A relação entre o desenvolvimento de formalismos gramaticais de base lexicalista e as exigências do PLN*

(The development of formal grammars under Lexicalist Hypothesis and its association to requirements of NLP systems)

Albano Dalla Pria
(Universidade do Estado de Mato Grosso, Campus de Alto Araguaia, Depto. de Letras)

Resumo: *O objetivo deste trabalho é mostrar que o desenvolvimento de formalismos gramaticais de base lexicalista está diretamente relacionado às exigências do PLN. Léxicos estruturados e o formalismo gramatical são centrais para um sistema de PLN. Este trabalho fundamenta-se no desenvolvimento do modelo HPSG, que se impôs no cenário lingüístico e lingüístico-tecnológico como uma teoria lingüística formal que, desde a sua origem, comprometeu-se com descrições lexicalistas, formalmente precisas e computacionalmente tratáveis dos fenômenos lingüísticos. Atualmente, conclui-se que as implementações do modelo HPSG reforçam seu comprometimento com o PLN.*
Palavras-chave: *Léxico; PLN; HPSG.*

Abstract: *This paper aims to associate the development of formal grammars under Lexicalist Hypothesis and requirements of NLP systems. Structured lexicons and formal grammars are required to NLP systems. This work is supported by the development of HPSG model. Since its inception HPSG has been conceived as a formal theory involved with formal descriptions and implementations of linguistic phenomena. Nowadays, HPSP engagement with NLP is reinforced by its implementations for many languages.*
Key-words: *Lexicon; NLP; HPSG.*

* Este trabalho teve o apoio financeiro da FAPESP [Processo nº 02/10650-8].

1. O campo de atuação do Processamento Automático das Línguas Naturais

O campo de atuação do Processamento Automático das Línguas Naturais (doravante PLN) visa à projeção e à implementação de sistemas computacionais para o processamento automático de línguas naturais[1] . Dada sua natureza teórica e aplicada, agrega diversas áreas do conhecimento, entre elas: a Lingüística, a Filosofia, a Psicologia, a Lógica, a Matemática, a Lingüística Computacional, a Ciência da Computação e a Inteligência Artificial.

Dois elementos são centrais para um sistema de PLN: o léxico e o formalismo gramatical[2] . Do ponto de vista do falante, o léxico constitui um módulo do processamento da linguagem, o léxico mental, estudado pela Lingüística e pela Psicolingüística; do ponto de vista do PLN, constitui um módulo do processamento lingüístico do sistema, o léxico do sistema, ou uma base de conhecimento lexical.

O léxico fornece ao sistema o conhecimento lingüístico necessário ao emprego das palavras da língua no processamento da linguagem e sua complexidade aumenta segundo a teoria lingüística subjacente ao sistema. Cabe à Lingüística, nesse contexto, a sistematização e a formalização do conhecimento lexical necessário para a implementação dos sistemas.

As reflexões sobre o tratamento do léxico, do modo como está sendo entendido aqui, serão aprofundadas da perspectiva da Lingüística, no item 2, e do PLN, no item 3. As reflexões sobre a relação entre o léxico e o PLN ampliam-se, no desenvolvimento do trabalho, mediante à apresentação de um formalismo gramatical, em particular, de descrição de léxicos, que se revela também apropriado para a utilização em sistemas de PLN: o modelo *Head-driven Phrase Structure Grammar* (HPSG). Para ser incorporado a um sistema de PLN, o conhecimento lexical, independentemente da sua natureza (morfológica, sintática ou semântica), deve ser necessariamente formalizado, para que seja eficientemente implementado.

---

1. Tais como sistemas de tradução automática (Arnold *et al*, 1994) e redes léxico-semânticas e lógico-conceptuais do tipo WordNet (Fellbaum, 1998)
2. Contudo, cabe ressaltar que o grau de eficiência dos sistemas varia em função do conhecimento lingüístico implementado e de outros expedientes, tais como a arquitetura adotada e os objetivos de sua aplicação.

2. A estruturação de um antigo repositório de idiossincrasias: o léxico

A base das línguas naturais é constituída essencialmente de um léxico e de uma gramática. Nos primórdios da Lingüística Estrutural, reconheciam-se três componentes da gramática: o sintático, o morfológico e o fonológico. Nesse contexto, considerava-se que todas as palavras de uma língua (o seu léxico) já existiam em potencial e não se atribuía, portanto, nenhum tipo de organização ou estruturação a esse conjunto. Nessa mesma época, Bloomfield (1933), ícone do estruturalismo americano, reconhecia o léxico como um repositório de fatos idiossincráticos e imprevisíveis sobre os itens lexicais (Briscoe, 1991).

Quando propôs a hipótese transformacional da gramática, Chomsky (1957) manteve a concepção de Bloomfield (1933) a respeito do léxico. Na primeira proposta de Gramática Transformacional (Chomsky, 1957), o léxico não era tratado dentro dos componentes da gramática; os itens lexicais já eram dados e convenientemente processados, entrando em uma estrutura sintática depois de ela ter passado por processos transformacionais a fim de substituir o símbolo terminal referente à sua categoria.

A influência exercida pelas Ciências Cognitivas (na década de 50) sobre a Lingüística e o deslocamento desta para a área das ciências naturais motivaram mudanças na configuração da Gramática Transformacional. As Ciências Cognitivas consideravam o cérebro um órgão modular e essa idéia de modularidade estendeu-se aos estudos da linguagem[3] . Passou-se a acreditar que, no processamento mental da linguagem, existem módulos específicos para a manipulação das diferentes instâncias lingüísticas (Chomsky, 1965, 1981).

Chomsky (1965), embora reconhecesse o léxico ainda como um repositório de idiossincrasias, identificou-o como uma dessas instâncias modulares, não-autônoma, integrada ao Componente de Base da gramática. Manteve-se a distinção entre o léxico e a gramática. Esta era o *locus* da regularidade, enquanto aquele era o *locus* das idiossincrasias e fornecia à sintaxe os elementos já convenientemente processados, não havendo qualquer processo de derivação no interior do léxico (relacionamento entre itens

3. A respeito da teoria da modularidade, consulte-se Fodor (1983).

lexicais) durante o processamento das sentenças. Acreditava-se na impossibilidade de atribuição de qualquer estruturação ou generalização (de base morfológica, sintática ou semântica) ao léxico (Di Sciullo & Williams, 1987).

Entretanto, Chomsky (1965) já demonstrava interesse em conferir ao léxico algum tipo de organização, tendo em vista reduzir o custo na seleção das palavras que preenchem as seqüências de símbolos terminais durante a derivação das sentenças. Esse custo era medido com base na quantidade de informações idiossincráticas que um indivíduo deveria saber a fim de processar uma sentença.

Chomsky (1965) começou por identificar traços específicos referentes às categorias às quais pertenciam às palavras do léxico, principalmente as chamadas categorias lexicais (Nome, Verbo, Adjetivo e Preposição). Essas categorias passaram a ser vistas como complexos estruturados de traços categoriais (referentes à classe da palavra), traços de subcategorização (referentes às outras classes com a qual uma determinada classe co-ocorre) e traços selecionais (referentes aos tipos semânticos dessas outras classes). A distinção entre classes, por exemplo, dava-se com base na oposição +/– Nome ("natureza nominal") e +/– Verbo ("natureza verbal"): nome [**+N, -V**], verbo [**-N, +V**], adjetivo [**+N, +V**] e preposição [**-N, -V**].

Essas investigações conduziram Chomsky (1970) à identificação de relacionamentos morfológicos, semânticos e sintáticos entre certos nomes (os deverbais) e os verbos dos quais eram derivados. Propôs que ambos tivessem uma única entrada lexical não marcada sintaticamente. Nessa abordagem, a entrada, processada dentro do Componente de Base da gramática, assumia a forma fonológica do verbo quando inserida sob o nó V (Verbo) e a forma fonológica do nome quando inserida sob o nó N (Nome). Com isso se lançaram as bases da Hipótese Lexicalista (Chomsky, 1970).

Essa hipótese nega a relação transformacional entre itens lexicais e institui um componente lexical a partir do qual os itens lexicais são gerados. Dentro desse componente são expressas generalizações lexicais (propriedades comuns a vários itens) através de **regras de redundância**, que servem para relacionar entradas lexicais e atribuir algum tipo de estruturação ao léxico (Briscoe, 1991). Jackendoff (1972) demonstrou ser possível a identificação não apenas de generalizações sintáticas no léxico, mas também de generalizações morfológicas e semânticas, atendendo, com isso, a expectativa de incorporação de informação semântica à gramática.

Jackendoff (1975), tentando explicar regras de redundância entre verbos e "nominais" relacionados, por exemplo, entre *decide* e *decision*, no inglês, conclui que uma das dificuldades para a gramática é o desenvolvimento de um formalismo que expresse tais regras de redundância entre entradas lexicais.

Na década de 1980, surgiram formalismos gramaticais com o intuito de capturar generalizações lingüísticas. Esses modelos atribuíram ao léxico o estatuto de componente essencial da gramática. Muitas das regras gramaticais passam a ser lexicalmente regidas, devendo, portanto, com base em um conjunto de informações morfológicas, sintáticas e semânticas (ou seja, para a maior parte dos fenômenos lingüísticos), restringir itens lexicais a classes refinadamente mais especificadas do que aquelas que se pode obter na classificação da gramática tradicional (Briscoe, 1991).

Nesse contexto, surgiram, por exemplo, a *Generalized Phrase Structure Grammar*, doravante GPSG, (Gazdar *et al*, 1985), a *Lexical Functional Grammar*, doravante LFG, (Bresnan, 1982) e a *Head-driven Phrase Structure Grammar*, doravante HPSG (Pollard & Sag, 1994). Nesses modelos, o léxico não só contém informações sobre as palavras, mas também é usado como uma unidade de controle para examinar a boa formação das sentenças geradas pelo próprio modelo (Handke, 1995).

Nas teorias lexicalistas, os itens lexicais são objetos complexos que contêm informações referentes à fonologia, morfologia, sintaxe e semântica (microestrutura lexical). Há também modelos que agregam outros tipos de informação ao léxico, tais como suas condições de uso (informação pragmática). Além disso, as generalizações feitas a partir dessas informações permitem que os itens lexicais sejam relacionados uns aos outros por meio de artifícios desenvolvidos pela gramática (macroestrutura lexical) (Gibbon, 2000).

A HPSG, por exemplo, adota uma postura radicalmente lexicalista, segundo a qual a maior parte da informação gramatical e semântica deve ser codificada nas entradas lexicais no processamento da linguagem. As entradas lexicais correspondem às palavras realizadas nas sentenças e podem ser vistas como elementos-chave que orientam a construção da estrutura sintática e semântica (Sag & Wasow, 1999). As informações lexicais são modeladas em "estruturas de atributo-valor" e as relações entre as entradas são expressas por meio de "regras lexicais", "hierarquia de tipos" e

"procedimentos de unificação". Esses construtos serão retomados no desenvolvimento do trabalho.

Essa estruturação em forma de atributo-valor permite à HPSG não se preocupar com a decomposição de categorias lexicais em traços primários do tipo [+/-N] e [+/-V]. Nome, verbo, adjetivo e preposição são listados como quatro possíveis valores para um traço de núcleo, mas o inventário de categorias fica em aberto, dependendo de um conjunto de cruzamentos de informações sintáticas, semânticas, morfológicas e fonológicas especificados pelas estruturas de traços dos elementos lexicais (Pollard & Sag, 1994; Malouf, 1998).

O papel de destaque atribuído ao léxico nos modelos gramaticais surgidos na segunda metade do século XX elevou-o a igual posição no âmbito dos estudos da linguagem. A Psicolingüística integrou o conhecimento lexical ao sistema de processamento mental da linguagem e a Lingüística Computacional integrou-o aos sistemas de PLN[4].

3. A centralidade do léxico em sistemas de PLN

Os sistemas de PLN que visam a emular o conhecimento lingüístico, por mais simples que sejam, exigem quantidade considerável de conhecimento lingüístico para sua aplicação. Parte do conhecimento exigido pelos sistemas refere-se ao conhecimento sobre o léxico das línguas naturais, i. e., o conjunto das palavras que as compõem. Surgem, então, questionamentos quanto ao tratamento do léxico nos sistemas (Briscoe, 1991; Grishman & Calzolari, 1995; Handke, 1995; Sanfilippo, 1995). Esses questionamentos referem-se à quantidade de informação especificada nesse componente nos sistemas e à especificação, nesse componente, de informações lingüísticas compartilhadas pelos itens lexicais. O léxico, dessa perspectiva, não configura mera lista de partes-do-discurso (especificadas pela Gramática Tradicional), muito menos um repositório de idiossincrasias (Bloomfield, 1933).

A especificação de informações lingüísticas no componente lexical contribui decisivamente para a redução de componentes individuais e, conse-

4. A respeito do processamento mental e computacional da linguagem, consulte-se Handke (1995).

quentemente, do sistema como um todo. No início da década de 1980, os sistemas de PLN já apresentavam componentes relativamente independentes. Acreditava-se que módulos adequadamente delimitados reduzissem o tamanho de componentes individuais do sistema e, conseqüentemente, o seu tamanho como um todo (Grishman, 1980).

SanFilippo (1995) e Handke (1995) ressaltam que, na última década, o PLN tem exigido léxicos estruturados, que forneçam informação morfológica, sintática e semântica das palavras e que possam ser eficientemente implementados em sistemas de PLN. Ressalta-se, do mesmo modo, que há uma tendência crescente na utilização de formalismos gramaticais que codificam informações lingüísticas em termos de estruturas de traços, tendo a herança e a unificação de traços como operações relacionais entre as estruturas.

A HPSG tem um objetivo comum a outras teorias lingüísticas, como a LFG (Bresnan, 1982, 2001) e a teoria de Princípios e Parâmetros, doravante P&P (Chomsky, 1981): delinear uma teoria da linguagem que investiga o uso da linguagem e propõe universais para explicar tais usos. O objetivo da HPSG, embora comum a outras teorias, é concebido sob uma perspectiva distinta. A teoria de Princípios e Parâmetros compromete-se em contribuir para a discussão do realismo psicológico da linguagem e a LFG, além desse comprometimento, pretende contribuir para a discussão de questões relacionadas ao processamento metal da linguagem (SELLS, 1985). A HPSG, embora reconheça que a linguagem constitui um conhecimento instanciado no cérebro, não se compromete com essas discussões. A HPSG, assim como outros formalismos gramaticais, como a Gramática Categorial, a LFG e a GPSG, compromete-se em atribuir um alto grau de modelagem matemática para a descrição das entidades lingüísticas (Pollard, 1997; Pollard & Sag, 1996; Steedman, 2000; Bresnan, 1982, 2001).

O rigor formal assumido pela HPSG contribui para a concepção de um modelo distinto de gramática. A HPSG e a LFG assumem uma abordagem não-derivacional baseada em restrições e unificação[5] (Bresnan, 2001), no qual os enunciados empiricamente observáveis são descritos em termos de estruturas de traços (Pollard, 1997; Pollard & Sag, 1996;). A HPSG codifica a informação lingüística – fonológica, sintática e semântica

5. Na literatura, o termo "unificação" é sinônimo de "complexo de traços" (Shieber, 1986).

– em um único nível de representação através de complexos estruturados de traços (Cooper, 1996). A teoria de P&P, além de assumir uma abordagem derivacional da gramática, no qual as estruturas gramaticais são descritas em termos de operações de deslocamento de constituintes a partir de estruturas subjacentes, codifica essa informação em múltiplos níveis de representação. A LFG, embora não postule estruturas gramaticais subjacentes, assim como a HPSG, descreve a informação lingüística em dois níveis centrais de representação: a estrutura-f, responsável por descrever relações gramaticais e a estrutura-c, responsável por descrever funções gramaticais. A LFG admite ainda outros dois níveis de representação: a estrutura-ó, responsável por descrever aspectos relevantes do significado, e a estrutura-a, responsável por relacionar argumentos sintáticos à descrição do seu significado (Bresnan, 2001).

Na seqüência, serão detalhados alguns dos elementos do formalismo HPSG, tais como a organização das estruturas de traços e os processos de herança e unificação.

4. A descrição e a formalização da informação lexical no modelo HPSG

A HPSG resultou de tentativas de modificação da GPSG no ambiente interdisciplinar do *Center for the Study of Language and Information* da Universidade de Stanford, na Califórnia, onde são também desenvolvidos sistemas de PLN. A HPSG tem sido desenvolvida como um esforço consciente de síntese de idéias de uma variedade de pesquisas recentes em sintaxe (principalmente abordagens não-derivacionais como a Gramática Categorial (Wood, 1993), a GPSG, a *Arc Pair Grammar* e a LFG), em semântica (especialmente a Semântica de Situações) e em Ciências da Computação (Representação do Conhecimento, Formalismos Baseados em Unificação, *Data Type Theory*). O desenvolvimento da teoria deu-se em um ambiente que pressupõe familiaridade com áreas do conhecimento como a Lógica, a Teoria dos Conjuntos, a Álgebra e a Teoria dos Grafos.

Desde o início da HPSG e da LFG, na metade da década de 1980, os pesquisadores estão envolvidos em implementações tecnológicas. A HPSG e a LFG se impõem no cenário lingüístico e lingüístico-tecnológico como teorias lingüísticas formais que, desde a sua origem, tiveram comprometi-

mento com descrições lexicalistas, formalmente precisas, computacionalmente tratáveis e monoestratais (i. e., não-transformacionais) dos fenômenos lingüísticos.

O nome "Head-driven Phrase Structure Grammar" reflete a importância atribuída pela teoria à informação codificada nos núcleos lexicais dos constituintes sintáticos. "Gramática de Estrutura de Constituintes Orientada para o Núcleo" parece ser uma tradução possível para o nome da teoria. (Neste trabalho, continuará sendo utilizada a sigla HPSG, mais difundida na comunidade acadêmica.) Trata-se de uma teoria que visa a especificar estruturas de constituintes a partir de um conjunto de informações especificadas sobre os itens lexicais que as integram, especialmente seus núcleos, ou seja, as relações de dependência estrutural são lexicalmente codificadas. Por exemplo, o argumento externo (sujeito) de um verbo é especificado (descrito), na HPSG, por um traço, chamado SUBJECT.

Os aspectos teóricos da HPSG foram desenvolvidos em detalhes em dois livros (Pollard & Sag, 1987, 1994) e em vários artigos[6] . Ginzburg & Sag (2000), além de descrever aspectos teóricos da HPSG, empregam o formalismo para a descrição de construções interrogativas. Algumas das idéias principais da teoria, e que serão desenvolvidas neste trabalho, são: a arquitetura baseada no signo lingüístico; a organização da informação lingüística através de tipos, hierarquias de tipos e heranças de restrições; a projeção de constituintes através de princípios gerais a partir de rica informação lexical; a organização da informação lexical através de sistemas de tipos lexicais; a fatoração de propriedades de constituintes em termos de restrições sobre construções específicas e gerais.

A HPSG pretende ser uma teoria explícita da competência lingüística humana, mas não se compromete em descrever como o conhecimento lingüístico está representado na mente do falante e nem como esse conhecimento é manipulado no processamento humano (mental) da linguagem. A HPSG foi concebida como uma teoria cujas descrições lingüísticas podem ser integradas a informações não-lingüísticas em diferentes modelos

6. Dois dos principais sites na *WEB* dedicados à divulgação da teoria estão sediados na *Stanford University*, com acesso através do endereço eletrônico <http://hpsg.stanford.edu/hpsg>, e na *Ohio State University*, com acesso através do endereço eletrônico <http://ling.ohio-state.edu/hpsg>.

de processamento da linguagem humana[7] (Pollard & Sag, 1994; Shieber, 1986; Cooper, 1996; Sag & Wasow, 1999), sejam esses modelos seriados, nos quais as informações percorrem uma seqüência de módulos, ou interativos, nos quais as informações não percorrem uma seqüência ordenada, podendo ser acessadas em qualquer ponto do processamento (Radford, 1999). Essa maleabilidade da HPSG tem contribuído para a elaboração de modelos interativos de processamento que entravam em conflito com teorias gramáticas, tal como a gramática de P&P, que assumem *a priori* a teoria da modularidade e a hipótese do processamento seqüencial da linguagem. Atualmente, resultados de experimentos psicolingüísticos[8] (Tanenhaus, 1995; Tanenhaus &Trueswell, 1995) e teorias gramaticais, tal como a HPSG, que não assumem *a priori* uma teoria do processamento têm contribuído para a construção de modelos interativos do processamento da linguagem humana com alto grau de refinamento e de consistência teórica (Sag & Wasow, no prelo).

A HPSG parte da hipótese de que a informação lingüística pode ser representada em termos de sistemas de tipos hierarquicamente organizados de objetos lingüísticos e a gramática das línguas pode ser representada em termos de sistemas de restrições especificadas por um conjunto de estruturas de traços.

A Figura 1 ilustra o conceito de classes, subclasses, tipos e hierarquias. Os elementos de alguns domínios de estudo podem ser organizados em classes, com base em similaridades de propriedades ou restrições que as caracterizam. Essas classes podem ser representadas formalmente em termos de tipos, i. e., entidades formais que descrevem explicitamente propriedades de objetos pertencentes a um determinado domínio e que são definidas em oposição a outros objetos num mesmo domínio. Algumas classes podem ser ainda subdivididas em subclasses que também compartilham similaridades ou restrições que as caracterizam. As subclasses também podem ser representadas formalmente em termos de tipos. No entanto,

7. O processamento lingüístico envolve vários tipos de informações extralingüísticas que se mesclam com informação lingüística.

8. Trata-se de experimentos que empregam recursos tecnológicos para captar movimentos oculares dos falantes durante o processo de interpretação dos enunciados e demonstram que informações lingüísticas ou não-lingüísticas são rapidamente acessadas na medida em que contribuam para a interpretação dos enunciados (Sag & Wasow, no prelo).

os tipos que representam subclasses apresentam restrições ou propriedades adicionais que não são identificadas nas classes a partir das quais tais subclasses são derivadas. Na HPSG, objetos lingüísticos são descritos formalmente em termos de tipos e suas propriedades são descritas formalmente em termos de conjuntos de traços, ou seja, cada tipo deverá prever a especificação de traços que representem tais propriedades.

Uma vez que as classes e subclasses que compõem um domínio podem ser representadas formalmente em termos de tipos, o domínio como um todo pode ser representado formalmente em termos de uma **hierarquia de tipos**, ou seja, uma representação formal do compartilhamento de propriedades e restrições entre classes e subclasses de objetos que compõem um domínio.

A Figura 1 ilustra, portanto, uma hierarquia de tipos de obras literárias segundo o seu gênero e sua origem. Uma obra literária pode ser classificada quanto ao gênero. Essa propriedade comum a um grupo de obras é representada formalmente pelo tipo ***gênero***. Um gênero literário compreende subgêneros se a obra for escrita em prosa, uma subclasse de gênero, ou em verso, outra subclasse. Essas subclasses diferem uma da outra e cada uma reúne um grupo de obras com propriedades semelhantes. Essas duas subclasses são representadas formalmente por dois tipos: ***prosa*** e ***verso***. O primeiro caracteriza-se por representar formalmente obras literárias do gênero prosa, que não apresentam metrificação nem versificação. Já o segundo caracteriza-se por representar formalmente obras literárias do gênero verso, que apresentam metrificação e/ou versificação. O gênero prosa pode ainda ser sub-classificado como ficcional ou não-ficcional. Essas duas subclasses da prosa podem ser representar formalmente por dois tipos: ***ficção*** e ***não-ficção***. Esses dois tipos representam formalmente obras do gênero prosa que fazem referência, respectivamente, a uma realidade imaginária ou uma realidade imediata observada por seu autor. Uma obra ficcional pode ser subclassificada como um romance ou um drama. Essas subclasses do gênero prosa ficcional podem ser representadas formalmente por dois tipos: ***romance*** e ***drama***. O primeiro tipo representa formalmente obras que não apresentam metrificação nem versificação; que fazem referência a uma realidade imaginária; e que tratam de um assunto histórico, lendário ou moral, quase sempre da tradição popular. O segundo tipo representa formalmente obras que apresentam metrificação e/ou versificação; que fazem referência a uma realidade imediata; e que podem ser

representada teatralmente. Uma obra escrita em verso pode ser subclassificada como lírica ou épica. Essas duas subclasses podem ser representadas formalmente por dois tipos: ***épica*** e ***lírica***. O primeiro tipo representa formalmente obras que narram feitos e fatos heróicos de personagens reais, lendários ou mitológicos. O segundo tipo representa formalmente obras que nas quais o poeta apresenta seus sentimentos, estados de espírito e percepções ao leitor.

Quanto à origem, uma obra literária pode ser produzida na América do Sul, na Europa ou ainda em outras localizações. Se produzida na América do Sul, pode ser escrita em português, espanhol ou outra língua específica. Se produzida na Europa, pode ser escrita em inglês, grego ou ainda outra língua específica. Essas informações poderiam ser representadas formalmente em termos de tipos e hierarquia de tipos.

Na HPSG, as hierarquias de tipos[9] foram introduzidas para a especificação de diferentes objetos lingüísticos, com diferentes graus de estruturação e, juntamente com as regras de redundância lexical, contribuem para a especificação da informação gramatical de modo simplificado e eficiente[10] (Flickinger & Pollard & Wasow, 1985), reduzindo ao mínimo a

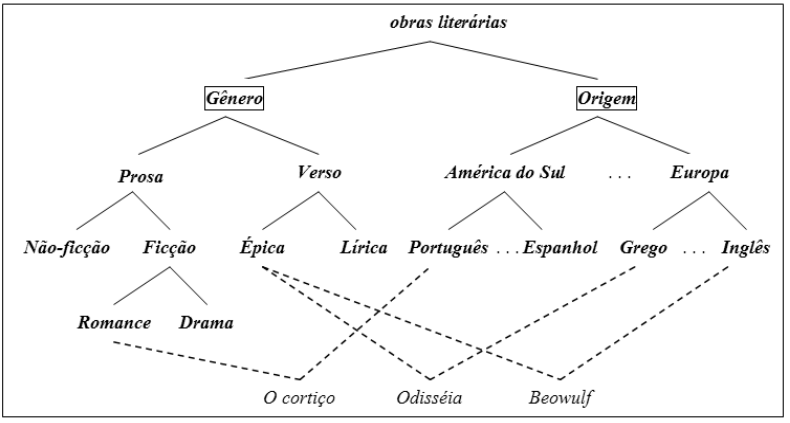

Figura 1: Descrição de obras literárias através
de uma hierarquia de tipos adaptada de Pollard e Sag (1987).

9. A respeito das hierarquias de tipos em lingüística, consulte-se Carpenter (1992).

10. A implementação desses mecanismos na HPSG contribuiu para uma redução considerável no número de regras gramaticais descritas pela sua antecessora, a GPSG (Flickinger & Pollard & Wasow, 1985).

especificação de regras gramaticais no modelo e impedindo que sejam especificadas regras redundantes, i. e., aplicáveis a vários itens lexicais.

A teoria especifica regras lexicais que utilizam informações de uma entrada lexical já existente no léxico como base para gerar outras entradas, para as quais são especificadas apenas propriedades particulares. As regras lexicais são utilizadas para gerar entradas lexicais preditivamente relacionadas segundo padrões recorrentes, tais como paradigmas flexionais (por exemplo, a flexão verbal), relações derivacionais (por exemplo, a nominalização) entre outros, como a derivação da forma passiva do verbo, ilustrada pela Figura 13 (Pollard & Sag, 1987).

Na Figura 2, por exemplo, os tipos descrevem três domínios ortogonais: (i) palavras (***word***) e constituintes sintagmáticos (***phrase***); (ii) partes-do-discurso (***part-of-speech***), tais como nome (***noun***), verbo (***verb***), adjetivo (***adjective***); e (iii) classes baseadas em valência, tais como transitividade (***transitive***) e intransitividade (***intransitive***).

As hierarquias de tipos caracterizam também uma **hierarquia de herança múltipla** de restrições. As restrições associadas às classes particulares são herdadas por suas subclasses e, em última instância, por seus membros individuais, ou seja, os subtipos herdam todas as restrições impostas sobre seus supertipos[11]. Em geral, os objetos lingüísticos perten-

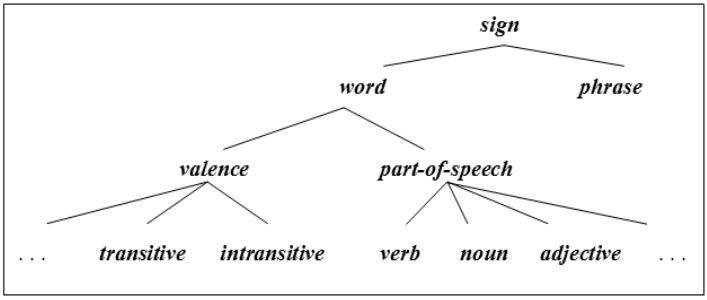

Figura 2: Representação de uma hierarquia de tipos.

11. Em Inglês, esse tipo de herança no qual as restrições associadas às classes particulares são herdadas por suas subclasses é chamado *default inheritance hierarchy* (Sag & Wasow, 1999). Neste trabalho, as hierarquias apresentadas são do tipo *default*, por isso não empregamos esse ou algum outro termo aportuguesado.

cem, ao mesmo tempo, a várias classes entre-cruzadas de tipos, daí o termo "herança múltipla". Por exemplo, uma palavra pode ser membro da classe dos verbos ou dos adjetivos e, ao mesmo, tempo ser membro das classes das palavras transitivas ou intransitivas, como demonstra a Figura 3.

As categorias lexicais[12] agregam generalizações com base em conjuntos de entradas lexicais que compartilham informações relacionadas, por exemplo, às suas propriedades semântica, morfológica, sintática. Com a utilização de hierarquias de heranças múltiplas, pretende-se evitar redundância no léxico, pois uma quantidade relativamente pequena de informações precisa ser especificada para cada entrada lexical dado que parte da informação é especificada através de compartilhamento de propriedades (Davis & Koenig, 2000). O verbo "amar", por exemplo, compartilha com outras expressões lingüísticas propriedades morfológicas comuns a objetos do tipo ***verb***[13] (que especifica parte-do-discurso) e propriedades sintáticas comuns a objetos do tipo ***transitive*** (que especifica informação valencial).

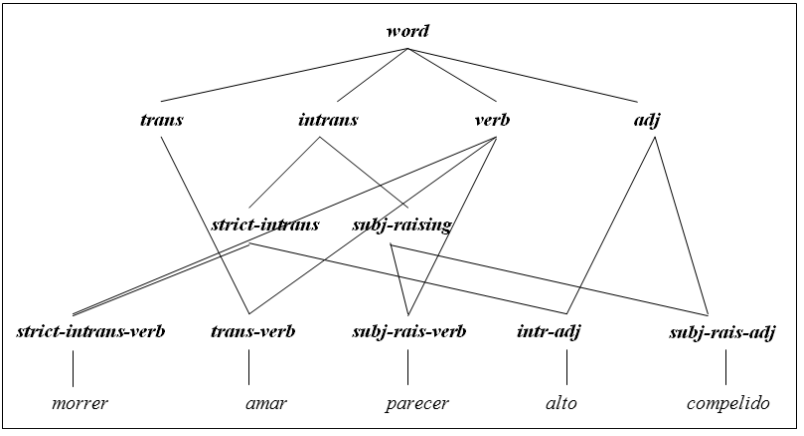

Figura 3: Hierarquia de herança adaptada de Davis e Koenig (2000).

12. Nomes, verbos e adjetivos incluem-se consensualmente no grupo das categorias lexicais. Determinantes, quantificadores e conjunções incluem-se consensualmente no grupo das categorias gramaticais. A inclusão das preposições e dos advérbios numa categoria ainda é pouco consensual (Neves, 2002).

13. Os tipos são representados pelo negrito itálico.

A entrada lexical de "amar", compartilhando informações de outras entradas, precisaria especificar apenas restrições sobre sua interpretação semântica, tal como está descrita na Figura 9.

Na HPSG, enquanto as hierarquias de tipos são utilizadas para a descrição de expressões lingüísticas, as **estruturas de traços**[14] são utilizadas para a descrição de restrições sobre essas expressões. Uma estrutura de traços é um objeto que descreve ou representa uma entidade qualquer através da especificação de *valores* para vários *atributos* da entidade descrita. Os traços são notações formais utilizadas para descrever formalmente um conjunto de informações lingüísticas. As estruturas de traços são anotadas tipicamente através de **matrizes de atributo-valor** (MAV). Um atributo pode ser pensado como um parâmetro mediante o qual as entidades podem ser diferenciadas e os valores identificam posicionamentos sobre os atributos. Por exemplo, um atributo OLHO poderia ser especificado para descrever a cor de olho. Dentre seus possíveis valores, estariam *azul*, *verde*, *castanho*, *preto*. Assim, na descrição de um conjunto de pessoas, algumas apresentariam o traço [OLHO *azul*], i. e., um atributo OLHO para o qual está especificado um valor *azul*, outras [OLHO *castanho*], e assim por diante. O termo traço designa o conjunto formado por um atributo, i.e., um parâmetro de diferenciação, e o valor a ele associado, p. ex., [OLHO *castanho*]. Os atributos são representados por maiúsculas, p. ex., OLHO, já os valores dos atributos são representados pelo itálico e ficam à direita do atributo.

A Figura 4 ilustra duas estruturas de traços. A Estrutura 1 fornece informação sobre alguém chamado João, cujo telefone é 9966-5544 e que trabalha na sala D27 de um departamento de vendas. Nessa estrutura de traços, os atributos são NOME, TELEFONE, SALA e DEPARTAMENTO, cujos valores são, respectivamente, *João*, *9966-5544*, *d27* e *vendas*. A Estrutura 2 fornece informação sobre alguém chamado Marcos, que trabalha como advogado, reside em Araraquara e cujo telefone é 5599-3322. Nessa estrutura de traços, os atributos são NOME, OLHOS, PROFISSÃO e CIDADE, cujos valores são, respectivamente, *Marcos*, *verde*, *advogado* e *Araraquara*.

14. A respeito das estruturas de traços, consulte-se Shieber (1986) e Pollard & Moshier (1990).

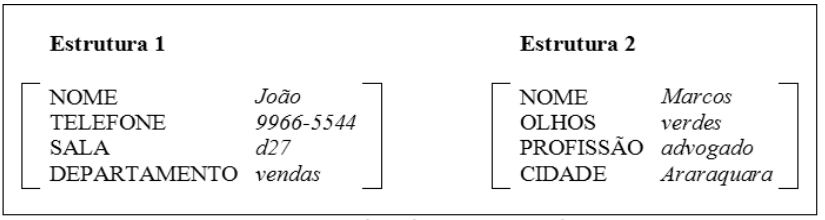

Figura 4: Exemplos de estruturas de traços.

Uma estrutura de traços fornece informação parcial sobre a entidade descrita. Na Estrutura 1 da Figura 4, por exemplo, são fornecidas informações sobre o nome, o telefone, a sala e o departamento de alguém chamado João, mas não são fornecidas informações sobre o seu sobrenome, o seu endereço ou sobre a cor de seus olhos.

O valor de um atributo em uma estrutura de traços pode ser atômico, tal como em [NOME *João*], na Figura 5, em que *João* é um valor atômico para o atributo NOME, ou ser especificado por outra estrutura de traços encaixada. Na Figura 5, o valor do atributo FÍSICO, especificado para a descrição de qualidades externas de uma pessoa (João), é outra estrutura de traços, que inclui os atributos PESO, ESTATURA e OLHOS, com seus respectivos valores, *70kg*, *1,80m* e *verde*.

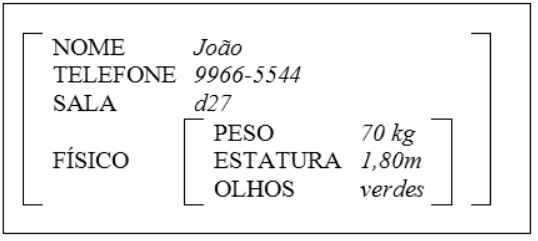

Figura 5: Exemplo de uma estrutura de traços encaixada.

Dois traços dentro de uma mesma estrutura de traços não podem se basear em um mesmo atributo, mas diferentes atributos podem compartilhar um mesmo valor, o que é representado por meio de uma etiqueta (numerada), por exemplo, **1, 2, 3, n,** permitindo o compartilhamento de

um mesmo valor entre atributos. Na Figura 6, o valor do atributo MANUTENÇÃO é outra estrutura de traços encaixada cujos atributos são NOVO e SEMINOVO, com seus respectivos valores, *gol* e *paraty*. A etiqueta **1** representa que o valor do atributo INDISPONÍVEL é o mesmo do atributo MANUTENÇÃO, evitando a repetição de informação redundante.

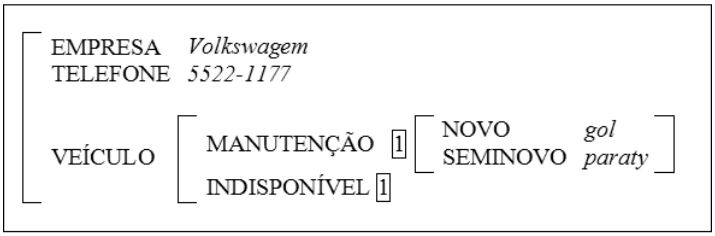

Figura 6: Exemplo de compartilhamento de estrutura.

A etiqueta representa um operador cuja função é copiar informações (i. e., valores atômicos ou estruturas de traços que funcionam como valores de atributos), que podem ser valores de vários atributos ao mesmo tempo, evitando que a mesma informação seja especificada várias vezes em uma estrutura de traços. Na descrição lingüística, esse recurso que permite representar a informação redundante de modo sucinto e econômico.

As estruturas de traços são parcialmente ordenadas por relações de inclusão, em um processo chamado **unificação**. Informalmente, a unificação é a operação de fusão de duas estruturas de traços para se obter uma terceira estrutura de traços que deve incluir toda a informação descrita pelas estruturas originais. A descrição de duas estruturas de traços pode unificar contanto que a informação descrita em ambas seja consistente, i. e, contanto que não haja atributos para o qual sejam especificados valores conflitantes (Shieber, 1986; Cooper, 1996; Sag & Wasow, 1999). O único impedimento para a unificação das estruturas 1 e 2 da Figura 4 são os valores conflitantes para o atributo NOME, ou seja, *João*, na Estrutura 1, e *Marcos*, na Estrutura 2. Havendo coincidência desses valores e, considerando-se que os demais traços descritos não causariam restrições à unificação, porque especificam atributos diferentes, seria possível a unificação das estruturas 1 e 2. Havendo vários atributos coincidentes entre estruturas de traços, só será possível a unificação dessas estruturas se os valores dos atributos coincidentes também forem coincidentes.

A informação descrita na Estrutura 1 da Figura 4, por exemplo, é consistente com a informação descrita na Figura 5, permitindo sua unificação, representada na Figura 7. A unificação consiste simplesmente de todos os atributos e valores especificados nas descrições das duas estruturas de traços.

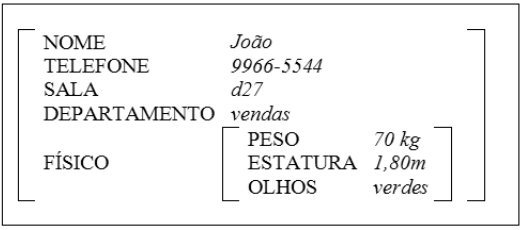

Figura 7: Unificação da estrutura 1 da Figura 4 com a estrutura da Figura 5.

Na HPSG, todo constituinte, seja lexical ou sintagmático, deve ser descrito por uma estrutura de traços classificada segundo os mecanismos que compõem a teoria: princípios gramaticais (o constituinte deve satisfazer cada constituinte gramatical), regras gramaticais (se o constituinte for sintagmático) e entradas lexicais (se o constituinte for lexical).

A HPSG, ampliando o conceito de *signo* de Saussure (1972), i. e., unidade mínima independente que relaciona, de modo arbitrário, forma e significado às palavras, sintagmas e pretende descrever um complexo estruturado de informações fonéticas, sintáticas, semânticas e restrições contextuais em termos de matrizes de estruturas de traços (Pollard & Sag, 1994). Na HPSG, as estruturas de traços são tipadas, porque descrevem diferentes grupos de estruturas lingüísticas que compartilham das mesmas restrições (cf. hierarquias de tipos). Além disso, essas estruturas de traços são representadas como MAVs, como ilustra a Figura 8.

A descrição de um objeto lingüístico do tipo ***word*** deve ter, no mínimo, dois atributos, PHON (PHONOLOGY) e SYNSEM (SYNTAX-SEMANTICS). O valor de PHON são suas descrições fonológicas e o valor de SYNSEM é outro objeto estruturado (i. e., uma estrutura de traços) do tipo ***synsem*** que descreve um conjunto de informações sintáticas e semânticas. Signos do tipo ***phrase***, que descrevem estruturas sintagmáticas, além de PHON e SYNSEM também possuem atributos DAUGHTER (DTRS),

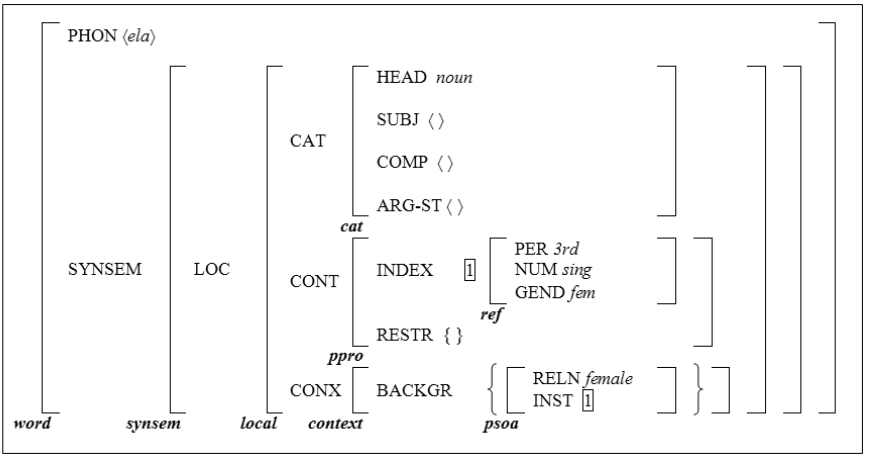

Figura 8: Exemplo de uma matriz de atributo-valor (MAV) representando o signo lexical *ela* (POLLARD; SAG, 1994).

cujo valor é uma estrutura de traços que representa a estrutura de constituintes imediatos da frase. O valor do atributo SYNSEM codifica a sintaxe e a semântica de um constituinte. As categorias sintáticas na HPSG também são representadas em termos de estruturas de traços complexas, cujo atributo é LOCAL (LOC). O atributo LOCAL descreve construções que não apresentam deslocamentos de constituintes[15] . O valor de LOC também é um objeto estruturado chamado ***local***, cujos atributos são CATEGORY (CAT), CONTENT (CONT) e CONTEXT (CONX). CAT é um objeto estruturado do tipo ***category*** (***cat***) que possui os atributos HEAD, SUBJECT (SUBJ), COMPLEMENTS (COMPS) e ARGUMENT-STRUCTURE (ARG-ST). O valor HEAD de um signo especifica sua parte-do-discurso, i. e., a categoria da palavra. O atributo ARG-ST especifica a estrutura de argumentos da palavra, ou seja, informação sobre dependentes que são mais ou menos idiossincraticamente requeridos pela palavra, já os atributos SUBJ e COMPS especificam a realização sintática desses dependentes na função de sujeito ou complemento, respectivamente.

15. O atributo NONLOCAL, a contrapartida do atributo LOCAL e irrelevante para a descrição do pronome *ela* descreve construções analisadas pelas gramáticas transformacionais em termos de movimentos-*qu* (p. ex., *I wonder* [who_i *Mary loves* $\_\_\__i$]) ou movimentos para a posição não-argumental de especificador do sintagma complementizador (p. ex., Who_i *have they kissed* $\_\_\__i$?) (Pollard & Sag, 1994).

O pronome *ela*, descrito na Figura 8, por exemplo, é representado por uma matriz de atributo-valor introduzida por um objeto do tipo ***word*** que especifica os atributos PHON e SYNSEM. O valor de PHON é a representação fonética[16] do pronome *ela* e o valor de SYNSEM é outra estrutura de traços introduzida por um objeto do tipo ***synsem*** que especifica o atributo LOCAL, cujo valor é uma estrutura de traços introduzida por um objeto do tipo ***local*** que especifica os atributos CAT, CONT e CONX. O valor do atributo CAT é outra estrutura de traços, nesse caso, introduzida por um objeto do tipo ***cat*** que especifica os atributos HEAD, SUBJ, COMP e ARG-ST. Os valores dos três últimos atributos SUBJ, COMP e ARG-ST são listas (representadas por colchetes angulares "á ñ") vazias e o valor atômico *noun* de HEAD especifica que a palavra pertence à categoria nome. O valor do atributo CONT é uma estrutura de traços introduzida por um objeto do tipo ***ppro*** que especifica os atributos INDEX e RESTR. O valor de RESTR é um conjunto (representado por chaves "{}") vazio. O valor de INDEX é uma estrutura de traços do tipo ***ref*** que introduz os atributos PER, NUM e GEND, cujos valores atômicos são, respectivamente, *3rd*, *sing* e *fem* que especificam, respectivamente, terceira pessoa do discurso, singular e gênero feminino. O valor do atributo CONX é uma estrutura de traços introduzida por um objeto do tipo ***context*** que especifica o atributo BACKGR, cujo valor também é uma estrutura de traços introduzida por um conjunto (representado por chaves "{}") de objetos do tipo ***psoa*** que introduz os atributos RELN e INST. O valor atômico de RELN, *female*, especifica que o pronome refere-se a alguém do sexo feminino e o valor de INST é um objeto do tipo ***ref*** que introduz os atributos PER, NUM e GEND, cujos valores atômicos são, respectivamente, *3rd*, *sing* e *fem*. A etiqueta **1** indica que o valor do atributo INDEX é uma estrutura de traços compartilhada que também funciona como valor do atributo INST. Uma estrutura de traços será adequada segundo os tipos e os traços que ela especifica. O traço NONLOCAL(cf. nota 14), p. ex., é inadequado para a descrição do pronome *ela*.

Na HPSG, uma vez que a sintaxe e a semântica são tratadas no mesmo nível de representação, e ambas estão sujeitas ao mesmo conjunto de restrições, não se postula um nível intermediário de representação semântica entre a análise sintática e a interpretação contextual, como a forma

16. Tendo em vista que neste trabalho não há preocupação com aspectos fonéticos, convencionou-se representar ortograficamente os valores de PHON.

lógica da teoria de P&P (Meulen, 1993), pois, na HPSG, as informações contextuais também são representadas no mesmo nível que a sintaxe e a semântica. A estrutura de traços é utilizada para representar todos os níveis de informação lingüística. A interpretação semântica de um constituinte está representada pelo valor do atributo CONT[17] . Esse valor é um objeto estruturado, chamado ***personal-pronoun*** (***ppro***) no caso do pronome *ela*, que inclui os atributos INDEX e RESTRICTION (RESTR), e representa a contribuição da palavra para a interpretação semântica do sintagma no qual ela se encontra. A Semântica de Situações[18] é a teoria semântica subjacente aos valores de CONT (Pollard & Sag, 1994; Sag & Wasow, 1999). O valor do atributo INDEX, ***reference*** (***ref***), é um objeto estruturado que especifica a contribuição do signo em relação a aspectos (semânticos) referenciais. O conteúdo de um signo pode introduzir uma restrição semântica sobre o seu índice, que corresponde a uma situação ou a um indivíduo a que se refere uma restrição. Quando presente (no caso do pronome "ela", não há esse tipo de restrição), essa restrição será representada pelo valor do atributo RESTR, caracterizado por uma lista (entre chaves) de condições que a situação ou indivíduo deve satisfazer. RESTR, na Figura 9, corresponde a uma predicação no cálculo de predicado, por exemplo **amar'(x,y)**.

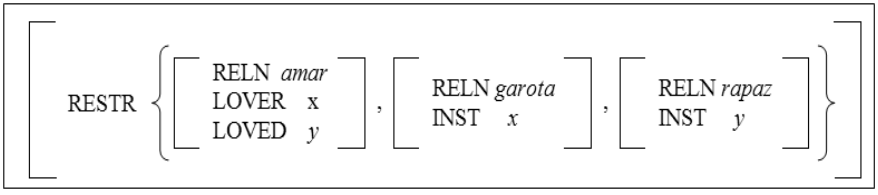

Figura 9: Representação de uma lista de condições especificadas pelo atributo RESTRICTION.

O valor do atributo CONX contém determinadas informações lingüísticas referentes a aspectos contextuais, tais como pressuposição, impli-

17. O traço CONT fornece informação que constitui uma espécie de representação de uma forma lógica.

18. A respeito da Semântica de Situações, consulte-se Barwise & Perry (1983).

cação etc.[19] . O atributo RELATION (RELN), na Figura 10, especifica o conteúdo do constituinte em questão ou o tipo de relação envolvida em uma predicação. Esse atributo corresponde a um predicado na lógica de predicados. Um predicado denota um indivíduo ou objeto, uma propriedade ou uma relação, por exemplo, à palavra *grande* corresponde o predicado **grande'(x)**, que denota a propriedade "ser grande"; à palavra *amar* corresponde o predicado **amar'(x,y)**, que denota uma relação em que "x ama y"[20] .

As restrições associadas a muitos nomes (que se referem a indivíduos ou objetos, como *homem, livro, chapéu*) e adjetivos (que se referem a propriedades de indivíduos ou objetos, como *grande, alto, feliz*) incluem a predicação de apenas um argumento. Por exemplo, na função **grande'(x)**, que gera proposições como *o homem é grande*, *a casa é grande*, e assim por diante, a variável x, que pode equivaler à denotação de *homem* ou à denotação de *casa*, é especificada, na HPSG, através do atributo INSTANCE (Figura 10). As restrições associadas a verbos (que descrevem situações *amar, dar*) incluem outros tipos de predicação. Para essas predicações são especificados os papéis de quem ou do que participa da relação de predicação, i. e., são especificados os papéis dos argumentos, na predicação. A função proposicional **amar'(x,y)** gera proposições como *o pai ama o filho*, *o marido ama sua mulher*, e assim por diante. As variáveis x e y que podem equivaler, respectivamente, às denotações de *pai* ou *marido* e à denotação de *filho* ou *mulher*, são especificadas através dos atributos LOVER e LOVED, distinguindo, respectivamente, quem ama de quem é amado (Figura 10).

Como observa Cooper (1996), embora a HPSG seja uma gramática de estrutura de constituintes, os constituintes descritos pela teoria não são licenciados por regras gramaticais de reescrita (Gazdar *et al*, 1985) ou transformacionais (Chomsky, 1957). Na HPSG, as estruturas de traços tipadas, através do processo de unificação de estruturas, descrevem constituintes com diferentes graus de estruturação. As poucas regras gramaticais da teoria são definidas em termos de Esquemas de Dominância Imediata (EDI), tal como o EDI para estruturas envolvendo adjuntos, na Figura 11. Esse

19. O atributo CONTEXT também pode conter informação lingüística que se liga a aspectos dependentes de contexto da interpretação semântica (Pollard & Sag, 1996).

20. O apóstrofo indica que se trata de um predicado.

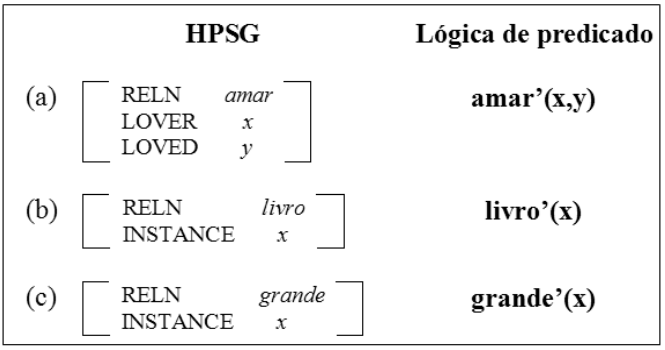

Figura 10: Representação de restrições semânticas
na HPSG e na lógica de predicados.

EDI equivale aproximadamente à regra de reescrita **XP → X(núcleo), X-ADJUNTO**.

O EDI da Figura 11 descreve uma estrutura do tipo ***headed-adjunct-structure***. Segundo esse EDI, o adjunto (X-ADJUNTO), o núcleo do sintagma (X) e sua projeção máxima (XP) são descritos por estruturas de traços e a estrutura que descreve o adjunto deve especificar que ele modifica o núcleo[21] . Essa especificação se dá através do atributo MODIFIER (MOD) e seu valor, um objeto estruturado do tipo ***synsem*** que descreve o núcleo. Esse compartilhamento está representado na Figura 11[22] através da etiqueta **3**.

A Figura 11 também formaliza dois princípios gramaticais em termos de estruturas de traços que devem ser satisfeitos por uma estrutura de adjunção: o Princípio de Traço de Núcleo (PTN) e o Princípio Semântico (PS). O PTN está baseado no fato de que as propriedades sintáticas de um sintagma são determinadas por seu núcleo e que, por isso, ambos devem pertencer à mesma categoria. A finalidade do princípio é garantir que os sintagmas sejam projeções de seus núcleos. O valor do atributo HEAD

21. Essa análise baseia-se no tratamento dado aos adjuntos pela Gramática Categorial, segundo a qual, os adjuntos são funções que tomam núcleos como seus argumentos (Pollard & Sag, 1994).
22. As setas e as indicações **XP**, **X (núcleo)** e **X-ADJUNTO** não são parte formal da árvore. Estão sendo empregadas apenas para indicar como os constituintes estão relacionados.

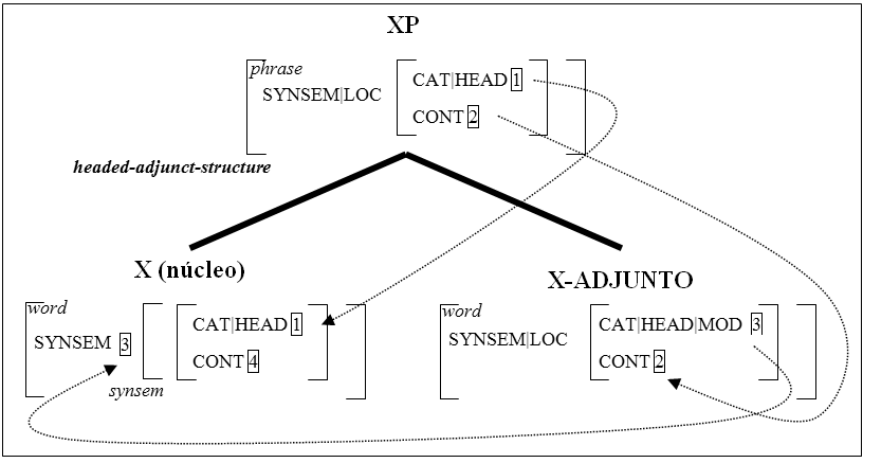

Figura 11: EDI de estruturas de adjunção segundo Pollard e Sag (1994).

que especifica a categoria do núcleo deverá ser o mesmo do atributo HEAD de sua projeção máxima. Na Figura 11, esse compartilhamento de valores está representado pela etiqueta **1**. Se o valor do atributo HEAD do núcleo for, por exemplo, *noun*, o valor do atributo HEAD de sua projeção máxima também o será. Segundo o PS, em se tratando de uma estrutura de adjunção, as propriedades semânticas de um sintagma são determinadas pelo adjunto. O valor do atributo CONT que especifica o conteúdo semântico do adjunto deverá ser o mesmo do atributo CONT do sintagma, i. e., da estrutura do tipo ***headed-adjunct-structure***. Na Figura 11, esse compartilhamento de valores está representado pela etiqueta **2**. Se o valor do atributo CONT do adjunto for, por exemplo, a estrutura (c) da Figura 10, o valor do atributo CONT do sintagma também o será[23] .

Na HPSG, as regras gramaticais (= estruturas de traços) não restringem a precedência linear dos constituintes, estabelecem apenas relações de dominância imediata entre eles. Pollard & Sag (1987)[24], reconhecendo que as línguas apresentam restrições específicas sobre a ordenação linear

23. Pollard & Sag (1994) argumentam que, em outros tipos de sintagmas, o núcleo semântico e o núcleo sintático coincidem, ou seja, o valor do atributo HEAD e o valor do atributo CONT do sintagma são idênticos aos valores dos atributos HEAD e CONT do núcleo.

24. Pollard & Sag (1987) discutem algumas questões relacionadas à ordenação de constituintes, já Pollard & Sag (1994) assumem que não tratam dessas questões.

dos constituintes, propõem um Princípio de Ordenação de Constituintes (POC) específico para cada língua[25] . No inglês, por exemplo, um princípio geral de linearização de constituintes prevê que os núcleos lexicais precedem seus complementos, o que está formalmente expresso na Figura 12 ("<" indica "precede"). O atributo LEXICAL (LEX) e seus valores booleanos[26] + e – especificam o caráter lexical ou sintagmático de um *signo*[27] .

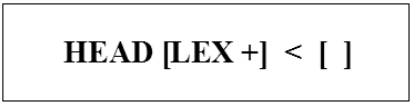

Figura 12: Princípio geral de linearização para o inglês.

Ressalte-se que o POC da Figura 12 coloca um problema para a representação do adjetivo, no inglês. Nessa língua, a ocorrência dos adjetivos é predominantemente anteposta ao núcleo nominal, contrariando o POC da Figura 12, requerendo um POC específico para a descrição da ordenação dos adjetivos no sintagma nominal nessa língua (Pria, 2005).

A abordagem lexicalista da HPSG orienta-se por poucas regras sintagmáticas e por um léxico rico e complexo. Além de ser tratada em termos de uma estrutura de traços e de hierarquias de heranças múltiplas, a informação lexical também é organizada em termos de **regras de redundância lexical**, com a finalidade de que sejam feitas generalizações ade-

25. A distinção de esquemas de dominância imediata e de informações sobre precedência linear em gramáticas de estrutura de frase, amplamente aceita por teorias monoestratais como a HPSG, tem sua origem nas primeiras versões da GPSG (Gazdar *et al*, 1985). Segundo Kasper & Kathol & Pollard (1995), o motivo para tal distinção é expressar generalizações sobre ordem linear presentes em diferentes esquemas de dominância imediata.

26. As álgebras booleanas foram definidas pelo matemático inglês George Boole (1915-1864) e constituem uma tentativa de manipular expressões no cálculo proposicional. A álgebra booleana trabalha com apenas duas grandezas e os símbolos "+" e "-" são utilizados respectivamente para representar presença e ausência de uma propriedade. Na Figura 12, a álgebra booleana é utilizada para representar presença ou ausência do caráter lexical ("+") ou sintagmático ("-") de um *signo*.

27. Esse traço não aparece nas estruturas de traços de Pollard & Sag (1994) porque os autores não tratam de questões de ordenação de constituintes. Apenas sugerem que o traço LEXICAL deve ser um dos valores de ***category***.

quadas e previsíveis sobre propriedades lingüísticas. A Figura 13 ilustra (Pollard & Sag, 1987) a regra lexical utilizada na geração de uma forma verbal passiva a partir da entrada lexical da forma ativa de um verbo existente no léxico. A regra descrita na Figura 13 aplica-se de modo redundante e generalizado na geração de formas passivas a partir de verbos transitivos.

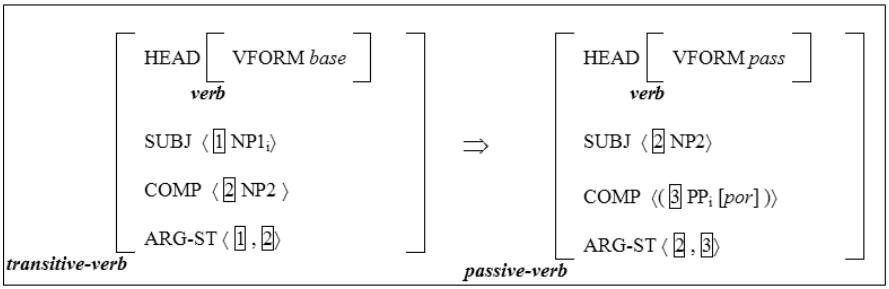

Figura 13: Regra lexical da forma passiva adaptada de Pollard e Sag (1996).

A regra aplica-se a um objeto do tipo ***transitive-verb*** e deriva um objeto do tipo ***passive-verb***. O primeiro especifica uma estrutura de traços cujos atributos são HEAD, SUBJ, COMP e ARG-ST[28] . O valor de HEAD também é um objeto estruturado do tipo *verb* que especifica o atributo VFORM, cuja finalidade é descrever a forma do verbo. O segundo especifica uma estrutura de traços cujos atributos são HEAD, SUBJ, COMP e ARG-ST. O valor de HEAD também é um objeto estruturado do tipo *verb* que especifica o atributo VFORM, cujo valor é *passive* (*pass*), indicando que o verbo deve estar na sua forma passiva. A regra lexical da forma passiva especifica que, em uma frase como [*O menino*]$_{NP1}$ *comeu* [*o bolo*]$_{NP2}$, o sintagma nominal sujeito (NP_1) da forma básica torna-se um sintagma (complemento) preposicionado (PP) introduzido pela preposição *por*, enquanto o sintagma nominal complemento (NP_2) da forma básica torna-se o sujeito da forma passiva[29] . Como resultado, obtém-se [*O bolo*]$_{NP2}$ *foi comido* [*pelo menino*]$_{PP}$.

As regras de redundância, associadas às hierarquias de heranças, não só contrariam a impossibilidade de atribuição estruturação ao léxico (Di

28. Na Figura 13, não estão especificados os traços que não sofrem alteração pela regra lexical.
29. NP e PP estão sendo utilizados para abreviar descrições em termos de estruturas de traços.

Sciullo & Williams, 1987), mas também descreve a informação lexical de modo estruturado e elegante, como se almejava na década de 1970 (Jackendoff, 1975), atendendo ainda requisitos formais, necessários à sua implementação em sistemas de PLN (Sanfilippo, 1995; Handke, 1995; Pollard, 1997; Pollard & Sag, 1996).

Embora não seja objetivo deste trabalho apresentar uma revisão crítica do modelo HPSG, e sim apresentá-lo como uma alternativa lexicalista para a descrição gramatical, mencionem-se as críticas à capacidade gerativa irrestrita das regras de redundância em alguns casos (Hinrichs & Nakazawa, 1996; Briscoe & Copestake, 1999). Se, por um lado, esse problema coloca um desafio pra a teoria, por outro, a busca pela sua solução tem motivado grande parte das pesquisas no âmbito da HPSG, atualmente.

5. Considerações finais

A construção de sistemas de PLN lingüisticamente motivados requerer, além da investigação e sistematização das diversas diversões (fonológica, morfológica, sintática, semântica e discursiva) do conhecimento lingüístico, a modelagem formal desse conhecimento através de modelos matemáticos e passíveis de implementação computacional.

Neste trabalho, a HPSG foi apresentada (i) como uma teoria lingüística de base lexicalista que se propõe a investigação e a sistematização do conhecimento necessário ao uso das línguas naturais e, ao mesmo tempo, (ii) como um modelo formal que atende às exigências de modelagem formal (através da lógica de traços e da hierarquia de heranças) necessária à interpretação e síntese do conhecimento lingüístico em sistemas de PLN.

Atualmente, as implementações disponíveis com base na HPSG compreendem analisadores morfossintáticos (*parsers*) para o inglês (Carpenter & Penn, 1998), alemão (Müller, 1996, 2004), koreano (Kim & Yang, 2004), japonês (Fujinami, 1996) e búlgaro (Simov *et al*, 2005). Uma das limitações da teoria/modelo ainda é descrever regras gerais o bastante que possibilitem a sistematização, formalização e, conseqüentemente, a implementação de um grande número de línguas. Embora a teoria tenha ultrapassado os limites da *Stanford University*, nos EUA, e seja uma das teorias/modelos voltados para PLN mais difundidos na Europa e no Orien-

te, sua divulgação e emprego ainda são incipientes no Brasil. Assim, este trabalho pretende dar as boas novas para pesquisadores que trabalham no âmbito do PLN.

Recebido em setembro de 2006
Aprovado em novembro de 2007
E-mail: adallapria@yahoo.com

Referências Bibliográficas

Arnold, D.J. & L. Balkan & S. Meijer & R. L. Humphreys & L. Sadler. 1994. *Machine Translation*: an Introductory Guide. London: Blackwells. Disponível em <http://www.essex.ac.uk/linguistics/clmt/MTbook/>. Acesso em: 02 agosto 2007.

Barwise, J. & J. Perry.1983. *Situations and attitudes*. Cambridge: The MIT Press.

Bloomfield, L. 1933. *Language*. Nova York: Holt, Rinehart e Winston.

Bouma, G. & F. V. Eynde & D. Flickinger. 2000. Constraint-based Lexica. IN: F. V. Eynde & D. Gibbon. *Lexicon Development for Speech and Language Processing*. Dordrecht-Boston-London: Kluwer. Disponível em <http://www.essex.ac.uk/linguistics/clmt/papers/hpsg/bouma-etal.pdf>. Acesso em: 02 agosto 2007.

Bresnan, J. Ed. 1982. *The mental representation of grammatical relations*. Cambridge: The MIT Press.

Bresnan, J. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell Publishers.

Briscoe, T. & A. Copestake. 1999. Lexical rules in constraint-based grammar. *Computational Linguistics,* 25(4): 487-526. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/27853/http:zSzzSzacl.ldc.upenn.eduzSzJzSzJ99zSzJ99-4002.pdf/briscoe99lexical.pdf>. Acesso em: 02 agosto 2007.

Briscoe, T. 1991. Lexical issues in natural language processing. IN: E. Klein & F. Veltman Eds.. *Natural language and speech*. Berlim: Springer-Verlag. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/1084/http:zSzzSzwww.cl.cam.ac.ukzSzftpzSzpaperszSzacquilexzSzacq1wp41.pdf/briscoe91lexical.pdf>. Acesso em: 02 agosto 2007.

Carpenter, B. & G. Penn. 1998. *ALE 3.1 User'sManual*. Disponível em <http://www.sfs.nphil.uni-tuebingen.de/_gpenn/ale.html>. Acesso em: 02 agosto 2007.

CARPENTER, B. 1992. *The logic of typed feature structure*. Cambridge: Cambridge University Press.

CHOMSKY, N. 1965. *Aspects of the theory of syntax*. Cambridge: The MIT Press.

_____. 1981. *Lectures on government and binding*. Dordrecht: Foris.

_____. 1970. Remarks on nominalizations. IN: R. A. JACOBS & P. S. ROSENBAUM. *Readings in English transformational grammar*. Waltham: Ginn and Company.

_____. 1957. *Syntactic structures*. Haia: Mouton.

COOPER, R. 1996. Head-driven phrase structure grammar. IN: K. BROWN & J. MILLER. *Concise encyclopedia of syntactic theories*. Oxford: Elsevier Science. Disponível em <http://www.psyc.bbk.ac.uk/people/academic/cooper_r/publications/encyc2.pdf>. Acesso em: 02 agosto 2007.

DAVIS, A. R. & J. KOENIG. 2000. Linking as constraints on word classes in a hierarchical lexicon. *Language*, 76 (1): 56-91. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/6054/http:zSzzSzwings.buffalo.eduzSzsoc-scizSzlinguisticszSzkoenigzSzlinking.pdf/davis99linking.pdf>. Acesso em: 02 agosto 2007.

DI SCIULLO, A. M. & E. WILLIAM. 1987. *On the definition of word*. Cambridge: The MIT Press.

FLICKINGER, D. & C. POLLARD & T. WASOW. 1985. Structure sharing in lexical representation. IN: *Proceedings of the 23rd annual meeting of the association for computational linguistics*. Morristown, N.J.: Association for computational linguistics. Disponível em <http://acl.ldc.upenn.edu/P/P85/P85-1032.pdf>. Acesso em: 02 agosto 2007.

FODOR, J. 1983. *The modularity of mind*. Cambridge: The MIT Press.

FUJINAMI, Y. 1996. An implementation of Japanese Grammar based on HPSG. Master's thesis. Department of Artificial Intelligence. Edinburgh University. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/4775/ftp:zSzzSzftp.ims.uni-stuttgart.dezSzpubzSzcufzSzenglish_paperszSzyoshiko.pdf/fujinami96implementation.pdf>. Acesso em: 02 agosto 2007.

GAZDAR, G. et al. 1985. *Generalized phrase structure grammar*. Cambridge: Harvard University Press.

GIBBON, D. 2000. Computational lexicography. IN: F. VAN EYNDE & D. GIBBON. Eds.. *Lexicon development for speech and language processing*. Dordrecht: Kluwer Academic Publishers.

GINZBURG, J. & I. SAG. 2000. *Interrogative Investigations*: the form, meaning, and use of English interrogatives. Stanford: CSLI.

GRISHMAN, R. 1980. Conjunctions and modularity in language analysis procedures. IN: *Proceedings of the 8th international conference on computational linguistics*. Tokyo: Association for Computational Linguistics, p. 500-503.

_____. & N. CALZOLARI. Lexicons. IN: R. A. Cole. Ed.. 1995. *Survey of the State of the Art in Human Language Technology*. Cambridge: Cambridge University Press.

HANDKE, J. 1995. *The structure of the lexicon*: human versus machine. Berlim: Mouton de Gruyter.

HINRICHS, E. W. & T. NAKAZAWA. 1996. Applying lexical rules under subsumption. IN: *Proceedings of the 16th conference on Computational linguistics – volume 1*. Copenhagen: Association for Computational Linguistics, p. 543–549. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/27847/http:zSzzSzacl.ldc.upenn.eduzSzCzSzC96zSzC96-1092.pdf/hinrichs96applying.pdf>. Acesso em: 2 agosto 2007.

JACKENDOFF, R. The base rules for prepositional phrases. IN: S. ANDERSON & P. KIPARSKY. Eds.. 1972. *A festschrift for Morris Halle*. New York: Hol, Rinchart e Winston.

JACKENDOFF, R. 1975. Morphological and semantic regularities in the lexicon. *Language*, **51** (3): 639-675.

KASPER, R. T. & A. KATHOL & C. J. POLLARD. 2003. Linear precedence constraints and reentrancy. IN: J. KILBURY & R. WIESE. Eds.. *Integrative Ans atze in der Computerlinguistik*. 1995. Disponível em: <http://citeseer.nj.nec.com/kasper95linear.html>. Acesso em: 10 setembro 2003.

KIM, J. & J. YANG. 2004. Projections from Morphology to Syntax in the Korean Resource Grammar: Implementing Typed Feature Structures. IN: *Lecture Notes in Computer Science*, v. 2945: 13-24. Disponível em <http://web.khu.ac.kr/~jongbok/research/final-papers/projecdtion.pdf>. Acesso em: 2 agosto 2007.

MALOUF, R. 2003. Categories, prototypes, and default inheritance. IN: *Proceedings of the Joint Conference on Formal Grammar, Head-Driven Phrase Structure Grammar, and Categorial Grammar*. 1998. p. 207-216. Disponível em: <http://odur.let.rug.nl/~malouf/papers/fhcg.pdf >. Acesso em: 10 junho 2003.

Fellbaum, C. Ed. 1998. *Wordnet*: a lexical reference system and its applications. Cambridge, Massachussets: The MIT Press.

Meulen, A. 1993. Linguistics and philosophy of language. IN: F. J. Newmeyer. Ed.. *Linguistics*: the Cambridge survey. I Linguistic theory: foundations. Cambridge: Cambridge University Press.

Müller, S. 1996. The Babel-system: An HPSG Prolog implementation. IN: *Proceedings of the 4th International Conference on the Practical Application of Prolog*. London: Practical Application Company Limited, p. 263–277.

_____. 2004.The Babel-System: a parser for an HPSG fragment of German. Disponível em <http://www.cl.uni-bremen.de/stefan/PS/babel.pdf>. Acesso em: 02 agosto 2007.

Neves, M. H. de M. 2002. *A gramática*: história, teoria e análise, ensino. São Paulo: Editora UNESP.

Pollard, C. 2003. *Lectures on the foundations of HPSG*. Ohio: Ohio State University, 1997. Disponível em: <http://www-csli.stanford.edu/~sag/L221a/cp-lec-notes.pdf>. Acesso em: 10 junho 2003.

_____. & D. Moshier.1990. Unifying partial descriptions of sets. *Information, language and cognition. Vancouver studies in cognitive science*, v. 1. Vancouver: University of British Columbia Press.

_____. & I. Sag. 1994. *Head-driven phrase structure grammar.* Stanford: CSLI/ The University of Chicago Press.

_____. 1996. *HPSG*: background and basics. 22p. Trabalho não publicado. Disponível em <http://citeseer.ist.psu.edu/cache/papers/cs/6350/http:zSzzSzwww.ling.helsinki.fizSzcourseszSzctl250zSzctl250_ht97zSzhpsg-overview.pdf/sag96hpsg.pdf>. Acesso em: 2 agosto 2007

_____. 1987. *Information-based syntax and semantics, volume. 1*. Stanford: CSLI Stanford University/University of Chicago Press.

Pria, A. D. 2005. *Uma proposta de representação lingüístico-computacional do comportamento sintático e semântico de adjetivos no sintagma nominal do inglês e do português*. Dissertação (Mestrado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras da Universidade Estadual Paulista, Araraquara.

Radford, A. et al. 1999. *Linguistics*: na in troduction. Cambridge: Cambridge University Press.

Sag, I. A. & T. Wasow. 1999. *Syntactic theory*: a formal introduction. Stanford: CSLI.

Sag, I. & T. Wasow. No prelo. Performance-Compatible Competence Grammar. IN: R. D. Borsley & K. Börjars. Eds.. *Non-Transformational Syntax*. Oxford: Blackwell.

Sanfilippo, A. 1995. Lexicons for constraint-based grammars. IN: R. A. Cole. Ed. *Survey of the state of the art in human language technology*. Oregon: Graduate Institute. Disponível em <http://cslu.cse.ogi.edu/HLTsurvey/ch3node6.html>. Acesso em: 02 agosto 2007.

Saussure, F. de. 1972. *Curso de lingüística geral*. 2. ed. São Paulo: Cultrix/ Editora da USP.

Sells, P. 1985. *Lectures on contemporary syntactic theories*. Stanford: CSLI Publications.

Shieber, S. M. 1986. *An introduction to unification-based approaches to grammar*. Stanford: CSLI.

Simov, K. & P. Osenova & A. Simov & M. Kouylekov. 2005. Design and Implementation of the Bulgarian HPSG-based Treebank. *Research on Language & Computation*, 2(4): 495-522.

Steedman, M. 2000. *The Syntactic Process*. Cambridge: MIT Press.

Tanenhaus, M.K. & J.C. Trueswell. 1995. Sentence comprehension. IN: J. Miller & P. Eimas. Eds. *Handbook of Perception and Cognition Vol. 11*: Speech and Language. New York: Academic Press.

Tanenhaus, M. & M. Spivey-Knowlton & K. Eberhard & J. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*. 268: 1632-1634.

Wood, M. M. 1993.*Categorial Grammars*. Routledge: London.