Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista – UNESP/Ar.

Modelo de entrada lexical (psico)lingüístico-computacional

Ariani Di Felippo
Bento Carlos Dias-da-Silva

NILC-TR-04-05

Maio, 2004

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste trabalho, apresenta-se um modelo de entrada lexical (psico)lingüístico-computacional. Mais especificamente, propõe-se um modelo de entrada lexical, com vistas ao Processamento Automático das Línguas Naturais (PLN), baseado em hipóteses sobre a estruturação interna das entradas no *léxico mental*.

Este trabalho contou com o apoio
financeiro da CAPES (2001-2003).

Centros de Estudos Lingüísticos e da Computação - **CELiC**

unesp
UNIVERSIDADE ESTADUAL
PAULISTA

Sumário

1. Introdução

Atualmente, o "léxico" tem ocupado lugar de centralidade nos estudos do domínio da Psicolingüística, da Lingüística e do Processamento Automático das Línguas Naturais (NLP) (Handke, 1995).

Nos estudos psicolingüísticos, vários modelos cognitivos têm afirmado a centralidade do *léxico mental* no processamento da linguagem. Entende-se por **léxico mental** (LM) a parte do conhecimento lexical do falante determinada pela estrutura da língua (Bierwisch e Schreuder, 1992). Dentre os principais tópicos investigados sobre o "léxico mental", estão: a estrutura interna (ou microestrutura) e global (ou macroestrutura) do LM e as estratégias de acesso aos itens (no LM).

Nos estudos lingüísticos de orientação gerativa, em específico, o "léxico" deixa de ser o vocabulário da língua como realidade externa; o objeto de estudo do lingüista é o *léxico mental*. Para alguns teóricos, o *léxico mental* é um "objeto" altamente estruturado, tanto do ponto de vista de sua estrutura interna (estruturação interna das entradas lexicais) quanto global (relações entre as entradas) (Mel'čuk, 1988; Briscoe 1991).

No domínio do PLN, o "léxico" é um dos componentes centrais dos sistemas de processamento das línguas naturais. Essa centralidade deve-se ao fato de que o "léxico" armazena, além das unidades lexicais, um complexo conjunto de informações que, de acordo com a especificidade do sistema, pode englobar propriedades fonológicas, morfológicas, sintáticas, semânticas e pragmático-discursivas dos itens lexicais da(s) língua(a). Tais informações fornecidas pelo léxico são manipuladas pelo sistema durante os processos de interpretação e/ou geração de língua natural (Handke, 1995; Palmer, 2001). Em função das aplicações e situações reais para as quais os programas de PLN são escritos, é premente, na construção de sistemas de tradução automática, por exemplo, a compilação de léxicos (monolíngües e multilíngües), que sejam (i) *manipuláveis* pelos programas que compõem o sistema e, sobretudo, (ii) *lingüisticamente motivados*, tanto do ponto de vista da quantidade de formas quanto do ponto de vista da qualidade e precisão das informações associadas às unidades lexicais.

Quanto ao item (ii), especificamente, são várias as questões que se colocam. Quais são, por exemplo, os itens lexicais de uma língua que devem estar em um léxico "computacional"? Ou ainda, quais os tipos de informação lexical lingüisticamente relevantes para o processamento automático das línguas naturais (Handke, 1995; Palmer, 2001)? Para tentar responder a essas e a outras questões, pesquisadores do PLN têm buscado, nos domínios lingüístico e psicolingüístico, os subsídios para a construção de léxicos lingüístico-computacionais. No caso deste trabalho, busca-se contribuir para a delimitação das informações lexicais que devem ser armazenadas em um léxico computacional.

Tal contribuição é feita por meio da proposição de um modelo de entrada lexical (psico)lingüístico-computacional, resultante da investigação das seguintes questões: (i) o papel do léxico (mental) no processamento cognitivo da linguagem; (ii) o acesso (entendido aqui como sinônimo de "fazer uso do léxico" ou "preparar o léxico para ser usado") às unidades lexicais na mente do falante; (iii) a organização da *microestrutura* (isto é, estruturação interna das entradas) do *léxico mental*.

Nas seções subseqüentes, apresenta-se a (i) concepção e (ii) as principais características do "léxico" sob o ponto de vista dos estudos do PLN e da Psicolingüística. Na seção 2, especificamente, apresenta-se o "léxico" sob o ponto de vista do PLN,

enfatizando sua *função* no sistema computacional e suas principais características. Na seção 3, apresenta-se a concepção e as características do "léxico" do ponto de vista dos estudos psicolingüísticos, enfatizando a *estruturação interna* das entradas no "léxico mental". Na seção 4, apresentam-se algumas considerações finais.

2. O olhar do PLN sobre o "léxico"

2.1. Da concepção computacional de "léxico"

Teoricamente, as arquiteturas propostas para sistemas de PLN acabam por espelhar a arquitetura proposta para o sistema lingüístico (Allen, 1987; Frazier, 1989). Como decorrência, um sistema de PLN deve possuir módulos autômatos, que realizam tarefas específicas e especializadas, e módulos que armazenam um modelo de conhecimento proposicional, que visa a criar simulacros de parcelas de mundo que lhe servem de referencial para interpretar os enunciados lingüísticos. Apesar da arquitetura de um sistema de PLN variar de acordo com as especificidades da aplicação, dois grupos de componentes são imprescindíveis para a implementação de qualquer sistema desse tipo: as *bases de conhecimento* e os *módulos de processamento* que atuam sobre essas bases (Dias-da-Silva, 1996). A Figura 1 ilustra esses dois grupos de componentes.

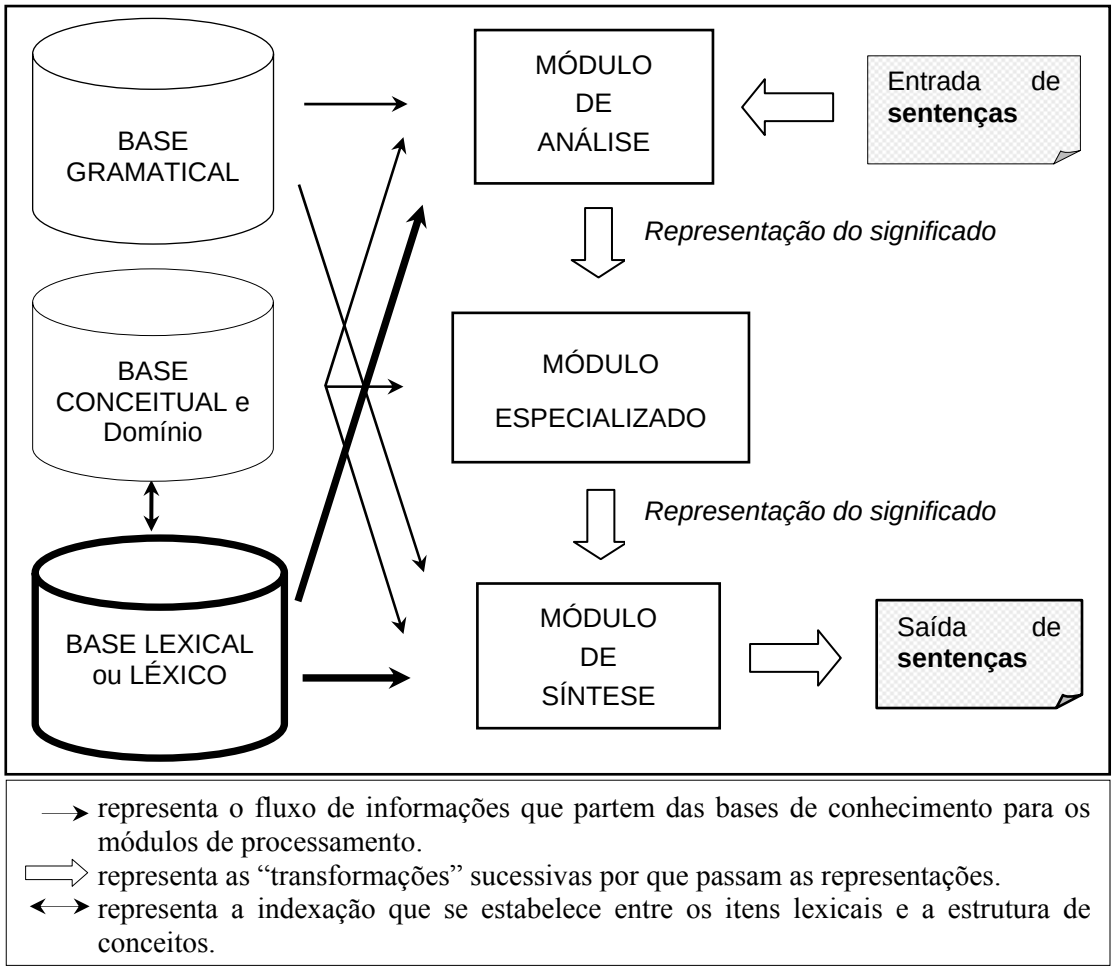

Figura 1: Arquitetura de um sistema de PLN genérico de Dias-da-Silva (1996)

Os módulos de conhecimento podem ser divididos em três módulos: o de *análise,* o *especializado* e o de *síntese*. As bases de conhecimento podem ser dividas em três bases: *gramatical, conceitual* e *lexical*. Com exceção do *módulo especializado*, os demais módulos de processamento e as bases de conhecimentos, embora os conteúdos possam variar em função da especificidade do sistema, possuem estrutura e funcionamento semelhantes. Toda a especificação dos módulos descrita a seguir foi extraída de Dias-da-Silva (1996).

O **módulo de análise** (MA) é geralmente formado pelo analisador morfológico e pelo analisador sintático (também denominado *parser*), além dos interpretadores semântico e pragmático-discursivo. Esse módulo é responsável pela construção de uma representação interna do significado das sentenças de entrada (no caso, digitadas via teclado).

O **módulo de síntese** (MS), por sua vez, transforma a representação abstrata gerada pelo MA em uma seqüência de "frases contextualizadas". Ao realizar a tarefa de construção de uma representação semântica, por exemplo, o MA utiliza-se, dependendo da sofisticação do sistema de que é parte, das bases gramatical, conceitual e lexical para executar todas ou parte das análises: morfológica, sintática, semântica e, até mesmo, pragmática. Assim, cada base de conhecimento, por sua vez, fornece ao MA informações de natureza diferente (cf. também Hutchins e Somers, 1997).

A **base gramatical** fornece a representação das regras sintáticas da língua, que podem ser vistas como condições de admissibilidade de estruturas sintáticas bem-formadas; condições que servirão de referência para o módulo de análise – responsável pela construção das representações sintáticas, semânticas e pragmático-discursivas.

A **base conceitual** fornece um modelo do mundo físico e conceitual, descrevendo tipos básicos de objetos, eventos, propriedades, relações e atributos em termos de representações hierarquicamente estruturadas, isto é, a sua estrutura consiste em uma rede de unidades conceituais interligadas. Essa base também pode fornecer conceitos mais específicos, ou seja, conceitos referentes a domínios particulares do conhecimento ou conceitos relacionadas a tarefas específicas para a qual o módulo esteja sendo projetado.

Em particular, à **base lexical**, fica a tarefa de fornecer, aos MA e MS, uma coleção de unidades lexicais, para as quais se faz necessária a especificação de conjuntos de traços morfológicos, sintáticos, semânticos e, até mesmo, pragmático-discursivos (cf. também Boguraev e Briscoe, 1989; Briscoe, 1991; Sanfilippo, 1995; Palmer, 2001). Esse tipo de base de dados, no domínio do PLN, é definido como sendo o "léxico" do sistema e recebe a denominação de **léxico tratável por máquina** ("*machine tractable dictionary*") (Wilks, 1988).

Todos os sistemas de PLN, desenvolvidos para serem aplicações reais, necessitam de léxicos que sejam *lingüisticamente motivados*, tanto do ponto de vista da (i) robustez (isto é, léxicos que contenham uma quantidade de unidades compatível com o "léxico" de uma língua natural) quanto da (ii) pertinência ("qualidade") das informações associadas às entradas. Isso acontece porque o desempenho de um sistema de PLN depende diretamente do número de entradas do léxico e da qualidade das informações associadas a essas entradas (Dorr, 1993; Saint-Dizier e Viegas, 1995; Palmer, 2001).

A seguir, são feitas algumas considerações a respeito do "léxico" sob o ponto de vista dos estudos (psico)lingüísticos.

3. Do olhar da Psicolingüística sobre o "léxico"

3.1. Do processamento cognitivo da linguagem: o léxico mental

Unindo pressupostos da Lingüística e da Psicologia, a Psicolingüística estuda a existência e o funcionamento de mecanismos mentais envolvidos no processamento da linguagem humana (Saint-Dizier e Viegas, 1995). Com o intuito de compreender, entre outras questões, como ocorre o armazenamento e o acesso aos itens lexicais de uma determinada língua, os psicolingüistas postulam a existência de um léxico mental (LM), definido como a parte do conhecimento lexical do indivíduo delimitada pela sua língua (Bierwisch e Schreuder, 1992; Levelt, 1992). De acordo com Bock (1982), Bierwisch e Schreuder (1992) e Levelt (1992), o léxico mental ocupa lugar de centralidade no processamento cognitivo da linguagem, o qual envolve três tipos de processos: (i) **conceitualização** (especificação de conceitos); (ii) **formulação** (seleção de palavras e construção de representações sintáticas e fonéticas); (iii) **articulação** (produção da fala). Esses três processos – conceitualização, formulação e articulação – e o papel desempenhado pelo LM estão ilustrados na Figura 2, baseada em Handke (1995) e Lowie (1998).

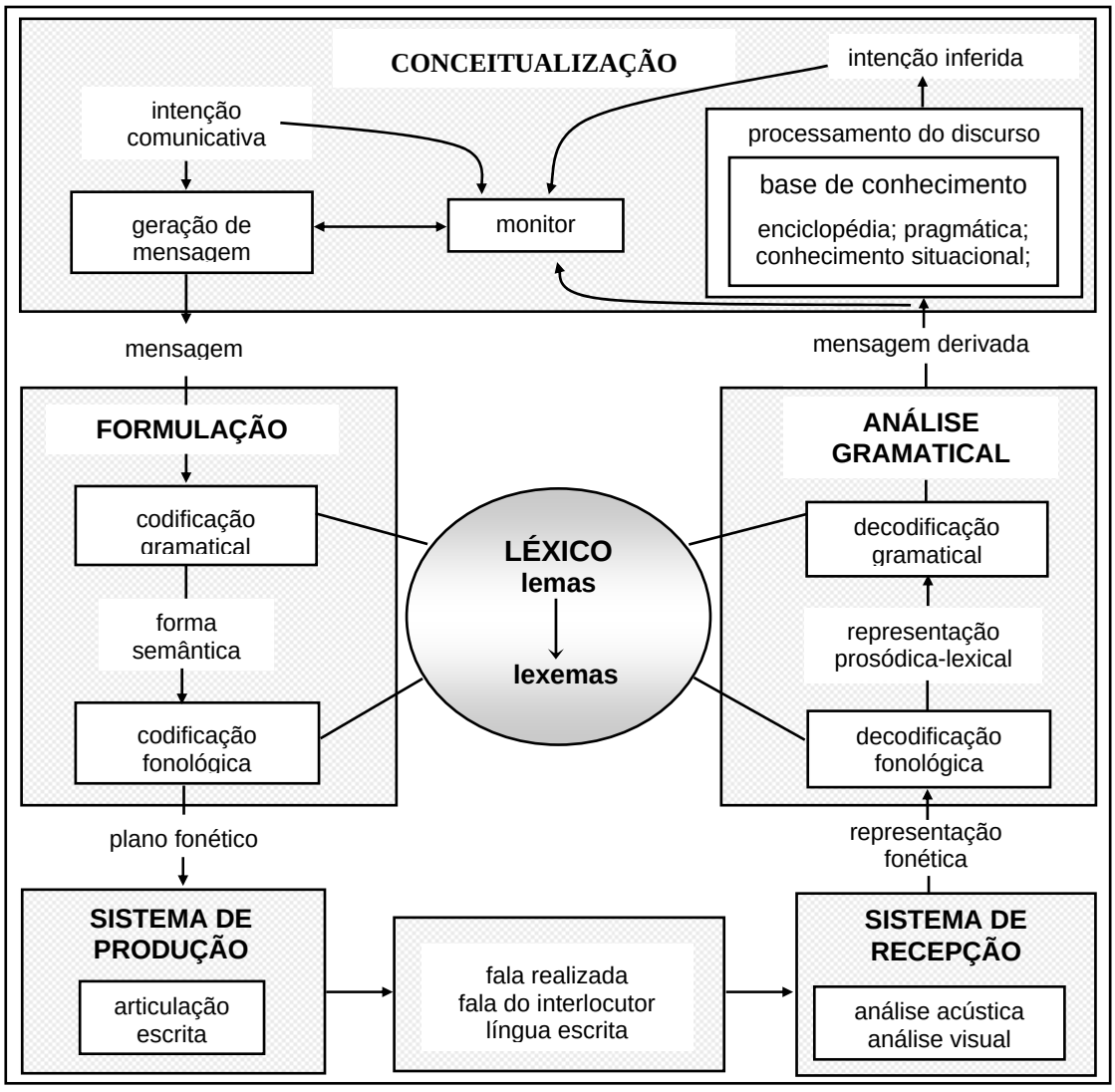

Figura 2: Modelo de processamento cognitivo da linguagem.

Para explicar o funcionamento desses três processos e o papel central desempenhado pelo LM no processamento da linguagem, é descrito, com mais detalhes, o sistema ou processo de **produção** de enunciados (lado esquerdo da Figura 2).

Nesse processo, a nomeação de um objeto perceptível envolve: (i) a identificação do objeto (conceitualização); (ii) a seleção de uma representação sintático-semântica do objeto, assim como a codificação dessa representação em termos fonológicos (formulação); (iii) a transformação da representação fonológica em realização fonética, que constitui o nome do objeto (articulação).

Mais especificamente, o processo de identificação do objeto ou **conceitualização** ativa uma robusta *base de conhecimento* que contém informações extralingüísticas provenientes de diversas fontes (visual, auditiva, motora, emotiva, conceitual, entre outras), além de princípios gerais de organização conceitual (ontologia do senso comum, conceitualizações do espaço e tempo, condições gerais subjacentes ao conhecimento enciclopédico ou a sistemas de crença, etc). O processo de conceitualização gera uma *estrutura conceitual* (EC) (pré-lingüística), que é a mensagem a ser verbalizada e organizada gramaticalmente pela *formulação*, no caso, essa mensagem será o *nome* do objeto.

A **formulação**, em específico, é responsável por transformar a *estrutura conceitual* gerada pelo processo de conceitualização em um enunciado lingüístico. Essa transformação, em específico, é mediada pelo LM, que é a parte do conhecimento lexical delimitada pela língua do falante. Ou seja, a língua do indivíduo delimita o LM, que, por sua vez, media a transformação da *estrutura conceitual* em *enunciado lingüístico*. Em outras palavras, pode-se dizer que os estímulos recebidos por um indivíduo (= estrutura conceitual) são traduzidos em itens lexicais de acordo com regras e princípios de cada língua. A essa hipótese, Glanzer e Clark (*apud* Biderman, 1981) deram a denominação de **elo verbal** (do inglês, "*verbal-loop hypothesis*").

De acordo com Bierwisch e Schreuder (1992), a conversão ou transformação da *estrutura conceitual* em enunciado lingüístico é feita em dois estágios e, para tanto, postula-se a existência do LM no nível lingüístico. O LM é central a todo o processamento da linguagem e contém todas as informações sobre os itens lexicais de uma língua, isto é, os *lemas* e os *lexemas*[1]. Dessa forma, o primeiro estágio da formulação é responsável pela seleção de uma representação sintático-semântica do objeto. Para tanto, é ativado, no LM, o **lema** do objeto, ao qual estão associadas informações sintáticas e semânticas que determinam, por exemplo, sua forma semântica, categoria sintática e estrutura de argumentos. De acordo com Bierwisch e Schreuder (1992), o resultado do primeiro estágio, a *formulação*, é uma *forma semântica* (FS)[2]. No segundo estágio, essa FS é transformada em uma *forma fonológica* (FF). Para que essa transformação seja possível, é ativado, no LM, o **lexema** do objeto, ao qual estão associadas informações fonológicas e morfológicas.

---

[1] Cabe ressaltar que os termos "lema" e "lexema" não estão sendo empregados no sentido típico do campo da lexicografia, isto é, representação canônica das unidades lexicais no dicionário e unidade lexical virtual que compõe o léxico, respectivamente (Biderman, 1999). Para Bierwisch e Schreuder (1992), "lema" é a representação das propriedades sintático-semânticas de um item lexical e "lexema" é a representação das estruturas morfológica e fonológica de um item.

[2] Vale ressaltar que há divergências quanto à postulação dos níveis EC e FS. A Bierwisch e Schreuder (1992), que defendem essa proposta, opõe-se, por exemplo, Jackendoff (1991, 1997), que propõe um nível único denominado nível da *estrutura léxico-conceitual* (ELC). Para Jackendoff, a FS não é concebida como uma entidade distinta da *estrutura conceitual*, mas sim um de seus subconjuntos.

Por fim, a **articulação** opera sobre a FF, ativando programas articulatórios que produzem a realização fonética do objeto.

3.1.1. Do *acesso* aos itens no LM

Partindo do modelo em que o processamento mental da linguagem é composto pelos três processos mencionados, o acesso a um item no LM, durante o processo de nomeação de um objeto, é, segundo Bierwisch e Schreuder (1992), realizado em duas etapas: seleção do lema e codificação da forma da palavra (ou *lexema*) a ser verbalizada. A **seleção do lema** consiste na ativação e seleção de um *lema* a partir da estrutura conceitual (EC). Já a **codificação da forma** consiste na construção de um programa articulatório que envolve a seleção de morfemas e de segmentos da forma da palavra e na ligação desses elementos às suas respectivas posições em uma estrutura denominada *esqueleto da forma da palavra* (Esq)[3].

3.1.2. Da microestrutura do LM: a estrutura interna das entradas

Tendo em vista que o acesso aos itens lexicais realiza-se nas etapas de *seleção do lema* e *codificação da forma* da palavra, a representação dos itens lexicais no LM, denominada de **entrada lexical**, está subdividida em *unidade de acesso* e *especificação lexical*. A Figura 3 ilustra essa bipartição das entradas.

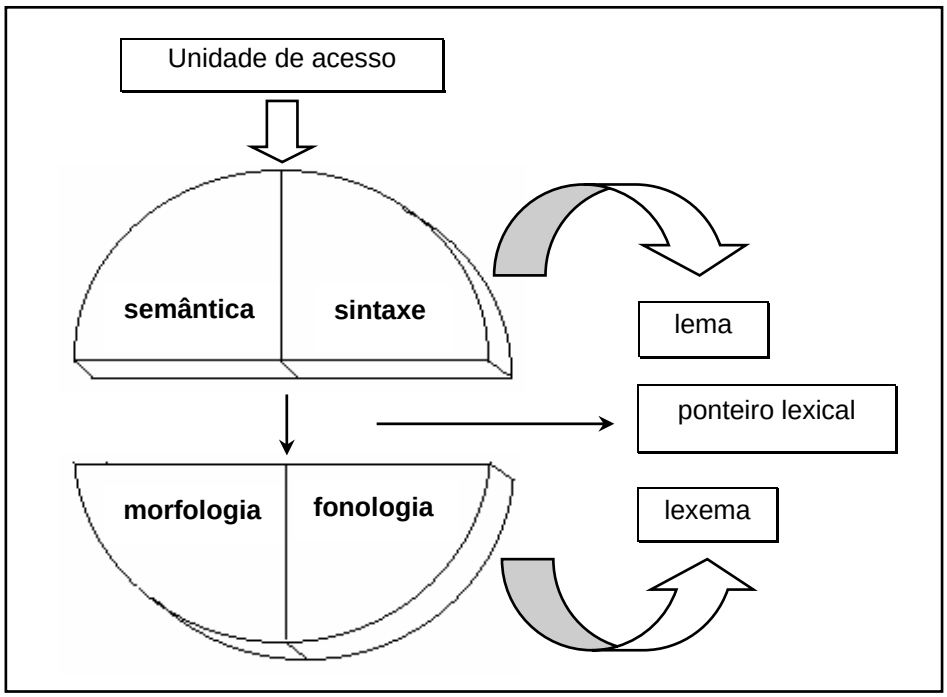

Figura 3: Proposta de entrada lexical bipartida de Handke (1995) e Lowie (1998).

[3] Por exemplo, o Esq da estrutura silábica de *gato*, [σ' σ], é preenchida pelos segmentos [/g/ /a/$_{\mu}$]$_{\sigma'}$ e [/t/ /u/$_{\mu}$]$_{\sigma}$. Os símbolos gregos **σ** (sigma) e **μ** (mi) são usados para representar, em fonologia métrica, a estrutura métrica de uma palavra em termos dos elementos abstratos: **sílaba** (**σ**) e **mora** (**μ**). O número de moras de uma sílaba caracteriza-a como *pesada* ou *leve*, p.ex.: sílabas que contêm vogal longa ou consoante final apresentam duas moras e são consideradas sílabas *pesadas*; sílabas que contêm vogais curtas (como as de *gato*) apresentam uma mora e são consideradas sílabas *leves* (Levelt, 1993).

(1) Da unidade de acesso

A **unidade de acesso** é o item lexical propriamente dito. Como bem salienta Langacker (1972) e Basílio (1999), a questão da delimitação das unidades que devem ser consideradas como pertencentes ao léxico é antiga e tem sido discutida sob diferentes perspectivas.

Com relação à forma dessas unidades, os especialistas divergem quanto a se a representação lexical delas se faz por *palavras* ou por *morfemas* (radicais ou raízes), seja na Lingüística, na Psicolingüística ou no PLN. De um modo geral, pode-se dizer que há três concepções de unidade ou item lexical (Butterworth, 1983; Handke, 1995).

Na primeira, listam-se, no léxico, as formas que servem de base para a formação de outras formas, isto é, as raízes e radicais, os afixos e as palavras funcionais; a esse tipo de léxico é dado o nome de "*root lexicon*" ("léxico de raízes" ou "léxico de morfema").

Na segunda, são listadas todas as formas possíveis da língua, inclusive as flexionadas; a esse tipo de léxico é dado o nome de "*full-form lexicon*" ("léxico de formas plenas", em oposição, portanto, à "léxico de morfemas").

Na terceira concepção, que é puramente computacional e independente de modelos (psico)lingüísticos, listam-se a raiz e outras formas-base idiossincrásicas, isto é, que são empregadas em processos morfológicos não regulares (Handke, 1995). Por exemplo, na entrada do verbo *agir*, seriam listadas as formas AG-, base para a maioria das formas do verbo *agir*, e AJ-, para as formas seguidas das letras *a* e *o* (ajo, aja, ajas, ajamos, ajais, ajam).

Além dessas três concepções, uma quarta pode ainda ser identificada. Nessa concepção as formas pertencentes ao paradigma flexional são realizações discursivas (isto é, palavras como *embalar, embalou, embalando*) do mesmo *item lexical* (EMBALAR). Por outro lado, as formas pertencentes ao paradigma derivacional (p.ex.: *embalar, embaladeira, embalado*) são itens lexicais distintos e, conseqüentemente, possuem entradas lexicais também distintas (Lyons, 1979). Dessa forma, observa-se que o termo *item lexical* refere-se ao sistema, isto é, à língua em si como entidade abstrata e supra-individual (*langue*), opondo-se, portanto, a *palavra*, que se aplica ao discurso (*parole*). Vale ressaltar que há outras denominações que são comumente aplicadas à unidade da *langue* e à unidade da *parole*. Por exemplo, Muller (1964) emprega a denominação *vocable* para a unidade do léxico e *mot* para a unidade ocorrente no texto; Biderman (1999), por sua vez, emprega o termo *lexema* para designar a unidade virtual que compõe o léxico e o termo *lexia* para designar as realizações discursivas dos lexemas.

(2) Da especificação lexical

A **especificação lexical** de um item é a representação do *lema* e do *lexema* do mesmo, os quais estão interligados por um ponteiro lexical (isto é, cada *lema* "aponta" para um *lexema* correspondente). O **lema** é a representação das propriedades semânticas e sintáticas de um item lexical; especifica as condições conceituais que garantem o uso apropriado do item, indicando, entre outras coisas, sua classe gramatical e seus argumentos. O **lexema**, por sua vez, é a representação das estruturas morfológica e fonológica de um item lexical.

3.2. Do modelo de entrada (psico)lingüístico-computacional

Partindo-se, então, da investigação sobre as questões relativas ao léxico sob o ponto de vista (psico)lingüístico, elaborou-se um esquema de representação em que estão especificados os tipos de informação lexical lingüisticamente relevantes para o processamento das línguas naturais. Dessa investigação, pôde-se conceber uma entrada lexical canônica como uma estrutura de dados, no sentido computacional desse termo, contendo quatro *componentes* lingüísticos, os quais estão especificados no Quadro 2.

Cada componente do Quadro 2 especifica tipos distintos de informação lexical, responsáveis pela *microestrutura* do léxico lingüístico-computacional, mais especificamente, pela estrutura interna das entradas.

Dimensão Lexemática	**FG** (E) (Leia-se: *Forma gráfica* de E) Esse componente especifica a forma gráfica e a estrutura morfológica de E.
Dimensão Lemática	**TG** (E) (Leia-se: *Traços gramaticais* de E) Esse componente determina as propriedades sintáticas de E e dos constituintes de hierarquia superior dos quais E é núcleo.
	EA (E) (Leia-se: *Estrutura de argumentos* de E) Esse componente especifica a seqüência de uma ou mais posições argumentais que corresponde aos argumentos exigidas por E.
	FS (E) (Leia-se: *Forma semântica* de E) Esse componente especifica o conteúdo proposicional da expressão contendo E, restringindo-a.

Quadro 1 - Os componentes da entrada lexical e as informações léxico-gramaticais correspondentes

3.2.1. Do aprofundamento teórico do modelo

3.2.1.1. Do componente FG

Como mencionado, o componente FG especificação a forma *gráfica* (ortográfica) e a estrutura *morfológica* das entradas lexicais. As características fonético-fonológicas das entradas, que também podem ser especificadas nesse componente, não foram previstas no esquema porque este trabalho concentra-se exclusivamente no tratamento computacional

da língua escrita[4]. A especificação da estrutura morfológica dos itens lexicais, no entanto, não será comentada neste relatório. Quanto à representação *gráfica*, isto é, *ortográfica* dos itens lexicais, salienta-se que esta pode ser entendida como a seqüência de *grafemas* que constitui um item lexical. Os *grafemas* são, na verdade, os símbolos gráficos unos, constituídos por traços gráficos distintivos, que permitem ao falante entender visualmente os itens da língua, da mesma forma que os *fonemas*[5] permitem ao falante entender esses itens auditivamente na língua oral. Na língua portuguesa, há símbolos gráficos (p.ex.: <c>, <s>) que podem representar em certos contextos um mesmo fonema (p.ex.: os símbolos gráficos <c>, <s> podem representar o mesmo fonema /s/), mas como grafemas podem distinguir, na língua escrita, os homófonos da língua oral (p.ex.: *cela* = tipo de aposento/ *sela* = arreio de cavalgadura).

Do ponto de vista do PLN, a especificação da forma gráfica dos itens lexicais é essencial para os sistemas computacionais que processam língua natural escrita porque permite, entre outras coisas, que tais itens sejam reconhecidos, pelo *analisador morfológico* ou *léxico*, como pertencentes à língua em questão.

3.2.1.2. Do componente TG

No componente TG da entrada de um item lexical, especificam-se as propriedades sintáticas desse item. Presume-se, aqui, que as propriedades sintáticas realizam-se por meio das **categorias gramaticais primárias** e das **categorias gramaticais secundárias**. Por categoria gramatical primária, entendem-se a classe gramatical a que o item pertence, por exemplo, s*ubstantivo, verbo* e *adjetivo*. A especificação da categoria primária do item engloba, além da classe, a descrição dos traços subcategoriais (isto é, o quadro de subcategorização). Por categorias gramaticais secundárias, presumem-se as categorias de *gênero, número, modo, caso, aspecto, voz*, entre outras (Lyons, 1979).

Do ponto de vista do PLN, a especificação de informações sobre as *categorias primárias* e *secundárias* dos itens lexicais é essencial para que os analisadores morfológico (ou léxico) e sintático possam atribuir categorias gramaticais a um item ***x*** e verificar a validade do relacionamento sintático do mesmo com os demais elementos da sentença, construindo, assim, uma estrutura abstrata da sentença que contenha o item ***x***.

3.2.1.3. Do componente EA

A todo **predicador** (PR) está associada uma **estrutura de argumentos** (estrutura-a) ou *valência*. Por predicador, entende-se todo elemento que atribui uma determinada propriedade a um certo termo ou estabelece uma relação entre termos, ou seja, uma *predicação* (Mira Mateus, et al., 1994; Neves, 1997). Os *predicadores* são itens lexicais semanticamente incompletos que, por isso, precisam necessariamente ligar-se a outros

---

[4] O destaque dado à língua escrita pauta-se no fato de que o tratamento dos aspectos ligados à fonética, fonologia e prosódia, essencial para o desenvolvimento de sistemas de fala, necessita de investigação à parte (Dias-da-Silva, 1996; Hutchins, Somers, 1997).

[5] Por se tratar de um trabalho multidisciplinar, lembra-se que "*Fonema* é a menor unidade destituída de sentido, passível de delimitação na cadeia da fala. Cada língua apresenta, em seu código, um número limitado e restrito de fonemas [...] que se combinam sucessivamente, ao longo da cadeia da fala, para constituir os significantes das mensagens, e se opõem, segmentalmente, em diferentes pontos da cadeia da fala, para distinguir as mensagens umas das outras. Sendo esta sua função essencial, o fonema é seguidamente definido como a unidade distintiva mínima" (Dubois, 1973, p. 280).

elementos – seus argumentos (As) - para adquirir um valor semântico completo (Dik, 1997; Neves, 2000).

As relações semânticas estabelecidas entre um predicador e seus argumentos são representadas por *papéis temáticos* (ou funções semânticas) (Gruber, 1965; Fillmore, 1968; Palmer, 1994; Davis, 2001).

Há dez anos, o construto denominado de *estrutura-a* foi equacionado como o número de argumentos (A) requerido ou projetado por um P (Grimshaw, 1992). De acordo com autores como Williams (1981) e Marantz (1984), a estrutura-a é um conjunto de As marcados como **internos** (subcategorizados) ou **externos** (não subcategorizados)[6]. Entretanto, com a crescente importância de princípios como o Princípio da Projeção[7] e Critério-θ[8] na Teoria da Regência e Ligação ("*Government-Binding Theory*" – GB) (cf. Chomsky, 1981) e com o desenvolvimento das teorias *lexicalistas*[9], um novo ponto de vista sobre a estrutura-a emergiu, segundo o qual ela representa uma interface entre a semântica e a sintaxe (Grimshaw, 1992; Levin, Pinker, 1991; Bresnan, 1981, 1982, 2000; Sag, Wasow, 1999; Sag, 1997; Sells, 1985).

Além dos *papéis temáticos*, também são empregadas as chamadas *restrições selecionais* na descrição da estrutura de argumentos de um predicador. Tais restrições são traços de natureza semântica caracterizadores dos argumentos selecionados pelo predicador. O princípio básico do emprego dessas *restrições* é associar a cada argumento do predicador uma lista de *traços* (F_i) que restringem o conteúdo semântico dos argumentos. Essa lista de restrições pode ter diferentes formatos (Saint-Dizier, Viegas, 1995):

(i) $[F_i]$: traço semântico único, p.ex.: *humano, animado*, etc.;
(ii) $[F_1 \wedge F_2 \ldots \wedge F_n]$: uma conjunção de elementos que expressa um conjunto de restrições que devem ser satisfeitas;
(iii) $[F_1 \vee F_2 \ldots F_m]$: uma disjunção de restrições: uma das restrições deve ser satisfeita;
(iv) uma combinação de conjunções e disjunções.

O verbo *comer,* por exemplo, projeta dois argumentos, A1 e A2, cujos papéis temáticos são: Agente e Objetivo, respectivamente. O argumento Agente requer o traço [animado] e o argumento Objetivo, por sua vez, requer no mínimo o traço [concreto]. Em outras palavras, os traços *animado* e *concreto* restringem o conteúdo semântico do A1 e A2 do verbo *beber*, respectivamente. O exemplo em (1), elaborado com base na Gramática Funcional - FG (Dik, 1997), ilustra o uso de *papéis temáticos* e de *restrições selecionais* na representação da estrutura de argumentos:

(1) **comer** [V] $(x_1$: <*anim*> $(x_1))_{Ag}$ $(x_2$: <*conc*> $(x_2))_{Objetivo}$

---

[6] Williams (1981) e Marantz (1984) definem os *papéis temáticos* de acordo com a concepção tradicional de Gruber (1965) e Fillmore (1968), ou seja, como um conjunto de rótulos conceituais que definem a participação dos argumentos na "cena" projetada pelo predicador.

[7] Segundo esse critério, as representações em cada nível sintático (estrutura profunda, estrutura superficial e forma lógica) são projetadas do léxico, isto é, observam as propriedades temáticas e de subcategorização dos itens lexicais (Raposo, 1992).

[8] O Critério-θ tem como finalidade assegurar que as posições projetadas pelo do Princípio de Projeção sejam devidamente preenchidas por argumentos (Raposo, 1992).

[9] A LFG (Kaplan, Bresnan, 1982), a GPSG (Gazdar et al., 1985) e a HPSG (Pollard, Sag, 1987; 1994) são exemplos paradigmáticos de modelos lexicalistas. Nesses modelos, o léxico "projeta" a sintaxe (Wasow, 1985).

Em (1), a **estrutura de predicado**[10] especifica a forma ortográfica (*comer*), a categoria gramatical (V) e a valência ou estrutura de argumentos de *comer*. Esta, por sua vez, consiste em duas posições, indicadas pelas variáveis x_1 e x_2, cujas funções semânticas ou papéis temáticos são, respectivamente, Agente (Ag) e Objetivo. Os argumentos indicados pelas variáveis x_1 e x_2 apresentam as *restrições selecionais* <animado> e <concreto>, respectivamente.

3.2.1.4. Do componente FS

Como mencionado, especifica-se, neste componente, a *forma semântica* (FS) de um item lexical. Segundo Bierwisch e Schreuder (1992), a FS de um item representa a contribuição que esse item faz para o significado das expressões que o contêm. A natureza dessa contribuição é assunto controverso nos estudos lingüísticos. Apesar da controvérsia, uma hipótese geral parece ser a de que a FS de um item restringe o conteúdo proposicional das expressões que o contêm.

Na Teoria do Léxico Gerativo de Pustejovsky (1996), por exemplo, a *forma semântica* de um predicador é especificada em termos de quatro estruturas ou níveis de representação sobre as quais operam mecanismos gerativos. São eles:

(a) a estrutura de argumentos: responsável pela relação entre o léxico e a sintaxe; especifica o número e o tipo de argumentos lógicos, além de especificar o modo como esses argumentos são realizados sintaticamente;

(b) a estrutura qualia: responsável pela especificação dos "modos de significação", apresenta os atributos e valores de um objeto em função dos *qualia*: FORMAL (de que ***x*** feito); CONSTITUTIVO (as partes de ***x***); TÉLICO (a função ou finalidade de ***x***) e AGENTIVO (como ***x*** origina-se);

(c) a estrutura de eventos: responsável pela descrição dos eventos, estados e transições, fornece os elementos para a representação semântica dos predicados;

(d) a estrutura de herança: responsável, do ponto de vista das categorias léxico-conceituais, pela hierarquização dos itens lexicais, em termos de relações de semelhança, oposição ou inclusão dos itens, imprime ao léxico uma organização global.

4. Considerações finais

Partindo-se, então, da investigação sobre as questões relativas ao léxico sob o ponto de vista (psico)lingüístico, elaborou-se um modelo de entrada lexical em que estão especificados os tipos de informação lexical lingüisticamente relevantes para o processamento das línguas naturais. Esse modelo é composto pelas informações: FG, TG, EA e FS. Com isso, pretende-se contribuir para o desenvolvimento de léxicos lingüístico-computacionais.

---

[10] Na FG, os predicados da língua estão armazenados no léxico em *estruturas de predicados*, que especificam um predicado juntamente com um "esqueleto" das estruturas nas quais ele pode aparecer.

5. Referências bibliográficas

Allen, J. F. (1987). *Natural language understanding.* Menlo Park: Benjamin Cummings

Biderman, M. T. C. (1981). A estrutural mental do léxico. In: *Estudos de Filologia e Lingüística* - Homenagem a Isaac Nicolau Salum. São Paulo: Editora da USP; T. A. Queiroz, p. 131-45.

____. Conceito Lingüístico de Palavra. (1999). In: Basílio, M. (ed.) *Palavra.* Departamento de Letras da PUC-Rio, pp. 81-97.

Bierwisch, M., Schreuder, R. (1992). From concepts to lexical items. *Cognition,* 42, p.23-60.

Bock, J. K. (1982). Towards a Cognitive Psychology of Syntax. *Psychological Review,* 89, p. 1-47.

Bresnan, J. (1981). An approach to Universal Grammar and the mental representation of language. *Cognition,* 10, p. 39-52.

____ (ed.). (1982). *The mental representation of grammatical relations*. Cambridge, Mass: The MIT Press.

____. *Lexical-functional syntax*. (2000). Stanford: University of Stanford.

Briscoe, T. (1991). Lexical issues in natural language processing. In: Klein, E.; Veltman, F. (Eds.). *Natural language and speech.* Springer-Verlag, p.39-68.

Butterworth, B. (Ed.) (1983). *Language production volume 2: development, writing and other language processes.* London: Academic Press.

Davis, A. R. (2001). *Linking by types in the hierarchical lexicon*. Stanford: CSLI Publications, 2001. Disponível em <http://www-linguistics.stanford.edu/~tdavis/thesis-ps.html>. Acesso em 1 agosto de 2002.

Dias-da-Silva, B. C. (1996). *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. Araraquara, 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara.

Dik, S. C. (1997). *The theory of functional grammar*. Berlin, New York: Mouton de Gruyter.

Dorr, B. J. (1993). *Machine translation: a view from the lexicon*. Massachussetts: Massachussetts Institute of Technology.

Dubois, J. et al. (1973). *Dicionário de Lingüística.* São Paulo, Cultrix.

Fillmore, C. J. (1968). The case for case. In: Bach, E., Harms, R. T. (Eds.). *Universals in linguistic theory.* Holt, Rinehart and Winston, Inc., p.1-88.

Frazier, L. (1989). Grammar and Language Processing. In: Newmeyer, F. (Ed.). *Linguistics: the Cambridge survey II: linguistic theory: extensions and implication.* Cambridge: CUP, p.15-34.

Gazdar, G. et al. (1985). *Generalized phrase structure grammar*. Cambridge, Mass.: Harvard University Press.

Gruber, J. (1965). *Studies in lexical relations*. Tese de Doutorado, Cambridge.

Handke, J. (1995). *The structure of the Lexicon: human versus machine*. Berlin: Mouton de Gruyter.

Hutchins, W. J.; Somers, H. L. (1997). *An introduction to machine translation*. London: Academic Press.

Jackendoff, R. (1991). *Semantic structures*. Cambridge: Mass.: The MIT Press.

____. (1997). *The architecture of the language faculty*. Cambridge: Mass.: The MIT Press.

Kaplan, R. e Bresnan, J. (1982). Lexical-functional grammar: a formal system for grammatical representation. In: Bresnan, J.(Ed.). *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Langacker, R. (1972). *Fundamentals of Linguistic Analysis*. New York: Harcort, Brace, Jovanovich.

Levelt, W .J. M. (1992). Accessing words in speech production: stages, processes and representations. *Cognition*, 42, p.1-22.

____. (1993). *Speaking: to intention to articulation*. Cambridge, Mass.: The MIT Press.

Levin, B., Pinker, S. (1991). Introduction to special issue of *Cognition* on lexical and conceptual semantics. *Cognition*, 41, p.1-7.

Lowie, W. (1998). *The acquisition of interlanguage morphology: a study into the role of morphology in the L2 learner's mental lexicon*. Groningen: University Library Groningen.

Lyons, J. (1979). *Introdução à lingüística teórica*. Tradução de Rosa Virgínia Mattos e Silva e Hélio Pimentel; revisão e supervisão de Isaac Nicolau Salum – São Paulo: Ed. Nacional, Editora da Universidade de São Paulo.

Marantz, A. (1984). On the nature of gramatical relations. *Linguistic Inquiry*. Monography 10. Cambridge: The MIT Press.

Mel'čuk, I. (1988). *Dependency Syntax: theory and practice***.** The SUNY Press, Albany, N.Y, 428p.

Mira Mateus, M. H. et al. (1994). *Gramática da Língua Portuguesa***.** 4ª ed., Lisboa: Caminho.

Muller, C. (1964). *Essai de statistique lexicale***.** L'IC, Klincksieck.

Neves, M. H. M. (2000). *Gramática de usos do português*. São Paulo: Editora UNESP.

Palmer, F. R. (1994). *Grammatical roles and relations*. Cambridge: Cambridge University Press.

Palmer, M (2001). Miltilingual resources – Chapter 1. In: Hovy, E., et al. (Eds.). *Linguistica Computazionale,* v.14-15.

Pollard, C., Sag, I. A. (1987). *Information-based Syntax and Semantics*. Volume I: Fundamentals. Stanford: CSLI Publications.

____. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Pustejovsky, J. (1996). *The generative lexicon*. 2ª ed. Cambridge: Mass.: The MIT Press.

Raposo, E.P. (1992). *Teoria da gramática: a faculdade da linguagem*. Lisboa: Caminho.

Sag, I. A. (1997). English relative clause constructions. Jo*urnal of Linguistics*, v.33, p.431-484.

Sag, I., Wasow, T. (1999). *Syntactic theory: a formal introduction*. Stanford: CSLI Publication.

Saint-Dizier, P., Viegas, (1995). E. *Computational lexical semantics*. Cambridge: Cambridge University Press.

Sanfilippo, A. (1995). Lexicons for constraint-based grammars. In: Cole, R. A (Ed). *Survey of the state of the art in human language technology*. Oregon: Graduate Institute, p. 118-121.

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

Wasow, T. Postscript. (1985). In: Sells, P. *Lectures on contemporary syntactic theories*. Chicago: The University of Chicago Press.

Williams, E. (1981). Argument structure and morphology. *Linguistic Review*, 1, p.81-114.

Wilks, Y. D. et al. (1988). Machine tractable dictionary as tools and resources for natural language processing. In: *Proceedings of Colling´88*, p. 750-55.