipo *phrase*:

$$\left[ \begin{array}{l} \textit{phrase} \\ \text{PHON} \quad \square \oplus \dots \oplus \square \\ \text{DAUGHTERS} \quad \textit{ne-list} \left\langle \left[ \begin{array}{l} \textit{sign} \\ \text{PHON} \quad \square \end{array} \right], \dots, \left[ \begin{array}{l} \textit{sign} \\ \text{PHON} \quad \square \end{array} \right] \right\rangle \end{array} \right]$$

A relação entre os elementos de PHON e os de DAUGHTERS pode porém não ser esta. É possível estabelecer as relações de constituição dissociando as restrições de precedência linear das restrições de dominância. Esta estratégia é sobretudo utilizada para línguas a que alguns autores se referem como não configuracionais, as quais apresentam uma ordem de palavras bastante livre.

Por razões de espaço não é possível apresentar estas abordagens neste capítulo. O leitor interessado pode consultar as obras mais proeminentes sobre este tópico, como (Reape, 1994) e (Kathol, 1995).

## 2.8 Leituras Subsequentes

O estudo do material introdutório apresentado neste capítulo pode ser continuado com o auxílio dos dois volumes fundacionais de Pollard e Sag (1987, 1994). Para o aprofundamento da discussão e análise dos diversos fenómenos sintáticos, serve como apoio fundamental o repositório bibliográfico *online* de HPSG, com acesso livre à maior parte das publicações aí listadas, que se encontra em (HPSG, Página Web).

Têm sido desenvolvidos vários trabalhos de investigação que tomam em consideração dados empíricos do Português e em que são focados aspectos diversos da gramática do Português ou da gramática universal.

As estruturas nominais com elipse foram estudadas por Branco e Costa (2006), tendo a estrutura do sintagma nominal sido aprofundada por Costa (2007).

A coordenação recebeu atenção de Villavicencio *et al.* (2005); Chaves (2005, 2006, 2007).

Os clíticos foram alvos de análise por Crysmann (1998, 1999, 2000b,a, 2002).

A anáfora e a teoria da ligação foram exploradas por Branco (1996); Branco e Marrafa (1997, 1999); Branco (2000, 2005b).

O fenómeno da concordância negativa foi estudado por Branco e Crysmann (2001).

A acomodação de diferenças dialectais, em particular entre a variante americana e europeia do Português, foi trabalhada por Branco e Costa (2007a,b,c).

Há resultados de investigação sobre outras línguas românicas, de fácil acesso e eventualmente relevante para o Português. A compilação em Balari e Dini (1997) constitui leitura inicial interessante.





# Capítulo 3

## Semântica

### 3.1 Introdução

No estudo da semântica, é crucial associar às expressões de uma língua natural uma representação do seu significado.

Por um lado, tal representação visa caracterizar as propriedades semânticas de cada expressão e prever o seu comportamento semântico. Por exemplo, no caso de se tratar de uma expressão oracional, procura prever de que outras expressões ela é sinónima, ou de que outras expressões ela pode ser inferida, etc; ou no caso de se tratar de uma expressão sub-oracional, a representação semântica pode permitir prever que outras expressões podem veicular subconceitos do conceito que ela exprime, etc.

Por outro lado, as representações do significado constituem uma base importante para diversas aplicações computacionais como a tradução automática, as interfaces em linguagem natural com agentes artificiais ou a sumarização automática, entre várias outras.

Neste capítulo apresentam-se os dispositivos básicos da representação semântica. Cabe lembrar que na HPSG, a representação gramatical é uma descrição integrada das diversas dimensões linguísticas. A apresentação separada da componente de representação semântica no presente capítulo faz-se apenas por conveniência de exposição.

Na Secção 3.2, apresenta-se a linguagem lógica em que se escrevem as representações semânticas, para na Secção 3.3 se indicar como se constrói a representação semântica de uma expressão complexa a partir das representações semânticas dos seus componentes. Nessas secções apresentam-se muito sumariamente os aspectos básicos da abordagem adoptada na Semântica de Montague (Montague, 1974), na qual se baseiam as soluções de composição semântica utilizadas na HPSG.

De seguida, na Secção 3.4 é apresentado o formalismo de representação semântica que permite representar a ambiguidade presente em expressões das línguas naturais sem a resolver. É focado um formato em particular—*Minimal Recursion Semantics* (MRS)—, que tem a vantagem de poder ser utilizado em implementações computacionais. A sua integração com a HPSG é apresentada na Secção 3.5.

Na Secção 3.6, são mostrados mais alguns exemplos de análise semântica. Este capítulo termina com a referência a leituras subsequentes na Secção 3.7.

## 3.2 Representação do Significado

A obtenção da representação do significado de uma frase consiste na determinação de uma fórmula de uma linguagem lógica apropriada. Quando interpretada num dado modelo formal, essa fórmula apresenta o mesmo significado da frase de partida.

### Forma lógica

Por exemplo, à frase “Um homem mordeu um cão” pode associar-se a fórmula da lógica de primeira ordem  $\exists x[\text{homem}(x) \wedge \exists y[\text{cão}(y) \wedge \text{morder}(x, y)]]$ .<sup>1</sup> Esta fórmula é verdadeira, por exemplo, no seguinte modelo, em que  $H$  é o conjunto dos homens,  $C$  é o conjunto dos cães e  $M$  é a relação constituída pelos pares de indivíduos em que o indivíduo na primeira ordenada morde o indivíduo na segunda ordenada:

$$\begin{aligned} H &= \{h_1, h_2\} \\ C &= \{c_1, c_2, c_3\} \\ M &= \{(h_1, c_1), (c_1, c_2)\} \end{aligned}$$

A fórmula é verdadeira neste modelo porque descreve uma situação que se verifica neste modelo, nomeadamente a situação em que há pelo menos um indivíduo que pertence ao conjunto dos homens e pelo menos um indivíduo que pertence ao conjunto dos cães tal que acontece formarem um par que pertence à relação *morder*, com um indivíduo homem relevante na primeira ordenada.

Um dos objectivos da investigação em semântica consiste em elucidar qual a linguagem lógica mais apropriada para ser usada na representação do significado das línguas naturais. Neste contexto, veio a demonstrar-se que a lógica de primeira ordem não é suficientemente expressiva para traduzir todas as expressões das línguas naturais (Barwise e Cooper, 1981). Por exemplo, para frases que incluam expressões quantificacionais como “a maioria de”, não é possível de forma geral encontrar fórmulas equivalentes em lógica de primeira ordem.

### Quantificadores Generalizados

Esta limitação não se verifica porém para a lógica de segunda ordem. Nesta última são possíveis relações que têm outras relações como argumentos. Em particular, é possível um tipo específico de relações conhecidas por quantificadores generalizados, para as

---

<sup>1</sup> Para facilidade de exposição, as representações semânticas apresentadas neste capítulo apresentam simplificações. Por exemplo, omitimos informação sobre o tempo e o número gramaticais, como aconteceu neste exemplo.

quais se pode traduzir “a maioria de” e outras expressões da linguagem natural que não recebem tradução em lógica de primeira ordem.

Para efeitos de ilustração, a terceira coluna apresenta as fórmulas com quantificadores generalizados correspondentes às frases na primeira coluna, com o equivalente em lógica de primeira ordem na segunda coluna:

“Alguns gatos miam”	$\exists x[gato(x) \wedge miar(x)]$	$existe(x, gato(x), miar(x))$
“Todos os gatos miam”	$\forall x[gato(x) \rightarrow miar(x)]$	$todo(x, gato(x), miar(x))$
“A maioria dos gatos mia”		$maioria(x, gato(x), miar(x))$

Um quantificador generalizado tem duas relações como argumento e prende ocorrências de variáveis que servem de argumento a essas relações. A primeira relação designa-se por restritor e a segunda por âmbito ou âmbito nuclear. Por exemplo, na fórmula  $existe(x, gato(x), miar(x))$ , o quantificador generalizado  $existe$  toma como argumentos  $gato(x)$  e  $miar(x)$ , em que a variável cujas ocorrências devem ser presas é indicada por  $x$ .

A fórmula  $maioria(x, P(x), Q(x))$ , por exemplo, é verdadeira se e somente se o cardinal da intersecção entre os conjuntos denotados por  $P$  e  $Q$  é superior a metade do cardinal do conjunto denotado por  $P$ .

Quanto ao quantificador generalizado  $existe$ , a sua semântica é tal que, para a fórmula que ele integra ser verdadeira, a intersecção dos dois conjuntos correspondentes às duas relações que lhe servem de argumentos é um conjunto que contém pelo menos um elemento.<sup>2</sup>

Assim, o nosso exemplo de trabalho “Um homem mordeu um cão” recebe a seguinte representação semântica:  $existe(x, homem(x), existe(y, cão(y), morder(x, y)))$ . Neste caso, o quantificador  $existe$ , que liga as ocorrências da variável  $x$ , tem a relação de se ser homem como restritor, dada por  $homem(x)$ , e a relação de se morder um cão como âmbito nuclear, dada por  $existe(y, cão(y), morder(x, y))$ . O quantificador que liga a variável  $y$  tem como restritor a relação de se ser cão, dada por  $cão(y)$ , e como âmbito a relação de se ser mordido por  $x$ , dada por  $morder(x, y)$ .

### 3.3 Composição do Significado

Sendo o número de frases de um língua natural ilimitado, há que estabelecer um dispositivo que permite obter a representação semântica de uma qualquer expressão a partir da representação semântica das suas expressões componentes.

Esse dispositivo baseia-se no cálculo lambda e permite que a construção da representação semântica de uma expressão resulte da aplicação funcional da representação de uma expressão sua componente à representação de outra expressão sua componente.

---

<sup>2</sup>Para uma introdução em português à noção de quantificador generalizado, vd. (Branco, 2001)

## Tipos Semânticos

Assim, se a representação semântica  $\alpha$  de uma expressão componente  $A$  é do tipo semântico  $\langle b, a \rangle$  (é uma função de  $bs$  para  $as$ ) e a representação semântica  $\beta$  da outra expressão componente  $B$  é do tipo semântico  $b$ , então a representação semântica da expressão  $AB$  resultante pode ser  $\alpha(\beta)$  (em que  $\beta$  é o argumento de  $\alpha$ ) e é do tipo semântico  $a$ .

Eventuais expressões alternativas como  $\alpha(\alpha)$  e  $\beta(\alpha)$  seriam mal formadas porque não respeitam o sistema de tipos semânticos— $\beta(\alpha)$  é mal formada porque  $\beta$  não é uma função e como tal não pode ter argumentos, e  $\alpha(\alpha)$  também é mal formada porque uma expressão do tipo  $\langle b, a \rangle$  requer um argumento do tipo  $b$  e não do tipo  $\langle b, a \rangle$ .

Para efeitos de exemplificação, considere-se um nome próprio, como “Maria”. Este denota uma entidade pelo que é comum assumir-se que um nome próprio tem o tipo semântico  $e$  (do inglês *entity*, entidade).

Uma frase declarativa, por sua vez, tem o tipo semântico  $t$  (do inglês *truth value*, valor de verdade) visto denotar um dos dois valores de verdade, verdadeiro ou falso.

Estes serão então os dois tipos semânticos básicos:

- $e$ —entidade;
- $t$ —valor de verdade;

A partir deles obtém-se os restantes tipos semânticos para outras expressões, da seguinte forma:

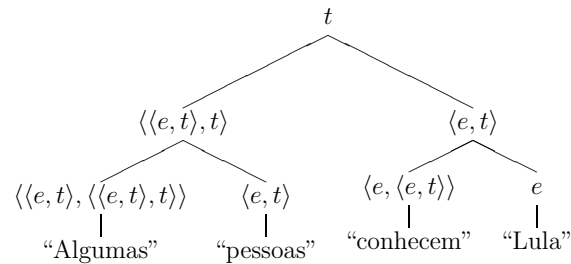
- se  $a$  é um tipo semântico e  $b$  é um tipo semântico, então  $\langle a, b \rangle$  é um tipo semântico e é uma função de  $as$  para  $bs$ .

Considere-se, por exemplo, os verbos intransitivos. Tome-se uma frase como “Maria saiu”. Dado que a frase é do tipo semântico  $t$  e que o nome próprio “Maria” é do tipo semântico  $e$ , o tipo semântico do verbo intransitivo será tal que quando é combinado com uma expressão de tipo  $e$  dê origem a uma expressão de tipo  $t$ . Semanticamente, os verbos intransitivos são portanto funções de entidades para valores de verdade, expressões do tipo semântico  $\langle e, t \rangle$ .

Através de elaborações análogas, determina-se os tipos semânticas de várias outras classes de expressões:

- verbos intransitivos e nomes comuns, como “sair”, “homem”, “gato”—tipo  $\langle e, t \rangle$
- verbos transitivos, como “morder”, “obedecer a”—tipo  $\langle e, \langle e, t \rangle \rangle$
- determinantes, como “algum”, “todo”, “bastantes”—tipo  $\langle \langle e, t \rangle, \langle \langle e, t \rangle, t \rangle \rangle$

O seguinte exemplo mostra os tipos das representações semânticas das várias expressões que integram a frase “Algumas pessoas conhecem Lula”, permitindo acompanhar o resultado da aplicação funcional entre si:



## Abstracção Lambda

Estabelecido o sistema de tipos semânticos que permite suportar a composicionalidade da representação semântica, há que apresentar a abstracção lambda, o dispositivo do cálculo lambda que permite construir e combinar as representações semânticas.

Um verbo intransitivo como “ressonar” pode ter como representação semântica uma expressão da linguagem lógica, por exemplo, o predicado *ressonar*, de tipo  $\langle e, t \rangle$ , cuja denotação é o conjunto das entidades que ressonam.

Assim sendo, pode aplicar-se funcionalmente a uma variável, de tipo  $e$ , por exemplo *ressonar*( $x$ ), obtendo-se uma fórmula, de tipo  $t$ .

Esta fórmula pode ser prefixada pelo operador  $\lambda x$ , daí resultando a expressão  $\lambda x. \textit{ressonar}(x)$ , de tipo  $\langle e, t \rangle$ . Esta expressão tem a mesma denotação de *ressonar*, e o processo da sua obtenção por via de a variável  $x$  ser ligada pelo operador lambda é conhecido como abstracção lambda.

Permita-se por agora que o nome próprio “Maria” tenha o símbolo  $m$  da linguagem lógica como representação semântica, o qual denota o mesmo indivíduo que é denotado pela expressão “Maria”.

## Conversão Lambda

Pode-se combinar funcionalmente estas duas representações semânticas. A aplicação funcional da representação semântica de “ressonar”,  $\lambda x. \textit{ressonar}(x)$ , à representação semântica de “Maria”,  $m$ , resulta em  $(\lambda x. \textit{ressonar}(x))(m)$ . Esta é uma representação semântica da frase “A Maria ressona”.

Esta representação pode ser simplificada por via da execução de um outro dispositivo do cálculo lambda chamado de conversão lambda. A conversão lambda permite eliminar o operador lambda desde que as ocorrências da variável que ele prende sejam substituídas pelo argumento da função criada pela abstracção lambda que ele operou.

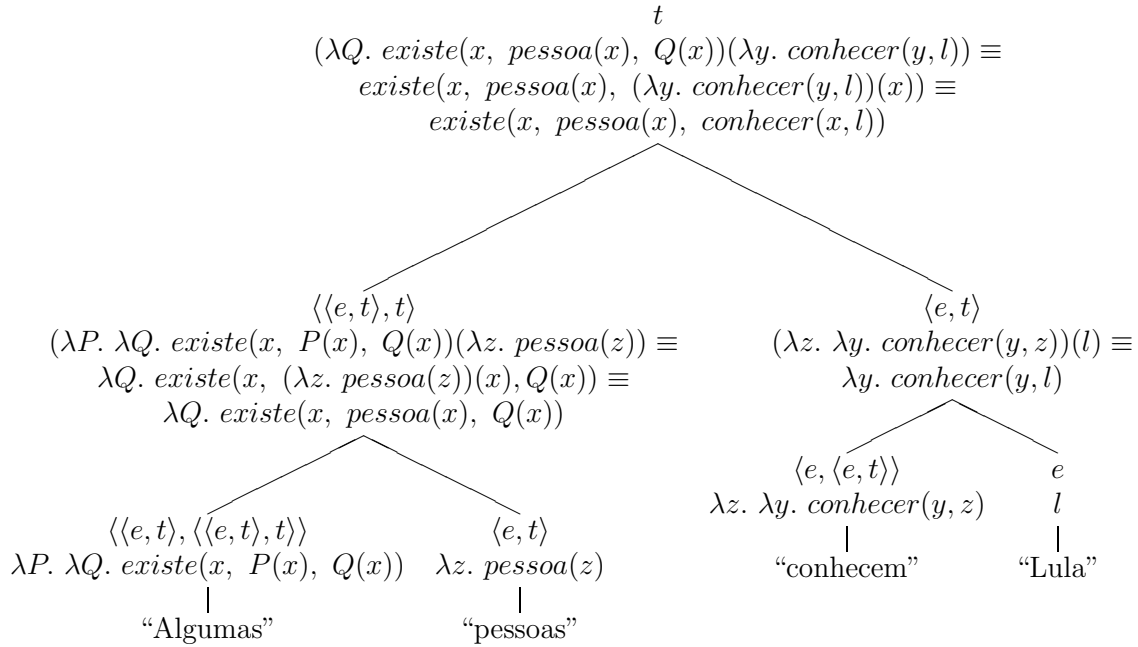
Por exemplo, no caso  $(\lambda x. \textit{ressonar}(x))(m)$  em discussão, o operador  $\lambda x.$  é eliminado e a ocorrência de  $x$  é substituída pelo argumento  $m$  da função, resultando a expressão simplificada *ressonar*( $m$ ). Esta expressão é uma proposição (do tipo  $t$ ), tal como a expressão equivalente  $(\lambda x. \textit{ressonar}(x))(m)$ .

Pode-se agora retomar o exemplo anterior, “Algumas pessoas conhecem Lula”, para o qual temos os tipos semânticos das representações semânticas das expressões componentes, mas ainda não as representações semânticas elas próprias.

## Composição Semântica

De acordo com a opção pela lógica de quantificadores generalizados como linguagem de representação do significado, a representação dessa frase pode ser  $existe(x, pessoa(x), conhecer(x, l))$ . Os dispositivos de abstracção e conversão lambda permitem agora estabelecer as representações semânticas das expressões componentes, com o tipo semântico conveniente, e de cuja aplicação funcional resulta essa representação pretendida para a frase.<sup>3</sup>

Na seguinte árvore de composição da representação semântica, cada nó exhibe o tipo semântico da expressão em causa, assim como a sua representação semântica antes e após as conversões lambda convenientes para se obter a sua simplificação:



Nos nós pré-terminais encontram-se as representações lexicais: o nome próprio “Lula” está associado ao símbolo  $l$  na sua entrada lexical, o nome “pessoa” está associado à expressão  $\lambda z. pessoa(z)$ , etc.

Para cada estrutura sintáctica mínima, é necessário estabelecer a forma de compor a representação semântica do constituinte em questão a partir da representação semântica dos nós filhos. Neste exemplo, a representação semântica do nó mãe é sempre o resultado de aplicar funcionalmente a representação semântica do nó filho à esquerda à representação semântica do nó filho à direita—o filho da direita contribui com o argumento e o filho da esquerda contribui com a função.

Pode não ser sempre esse o caso, sendo necessário especificar para cada regra sintáctica a informação acerca da forma de obter a representação semântica do constituinte resultante a partir da representação semântica dos seus componentes.

<sup>3</sup> Por convenção, vamos utilizar letras minúsculas do final do alfabeto para variáveis do tipo  $e$  e letras maiúsculas para variáveis do tipo  $\langle e, t \rangle$ .

## Tipo Semântico dos SNs

O que foi apresentado acima é apenas um esboço simplificado dos dispositivos de representação e composição semântica. Há muitas questões que, por limitação de espaço, não é possível abordar aqui. Para melhor se ilustrar estes dispositivos, porém, vai-se proceder ainda a um ajuste ao que foi apresentado até agora.

Partimos da observação de que a denotação dos dois SNs do exemplo acima é diferente. Ao SN com o nome próprio “Lula” foi dado o tipo semântico  $e$  e ao SN “algumas pessoas” foi dado o tipo semântico  $\langle\langle e, t \rangle, t\rangle$ .

Para além disto, estabeleceu-se que a denotação da frase é a aplicação da denotação do SN Sujeito à denotação do SV.

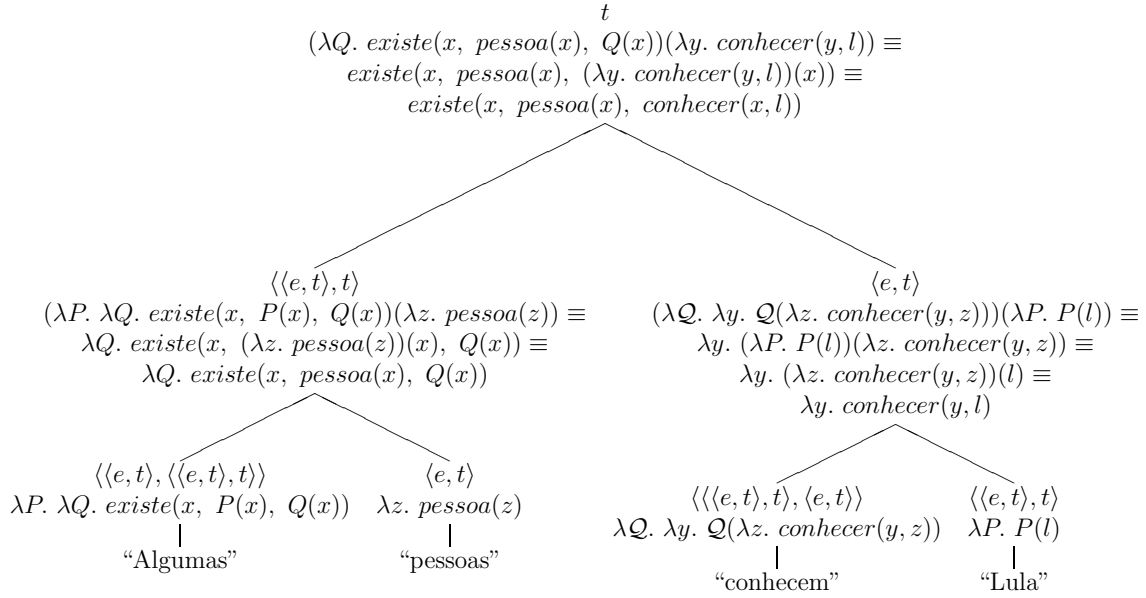
Uma questão que se coloca é o que acontece se um nome próprio ocorrer na posição de Sujeito. Caso se mantenha que os SNs com nomes próprios denotam entidades, a representação semântica da frase terá de resultar da aplicação da representação do SV à representação do SN Sujeito, uma situação em que os papéis de argumento e função se invertem em relação ao que se passa no exemplo acima, em que o Sujeito é um SN com quantificador.

Este problema dissolve-se, e uma solução uniforme é encontrada, assumindo que os SNs encabeçados por nomes próprios não são do tipo semântico  $e$ , mas também do tipo  $\langle\langle e, t \rangle, t\rangle$ , à semelhança dos SNs quantificados. Nesta perspectiva, a representação do SN “Lula” é  $\lambda P. P(l)$ , em vez de  $l$ , denotando não uma entidade mas sim uma função. Dado um conjunto de entidades, esta função deve devolver verdade se e somente se a entidade relevante, denotada por  $l$  neste exemplo, pertence a esse conjunto.

A composição da representação semântica da frase “Lula saiu” a partir das representações semânticas do SN sujeito e do SV é assim análoga à do exemplo anterior e resulta da aplicação da representação do SN Sujeito à representação do SV:

$$\begin{array}{c}
 t \\
 (\lambda P. P(l))(\lambda x. sair(x)) \equiv \\
 (\lambda x. sair(x))(l) \equiv \\
 sair(l) \\
 \swarrow \quad \searrow \\
 \langle\langle e, t \rangle, t\rangle \quad \langle e, t \rangle \\
 \lambda P. P(l) \quad \lambda x. sair(x) \\
 | \quad \quad | \\
 \text{“Lula”} \quad \text{“saiu”}
 \end{array}$$

Este ajuste requer também que agora se dê aos verbos transitivos um outro tipo semântico, nomeadamente  $\langle\langle\langle e, t \rangle, t\rangle, \langle e, t \rangle\rangle$ . A expressão lambda para o verbo “conhecer” é  $\lambda Q. \lambda x. Q(\lambda y. conhecer(x, y))$ :



### 3.4 Ambiguidade e Sub-especificação

Algumas frases das línguas naturais são ambíguas, podendo descrever mais de um tipo de situação. Considere-se, por exemplo, as seguintes frases do português:

- (1) a. Todos os homens leram um livro.  
 b. A Ana não atendeu uma chamada.

A frase em (1a) admite duas interpretações. Numa delas descreve-se a situação em que há um mesmo livro (e.g. o Código da Estrada) que é lido por todos os homens. Mas a frase também pode também ser verdadeira mesmo que não haja um mesmo livro lido por todos os homens. Nesse caso, estamos perante uma outra interpretação da mesma frase em que se descreve a situação em que diferentes pessoas leram diferentes livros (e.g. o Pedro leu Guerra e Paz, o Rui leu Os Miseráveis, ...).

A frase em (1b), por sua vez, admite também duas interpretações. Numa leitura, essa frase pode ser parafraseada por “A Ana não atendeu nenhuma chamada” ou por “A Ana não atendeu uma única chamada”. Noutra leitura, significa o mesmo que “Houve uma chamada que a Ana não atendeu”.

#### Representação das Expressões Ambíguas

As expressões lógicas usadas como representações semânticas não admitem ambiguidade pelo que uma frase ambígua se encontrará associada a mais de uma representação semântica. Cada uma dessas expressões identificará cada uma das diferentes acepções da frase ambígua em questão.

Deste modo, as frases dos exemplos acima recebem várias representações:



- (2) a. Todos os homens leram um livro.  
 $todo(x, homem(x), existe(y, livro(y), ler(x, y)))$   
 $existe(y, livro(y), todo(x, homem(x), ler(x, y)))$
- b. A Ana não atendeu uma chamada.  
 $neg(existe(x, chamada(x), atender(a, x)))$   
 $existe(x, chamada(x), neg(atender(a, x)))$

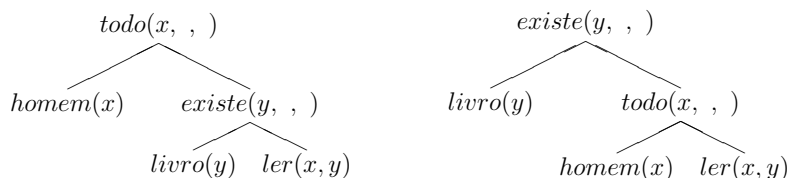
Nestes casos, as diferentes leituras devem-se a diferenças entre o âmbito dos operadores relevantes, entre os quantificadores e outros quantificadores ou a negação (aqui representada pelo predicado *neg* em vez do símbolo mais usual  $\neg$ ).

O formalismo de representação semântica pode porém ser ajustado de modo a que, mesmo para uma expressão da linguagem natural ambígua, esta receba uma única representação semântica. Nesse sentido, o formalismo é alargado para que numa sua fórmula as relações de âmbito entre os operadores permaneçam sub-especificadas. Essa fórmula tem porém de poder ser completamente especificada para cada uma das interpretações relevantes e apenas essas.

## Representação Semântica Sub-especificada

A Semântica com Recursão Mínima (MRS—*Minimal Recursion Semantics* (Copestake *et al.*, 2005)) é um formalismo desenhado para permitir a representação semântica sub-especificada nos termos agora indicados. A MRS não é propriamente uma teoria semântica, mas um formato de representação semântica: uma representação em MRS para uma determinada frase permite derivar várias fórmulas que são as representações das leituras possíveis dessa frase.

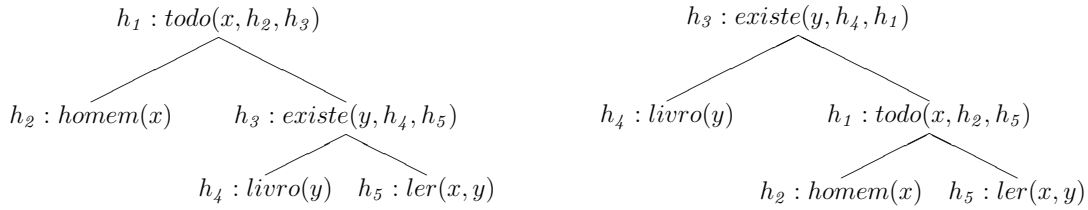
Para exemplificar a utilização da MRS na representação do significado, retome-se como exemplo as duas leituras para a frase em (2a). Pode-se associar a cada uma delas uma representação, onde as relações de âmbito entre as várias sub-expressões são postas em destaque de forma diagramática:



Por exemplo, na árvore da esquerda,  $homem(x)$  é um dos nós filhos do nó com  $todo(x, , )$ , uma vez que a relação *homem* é um argumento da relação *todo*. O facto de ser o nó filho mais à esquerda indica que é o primeiro argumento.

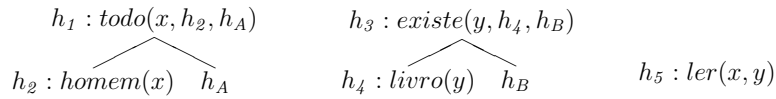
Etiquete-se agora cada um dos nós destas árvores com uma pega (*handle*) com um designador da forma  $h_n$ , em que  $n \geq 0$ :<sup>4</sup>

<sup>4</sup>Começamos porém com 1 nos primeiros exemplos, para manter os nomes das pegas nos exemplos mais à frente.



Com estes dispositivos à mão, em alternativa a esta apresentação diagramática, cada uma destas duas fórmulas pode ser estabelecida como um conjunto de pares de etiquetas e relações. Por exemplo, a fórmula da esquerda pode ser escrita como:  $h_1 : \text{todo}(x, h_2, h_3)$ ,  $h_2 : \text{homem}(x)$ ,  $h_3 : \text{existe}(x, h_4, h_5)$ ,  $h_4 : \text{livro}(x)$ ,  $h_5 : \text{ler}(x, y)$ .

No sentido de se obter uma única representação em que estas duas representações se encontram sub-especificadas, cabe notar que há fragmentos comuns às duas (aqui  $h_A$  e  $h_B$  representam variáveis sobre pegadas, que têm de ser instanciadas com pegadas existentes):



Quando se der a seguinte instanciação das variáveis  $h_A$  e  $h_B$ ,  $h_A = h_3$  e  $h_B = h_5$ , obtém-se a representação da esquerda e portanto a leitura que resulta de o quantificador universal ter âmbito sobre o quantificador existencial. Quando se der a instanciação  $h_A = h_5$  e  $h_B = h_1$  obtém-se a representação da direita, com o quantificador existencial a ter âmbito sobre o quantificador universal. É de notar que nenhuma outra leitura é uma solução possível para a especificação desta MRS em virtude dos requisitos de que (i) todas as variáveis sobre pegadas têm de ser instanciadas e (ii) todas as relações têm que estar presentes em qualquer fórmula derivada desta fórmula sub-especificada.

Por conseguinte, a representação  $h_1 : \text{todo}(x, h_2, h_A)$ ,  $h_2 : \text{homem}(x)$ ,  $h_3 : \text{existe}(y, h_4, h_B)$ ,  $h_4 : \text{livro}(y)$ ,  $h_5 : \text{ler}(x, y)$  permite assim captar as duas leituras desta frase.

Para maior simplificação descritiva, é porém costume não fazer uma distinção explícita entre nomes de pegadas e variáveis sobre pegadas. A fórmula MRS anterior pode ser representada por  $h_1 : \text{todo}(x, h_2, h_6)$ ,  $h_2 : \text{homem}(x)$ ,  $h_3 : \text{existe}(y, h_4, h_7)$ ,  $h_4 : \text{livro}(y)$ ,  $h_5 : \text{ler}(x, y)$ . As pegadas  $h_6$  e  $h_7$  (as anteriores  $h_A$  e  $h_B$ ) não etiquetam nenhuma relação, portanto são variáveis e podem ser instanciadas com outras pegadas.

De forma a dar conta de outros exemplos mais complexos envolvendo ambiguidade resultante de diferenças de âmbito entre quantificadores, a MRS recorre a restrições sobre pegadas.

A restrição entre pegadas mais importante é a relação de igualdade módulo quantificadores, representada pelo símbolo  $=_q$ , também designada como relação *geq*. Se uma pega  $h_A$  está nesta relação com outra pega  $h_B$  ( $h_A =_q h_B$ ), então  $h_A$  é uma variável que tem de ser instanciada de uma das seguintes formas em qualquer fórmula totalmente especificada:  $h_A$  é instanciada como  $h_B$ ; ou  $h_A$  é instanciada com a pega de uma

relação de quantificação  $h_C$  cuja pega do seu âmbito nuclear  $h_D$  se encontra em relação de igualdade módulo quantificadores com  $h_B$  ( $h_D =_q h_B$ ).

Com a introdução das restrições sobre pegas, uma representação do significado (i.e. uma fórmula em MRS) é agora um triplo.

O primeiro elemento deste triplo é a pega correspondente ao nó superior. A esta pega chama-se o topo global (*global top*).

O segundo elemento é um multi-conjunto de predicacões elementares. Uma predicacão elementar é uma relação etiquetada com uma pega.

O terceiro elemento é um conjunto de restrições sobre pegas.

## Especificação de Leituras

A representação semântica sub-especificada para o exemplo anterior pode ser agora escrita assim:

$$\langle h_0, \{ h_1 : \text{todo}(x, h_2, h_6), h_2 : \text{homem}(x), h_3 : \text{existe}(y, h_4, h_7), h_4 : \text{livro}(y), h_5 : \text{ler}(x, y) \}, \{ h_0 =_q h_5 \} \rangle$$

Esta representação pode ser especificada para cada uma das duas fórmulas da lógica de quantificadores generalizados que representa cada uma das duas possíveis interpretações da frase “Todos os homens leram um livro”.

A aceção codificada por  $\text{todo}(x, \text{homem}(x), \text{existe}(y, \text{livro}(y), \text{ler}(x, y)))$ , obtém-se com a seguinte instanciação das pegas:  $h_0 = h_1$ ,  $h_6 = h_3$ ,  $h_7 = h_5$ .

A leitura dada por  $\text{existe}(y, \text{livro}(y), \text{todo}(x, \text{homem}(x), \text{ler}(x, y)))$ , por sua vez, obtém-se com a seguinte instanciação:  $h_0 = h_3$ ,  $h_7 = h_1$ ,  $h_6 = h_5$ .

Com estes dispositivos de representação disponíveis, é possível estabelecer a representação sub-especificada para exemplos com negação, como o apresentado em (2b), “A Ana não atendeu uma chamada”:

$$\langle h_0, \{ h_1 : \text{neg}(h_2), h_3 : \text{existe}(x, h_4, h_5), h_4 : \text{chamada}(x), h_6 : \text{atender}(a, x) \}, \{ h_0 =_q h_1, h_2 =_q h_6 \} \rangle$$

Numa leitura, a negação toma âmbito sobre todo o resto da fórmula, enquanto na outra leitura, só sobre a relação *atender*.

A representação da primeira leitura resulta da seguinte especificação:  $h_0 = h_1$ ,  $h_2 = h_3$ ,  $h_5 = h_6$ .

A representação da segunda leitura de:  $h_0 = h_3$ ,  $h_5 = h_1$ ,  $h_2 = h_6$ .

Um último ajuste ao formato das representações em MRS apresentado aqui é motivado por exemplos como o seguinte:

- (3) Todos os filhos de um peixe nadam.  
 $\text{todo}(x, \text{existe}(y, \text{peixe}(y), \text{filho}(x, y)), \text{nadar}(x))$   
 $\text{existe}(y, \text{peixe}(y), \text{todo}(x, \text{filho}(x, y), \text{nadar}(x)))$

A primeira leitura é aquela em que cada um que é filho de peixe nada. A segunda diz-nos que há pelo menos um peixe tal que todos os seus filhos nadam.

A diferença entre as duas fórmulas que lhes correspondem é que na primeira o quantificador existencial deve aparecer no restritor do quantificador universal (com a relação *filho* a ocorrer no âmbito nuclear do quantificador existencial), mas na segunda é o quantificador universal que ocorre no âmbito nuclear do quantificador existencial. Em ambos os casos, a relação *filho* ocorre dentro do restritor de *todo*, mas pode não ser directamente, já que na primeira fórmula a relação *filho* ocorre dentro do restritor de *existe*, o qual ocorre dentro do restritor de *todo*.

Este exemplo mostra que, em vez de instanciar a pega da posição de restritor de um quantificador com a pega do material que ocorre nesta posição, é necessário estabelecer uma relação de igualdade módulo quantificadores entre ambas. Portanto, numa representação sub-especificada, todas as relações de quantificadores estão ligadas ao seu restritor por meio de uma relação *peq* entre as pegas em questão.

Concretizando para o exemplo em apreço, tem-se:

$$\langle h_0, \{ h_1 : \text{todo}(x, h_2, h_3), h_4 : \text{filho}(x, y), h_5 : \text{existe}(y, h_6, h_7), h_8 : \text{peixe}(y), h_9 : \text{nadar}(x) \}, \{ h_0 =_q h_9, h_2 =_q h_4, h_6 =_q h_8 \} \rangle$$

A primeira leitura corresponde à solução em que  $h_0 = h_1$ ,  $h_2 = h_5$ ,  $h_6 = h_8$ ,  $h_7 = h_4$ ,  $h_3 = h_9$ .

A segunda leitura obtém-se com  $h_0 = h_5$ ,  $h_6 = h_8$ ,  $h_7 = h_1$ ,  $h_2 = h_4$ ,  $h_3 = h_9$ .

Não há mais leituras possíveis. A relação correspondente ao verbo, *nadar*, não pode aparecer no restritor de nenhum quantificador, uma vez que a sua pega ( $h_9$ ) não está numa relação *peq* com nenhuma das pegas dos restritores dos quantificadores ( $h_2$  e  $h_6$ ) e porque não é uma relação de quantificador (só quantificadores podem intervir). A relação correspondente ao verbo principal de uma frase estará sempre no âmbito nuclear do quantificador com âmbito mais curto. A pega da relação *peixe* ( $h_8$ ) também não pode ser identificada com a pega do restritor de *todo* ( $h_2$ ), uma vez que não estão numa relação *peq* e *peixe* não é um quantificador.

É de notar que a solução  $\text{todo}(x, \text{filho}(x, y), \text{existe}(y, \text{peixe}(x), \text{nadar}(x)))$  é bloqueada simplesmente porque a variável  $y$  ocorre livre (fora do âmbito do quantificador *existe*, que a prende) e, como assinalámos, as soluções aceitáveis não podem ter variáveis livres.

### 3.5 Representações Semânticas em HPSG

É possível utilizar o formalismo de descrição da HPSG, baseado em estruturas de traços, para codificar representações semânticas, e em particular representações no formato MRS.

## Representação com Estruturas de Traços

Toda a informação semântica encontra-se sob o atributo SYNSEM|LOCAL|CONT, o qual contém os seguintes traços:

$$\left[ \begin{array}{l} \text{SYNSEM|LOCAL|CONT} \end{array} \left[ \begin{array}{l} \text{mrs} \\ \text{HOOK} \left[ \begin{array}{ll} \text{LTOP} & \text{handle} \\ \text{INDEX} & \text{index} \end{array} \right] \\ \text{RELS} & \text{set}(\text{relation}) \\ \text{HCONS} & \text{set}(\text{handle-constraint}) \end{array} \right] \right]$$

O atributo HOOK agrupa os atributos LTOP (LOCAL TOP) e INDEX, utilizados para a composição das representações semânticas.

O atributo LTOP guarda o valor do topo local. O topo local de uma frase matriz é o topo global de toda a representação semântica dessa frase, enquanto o topo local de constituintes mais pequenos é apenas a pega da relação com maior âmbito na representação desse constituinte.

O atributo INDEX, por sua vez, toma como valores instâncias do tipo *index*, para variáveis sobre entidades, e é utilizado também na composição da semântica.

O traço RELS (RELATIONS) toma como valor um conjunto de predicções elementares, sendo os seus elementos do tipo *relation*.

Por último, o traço HCONS (HANDLE CONSTRAINTS) é onde as restrições sobre as pegas são representadas, as quais são um conjunto de instâncias do tipo *handle-constraint*.

Considere-se como exemplo a frase “Algumas pessoas saíram”. A sua representação semântica é a proposição  $\text{existe}(x, \text{pessoa}(x), \text{sair}(x))$ , a qual no formato MRS se escreve:

$$\langle h_0, \{ h_1 : \text{existe}(x, h_2, h_3), h_4 : \text{pessoa}(x), h_5 : \text{sair}(x) \}, \{ h_0 =_q h_5, h_2 =_q h_4 \} \rangle$$

Em termos de estruturas de traços, esta fórmula MRS deixa-se escrever da seguinte forma:

$$\left[ \begin{array}{l} \text{HOOK|LTOP} \left[ \begin{array}{l} \boxed{h_0} \text{ handle} \end{array} \right] \\ \text{RELS} \left\{ \left[ \begin{array}{l} \text{quantifier-relation} \\ \text{LBL} \quad \boxed{h_1} \text{ handle} \\ \text{PRED} \quad \text{existe} \\ \text{ARG0} \quad \boxed{x} \text{ index} \\ \text{RSTR} \quad \boxed{h_2} \text{ handle} \\ \text{BODY} \quad \boxed{h_3} \text{ handle} \end{array} \right], \left[ \begin{array}{l} \text{relation} \\ \text{LBL} \quad \boxed{h_4} \text{ handle} \\ \text{PRED} \quad \text{pessoa} \\ \text{ARG0} \quad \boxed{x} \end{array} \right], \left[ \begin{array}{l} \text{relation} \\ \text{LBL} \quad \boxed{h_5} \\ \text{PRED} \quad \text{sair} \\ \text{ARG0} \quad \boxed{x} \end{array} \right] \right\} \\ \text{HCONS} \left\{ \left[ \begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{h_0} \\ \text{LARG} \quad \boxed{h_5} \end{array} \right], \left[ \begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{h_2} \\ \text{LARG} \quad \boxed{h_4} \end{array} \right] \right\} \end{array} \right]$$

Os objectos do tipo *relation* têm os atributos LBL (LABEL), PRED (PREDICATE) e ARG0 (ARGUMENT 0). O traço LBL representa a pega da relação em questão, PRED o seu nome, e ARG0 é o primeiro argumento.

O tipo *quantifier-relation* é um subtipo de *relation*. É o tipo dos quantificadores e contém adicionalmente os atributos RSTR (RESTRICTOR), para a pega do restritor do quantificador, e BODY, para o seu âmbito nuclear.

As relações  $=_q$  são do tipo *qeq*, o qual é um subtipo de *handle-constraint*,<sup>5</sup> para o qual são apropriados os traços HARG (HIGH ARGUMENT) e LARG (LOW ARGUMENT). O primeiro representa o argumento esquerdo destas relações, i.e. a variável sobre pegas a instanciar, e LARG o segundo argumento com a pega instanciadora.

Para se obter esta representação composicionalmente para o nó frase, há que associar às regras sintácticas dos constituintes (vd. Capítulo 2) restrições que controlam a composição da representação semântica desses constituintes. De igual modo, há que associar representações semânticas às entradas lexicais.

## Entradas Lexicais

Para este nosso exemplo de trabalho, as regras relevantes são a *specifier-head-phrase* e a *subject-head-phrase*. Também é preciso especificar o valor semântico de cada item lexical, nomeadamente “pessoas”, “saíram” e “algumas”.

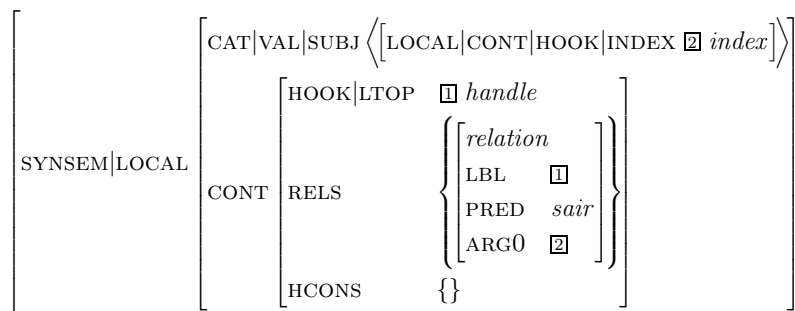
Para além das restrições que definem as suas propriedades sintácticas, aqui omitidas para efeito de simplificação da exposição, a entrada lexical de um nome como “pessoa” tem as seguintes restrições:

$$\left[ \begin{array}{l} \text{SYNSEM|LOCAL|CONT} \\ \text{HOOK} \\ \text{RELS} \\ \text{HCONS} \end{array} \left[ \begin{array}{l} \left[ \begin{array}{l} \text{LTOP} \quad \boxed{1} \textit{ handle} \\ \text{INDEX} \quad \boxed{2} \textit{ index} \end{array} \right] \\ \left\{ \begin{array}{l} \textit{ relation} \\ \text{LBL} \quad \boxed{1} \\ \text{PRED} \quad \textit{ pessoa} \\ \text{ARG0} \quad \boxed{2} \end{array} \right\} \\ \{ \} \end{array} \right] \right]$$

A este nível lexical, não é necessário introduzir nenhuma restrição sobre pegas, pelo que o atributo HCONS tem aqui um conjunto vazio como valor. Nos atributos LTOP e INDEX representa-se o valor da pega para a relação *pessoa* e a variável que lhe serve de argumento, daí eles estarem unificados com os traços relevantes que estão dentro da estrutura de traços para a relação.

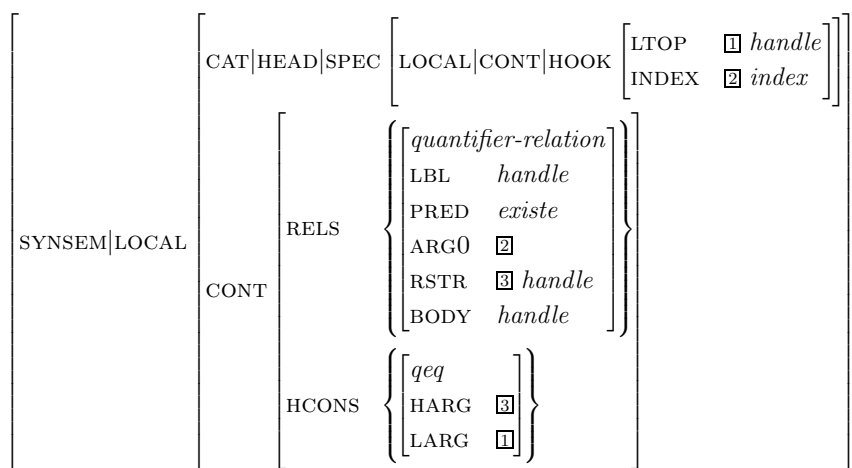
Para um verbo como “sair”, por sua vez, a entrada lexical inclui a seguinte informação:

<sup>5</sup>Há mais tipos de restrições sobre pegas, mas aqui utilizamos apenas a relação de igualdade módulo quantificadores (*qeq*).



O ARG0 da relação é identificado com o INDEX do Sujeito subcategorizado pelo verbo, de forma a se ter a mesma variável como argumento da relação *pessoa* e da relação *sair*. Mais uma vez, torna-se a pega da relação visível no traço LTOP.

Finalmente, as representações lexicais dos determinantes têm restrições análogas às do seguinte exemplo para o determinante “algun”:



O traço SPEC permite aos determinantes aceder ao seu nó irmão uma vez que a regra que projecta especificadores unifica o SPEC do especificador com o SYNSEM do núcleo irmão (vd. Secção 2.4). Neste exemplo, há acesso ao INDEX do nome através da partilha assinalada por  $\boxed{2}$ , onde se encontra a variável a ser presa por este quantificador. Há também acesso à pega que está no seu traço LTOP através da partilha assinalada por  $\boxed{1}$ , de forma a colocá-la numa relação *qeq* com a pega do restritor do quantificador.

## Regras Sintácticas e Composição Semântica

No que toca às regras sintácticas para a formação de constituintes, em geral, a composição da representação semântica resulta da união das representações componentes: o valor de RELS e HCONS do nó mãe é a união dos valores que esses atributos têm nos nós filhos.

Pode utilizar-se ainda o atributo C-CONT (CONSTRUCTIONAL CONTENT). Aqui é incluída a informação semântica que é específica da própria construção e não resulta de

nenhum dos constituintes, e que contribui para a representação semântica da construção também por via da união com as representações dos constituintes.

As construções discutidas até agora têm portanto as seguintes restrições:

$$\left[ \begin{array}{l} \textit{headed-phrase} \\ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{A} \cup \boxed{B_1} \cup \boxed{B_n} \\ \text{HCONS} \quad \boxed{C} \cup \boxed{D_1} \cup \boxed{D_n} \end{array} \right] \\ \text{C-CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{A} \\ \text{HCONS} \quad \boxed{C} \end{array} \right] \\ \text{DAUGHTERS} \left\langle \left[ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{B_1} \\ \text{HCONS} \quad \boxed{D_1} \end{array} \right] \right], \dots, \left[ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{B_n} \\ \text{HCONS} \quad \boxed{D_n} \end{array} \right] \right] \right\rangle \end{array} \right]$$

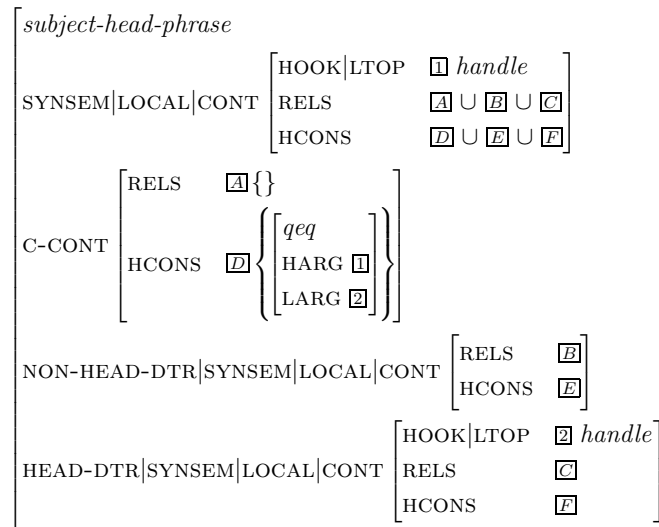
O atributo RELS dentro de C-CONT inclui as relações contribuídas especificamente pela construção, e o atributo HCONS as restrições sobre pegadas associadas a essas relações.

Considerando agora a regra para projectar especificadores, cabe notar que aqui é necessário percolar o INDEX do núcleo para o nó SN de topo, através de  $\boxed{1}$ , uma vez que através do seu quadro de subcategorização o verbo não tem acesso directo ao INDEX do nome mas apenas ao do SN. Esta regra não acrescenta nenhuma informação semântica pelo que os atributos sob C-CONT têm conjuntos vazios como valores:

$$\left[ \begin{array}{l} \textit{specifier-head-phrase} \\ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|INDEX} \quad \boxed{1} \textit{ index} \\ \text{RELS} \quad \boxed{A} \cup \boxed{B} \cup \boxed{C} \\ \text{HCONS} \quad \boxed{D} \cup \boxed{E} \cup \boxed{F} \end{array} \right] \\ \text{C-CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{A} \{\} \\ \text{HCONS} \quad \boxed{D} \{\} \end{array} \right] \\ \text{NON-HEAD-DTR|SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{B} \\ \text{HCONS} \quad \boxed{E} \end{array} \right] \\ \text{HEAD-DTR|SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|INDEX} \quad \boxed{1} \\ \text{RELS} \quad \boxed{C} \\ \text{HCONS} \quad \boxed{F} \end{array} \right] \end{array} \right]$$

Por último, nas construções responsáveis por projectar o Sujeito, os valores de RELS e HCONS obtêm-se de forma análoga. Uma vez que estas construções dão origem a nós raiz, o seu LTOP é o topo global. Por isso, é preciso acrescentar aqui uma restrição *req* entre o topo global e a pega da relação do verbo, que é o LTOP do núcleo, através de  $\boxed{D}$ :





O nosso exemplo de trabalho “Algumas pessoas saíram” tem a representação semântica  $existe(x, pessoa(x), sair(x))$  na lógica de quantificadores generalizados, que se deixa codificar em MRS da seguinte forma:

$$\langle h_0, \{ h_1 : existe(x, h_2, h_3), h_4 : pessoa(x), h_5 : sair(x) \}, \{ h_0 =_q h_5, h_2 =_q h_4 \} \rangle$$

Em resultado dos dispositivos acabados de apresentar, esta representação é transposta para o formalismo de estrutura de traços como se apresenta na Figura 3.1.

## 3.6 Outros Exemplos

Considerem-se agora alguns outros exemplos de frases ambíguas e suas representações semânticas.

### Ambiguidade com Âmbito de Quantificadores

Considere-se de novo a frase em (2a) “Todos os homens leram um livro”. Para se analisar uma frase com um verbo transitivo como esta, há que estabelecer o modo de compor a representação semântica das construções em que os complementos são projectados. As restrições para estas construções são as seguintes:

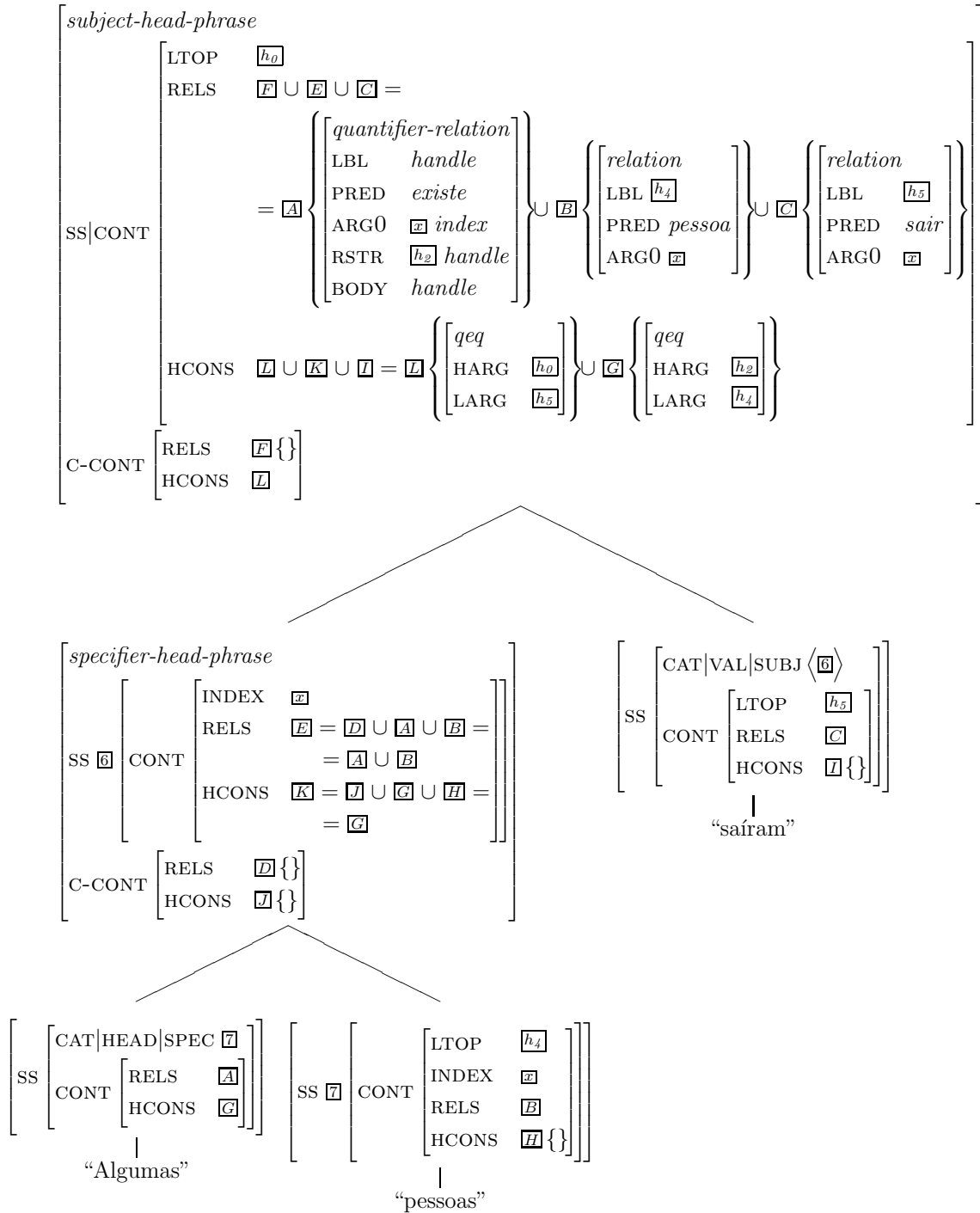


Figura 3.1: Composição da representação semântica para o exemplo de trabalho “Algumas pessoas saíram”. SS abrevia SYNSEM. Os atributos LOCAL e HOOK encontram-se omitidos, com CAT e CONT directamente sob SYNSEM e LTOP e INDEX directamente sob CONT, por razões de espaço. Não se apresentam etiquetas para os traços LBL e BODY do quantificador ( $h_1$  e  $h_4$  respectivamente) apenas porque não são unificados com nenhum outro atributo nesta estrutura.

$$\left[ \begin{array}{l} \text{head-complements-phrase} \\ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|LTOP} \boxed{1} \text{ handle} \\ \text{RELS} \quad \boxed{A} \cup \boxed{B} \cup \boxed{C_1} \cup \dots \cup \boxed{C_n} \\ \text{HCONS} \quad \boxed{D} \cup \boxed{E} \cup \boxed{F_1} \cup \dots \cup \boxed{F_n} \end{array} \right] \\ \text{C-CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{A} \{\} \\ \text{HCONS} \quad \boxed{D} \{\} \end{array} \right] \\ \text{HEAD-DTR|SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|LTOP} \boxed{1} \\ \text{RELS} \quad \boxed{B} \\ \text{HCONS} \quad \boxed{E} \end{array} \right] \\ \text{NON-HEAD-DTRS} \left\langle \left[ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{C_1} \\ \text{HCONS} \quad \boxed{F_1} \end{array} \right] \right], \dots, \left[ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{C_n} \\ \text{HCONS} \quad \boxed{F_n} \end{array} \right] \right] \right\rangle \end{array} \right]$$

A identificação do LTOP do nó mãe, correspondente ao SV, com o LTOP do nó correspondente ao verbo (através de  $\boxed{1}$ ) garante que a pega do verbo é visível no ponto em que a relação *geq* entre o topo global e esta pega é introduzida na representação semântica, ou seja na construção que projecta o Sujeito à esquerda do SV.

Na representação lexical dos verbos transitivos, por sua vez, o traço ARG1, que corresponde ao segundo argumento da relação expressa pelo verbo, é unificado com o INDEX do seu complemento. Neste caso, da entrada lexical de “ler”, isso é feito através da partilha assinalada por  $\boxed{3}$ :

$$\left[ \begin{array}{l} \text{SYNSEM|LOCAL} \\ \text{CAT|VAL} \left[ \begin{array}{l} \text{SUBJ} \left\langle \left[ \text{LOCAL|CONT|HOOK|INDEX} \quad \boxed{2} \text{ index} \right] \right\rangle \\ \text{COMPS} \left\langle \left[ \text{LOCAL|CONT|HOOK|INDEX} \quad \boxed{3} \text{ index} \right] \right\rangle \end{array} \right] \\ \text{CONT} \left[ \begin{array}{l} \text{HOOK|LTOP} \quad \boxed{1} \text{ handle} \\ \left\{ \begin{array}{l} \text{binary-relation} \\ \text{LBL} \quad \boxed{1} \\ \text{PRED} \quad \text{ler} \\ \text{ARG0} \quad \boxed{2} \\ \text{ARG1} \quad \boxed{3} \end{array} \right\} \\ \text{HCONS} \quad \{\} \end{array} \right] \end{array} \right]$$

Com estes dispositivos, a obtenção da representação da frase “Todos os homens leram um livro” é no essencial semelhante à do exemplo anterior, apenas com mais duas relações e mais uma restrição sobre pegadas, contribuídas pelo SN complemento do verbo:

$$\left[ \begin{array}{l} \text{LTOP} \\ \text{RELS} \\ \text{HCONS} \end{array} \left[ \begin{array}{l} \boxed{h_0} \\ \left\{ \begin{array}{l} \text{quantifier-relation} \\ \text{LBL } \boxed{h_1} \\ \text{PRED } \textit{todo} \\ \text{ARG0 } \boxed{x_5} \\ \text{RSTR } \boxed{h_2} \\ \text{BODY } \boxed{h_3} \end{array} \right\}, \left\{ \begin{array}{l} \text{relation} \\ \text{LBL } \boxed{h_4} \\ \text{PRED } \textit{homem} \\ \text{ARG0 } \boxed{x_5} \end{array} \right\}, \left\{ \begin{array}{l} \text{binary-relation} \\ \text{LBL } \boxed{h_6} \\ \text{PRED } \textit{ler} \\ \text{ARG0 } \boxed{x_5} \\ \text{ARG1 } \boxed{x_{11}} \end{array} \right\}, \left\{ \begin{array}{l} \text{quantifier-relation} \\ \text{LBL } \boxed{h_7} \\ \text{PRED } \textit{existe} \\ \text{ARG0 } \boxed{x_{11}} \\ \text{RSTR } \boxed{h_8} \\ \text{BODY } \boxed{h_9} \end{array} \right\}, \left\{ \begin{array}{l} \text{relation} \\ \text{LBL } \boxed{h_{10}} \\ \text{PRED } \textit{livro} \\ \text{ARG0 } \boxed{x_{11}} \end{array} \right\} \right\} \\ \left\{ \begin{array}{l} \text{qeq} \\ \text{HARG } \boxed{h_0} \\ \text{LARG } \boxed{h_6} \end{array} \right\}, \left\{ \begin{array}{l} \text{qeq} \\ \text{HARG } \boxed{h_2} \\ \text{LARG } \boxed{h_4} \end{array} \right\}, \left\{ \begin{array}{l} \text{qeq} \\ \text{HARG } \boxed{h_8} \\ \text{LARG } \boxed{h_{10}} \end{array} \right\} \right\} \end{array} \right]$$

Esta representação semântica, escrita no formalismo de representação da HPSG, é equivalente à seguinte representação, escrita no formato MRS:

$$\langle h_0, \\
 \{h_1 : \textit{todo}(x_5, h_2, h_3), h_4 : \textit{homem}(x_5), h_6 : \textit{ler}(x_5, x_{11}), h_7 : \textit{existe}(x_{11}, h_8, h_9), h_{10} : \textit{livro}(x_{11})\} \\
 \{h_0 =_q h_6, h_2 =_q h_4, h_8 =_q h_{10}\} \rangle$$

Esta é uma representação sub-especificada do significado da frase, a qual vindo a ser eventualmente subsequentemente especificada, permite derivar cada uma das duas representações da frase escritas na linguagem da lógica de quantificadores generalizados:

$$\begin{array}{l} \textit{todo}(x, \textit{homem}(x), \textit{existe}(y, \textit{livro}(y), \textit{ler}(x, y))) \\ \textit{existe}(y, \textit{livro}(y), \textit{todo}(x, \textit{homem}(x), \textit{ler}(x, y))) \end{array}$$

No primeiro caso, a especificação resultará de se juntar  $h_3 = h_7$  e  $h_9 = h_6$  ao conjunto de restrições sobre as pegadas; no segundo, resultará de, em alternativa, se juntar  $h_8 = h_1$  e  $h_3 = h_6$ .

## Ambiguidade com a Negação

Considere-se agora um outro exemplo, como “Não chegou uma pessoa”, com ambiguidade envolvendo a negação. Note que por via de MOD, o modificador tem acesso ao seu argumento, que se encontra no seu âmbito, através de [3]. A entrada lexical para o advérbio “não” contém as seguintes restrições:

$$\left[ \begin{array}{l} \text{SYNSEM|LOCAL} \\ \text{CONT} \end{array} \left[ \begin{array}{l} \text{CAT|HEAD|MOD} \langle \left[ \text{LOCAL|CONT|HOOK|LTOP} \text{ [3]} \right] \rangle \\ \text{HOOK|LTOP} \text{ [1] } \textit{handle} \\ \text{RELS} \left\{ \begin{array}{l} \text{relation} \\ \text{LBL} \text{ [1]} \\ \text{PRED} \textit{neg} \\ \text{ARG0} \text{ [2] } \textit{handle} \end{array} \right\} \\ \text{HCONS} \left\{ \begin{array}{l} \text{qeq} \\ \text{HARG} \text{ [2]} \\ \text{LARG} \text{ [3] } \textit{handle} \end{array} \right\} \end{array} \right] \right]$$

Para as construções em que o modificador precede o modificado, as restrições relevantes para as construções envolvendo modificadores que tomam âmbito sobre o material que modificam é a seguinte:

$$\left[ \begin{array}{l} \text{modifier-head-phrase} \\ \text{SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|LTOP} \quad \boxed{1} \\ \text{RELS} \quad \boxed{A} \cup \boxed{B} \cup \boxed{C} \\ \text{HCONS} \quad \boxed{D} \cup \boxed{E} \cup \boxed{F} \end{array} \right] \\ \text{C-CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{A} \{ \} \\ \text{HCONS} \quad \boxed{D} \{ \} \end{array} \right] \\ \text{NON-HEAD-DTR|SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{HOOK|LTOP} \quad \boxed{1} \\ \text{RELS} \quad \boxed{B} \\ \text{HCONS} \quad \boxed{E} \end{array} \right] \\ \text{NON-HEAD-DTR|SYNSEM|LOCAL|CONT} \left[ \begin{array}{l} \text{RELS} \quad \boxed{C} \\ \text{HCONS} \quad \boxed{F} \end{array} \right] \end{array} \right]$$

O LTOP do nó mãe é o LTOP do modificador, através da partilha do valor assinalado com  $\boxed{1}$ . Assim, é este o valor do LTOP do nó irmão do Sujeito, o qual vai ser equacionado com o topo global por via da construção relevante para a frase matriz.

Desta forma, e assumindo que a construção que projecta sujeitos à direita do verbo tem restrições semânticas semelhantes às da regra para os sujeitos pré-verbais (*subject-head-phrase*), já apresentada atrás, à nossa frase de trabalho “Não chegou uma pessoa” corresponde a seguinte representação semântica:

$$\left[ \begin{array}{l} \text{LTOP} \quad \boxed{h_0} \\ \text{RELS} \quad \left\{ \left[ \begin{array}{l} \text{relation} \\ \text{LBL} \quad \boxed{h_1} \\ \text{PRED} \quad \text{neg} \\ \text{ARG0} \quad \boxed{h_2} \end{array} \right], \left[ \begin{array}{l} \text{relation} \\ \text{LBL} \quad \boxed{h_3} \\ \text{PRED} \quad \text{chegar} \\ \text{ARG0} \quad \boxed{x_4} \end{array} \right], \left[ \begin{array}{l} \text{quantifier-relation} \\ \text{LBL} \quad \boxed{h_5} \\ \text{PRED} \quad \text{existe} \\ \text{ARG0} \quad \boxed{x_4} \\ \text{RSTR} \quad \boxed{h_6} \\ \text{BODY} \quad \boxed{h_7} \end{array} \right], \left[ \begin{array}{l} \text{relation} \\ \text{LBL} \quad \boxed{h_8} \\ \text{PRED} \quad \text{pessoa} \\ \text{ARG0} \quad \boxed{x_4} \end{array} \right] \right\} \\ \text{HCONS} \quad \left\{ \left[ \begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{h_0} \\ \text{LARG} \quad \boxed{h_1} \end{array} \right], \left[ \begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{h_2} \\ \text{LARG} \quad \boxed{h_3} \end{array} \right], \left[ \begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{h_6} \\ \text{LARG} \quad \boxed{h_8} \end{array} \right] \right\} \end{array} \right]$$

No formato MRS, esta representação escreve-se da seguinte forma:

$$\langle h_0, \{ h_1 : \text{neg}(h_2), h_3 : \text{chegar}(x_4), h_5 : \text{existe}(x_4, h_6, h_7), h_8 : \text{pessoa}(x_4) \}, \{ h_0 =_q h_1, h_2 =_q h_3, h_6 =_q h_8 \} \rangle$$

Esta representação permite a especificação das duas leituras relevantes:

$$\begin{array}{l} \text{neg}(\text{existe}(x, \text{pessoa}(x), \text{chegar}(x))) \\ \text{existe}(x, \text{pessoa}(x), \text{neg}(\text{chegar}(x))) \end{array}$$

A primeira é sinónima da paráfrase “Não há ninguém que tenha chegado”, e a segunda por é sinónima de “Existe uma pessoa que não chegou”.

### 3.7 Leituras Subsequentes

Existe um leque razoável de manuais para se estudar a semântica composicional, desde um nível introdutório, como o clássico de Dowty *et al.* (1981), com uma apresentação que segue mais de perto os resultados originais de Richard Montague; os manuais por Chierchia e McConnell-Ginet (1990) e de Swart (1998), com uma apresentação mais centrada na associação entre representações semânticas e representações sintáticas; a obra de Gamut (1991), com uma apresentação mais centrada na linguagem lógica que suporta a representação semântica; e o compêndio de Partee *et al.* (1990), com uma apresentação de aspectos lógicos e matemáticos fundacionais.

Cabe ainda destacar o manual de Kamp e Reyle (1993), com uma introdução à semântica composicional através da apresentação da DRT (*Discourse Representation Theory*).

Haverá certamente vantagem em ter acesso à obras pioneiras como as de Montague (1973) e Barwise e Cooper (1981).

Há um universo de manuais sobre lógica formal úteis para acompanhar os manuais indicados acima, dos quais destacamos Barwise e Etchemendy (1999). Para uma introdução breve em português à noção de quantificador generalizado, veja-se Branco (2001).

Para a apresentação original e completa da MRS, deve-se consultar Copestake *et al.* (2005). Um desenvolvimento subsequente é a RMRS (*Robust Minimal Recursion Semantics*) (Copestake, 2004), que permite sub-especificar a aridade de predicados e tem sido utilizada sobretudo em processamento de linguagem natural dito de superficial, sem recurso a um léxico.

Há vários outros formatos de representação semântica que permitem sub-especificação. Veja-se, por exemplo, a *Quasi Logical Form* (Alshawi e Crouch, 1992), a *Under-specified Discourse Representation Theory* (Reyle, 1993), a *Hole Semantics* (Bos, 1996) ou a *Glue Semantics* (Asudeh e Crouch, 2002).

Os volumes de definição da HPSG ((Pollard e Sag, 1987), (Pollard e Sag, 1994) e (Sag *et al.*, 2003)) utilizam outros formatos para a representação semântica, que tendo sido preteridos em favor do formato MRS, sobretudo em aplicações computacionais, podem ser interessantes por motivos históricos.

Uma gramática HPSG permite o processamento linguístico profundo e é um recurso importante no desenvolvimento de aplicações que envolvem o processamento computacional da linguagem natural. Dos inúmeros casos de aplicação de gramáticas computacionais, nesta secção cabe também indicar alguns exemplos a título ilustrativo.

As representações semânticas abstraem de muitos pormenores linguísticos superficiais (e.g. flexão, ordem de palavras, etc), sendo mais fáceis de manipular automaticamente em aplicações de extracção de informação a partir de texto em linguagem natural.

Em (Herbelot e Copestake, 2006), por exemplo, procura-se por padrões em representações parecidas a MRS obtidas automaticamente a partir de texto na Wikipédia, de forma a extrair automaticamente relações de hiperonímia entre conceitos com vista

à construção de ontologias.

A tradução automática é um outro exemplo de aplicação. As representações semânticas podem ser usadas como representação intermédia entre a língua de partida e a língua de chegada. Uma gramática computacional pode ser usada para produzir representações do significado a partir de linguagem natural (análise) assim como a operação inversa de produzir frases da linguagem natural a partir de representações semânticas (geração). Uma estratégia frequente é usar uma gramática computacional da língua de partida para análise. Em seguida, a partir das representações semânticas obtidas gera-se a tradução com outra gramática para a língua de chegada—ver (Wahlster, 2000; Bond *et al.*, 2005); ver também (Copestake *et al.*, 1995).

As representações do significado são escritas numa linguagem lógica, o que abre a possibilidade de, pelo menos parcialmente, se automatizar processos de inferência. Uma gramática computacional que produza representações lógicas do significado pode ser ligada a motores de inferência automática (Blackburn e Bos, 2005) por forma a apoiar raciocínio a partir de premissas em linguagem natural.

De igual forma, uma gramática computacional pode ser usada para a obtenção automática da representação do conhecimento contido em repositórios textuais. Isto permite apoiar os processos ou aplicações de extracção de informação, preenchimento de bases de dados de conhecimento, ou anotação automática de meta-dados de páginas Web apoio à Web semântica.





# Bibliografia

- ALSHAWI, HIYAN E RICHARD S. CROUCH, 1992. Monotonic Semantic Interpretation. In *Proceedings of the 30<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL-92)*, pp. 32–39. Newark, NJ.
- ASUDEH, ASH E RICHARD CROUCH, 2002. Glue Semantics for HPSG. In FRANK VAN EYNDE, LARS HELLAN E DOROTHEE BEERMANN, orgs., *Proceedings of the 8<sup>th</sup> International Conference on Head-Driven Phrase Structure Grammar*, pp. 1–19. CSLI Publications, Stanford.
- BALARI, SERGIO E LUCA DINI, orgs., 1997. *Romance in HPSG*. CSLI Publications, Stanford.
- BARWISE, JON E ROBIN COOPER, 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4(1): 159–219.
- BARWISE, JON E JOHN ETCEHEMENDY, 1999. *Language, Proof and Logic*. CSLI Publications, Stanford.
- BLACKBURN, PATRICK E JOHAN BOS, 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. CSLI Publications, Stanford.
- BOND, FRANCIS, STEPHAN OEPEN, MELANIE SIEGEL, ANN COPESTAKE E DAN FLICKINGER, 2005. Open Source Machine Translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pp. 15–22. Phuket.
- BOS, JOHAN, 1996. Predicate Logic Unplugged. In *Proceedings of the 10<sup>th</sup> Amsterdam Colloquium*, pp. 133–143. ILLC, Amsterdão.
- BOUMA, GOSSE, ROB MALOUF E IVAN SAG, 2001. Satisfying Constraints on Extraction and Adjunction. *Natural Language and Linguistic Theory*, 1(19): 1–65.
- BRANCO, ANTÓNIO, 1996. Branching Split Obliqueness at the Syntax-Semantics Interface. In *Proceedings of COLING 96*, pp. 149–156. Center for Sprogteknologi, Copenhagen.

- BRANCO, ANTÓNIO, 2000. *Reference Processing and its Universal Constraints*. Edições Colibri, Lisboa.
- BRANCO, ANTÓNIO, 2001. Quantificação Generalizada. In JOÃO BRANQUINHO E DESIDÉRIO MURCHO, orgs., *Enciclopédia de Termos Lógico-Filosóficos*, pp. 590–591. Gradiva, Lisboa.
- BRANCO, ANTÓNIO, 2005a. Cognitive Science and its Computational Foundations: A Natural Language Perspective. In JOÃO BRANQUINHO, org., *Cognition and Content*. Centro de Filosofia da Universidade de Lisboa, Lisboa.
- BRANCO, ANTÓNIO, 2005b. Reflexives: Escaping Exemption via Domain Reshuffling. In STEFAN MÜLLER, org., *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pp. 467–481. CSLI Publications, Stanford.
- BRANCO, ANTÓNIO E FRANCISCO COSTA, 2006. Noun Ellipsis without Empty Categories. In STEFAN MÜLLER, org., *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pp. 81–101. CSLI Publications, Stanford.
- BRANCO, ANTÓNIO E FRANCISCO COSTA, 2007a. Accommodating Language Variation in Deep Processing. In TRACY HOLLOWAY KING E EMILY M. BENDER, orgs., *Proceedings of the Grammar Engineering Across Frameworks Workshop (GEAF07)*, pp. 67–86. CSLI Publications, Stanford.
- BRANCO, ANTÓNIO E FRANCISCO COSTA, 2007b. Identification and Handling of Dialectal Variation with a Single Grammar. In PETER DIRIX, INEKE SCHURMAN, VINCENT VANDEGHINSTE E FRANK VAN EYNDE, orgs., *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN17)*, pp. 5–19. LOT, Utrecht.
- BRANCO, ANTÓNIO E FRANCISCO COSTA, 2007c. Self- or Pre-Tuning? Deep Linguistic Processing of Language Variants. In *ACL 2007 Workshop on Deep Linguistic Processing*, pp. 57–64. Association for Computational Linguistics, Praga.
- BRANCO, ANTÓNIO E BERTHOLD CRYSMANN, 2001. Negative Concord and the Distribution of Quantifiers. In YVES D’HULST, JOHAN ROORYCK E JAN SCHROTEN, orgs., *Romance Languages and Linguistic Theory*. John Benjamins, Amsterdão.
- BRANCO, ANTÓNIO E PALMIRA MARRAFA, 1997. Subject-oriented and non Subject-oriented Long-distance Anaphora: an integrated approach. In BYUNG-SOO PARK E JONG-BOK KIM, orgs., *Proceedings of the 11th Pacific-Asia Conference on Language, Information and Computation (PACLIC’96)*, pp. 21–31. Language Education and Research Institute, Kyung Hee University, Seul.

- BRANCO, ANTÓNIO E PALMIRA MARRAFA, 1999. Long-Distance Reflexives and the Binding Square of Opposition. In GERT WEBELHUTH, JEAN-PIERRE KOENIG E ANDREAS KATHOL, orgs., *Lexical and Constructional Aspects of Linguistic Explanation*, pp. 111–123. CSLI Publications, Stanford.
- CARPENTER, BOB, 1992. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge.
- CHAVES, RUI PEDRO, 2005. A Linearization-Based Approach to Gapping. In G. JÄGER, P. MONACHESI, G. PENN E S. WINTNER, orgs., *FG-MOL 2005: The 10th conference on Formal Grammar and The 9th Meeting on Mathematics of Language*. CSLI Publications, Stanford.
- CHAVES, RUI PEDRO, 2006. Coordination of Unlikes without Unlike Categories. In STEFAN MÜLLER, org., *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pp. 102–122. CSLI Publications, Stanford.
- CHAVES, RUI PEDRO, 2007. *Coordinate Structures—Constraint-based Syntax-Semantics Processing*. Tese de doutoramento, Universidade de Lisboa, Lisboa.
- CHIERCHIA, GENNARO E SALLY MCCONNELL-GINET, 1990. *Meaning and Grammar: An Introduction to Semantics*. MIT Press, Cambridge.
- COPESTAKE, ANN, 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.
- COPESTAKE, ANN, 2004. Robust Minimal Recursion Semantics. Manuscrito.
- COPESTAKE, ANN, DAN FLICKINGER, ROB MALOUF, SUSANNE RIEHEMANN E IVAN SAG, 1995. Translation using Minimal Recursion Semantics. In *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*. Lovaina.
- COPESTAKE, ANN, DAN FLICKINGER, IVAN A. SAG E CARL POLLARD, 2005. Minimal Recursion Semantics: An Introduction. *Journal of Research on Language and Computation*, 3(2–3): 281–332.
- COSTA, FRANCISCO, 2007. *Deep Linguistic Processing of Portuguese Noun Phrases*. Tese de mestrado, Universidade de Lisboa, Lisboa.
- CRYSMANN, BERTHOLD, 1998. The Conspiracy of Quantification and Linear Precedence in European Portuguese Proclisis. In *Proceedings of the Eastern States Conference on Linguistics (ESCOL '97)*. CLC Publications, Ithaca.
- CRYSMANN, BERTHOLD, 1999. Licensing Proclisis in European Portuguese. In F. CORBLIN, J.-M. MARANDIN E C. DOBROVIE-SORIN, orgs., *Empirical Issues in Formal Syntax and Semantics. Selected papers from the Colloque de Syntaxe et de Sémantique de Paris (CSSP 1997)*, pp. 255–276. Thesus, Haia.

- CRYSMANN, BERTHOLD, 2000a. Clitics and Coordination in Linear Structure. In BIRGIT GERLACH E JANET GRIJZENHOUT, orgs., *Clitics in Phonology, Morphology, and Syntax*, pp. 121–159. John Benjamins, Amsterdão.
- CRYSMANN, BERTHOLD, 2000b. Syntactic Transparency of Pronominal Affixes. In RONNIE CANN, CLAIRE GROVER E PHILIP MILLER, orgs., *Grammatical Interfaces in HPSG*, pp. 77–96. CSLI Publications, Stanford.
- CRYSMANN, BERTHOLD, 2002. *Constraint-Based Coanalysis: Portuguese Cliticisation and Morphology-Syntax Interaction in HPSG*. Saarbrücken Dissertations in Computational Linguistics and Language Technology. Deutsches Forschungszentrum für Künstliche Intelligenz e Universität des Saarlandes, Saarbrücken.
- DE SWART, HENRIËTTE, 1998. *Introduction to Natural Language Semantics*. CSLI Publications, Stanford.
- DELPH-IN, Página Web. Página com a apresentação da iniciativa Deep Linguistic Processing with HPSG. <http://www.delph-in.net>.
- DOWTY, DAVID, ROBERT WALL E STANLEY PETERS, 1981. *Introduction to Montague Semantics*. D. Reidel, Dordrecht.
- FRIEDENBERG, JAY E GORDON SILVERMAN, 2006. *Cognitive Science: An Introduction to the Study of Mind*. Sage Publications, Thousand Oaks.
- FROMKIN, VICTORIA, 2000. *Linguistics: An Introduction to Linguistic Theory*. Blackwell, Oxford.
- GAMUT, L. T. F., 1991. *Logic, Language and Meaning*. The University of Chicago Press, Chicago e Londres. 2 volumes; Volume I: Introduction to Logic; Volume II: Intensional Logic and Logical Grammar.
- HERBELOT, AURELIE E ANN COPESTAKE, 2006. Acquiring Ontological Relationships from Wikipedia Using RMRS. In *Proceedings of the Workshop on Web content Mining with Human Language Technologies (ISWC'06)*.
- HPSG, Página Web. The HPSG Bibliography. Repositório de publicações de trabalhos de investigação realizados no quadro da HPSG. <http://hpsg.fu-berlin.de/HPSG-Bib>.
- KAMP, HANS E UWE REYLE, 1993. *From Discourse to Logic: An Introduction to Modeltheoretic Semantics, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.
- KATHOL, ANDREAS, 1995. *Linearization-Based German Syntax*. Tese de doutoramento, Ohio State University.

- KIM, JONG-BOK E PETER SELLS, 2008. *English Syntax: An Introduction*. CSLI Publications, Stanford.
- LKB, Página Web. Página que resume o procedimento para obter e instalar o software LKB. <http://wiki.delph-in.net/moin/LkbInstallation>.
- LXGram, Página Web. Sítio da Web em que se encontra disponível a gramática LXGram para o português. <http://nlx.di.fc.ul.pt>.
- MONTAGUE, RICHARD, 1973. The Proper Treatment of Quantification in Ordinary English. In K.J.J. HINTIKKA, J.M.E. MORAVCSIK E P. SUPPES, orgs., *Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pp. 212–242. Reidel, Dordrecht.
- MONTAGUE, RICHARD, 1974. The Proper Treatment of Quantification in Ordinary English. In R. THOMASON, org., *Formal Philosophy*, pp. 247–270. Yale University Press, New Haven.
- PARTEE, BARBARA, ALICE TER MEULEN E ROBERT E. WALL, 1990. *Mathematical Methods in Linguistics*. Kluwer, Dordrecht.
- POLLARD, CARL E IVAN SAG, 1987. *Information-Based Syntax and Semantics, Vol. 1*. CSLI Publications, Stanford.
- POLLARD, CARL E IVAN SAG, 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press e CSLI Publications, Stanford.
- REAPE, MIKE, 1994. Domain Union and Word Order Variation in German. In JOHN NERBONNE, KLAUS NETTER E CARL J. POLLARD, orgs., *German in Head-Driven Phrase Structure Grammar*, n. 46 de CSLI Lecture Notes, pp. 151–197. CSLI Publications, Stanford.
- REYLE, UWE, 1993. Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction. *Journal of Semantics*, 10: 123–179.
- SAG, IVAN A., THOMAS WASOW E EMILY M. BENDER, 2003. *Syntactic Theory – A Formal Introduction*. CSLI Publications, Stanford.
- SHIEBER, STUART, 1986. *An Introduction to Unification-Based Approaches to Grammars*. CSLI Publications, Stanford.
- TRALE, Página Web. Página que resume o procedimento para obter e instalar o software TRALE. <http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale>.
- USZKOREIT, HANS, ROLF BACKOFEN, JO CALDER, JOANNE CAPSTICK, LUCA DINI, JOCHEN DÖRRE, GREGOR ERBACH, DOMINIQUE ESTIVAL, SURESH MANANDHAR, ANNE-MARIE MINEUR E STEPHAN OEPEN, 1996.

- The EAGLES Formalisms Working Group — Final Report Expert Advisory Group on Language Engineering Standards. Relatório Técnico LRE 61–100, DFKI — Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken. [http://www.dfki.de/lt/publications\\_show.php?id=563](http://www.dfki.de/lt/publications_show.php?id=563).
- VILLAVICENCIO, ALINE, LOUISA SADLER E DOUG ARNOLD, 2005. An HPSG Account of Closest Conjunct Agreement in NP Coordination in Portuguese. In STEFAN MÜLLER, org., *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pp. 427–447. CSLI Publications, Stanford.
- WAHLSTER, WOLFGANG, org., 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.