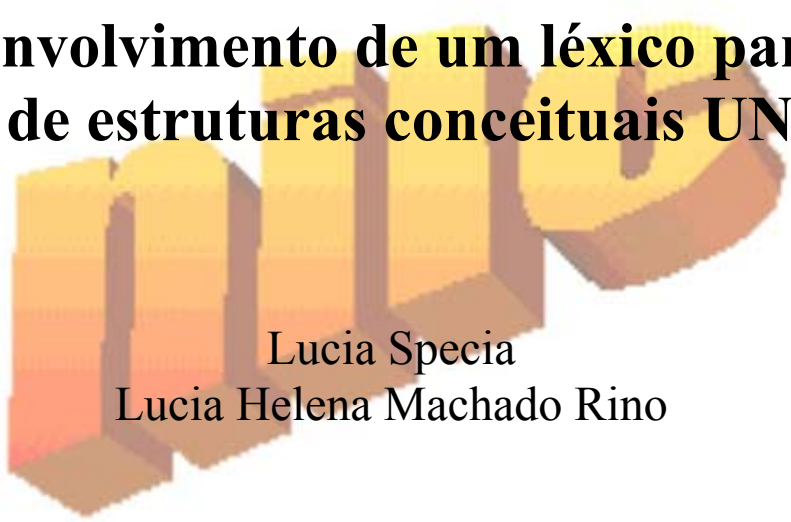
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

O desenvolvimento de um léxico para a geração de estruturas conceituais UNL



Lucia Specia
Lucia Helena Machado Rino

NILC-TR-02-14

Setembro, 2002

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

Resumo

Este relatório descreve o processo de desenvolvimento de um léxico da língua portuguesa a ser utilizado na tarefa de geração de estruturas conceituais UNL a partir de estruturas sintáticas, focalizando as questões relacionadas às informações semânticas necessárias a essa tarefa. São discutidas as principais características do léxico, a aquisição das informações pertinentes aos itens lexicais e a sua implementação.

Este trabalho conta com o apoio
financeiro da CAPES

Índice

Figuras

Tabelas

1 Introdução

O Léxico é um recurso lingüístico necessário a qualquer sistema de Processamento da Língua Natural (PLN), seja ele voltado para a interpretação (análise) ou para a geração (síntese) de LN(s), pois representa o vocabulário da(s) LN(s) desse sistema. Esse recurso geralmente consiste em um conjunto de palavras ou expressões da LN, chamadas de **itens lexicais**, associadas à sua descrição, ou seja, a um conjunto de **traços** (morfológicos, sintáticos e semânticos, por exemplo) cujos valores fornecem as informações necessárias para que tais palavras ou expressões sejam processadas pelo sistema. Cada item, juntamente com sua descrição, é chamado de **entrada lexical**. As informações de um léxico são recuperadas pelas diferentes etapas de um sistema de PLN (processamento morfológico, sintático, etc) e são utilizadas de acordo com os objetivos pertinentes a cada etapa.

Os léxicos de sistemas de PLN podem apresentar variações em diferentes aspectos. A natureza e a quantidade de informações armazenadas para cada item lexical, por exemplo, dependem da estruturação do léxico e do tipo de processamento a ser realizado pela aplicação a que se destina: alguns léxicos armazenam somente informações morfossintáticas, outros incluem informações semânticas, como aquelas que permitem projeções de estruturas sintáticas em semânticas e vice-versa. Da mesma forma, os tipos de itens lexicais podem variar: alguns léxicos armazenam somente as formas básicas das palavras, chamadas aqui de **formas canônicas**, utilizando-se de processos de derivação e/ou flexão para obter suas formas variantes, chamadas aqui de **formas analisadas** (flexões de gênero, número, grau, modo, tempo, etc.). Outros léxicos, no entanto, armazenam todas as possíveis variantes de uma dada forma básica. Ainda, a quantidade de itens lexicais depende da abrangência do sistema: alguns léxicos são limitados a um *corpus* de textos, outros, pretendem cobrir grande parte do vocabulário de uma LN.

O léxico em um sistema de PLN pode ser estruturado de acordo com diferentes técnicas de representação do conhecimento da Inteligência Artificial (lógica de predicados, regras de produção, redes semânticas, *frames*, etc.) ou de acordo com diferentes abordagens de representação lexical, propriamente ditas, por exemplo, o modelo teórico do Léxico Gerativo de Pustejovsky (1995) e as linguagens DATR (Evans & Gazdar, 1996) e LAUREL (Copestake, 1993). Mais restritivas, essas abordagens especificam uma estrutura formal para o léxico, com uma sintaxe bem definida para a descrição dos itens. Geralmente, essas abordagens delineiam um dado conjunto de traços, juntamente com o seu conjunto de possíveis valores. Dependendo da complexidade da representação lexical, são desenvolvidos, ainda, mecanismos que visam à otimização do léxico, como a herança de traços entre itens, a composicionalidade, entre outros. A maioria dessas abordagens é voltada para alguma aplicação específica do PLN, portanto, elas nem sempre são adequadas para tarefas gerais, seja porque não apresentam todos os traços necessários, porque apresentam traços para os quais não se dispõem de informações, etc.

Este relatório descreve o desenvolvimento de um léxico, aqui chamado de **Léxico Enriquecido**, utilizado na tarefa de geração de estruturas semânticas (ou conceituais) a partir de estruturas sintáticas de sentenças da língua portuguesa, tarefa esta realizada pelo sistema **Gerador Conceitual**, cuja arquitetura é apresentada na Seção 2. Em função das especificidades dessa tarefa, a estrutura do Léxico Enriquecido não se baseia completamente em nenhuma abordagem de representação lexical, conforme descrito na Seção 3. O processo de aquisição das informações do Léxico Enriquecido, com enfoque

nas questões relacionadas às informações semânticas, é discutido na Seção 4. A Seção 5 descreve a implementação desse recurso. A Seção 6 apresenta algumas considerações finais.

2 A arquitetura do Gerador Conceitual

O Gerador Conceitual, cuja arquitetura global é ilustrada na Figura 1, é um sistema que realiza o mapeamento de estruturas sintáticas de sentença do português em estruturas conceituais. Mesmo partindo de estruturas sintáticas e não de sentenças em LN, esse sistema pode ser classificado como de interpretação da LN. Dentre os recursos lingüísticos e processos utilizados, alguns já estavam disponíveis no NILC[1] (módulos em azul) e foram, de algum modo, reutilizados, enquanto outros foram desenvolvidos exclusivamente para o Gerador Conceitual (módulos em vermelho). Uma descrição geral dos recursos e processos envolvidos nessa arquitetura pode ser consultada em Specia (2002).

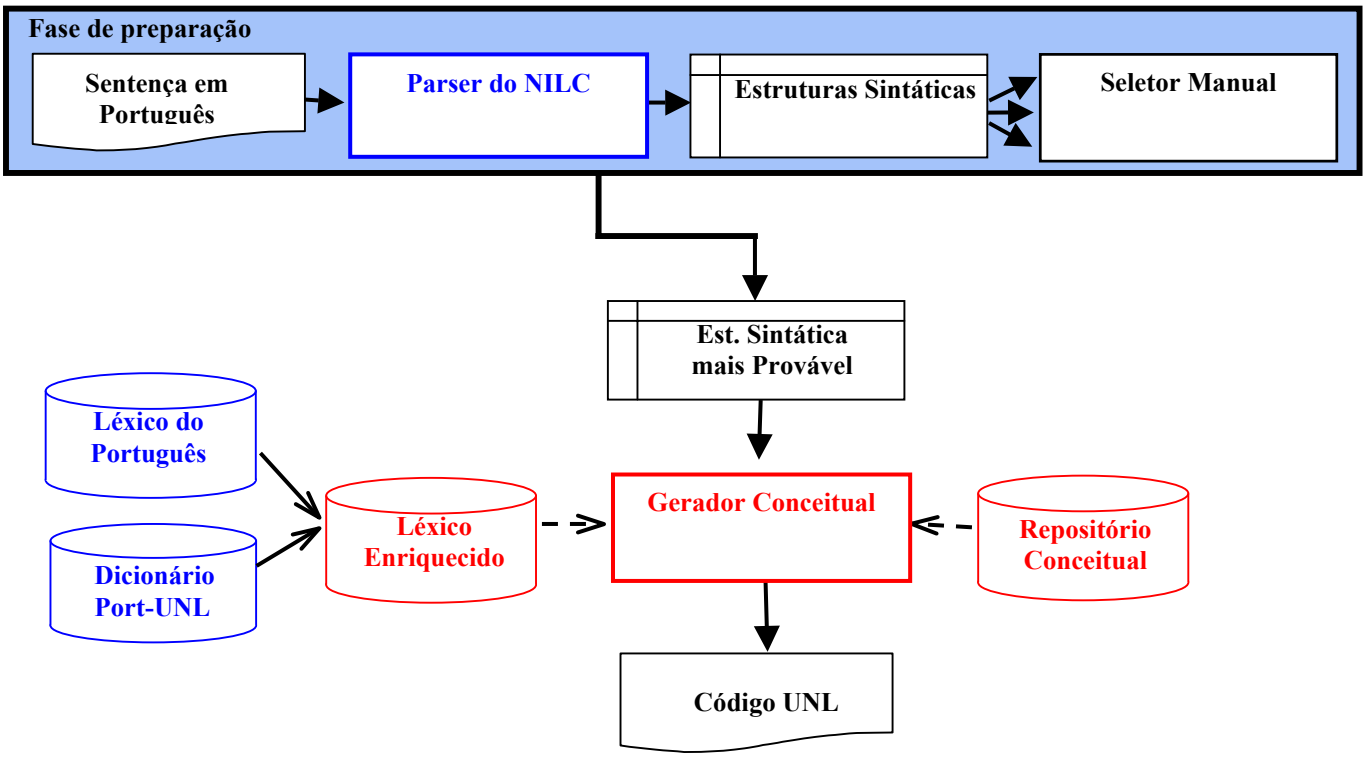


Figura 1 – Arquitetura do Gerador Conceitual.

Fundamentalmente, o processamento descrito nessa arquitetura consiste dos seguintes passos: 1) na **Fase de Preparação**, que utiliza processos e recursos independentes do Gerador Conceitual, disponíveis no NILC, o *parser* do NILC[2] gera todas as estruturas sintáticas possíveis para uma sentença do português; 2) ainda nessa fase, a estrutura sintática mais provável é escolhida manualmente; e 3) o Gerador Conceitual

[1] Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.sc.usp.br).
[2] Este *parser* é independente de aplicação e ainda está em desenvolvimento no NILC.

converte tal estrutura sintática em uma representação conceitual, aqui chamada de **Código UNL**.

Como entrada, o Gerador Conceitual contempla somente um subgrupo de construções gramaticais do português, que correspondem a sentenças de um determinado domínio, as quais constituem o *corpus* com base no qual o léxico (assim como os demais recursos lingüísticos e processos) foi especificado. O domínio desse *corpus* é o de textos de horóscopo, dos quais foram selecionadas 36 sentenças, com base em critérios descritos em Specia (2002). Essas sentenças são constituídas de 134 palavras diferentes (de classe fechada, ou seja, que possuem um número finito de representantes, como preposições, conjunções, etc. e de classe aberta, ou seja, cujos representantes não podem ser enumerados, como substantivos, verbos, etc.), considerando somente a forma na qual a palavra se apresenta no *corpus* (analisada ou canônica)[3].

Como saída, a linguagem de representação conceitual utilizada pelo Gerador Conceitual é a UNL (*Universal Networking Language*) (UNL, 2001), desenvolvida com o propósito de ser implementada em sistemas de comunicação interlingual, mais especificamente, para ser utilizada como representação intermediária no Projeto UNL (Uchida et al., 1999), de tradução automática por interlíngua. Uma descrição dessa linguagem pode ser consultada na sua especificação, em UNL (2001)[4] e também em Specia & Rino (2002).

Como exemplo do processamento do Gerador Conceitual, considere a sentença em (1) como entrada e o código UNL gerado em (2) como saída.

(1) O Sol em Sagitário ilumina seus relacionamentos.

(2)
```
agt(illuminate.@entry, sun.@def)
plc(sun.@def, sagittarius)
obj(illuminate.@entry, relationship.@pl)
pos(relationship.@pl, you)
```

A estrutura conceitual da UNL é uma representação lógica, baseada em relações semânticas (*Relation Labels – RLs*) entre dois conceitos (*Universal Words – UWs*), os quais podem apresentar certos atributos (*Attribute Labels – ALs*). No caso da sentença em (1), são utilizadas quatro relações semânticas (agt = agente; plc = lugar; obj = objeto; pos = posse) entre os conceitos *sun* (sol), *sagittarius* (sagitário), *illuminate* (ilumina) e *relationship* (relacionamentos), sendo que alguns desses conceitos têm atributos associados a eles, como *.@pl* em *relationship*, que indica que o conceito está no plural. Apesar de ser independente de LN, como se pode verificar no exemplo, a UNL utiliza palavras da língua inglesa como símbolos para expressar seus conceitos.

Como em qualquer sistema de interpretação profunda da LN, o processo de obtenção de estruturas conceituais, ou seja, de estruturas de representação do significado, exige recursos lingüísticos que forneçam informações de diversas naturezas sobre a LN, incluindo informações morfológicas, sintáticas e semânticas. Essa foi a principal motivação para o desenvolvimento do Léxico Enriquecido, pois não dispúnhamos de

[3] O número de palavras é pequeno em função da metodologia de desenvolvimento do Gerador Conceitual: de prototipagem. No seu estágio atual, o Léxico Enriquecido fornece informações à versão inicial do protótipo e, portanto, esse número tende a aumentar de acordo com a construção incremental dos recursos e processos do Gerador.

[4] Disponível em http://www.unl.ias.unu.edu.

recursos com todas as informações necessárias, mais especificamente, com informações de cunho semântico. Por isso, o foco do Léxico Enriquecido, bem como deste relatório, é a especificação dessas informações, conforme será descrito nas próximas seções.

3 As características do Léxico Enriquecido

3.1 Estrutura

Para a definição da estrutura do Léxico Enriquecido, além das técnicas de representação do conhecimento da Inteligência Artificial (lógica de predicados, regras de produção, redes semânticas, *frames*, etc.), foram investigadas diferentes categorias de modelos teóricos e/ou computacionais que permitem a representação do conhecimento lexical, incluindo informações semânticas:

1) Modelos de representação semântica (ou representação conceitual), como as teorias da Dependência Conceitual (Schank, 1975) e da Semântica Conceitual (Jackendoff, 1990) e a linguagem LCS (*Lexical Conceptual Structure*) (Dorr, 1992; Dorr, 1993);

2) Modelos de representação lexical independentes de aplicação, como a Teoria do Léxico Gerativo de Pustejovsky (1995) e a linguagem DATR (Evans & Gazdar, 1996);

3) Modelos de representação lexical desenvolvidos para alguma aplicação específica, como a linguagem LAUREL (Copestake, 1993), utilizada no projeto AQUILEX (Briscoe, 1991) de extração de informações de dicionários eletrônicos para a criação de bases de dados lexicais; o modelo de léxico semântico utilizado no projeto Mikrokosmos de tradução automática (Onyshkevych & Nirenburg, 1994; Viegas & Raskin, 1998); o modelo de estruturação lexical para o português de Dias (1994), baseado na Semântica Conceitual de Jackendoff (1990) e utilizado no projeto LINX (Garcia, 1995) de comunicação em língua natural entre usuários e bases de conhecimento lógicas; a adaptação do modelo do léxico gerativo de Pustejovsky (1995) para o português, desenvolvida por Abrahão (1997) e utilizada no projeto Nálamas (Lima et al., 1996); e a estrutura lexical do sistema Wordnet (Miller, 1990).

Em função da sua aplicação na arquitetura do Gerador Conceitual, o Léxico Enriquecido apresenta um conjunto diferenciado de informações e, consequentemente, sua estrutura não segue nenhum desses modelos, apesar de apresentar características que podem ser verificadas em alguns deles. De qualquer modo, a investigação desses modelos foi importante para a definição das informações semânticas do Léxico Enriquecido e, portanto, essa questão será retomada na Seção 4.2.

3.2 Itens

Os itens do Léxico Enriquecido são todas as palavras[5] das sentenças do *corpus* (134 palavras) e suas respectivas formas canônicas (quando as palavras das sentenças já não são as próprias canônicas), totalizando 200 entradas. Optou-se por essa forma de representação (isto é, canônicas e analisadas) porque alguns valores de traços são específicos das diferentes formas analisadas de uma mesma canônica (o tempo de um verbo ou o número de um substantivo, por exemplo), enquanto outros são válidos para todas as formas

[5] Uma única palavra ou, no máximo, palavras compostas hifenizadas.

analisadas de uma mesma canônica (os traços semânticos e o conceito UNL, por exemplo). Assim, para evitar a repetição de informações que são válidas tanto para uma canônica quanto para suas formas analisadas, elas são armazenadas apenas nas formas canônicas, sendo que cada forma analisada remete à sua respectiva canônica.

Em algumas classes de itens, as canônicas apresentam traços que são também especificados para suas formas analisadas, mas com valores diferentes. Por exemplo, o traço "número" dos substantivos, cujo valor poder "plural", em uma forma analisada ("relacionamentos", por exemplo) e "singular", em uma forma canônica ("relacionamento"). Nesses casos, o Gerador Conceitual utiliza o valor do traço para a forma na qual a palavra se encontra na sentença, analisada ou canônica.

Caso a mesma palavra ocorra no *corpus* em diferentes ambientes morfossintáticos ou semânticos, ou seja, tenha diferentes conjuntos de traços ou valores para o mesmo traço, todas as ocorrências são representadas como itens lexicais, sendo definida uma entrada lexical para cada um dos diferentes conjuntos de traços ou valores. Por exemplo, se uma sentença utiliza a palavra "controle" como substantivo (ex.: "mantenha o controle dos seus sentimentos") e outra utiliza essa mesma palavra como verbo (ex.: "controle suas emoções"), são criados duas entradas lexicais cujo item é "controle", cada qual com seus respectivos traços e valores.

3.3 Traços

O conjunto de informações representadas para os itens lexicais depende da categoria sintática de cada item. Ao todo, foram consideradas 8 categorias, com base na análise do *corpus* citado. As informações foram agrupadas em diferentes campos, sendo que alguns campos possuem mais de um traço. A Tabela 1 apresenta as informações especificadas para cada uma das categorias sintáticas, considerando suas formas canônica e analisada[6], agrupadas nos diferentes campos.

Tabela 1 – Categorias sintáticas dos itens lexicais e suas respectivas informações.

Categoria sintática	Campos para a forma analisada	Traços para o campo
Adjetivo	Canônica	Canônica
Advérbio	-	-
Artigo	Sintático	Tipo
	Canônica	Canônica
Conjunção	-	-
Preposição	Canônica	Canônica
Pronome	Sintático	Tipo
	Canônica	Canônica
Substantivo	Sintático	Número
	Canônica	Canônica
Verbo	Canônica	Canônica
	Sintático	Modo/tempo

[6] No caso das categorias "advérbio" e "conjunção", são apresentados somente os traços para a forma canônica, visto que são categorias invariáveis para o conjunto de traços considerado.

Categoria sintática	**Campos para a forma canônica**	**Traços para o campo**
Adjetivo	UNL	Conceito UNL
Advérbio	Sintático	Tipo
	UNL	Conceito UNL
Artigo	Sintático	Tipo
	UNL	Conceito UNL
Conjunção	Sintático	Tipo
	UNL	Conceito UNL
Preposição	UNL	Conceito UNL
Pronome	Sintático	Tipo
	UNL	Conceito UNL
Substantivo	Sintático	Número
	UNL	Conceito UNL
	Semântico	Traços semânticos
Verbo	Sintático	Modo/tempo Tipo
	UNL	Conceito UNL
	Classe	Classe verbal
	Subcategorização	Traços de Subcategorização
	Restrições	Restrições de seleção

O levantamento dos traços para cada categoria foi baseado exclusivamente nas necessidades identificadas no processo de mapeamento das estruturas sintáticas geradas pelo *parser* em estruturas conceituais. Na Tabela 2 são descritos os tipos de informação que esses traços representam e exemplos de valores para eles.

Tabela 2 – Informações que os traços representam.

Categoria sintática	**Traços**
Adjetivo	Canônica – forma básica do adjetivo, ex.: pessoal
	UNL – conceito UNL do adjetivo, ex.: *personal*
Advérbio	Tipo – tipo do advérbio, ex.: tempo, modo, negação, etc.
	UNL – conceito UNL do advérbio, ex.: *today*
Artigo	Tipo – tipo do artigo: definido ou indefinido
	Canônica – forma básica do artigo, ex.: um
	UNL – conceito UNL do artigo, ex.: *a*
Conjunção	Tipo – tipo da conjunção (coordenativa), ex.: aditiva, alternativa, adversativa
	UNL – conceito UNL da conjunção, ex.: *and*
Preposição	Canônica – forma básica da preposição, ex.: de
	UNL – conceito UNL da preposição, ex.: *of*
Pronome	Tipo – tipo do pronome, ex.: pessoal, de tratamento, reflexivo, etc.
	Canônica – forma básica do pronome, ex.: você
	UNL – conceito UNL do pronome, ex.: *you*

Substantivo	Número – número do substantivo: singular ou plural
	Canônica – forma básica do substantivo, ex.: emoção
	UNL – conceito UNL do substantivo, ex.: *emotion*
	Traços semânticos – traços do substantivo, ex.: animado, concreto, etc.
Verbo	Modo/tempo – modo e tempo verbal, ex.: presente do indicativo, futuro do subjuntivo, etc.
	Canônica – forma básica do verbo, ex.: prometer
	Tipo – tipo do verbo, ex.: transitivo direto, transitivo indireto, intransitivo, auxiliar, de ligação, etc.
	UNL – conceito UNL do verbo, ex.: *promise*
	Classe – classificação verbal, ex.: ação, estado, etc.
	Subcategorização – estrutura de argumentos do verbo, ex.: sujeito e objeto; sujeito, objeto e complemento, etc.
	Restrições de seleção – restrições semânticas dos argumentos do verbo, ex.: sujeito = animado, humano; objeto = concreto; etc.

Os traços "UNL" e "canônica" possuem valores em aberto, enquanto os demais traços possuem um conjunto limitado de alternativas para seus valores, definido em função de características da língua portuguesa, do *corpus* em questão e do modelo de mapeamento sintático-conceitual do Gerador Conceitual. A Tabela 3 apresenta essas alternativas, juntamente com sua sintaxe no Léxico Enriquecido. A especificação dessas opções de valores foi baseada no estudo de obras de referência e de outros recursos lingüísticos, conforme será descrito na Seção 4.

Tabela 3 – Opções de valores disponíveis para os traços restritos.

Categoria sintática	Traço	Opções	Sintaxe
Advérbio	Tipo	de Afirmação	afir
		de Dúvida	duv
		de Intensidade	int
		de Lugar	cir_lug
		de Modo	cir_mod
		de Tempo	cir_temp
		de Negação	neg
Artigo	Tipo	Indefinido	i
		Definido	de
Conjunções	Tipo	Coordenativa aditiva	coord,adit
		Coordenativa adversativa	coord,adve
		Coordenativa alternativa	coord,alter
		Coordenativa conclusiva	coord,concl
		Coordenativa explicativa	coord,expl
Pronomes	Tipo	Pessoal Oblíquo Tônico	obl_to
		Pessoal Oblíquo Átono	obl_at
		Pessoal Reto	ret
		Reflexivo	refl
		Tratamento	trat
		Possessivo	poss
		Demonstrativo	dem
		Interrogativo	inte
		Indefinido	inde

		Relativo	rel
Substantivo	Número	Plural Singular	pl si
	Traços semânticos	Concreto Abstrato Animado Inanimado Humano Não humano Lugar Tempo	concreto abstrato animado inanimado humano nao_humano lugar tempo
Verbo	Tempo/modo	Indicativo Presente Indicativo Pretérito Imperfeito Indicativo Pretérito Perfeito Indicativo Pretérito + que Perfeito Indicativo Futuro do Presente Indicativo Futuro do Pretérito Subjuntivo Presente Subjuntivo Pretérito Imperfeito Subjuntivo Pretérito Perfeito Subjuntivo Pretérito + que Perfeito Subjuntivo Futuro Imperativo Infinitivo Pessoal Gerúndio Particípio	Pres Pret_Imperf Pret Pret_M_Q_P Fut_Pres Fut_Pret Pres_Subj Pret_Imperf_Subj Pret_Subj Pret_M_Q_P_Subj Fut_Subj Imper_Afirm Inf_Pess Gerun Partic
	Tipo	Transitivo Direto Transitivo Indireto Bitransitivo Intransitivo Ligação Auxiliar Pronominal	td ti bi int lig aux pronom
	Classe	Ação Processo Ação-Processo Estado Modal	acao processo acao_proc estado modal
	Subcategorização	Sujeito Objeto Complemento 1 Complemento 2	suj obj comp comp1
	Restrições de seleção	Concreto Abstrato Animado Inanimado Humano Não humano Lugar Tempo	concreto abstrato animado inanimado humano nao_humano lugar tempo

4 O processo de aquisição de informações lingüísticas

Conforme mencionado, a definição do conjunto de traços necessários a cada categoria sintática foi feita com base nas necessidades do processo de mapeamento de estruturas sintáticas em conceituais. Após a identificação desses traços, partimos para a especificação de dois tipos de informação: 1) o conjunto das possíveis opções de valores para cada traço (no caso dos traços restritos ilustrados na Tabela 3); e 2) os valores de cada traço, para todos os itens lexicais.

Para auxiliar essa especificação, procuramos utilizar alguns recursos lingüísticos como fontes de informação. No entanto, isso só foi possível no caso dos traços morfossintáticos e do conceito UNL. No caso dos traços de natureza semântica, recorremos a obras de referência e recursos para outras línguas.

Para tornar clara a distinção entre as informações que puderam ser recuperadas de outros recursos lingüísticos e as informações que foram especificadas exclusivamente para o Léxico Enriquecido, bem como o processo para recuperar ou especificar essas informações, dividimos as questões referentes à aquisição de informações em duas partes, descritas nas seções seguintes.

4.1 Informações já disponíveis

A reutilização de recursos e módulos de processamento de interpretação do português disponíveis no NILC representa um papel importante no desenvolvimento do Gerador Conceitual como um todo, pois permite concentrar os esforços na obtenção da estrutura conceitual propriamente dita. No desenvolvimento do Léxico Enriquecido, essa metodologia de reutilização foi preservada, pois dispúnhamos de certos recursos bem elaborados e abrangentes, principalmente para o processamento morfossintático.

A partir desses recursos, portanto, recuperamos as informações de natureza morfossintática – tanto para a definição do conjunto de opções de valores para seus traços quanto para a definição dos valores desses traços – e o conceito UNL. Os recursos utilizados para fornecer essas informações e o modo como elas foram recuperadas são descritos nas Seções 4.1.1 e 4.1.2.

4.1.1 Traços morfossintáticos

Para fornecer os traços morfossintáticos ao Léxico Enriquecido foi utilizado o Léxico do Português do NILC (Nunes et al., 1996), doravante **Lex-Port**.

Cada item no Lex-Port é constituído de uma palavra, simples ou composta, nas suas formas analisada e canônica. Desse modo, o Lex-Port é uma lista de palavras contendo todas as conjugações de verbos (tempo, pessoa, modo) e flexões de gênero, número e grau, inclusive ênclises e mesóclises, totalizando cerca de 1.500.000 itens lexicais.

Como exemplo de entradas do Lex-Port, considere as descrições lexicais de algumas das palavras da sentença (1), repetida aqui, e suas respectivas canônicas, na Tabela 4.

(1) O Sol em Sagitário ilumina seus relacionamentos.

Tabela 4 – Exemplos de entradas lexicais do Lex-Port.

Palavra: relacionamentos	
Analisada	relacionamentos=<S.M.PL.N.[]?.?.[relacionamento]0.>
Canônica	relacionamento=<S.M.SI.N.[]?.?.[relacionamento]0.>
Palavra: ilumina	
Analisada	ilumina=<V.[][IMPER-AFIRM.TU.PRES.ELE.]N.[][iluminar]0.>
Canônica	iluminar=<V.[PRONOM.TD.][FUT-SUBJ.ELE.FUT-SUBJ.EU.INF-PESS.ELE.INF-PESS.EU.]N.[][iluminar]0.>

Na descrição de "ilumina", por exemplo, as informações morfossintáticas indicam que se trata de um verbo (V), 2ª pessoa singular do imperativo-afirmativo (IMPER-AFIRM.TU), ou no modo/tempo 3ª pessoa singular do presente do indicativo (PRES.ELE), cuja canônica é "iluminar". Já na descrição de "iluminar", as informações indicam que se trata também de um verbo (V), do tipo pronominal (PRONOM) ou transitivo direto (TD), modo/tempo 3ª ou 1ª pessoa singular do futuro do subjuntivo (FUT-SUBJ.ELE.FUT-SUBJ.EU), ou 3ª ou 1ª pessoa singular do infinitivo pessoal (INF-PESS.ELE.INF-PESS.EU), cuja canônica é o próprio "iluminar".

A Tabela 5 ilustra os traços do Lex-Port que foram incorporados ao Léxico Enriquecido. Nem todos os traços morfossintáticos do Lex-Port são necessários ao Léxico Enriquecido. Neste último, algumas categorias sintáticas nem sequer apresentam traços morfossintáticos (isto é, adjetivos e preposições) enquanto outras apresentam somente para a forma canônica (isto é, advérbios e conjunções).

Tabela 5 – Traços do Léxico Enriquecido importados do Lex-Port.

Categoria sintática	Traços da forma analisada	Traços da forma canônica
Advérbio	-	Tipo
Artigo	Tipo, canônica	Tipo
Conjunção	-	Tipo
Preposição	Canônica	
Pronome	Tipo, canônica	Tipo
Substantivo	Número, canônica	Número
Verbo	Modo/tempo	Modo/tempo, tipo

Para especificar o conjunto de opções para cada traço ilustrado na Tabela 3, além das informações provindas do Lex-Port, foram consideradas certas limitações em função do *corpus* e das restrições do Gerador Conceitual. Por exemplo, apesar de existirem no Lex-Port conjunções coordenativas e subordinativas, somente as coordenativas estão sendo consideradas no Léxico Enriquecido, pois as subordinativas, que introduzem um novo período (ou oração) em uma sentença, não são consideradas pelo Gerador Conceitual, que só trata de sentenças com um único verbo, ou, no máximo, locuções verbais.

Conforme mencionado, no Léxico Enriquecido, os valores dos traços de uma palavra foram definidos de acordo com o(s) uso(s) da palavra no *corpus* e, no caso da ocorrência de uma mesma palavra com diferentes usos, foram criados diferentes itens para ela. Para o verbo "ilumina", por exemplo, foi preciso escolher qual o valor correto para o traço que indica tempo/modo (IMPER-AFIRM.TU ou PRES.ELE). Se ambos tivessem sido necessários, teria sido preciso separar a entrada em outras duas, uma para cada valor de traço.

Com relação à sintaxe do Léxico Enriquecido, foram mantidos, por convenção, somente os símbolos que representam os valores para os traços, mas em letras minúsculas,

devido às restrições da linguagem na qual foi implementado (conforme será descrito na Seção 4).

A transferência dos traços do Lex-Port para o Léxico Enriquecido foi feita por meio de um processo de cópia manual, principalmente por dois motivos: as diferenças no formalismo de representação e na sintaxe dos dois recursos e a necessidade de escolha por uma entre várias opções de conjuntos de traços ou de valores de traços (ou a separação em dois ou mais entradas lexicais para o mesmo item).

4.1.2 Conceito UNL

Para especificar o valor do traço correspondente ao conceito UNL de cada item lexical foi utilizado Dicionário Português-UNL (Dias-da-Silva et al., 1998), doravante **Dic Port-UNL**, criado durante o desenvolvimento do módulo de decodificação da UNL para o português (Nunes et al., 1997) do Projeto UNL-Brasil (Oliveira et al., 2001).

O Dic Port-UNL faz a correspondência entre as palavras da língua portuguesa (formas canônicas, apenas) e os conceitos da interlíngua UNL. Uma palavra da língua portuguesa pode remeter a vários conceitos UNL, os quais correspondem às diferentes acepções do item. Para tanto, para cada canônica do português são criadas uma ou mais entradas lexicais, cada qual dispondo dos seguintes traços: 1) a UW que representa seu conceito, em termos de suas relações com os outros conceitos (caso existam)[7]; 2) a categoria gramatical; 3) informações morfológicas (gênero, número, tempo verbal, etc.); 4) informações sintáticas (regência); 5) tipo do verbo (ação ou estado); e 6) algumas informações adicionais (idioma, freqüência e prioridade). A Tabela 6 ilustra exemplos de algumas descrições para palavras da sentença em (1).

Tabela 6 – Exemplos de entradas lexicais do Dic Port-UNL.

Palavra: sol
[Sol] {} "sun(icl>solar system)" (s,masc,sing) <P,0,2>;
[sol] {} "sun" (s,masc,sing) <P,0,2>;
Palavra: iluminar
[ilumin] {} "stimulate" (v,5,stem,vtd,ação) <P,0,0>;
[ilumin] {} "illuminate" (v,5,stem,vtd,ação) <P,0,0>;
[ilumin] {} "enlighten" (v,5,stem,vtd,ação) <P,0,0>;
[ilumin] {} "inspire" (v,5,stem,vtd,ação) <P,0,0>;
[ilumin] {} "light up" (v,5,stem,vtd,ação) <P,0,0>;
Palavra: em
[em] {} "at" (prep) <P,0,1>;
[em] {} "in" (prep) <P,0,2>;
[em] {} "in(icl>manner)" (prep) <P,0,1>;
[em] {} "into" (prep) <P,0,1>;
[em] {} "on" (prep) <P,0,2>;

[7] A UNL oferece a opção de se utilizar conceitos restritos, que remetem a outros conceitos na sua hierarquia conceitual (ontologia), como é o caso da primeira descrição para "sol" na Tabela 6: "sun(icl>solar system)", que indica "sun" está numa relação de hiponímia (inclusão) com "solar system". No entanto, como a definição dessa hierarquia ainda não foi concluída pelo Projeto UNL, no Léxico Enriquecido foram utilizadas somente os conceitos genéricos da UNL, neste caso, a segunda descrição para "sol": "sun".

Como os demais traços morfossintáticos existentes no Dic Port-UNL foram recuperados, no Léxico Enriquecido, a partir do Lex-Port, a única informação do Dic Port-UNL utilizada no Léxico Enriquecido foi o conceito UNL para as entradas das formas canônicas. Para recuperar essa informação, no entanto, na maioria das vezes foi preciso escolher uma, entre as várias descrições de um mesmo item, seja para conceitos totalmente diferentes, seja para variações do mesmo conceito. Para evitar a ambigüidade semântica, somente um conceito para cada item foi representado no Léxico Enriquecido.

Como pode ser verificado na Tabela 6, o formato das entradas do Dic Port-UNL não é igual ao do Lex-Port e tampouco igual ao do Léxico Enriquecido, nem mesmo as canônicas são representadas da mesma forma: no Dic Port-UNL elas são apenas os radicais das palavras, as desinências nominais/verbais são representadas como traços do item. Por exemplo, uma das acepções de "iluminar", cujo conceito UNL é "illuminate", tem como item "illumin" e uma indicação de que se trata de um verbo cuja desinência se dá de acordo com um determinado paradigma de conjugação (5), que indica a terminação em "-ar".

Pelos mesmos motivos citados no caso do Lex-Port, ou seja, a não uniformidade dos formatos de representação e a necessidade de escolha entre dois ou mais conceitos para um mesmo item, a transferência do conceito UNL para o Léxico Enriquecido foi feita por meio de um processo de escolha e cópia manual.

Como uma alternativa aos dois recursos citados (isto é, Lex-Port e Dic Port-UNL), outro recurso, também disponível no NILC, foi investigado como possível fonte de informação lingüística: a DIADORIM (Greghi et al., 2001; Greghi, 2002), uma base de dados lexicais cujo objetivo é unificar e representar as informações provindas de outros recursos, incluindo esses dois. Nessa base, cada canônica do Lex-Port é associada a um conceito UNL, o que facilitaria a recuperação das informações necessárias, inclusive de forma automática. No entanto, optamos por não utilizar essa base, pois, no momento em que iniciamos o desenvolvimento do léxico, a inclusão do conceito UNL ainda não havia sido concluída.

4.2 Informações específicas para o Gerador Conceitual

Por se tratar da tarefa de geração conceitual, as informações morfossintáticas e o conceito UNL não são suficientes para descrever um item lexical; são necessárias também informações de natureza semântica. A definição das fontes lingüísticas para fornecer essas informações representou o maior problema no desenvolvimento do Léxico Enriquecido.

Com base na análise do processo de mapeamento, as seguintes informações semânticas foram identificadas como necessárias ao Léxico Enriquecido: 1) informações de subcategorização, que indicam o(s) ambiente(s) sintático(s) no(s) qual(is) o verbo pode ocorrer, ou seja, indicam quais os argumentos do verbo (sujeito, objeto, etc.); 2) restrições de seleção, que limitam as propriedades semânticas dos argumentos do verbo; 3) traços semânticos, que determinam propriedades semânticas dos componentes nominais da sentença (somente dos substantivos); e 4) a classe, que indica a categoria de um determinado verbo num dado contexto. A Tabela 7 ilustra a distribuição dessas informações nas categorias sintáticas em que foram incorporadas: verbos e substantivos. Como no caso do conceito UNL, essas informações foram adicionadas somente às entradas das formas canônicas do Léxico Enriquecido.

Tabela 7 – Traços semânticos do Léxico Enriquecido.

Categoria sintática	Traços da forma analisada	Traços da forma canônica
Substantivo	-	Traços semânticos
Verbo	-	Subcategorização, restrições de seleção, classe

Para a escolha do conjunto de opções para cada uma dessas informações semânticas e para a definição dos seus valores para os diferentes itens lexicais, foram investigadas os modelos de representação lexical citados na Seção 3, analisados agora sob o enfoque da sua capacidade de representação semântica e da sua adequabilidade à estrutura do Léxico Enriquecido e aos objetivos do Gerador Conceitual.

A primeira categoria citada, ou seja, os modelos completos de representação semântica como a Dependência Conceitual, a Semântica Conceitual e a LCS (*Lexical Conceptual Structure*), possuem uma estrutura de representação conceitual própria, diferente da UNL, portanto, alguns deles não fornecem as informações necessárias ao Léxico Enriquecido para o mapeamento no Gerador Conceitual, ou, quando fornecem, tais informações se apresentam num formato diferente e uma possível conversão entre esse formato e o da representação UNL seria muito complexa.

Dos modelos de representação lexical independentes de aplicação, a Teoria do Léxico Gerativo possui mecanismos eficazes de organização lexical, como a composicionalidade e a herança, e uma descrição lexical bastante rica e complexa. No entanto, apresenta muitas informações semânticas cuja especificação exige investigação lingüística avançada e que não são necessárias à estrutura simplificada do Léxico Enriquecido. Já a linguagem DATR, além de alguns mecanismos de organização lexical como herança, inferência e algumas técnicas de otimização, possui uma estrutura bastante simples, na qual as informações são codificadas em termos de atributos e valores. Contudo, essa linguagem não provê uma especificação de quais informações semânticas poderiam ser representadas, e, conseqüentemente, não fornece as informações necessárias ao Léxico Enriquecido.

Os modelos de representação lexical dependentes de aplicação, em geral, apresentam características específicas para as necessidades de tal aplicação, portanto, não fornecem todas as informações necessárias ao Léxico Enriquecido, ou fornecem informações desnecessárias. Além disso, exceto o modelo do léxico do projeto Mikrokosmos e a estrutura lexical do sistema Wordnet, os demais léxicos analisados eram fortemente baseados em outros modelos de representação lexical e/ou semântica, como a Teoria da Semântica Conceitual e a Teoria do Léxico Gerativo.

Por não se mostrarem adequados à estrutura e aos objetivos almejados, nenhum desses modelos foi integralmente empregado no desenvolvimento do Léxico Enriquecido. Assim, na falta de um modelo que pudesse prover todas as informações semânticas necessárias, resolvemos investigar fontes para essas informações separadamente, apesar de reconhecermos seu inter-relacionamento. Novamente, determinadas características de alguns desses modelos foram analisados, juntamente com outros modelos específicos para cada uma das informações.

O primeiro ponto analisado foi a classificação dos verbos, pois a escolha de um determinado modelo lingüístico para essa informação certamente influenciaria na especificação das demais informações.

Os modelos completos de representação semântica citados prevêem a classificação dos verbos em um conjunto de ações ou estados primitivos, representados por meio de

conceitos independentes de LN. A classificação da DC prevê um conjunto limitado dessas ações ou estados, no entanto, não fornece uma indicação clara de quais verbos pertenceriam a quais classes. Já nas classificações da Semântica Conceitual e da LCS (que é derivada da teoria da Semântica Conceitual) o conjunto dessas ações ou estados não é delimitado, vários conceitos podem ser especializados para representar novas ações ou estados primitivos. Com isso, a classificação se torna bastante detalhada e, como conseqüência, complexa.

Entre os modelos de representação lexical, a Teoria do Léxico Gerativo, o léxico do sistema Wordnet e o léxico do sistema Mikrokosmos apresentam uma classificação verbal própria. Nesses dois últimos sistemas, tal classificação faz parte de uma organização lexical mais ampla, que envolve todas as categorias de palavras, não apenas os verbos e é, portanto, bastante complexa e específica para a aplicação.

A estrutura de classificação verbal da Teoria do Léxico Gerativo, ou seja, a Estrutura de Eventos dessa teoria (Pustejovsky, 1991; Pustejovsky ,1995), foi analisada juntamente com outras propostas de mesma finalidade: o Esquema de Tempo de Vendler (1967), a Tipologia de Verbos de Chafe (1970), a Tipologia de Estados de Coisas de Dik (1997), as Classes Verbais de Levin (1994) – todos para a língua inglesa – e a Teoria da Valência de Borba (1990), que deriva, em parte, do modelo de Chafe. A seguir, apresentamos uma breve discussão sobre essas abordagens de classificação verbal, a fim de justificar a escolha pela Teoria da Valência de Borba.

A classificação de Levin, baseada nos diferentes sentidos dos verbos, foi desconsiderada em função do seu alto grau de especificidade e complexidade. As demais propostas apresentam algumas variações, as quais são sintetizadas na Tabela 8 (Moraes, 2002) e descritas em seguida.

Tabela 8 – Características dos modelos de Vendler, Dik, Pustejovsky e Chafe/Borba.

Autor	Proposta	Categoria	Tipos de Categorias
Vendler	Esquema Temporal	Baseadas em relações temporais	*Estado, Atividade, "Accomplishment", "Achievement"*
Dik	Tipologia de Estados de Coisas	Baseadas nos valores de parâmetros primitivos	*Situação, Evento, Processo e Ação.*
Pustejovsky	Estrutura de Eventos	Baseadas em permanência ou mudança de estado e causa	*Estados, Processos e Transições ("accomplishment" e "achievement")*
Chafe/Borba	Tipologia de Verbos	Definidas pela combinação do predicador e seus argumentos	*Estados, Ação, Ação-processo e Processo.*

Na sua proposta, Vendler apresenta um esquema de tempo (*time schemata*) pressuposto por vários verbos, com quatro classes de natureza estritamente temporal: estados, atividades, *achievements* e *accomplishments*. Os **estados** não possuem mudança durante o período de tempo em que são verdade (ex.: "Ana ama Pedro"). As **atividades** correspondem a eventos constituídos de fases sucessivas, sem um limite obrigatório (ex.: "Maria correu pela rua"). Já ***accomplishments*** são eventos com duração e final obrigatórios (ex.: "Ana caminhou até sua casa"), enquanto ***achievements*** são eventos com final instantâneo, sem duração (ex.: "Ana chegou em casa").

A Tipologia de Estados de Coisas de Dik é determinada pelos valores (+ ou –) de um grupo de parâmetros, sendo os seguintes os mais importantes: dinâmico, télico[8], momentâneo e de controle. Os três primeiros parâmetros correspondem aos tipos verbais de Vendler. O parâmetro [+ ou – dinâmico] diferencia os estados (que Dik chama de "situações") das demais categorias (chamadas de "eventos"). A diferença entre as atividades, *accomplishments* e *achievements* da teoria de Vendler é estabelecida pelo parâmetro [+ ou – télico], pois um evento télico tem um final temporal obrigatório, dessa forma *achievements* e *accomplishments* são [+ télico], enquanto atividades são [– télico]. A diferença entre *accomplishments* e *achievements* é feita com o uso do parâmetro [+ ou – momentâneo]. Um estado de coisas é [+ ou – controlado] se seu primeiro argumento determinar se o estado de coisas vai ocorrer ou não. A combinação de valores desses e de outros parâmetros gera uma tipologia com os seguintes tipos: **situação** e seus subtipos (**estado**, **posição**), **evento**, **processo** e seus subtipos (**dinamismo**, **mudança**), **ação** e seus subtipos (**atividade** e ***accomplishment)***. Por exemplo, para a sentença "Ana caminhou até a sua casa", o tipo de estado de coisas é *accomplishment*, definido pelos parâmetros [+ dinâmico], [+ controlado], [+ télico] e [– momentâneo].

Na Teoria do Léxico Gerativo, Pustejovsky propõe uma Estrutura de Eventos composta de três tipos diferentes de eventos: **estados**, que de indicam um evento único que é avaliado sem referência a nenhum outro evento (ex.: "João ama Maria"); **processos**, que indicam uma seqüência de eventos que identificam a mesma expressão semântica (ex.: " Ana corre pela rua"; e **transições**, que indicam um evento identificando uma expressão semântica, avaliado em relação a sua oposição (ex.: "Paulo fechou a porta"). Esse último tipo de evento engloba as categorias *achievements* e *accomplishments* da teoria de Vendler. Por exemplo, em "a porta fechou-se", assume-se um *achievement* e em "João fechou a porta", um *accomplishment.* Em vez de consideras apenas o fator de duração temporal inerente a verbos, Pustejovsky parte das relações de causa, permanência e mudança de estados que os eventos expressam.

A tipologia de verbos de Chafe, da qual o modelo de Borba é derivado, parte de expressões semânticas que são resultados da combinação de elementos predicativos, que expressam estados ou eventos (verbos), com elementos nominais. As categorias verbais são definidas por Chafe de acordo com os valores semânticos tanto do verbo como dos argumentos que o acompanham. Para a classificação verbal do Léxico Enriquecido, essa tipologia se mostrou o modelo mais aplicável, em função de uma série de motivos, dentre eles os listados abaixo:

- A estrutura do modelo é apropriada para a tarefa pretendida, devido à própria estrutura conceitual da UNL, que também é baseada na estrutura de argumentos (subcategorização), ou seja, ambas as estruturas de classificação de verbos são baseadas na composição semântica entre o verbo e seus argumentos.
- É o modelo mais adequado para a língua portuguesa, pois algumas questões consideradas pelos outros modelos, como os esquemas temporais de Vendler e Dik, não são aplicáveis ao português, considerando que o tratamento aspectual/temporal de alguns verbos é diferente nas duas línguas;
- É um modelo relativamente simples, com um conjunto limitado e pequeno de classes de verbos.
- É o único modelo adaptado para a língua portuguesa, com o dicionário de Borba (1990) contendo a classificação dos principais verbos da língua;

[8] O termo "télico", no PLN, é utilizado para denotar uma ação que visa atingir uma finalidade e que cessa quando tal finalidade é atingida, a exemplo dos verbos construir, consertar, adormecer, entrar, entre outros.

- Nesse dicionário, além da classificação dos verbos, são definidas a subcategorização e as restrições de seleção desses verbos. Ao definir as restrições de seleção, de forma indireta, Borba define também um conjunto de opções de traços semânticos para substantivos, que auxilia a identificação dos traços semânticos dos substantivos do *corpus*, por meio de um estudo desse *corpus*; e
- Esse modelo já foi aplicado em outro trabalho do NILC, o projeto TraSem (Rino et al., 2001), cujo objetivo era a definição dos traços semânticos dos itens do Lex-Port para melhorar o desempenho do revisor ortográfico e gramatical ReGra. Poderíamos, portanto, aproveitar a experiência adquirida nesse projeto.

Mantendo a classificação verbal de Chafe, escolhida por ser a mais apropriada pelos motivos citados acima, adotamos o modelo de Borba como principal referência para a obtenção das informações semânticas para o Léxico Enriquecido, visto que esse modelo é voltado para a língua portuguesa e que o referido dicionário fornece suporte para a definição das informações semânticas adicionais, além da classificação verbal. Portanto, apesar de algumas informações serem derivadas do modelo de Chafe, nos referiremos, a partir de agora, somente ao modelo de Borba.

Conforme ilustrado na Tabela 3, as quatro classes verbais utilizadas nesse modelo são:

1) Ação: atividade expressa pelo verbo e realizada pelo sujeito agente. O verbo de ação indica um fazer, por parte do sujeito. Ex.: "o pássaro voa"; "o homem pensa"; "o garoto brinca".

2) Ação-Processo: expressão de uma ação realizada por um sujeito agente e/ou de uma causação levada a efeito por um sujeito causativo, que afetam um complemento. A ação-processo sempre atinge um complemento que expressa uma mudança de estado, de condição ou de posição, ou, então, algo que passa a existir. Ex.: "José abriu a porta"; "José escreveu um romance".

3) Estado: expressão de uma propriedade, de uma condição ou de uma situação localizadas no sujeito. Ex.: "Ana é bonita"; "eu estou com fome".

4) Processo: evento ou sucessão de eventos que afetam um sujeito paciente, experimentador ou beneficiário. Um verbo de processo traduz algo que se passa com o sujeito ou que ele experimenta ou recebe. Ex.: "a chuva parou"; "o bebê acordou"; "Ana sente frio"; "Marta ouve música".

Além dessas classes originais do modelo de Chafe, consideramos também uma classe para expressar modalidade, à qual chamamos de "**Modal**", seguindo a nomenclatura de Borba, para classificar verbos cuja função é de auxiliar de outro verbo, ou seja, que modificam a relação existente entre o sujeito e o predicado. Ex.: "João *deve* ser bom", "O tempo *pode* mudar".

Ao escolher o modelo de Borba como fonte lingüística para a classificação dos verbos, além do traço "classe", as demais informações semânticas necessárias (subcategorização, restrições de seleção e traços semânticos) puderam ser especificadas com base no seu dicionário de verbos e no estudo do *corpus* em questão.

Por exemplo, segundo o dicionário, o verbo iluminar, na acepção adequada para a sentença (1), indica ação-processo, com primeiro o argumento da sentença ("o sol em sagitário") semanticamente especificado como um sujeito "agente" e o segundo argumento

(“seus relacionamentos”) semanticamente especificado como um complemento expresso por nome abstrato.

Com essas informações, o dicionário permite identificar que o verbo requer, como argumentos (**subcategorização**), um “sujeito” e um “complemento”, e que, como **restrição de seleção**, esse complemento deve ter o traço “abstrato”. Além dessa restrição de seleção explícita, identificamos, implicitamente, que o sujeito, por ser agente, deve ser “animado”. Com isso, foi possível especificar, também, os **traços semânticos** dos constituintes da sentença: o núcleo do constituinte “o sol em sagitário”, ou seja, “sol”, deve ter o traço “animado” e o núcleo do objeto “seus relacionamentos”, ou seja, “relacionamentos”, deve ter o traço “abstrato”.

Apesar de Borba não explicitar um conjunto fechado de traços semânticos para descrever os constituintes nominais da sentença (apenas os substantivos, neste caso), com base nos estudos realizados no projeto TraSem e no estudo do *corpus* considerado para o Gerados Conceitual, foram identificados os seguintes traços:

- **Baseados no projeto TraSem**: Concreto, Abstrato, Animado, Inanimado, Humano, Não humano.
- **Baseados na análise do *corpus***: Lugar, Tempo.

Como vimos no exemplo, ao identificar quais são os argumentos de um verbo, Borba define algumas restrições para esses argumentos. Essas restrições são definidas de três diferentes modos: 1) os argumentos do verbo são especificados de acordo com seus papéis semânticos, por exemplo, “ingressar” necessita de um sujeito “agente” e um complemento “locativo”; 2) os argumentos do verbo são especificados de acordo com seus traços semânticos, por exemplo, “acontecer” necessita de sujeito expresso por nome “abstrato”, “esperar” necessita de sujeito expresso por nome “animado” e complemento expresso por nome “abstrato”; e 3) os argumentos são especificados sem restrições semânticas, por exemplo, “ser” necessita de sujeito inativo e predicativo expresso por nome, adjetivo ou equivalente.

Para os casos em que utiliza os papéis semânticos dos argumentos para definir as restrições de seleção do verbo, Borba considera os seguintes papéis: Agente, Beneficiário, Experimentador, Objetivo, Locativo, Instrumental, Factitivo, Causativo, Meta, Origem, Temporal. Os constituintes sintáticos que exercem esses papéis devem apresentar determinados traços semânticos, dependendo da sua ocorrência no *corpus*. Portanto, não hás um mapeamento direto entre função sintática e semântica, por exemplo, a função **sujeito** pode ser mapeada em diferentes papéis semânticos, conforme ilustrado por Borba: **sujeito agente** (aquele que desencadeia uma atividade); **sujeito beneficiário** (sede da transferência de posse ou destinatário de um benefício); **sujeito causativo** (o que provoca um efeito ou, então, é o responsável pela realização do estado de coisas indicado no verbo); **sujeito experimentador** (aquele que expressa uma experiência ligada a uma disposição mental); **sujeito factitivo** (o que instiga ou estimula uma ação, isto é, o que comanda um agente); **sujeito paciente** (o afetado por aquilo que o verbo expressa); e **sujeito inativo** (suporte de uma propriedade, condição ou situação expressa pelo predicado).

Alguns autores, como Lobato (1986), afirmam que nem todas as restrições de seleção podem ser formuladas em termos de um conjunto finito de traços, visto que certos traços restringem muito a escolha dos elementos que podem ser selecionados como argumentos de um verbo. No entanto, no Léxico Enriquecido adotamos como restrições de seleção o conjunto finito de traços semânticos exigidos pelo argumento e não os papéis semânticos desses argumentos, pois não dispúnhamos dos papéis semânticos necessários

aos argumentos de todos os verbos (conforme mencionado nos três casos de restrições de seleção descritos acima), e pretendíamos descrever os traços semânticos dos substantivos. Além disso, a conversão das restrições de um papel semântico em restrições na forma de traços semânticos, com base na análise de *corpus*, é relativamente simples, o que não ocorre no caso da conversão de um conjunto de traços em um papel semântico. Além disso, seria mais simples designar traços que papéis semânticos como restrições para aqueles casos não especificados por Borba (o caso 3 descrito acima). Por exemplo, para o verbo "ingressar", que exige sujeito "agente", é possível identificar que esse sujeito deve ser animado e concreto, em uma sentença como "O sol ingressou em sagitário". Já no caso do verbo "prometer", que exige sujeito expresso por nome "não animado", não é trivial a identificação de qual seria o papel semântico desse sujeito. Borba atribui o papel "inativo" para esses casos, mas o uso de papel nada acrescentaria em termos de restrições de seleção.

Identificadas as fontes lingüísticas para todas as informações necessárias ao léxico (morfossintáticas, conceito UNL e semânticas), passamos ao processo de implementação do Léxico Enriquecido, que será descrito na seção seguinte.

5 A implementação do Léxico Enriquecido

Em termos computacionais, existem várias maneiras de representar um léxico, de acordo com seu tamanho e com a aplicação a que se destina. As formas mais comuns são os arquivos texto e os bancos de dados. No primeiro caso, a representação é simples, porém, não otimizada, principalmente para léxicos com número elevado de itens, pois a manipulação dos arquivos pode tornar-se muito complexa, onerosa e lenta. No segundo caso, o trabalho de armazenamento e indexação dos itens é feito pelo sistema gerenciador de banco de dados, o que facilita a manipulação do léxico, permitindo, por exemplo, a realização de consultas complexas e o seu compartilhamento entre diferentes usuários. Essa representação demanda, no entanto, um grande espaço de armazenamento. Além dessas formas comuns de representação, para aplicações que necessitam acessar léxicos grandes, de modo rápido e otimizado, pode-se utilizar estruturas de dados mais eficientes como árvores, tabelas *hash*, autômatos finitos, listas, redes neurais, etc.

Por se tratar de um léxico com poucos itens e pelas facilidades que a linguagem utilizada para implementar o Gerador Conceitual – Prolog – oferece para acessar bases de dados textuais (apesar de ser possível a interação com outros tipos de bases), foi escolhido o formato de arquivo texto, sendo as entradas especificadas de acordo com a sintaxe do Prolog, ou seja, a representação por cláusulas (regras ou fatos) no formalismo do Cálculo de Predicados de Primeira Ordem (ou Lógica de Primeira Ordem).

A linguagem Prolog foi utilizada para implementação do Gerador Conceitual por apresentar maior clareza de representação para PLN e por possuir um mecanismo próprio de inferência, que facilita o processo de unificação das entradas do léxico com as estruturas utilizadas na geração conceitual.

No Cálculo de Predicados de Primeira Ordem, uma cláusula consiste de um predicado e um conjunto ordenado de termos como seus argumentos – sendo que estes podem ser variáveis, constantes ou outros predicados –, podendo também apresentar conectivos lógicos (negação, conjunção, disjunção, implicação, etc.) e quantificadores. Por exemplo, as cláusulas na Tabela 9 são válidas em Prolog:

Tabela 9 – Exemplos de cláusulas do Prolog.

Proposição	Significado
humano(joão)	joão é humano
gosta(joão,vinho)	joão gosta de vinho
todo(X, homem(X) → humano(X))	para todo X, se X é homem, então X é humano
humano(joão) :- homem(João)	se joão é homem então joão é humano

Para tornar a declaração das entradas lexicais mais clara, em vez de Prolog puro, utilizamos o formalismo da DCG (*Definite Clause Grammar*), uma ferramenta específica de modelagem lingüística criada por Pereira & Warren (1980) e incorporada às diversas versões do Prolog, que permite a representação de informações gramaticais (sintáticas e/ou semânticas) e lexicais para sistemas de PLN.

O formalismo da DCG é bastante flexível, pois possibilita a representação de qualquer cláusula do Cálculo de Predicados de Primeira Ordem, desde que sejam circunscritas a cláusulas do Prolog puro, como as exemplificadas na Tabela 9. Por exemplo, uma regra em Prolog puro to tipo X:-Y, que é interpretada por "X se Y", é reescrita em DCG como X-->Y.

Para a representação de informações lexicais, uma cláusula da DCG tem a forma de uma regra de produção do tipo X-->[Y], que faz a correspondência de um símbolo não-terminal X com um item lexical Y. Por exemplo, na DCG, a palavra "relacionamentos", poderia ser descrita pela cláusula em (3):

(3) s(sin(pl),can(relacionamento)) --> [relacionamentos].

Essa cláusula é constituída de um predicado "s", que simboliza a categoria sintática do item (substantivo), com dois argumentos, ambos também predicados com outros argumentos: 1) "sin", que simboliza os traços sintáticos e tem como argumento o número do substantivo (pl), 2) "can", que simboliza a forma canônica e tem como argumento a forma canônica do substantivo (relacionamento). O item lexical é sempre representado do lado direito e na forma de lista.

É importante ressaltar que essa não é uma estrutura padrão para todos os léxicos, mesmo em DCG. Estruturas diferentes (como árvores, grafos, etc.), predicados e/ou argumentos diferentes poderiam ter sido definidos. Por exemplo, os argumentos que correspondem aos traços sintáticos e à canônica poderiam ser especificados como termos atômicos de "s", ou como um único argumento na forma de lista, não como outros predicados.

Outros exemplos de entradas do Léxico Enriquecido especificados dessa forma são ilustrados na Tabela 10.

Tabela 10 – Exemplos de entradas do Léxico Enriquecido.

Adjetivos	
Palavra: pessoal	
Analisada	adj(can(pessoal)) → [pessoais].
Canônica	adj(unl(personal)) → [pessoal].
Advérbios	
Palavra: hoje	
Canônica	adv(sin(cir_temp),unl(today)) → [hoje].

Artigos	
Palavra: o	
Analisada	art(sin(de),can(o)) --> [os].
Canônica	art(sin(de),unl(the)) --> [o].
Conjunções	
Palavra: e	
Canônica	conj(sin(coord,adit),unl(and)) → [e].
Preposições	
Palavra: do	
Analisada	prep(can(de)) --> [do].
Canônica	prep(unl(of)) --> [de].
Palavra: em	
Canônica	prep(unl(in)) --> [em].
Pronomes	
Palavra: você	
Analisada	pron(sin(trat),can(você)) → [vocês].
Canônica	pron(sin(trat),unl(you)) → [você].
Substantivos	
Palavra: relacionamentos	
Analisada	s(sin(pl),can(relacionamento)) --> [relacionamentos].
Canônica	s(sin(si),unl(relationship),sem([abstrato,inanimado,nao_humano])) --> [relacionamento].
Palavra: sol	
Canônica	s(sin(si),unl(sun),sem([concreto,animado,nao_humano])) --> [sol].
Palavra: sagitário	
Canônica	s(sin(si),unl(sagittarius),sem([abstrato,inanimado,nao_humano,lugar])) --> [sagitário].
Verbos	
Palavra: ilumina	
Analisada	v(sin(pres),can(iluminar)) --> [ilumina].
Canônica	v(sin(inf_pess,td),unl(illuminate),cl(acao_proc),sub([suj,obj]), rest([suj([concreto,animado]),obj([abstrato])])) --> [iluminar].

Uma vez que, quando interpretadas (ou compiladas), as cláusulas DCG são transformadas diretamente em cláusulas em Prolog, essa estrutura não é eficiente em termos de desempenho computacional. No entanto, como o Léxico Enriquecido possui poucos itens, o uso de estruturas DCG não implica um prejuízo significante ao desempenho geral do sistema. Além disso, o uso dessa estrutura não inviabiliza uma futura conversão em estruturas mais eficientes, que possam vir a otimizar o desempenho do Gerador Conceitual caso o número de itens lexicais cresça consideravelmente.

Conforme mencionado, a extração de informações dos recursos disponíveis (Lex-Port e Dic Port-UNL) foi feita de forma manual. Sempre que possível, a sintaxe das informações recuperadas foi mantida, em alguns casos, no entanto, ela teve de ser modificada em função de restrições da linguagem Prolog, principalmente restrições relativas à manipulação de *strings*, como concatenações. Alguns exemplos das modificações realizadas são:

- Expressões com espaço (itens na forma analisada, itens na forma canônica ou valores de traços), como "*go in*" e "*in tune*" tiveram o espaço substituído por *underscore*: "*go_in*" e "*in_tune*";

- Traços com caracteres especiais como acentos e cedilhas, como “não humano”, tiveram tais caracteres removidos: “nao_humano”. Os caracteres especiais foram mantidos apenas nos itens na forma analisada e na forma canônica que não são sofreram manipulações problemáticas;
- Expressões com hífens (itens na forma analisada, itens na forma canônica ou valores de traços), como “bem-vindo”, tiveram o hífen substituído por *underscore*: “bem_vindo”.

Para que o Gerador Conceitual não apresente como resultado palavras grafadas de forma incorreta, essas alterações serão revertidas, quando necessário, pelo programa responsável pela interface entre o Prolog e o usuário.

6 Considerações Finais

Apesar de reconhecida a importância da semântica para muitas aplicações do PLN, o nível de desenvolvimento das pesquisas nessa área não é tão avançado quanto nas demais. Isso se deve à própria complexidade da área e à escassez de fontes de informações lingüísticas dessa natureza.

Especialmente no caso da língua portuguesa, os sistemas que realizam o processamento semântico acabam, geralmente, definindo recursos lingüísticos bastante específicos para sua aplicação, como os léxicos desenvolvidos para os projetos Nalamas (Lima et al., 196) e Linx (Garcia, 195). Dessa forma, são poucos os casos em que se pode contar com a reutilização ou o compartilhamento de recursos lingüísticos que oferecem suporte ao processamento semântico.

A principal motivação para o desenvolvimento do Léxico Enriquecido foi justamente a falta de outros recursos que pudessem fornecer as informações semânticas necessárias ao processamento do Gerador Conceitual, uma vez que, para as informações morfossintáticas, já existem outras fontes de referência, incluindo recursos lingüísticos disponíveis no NILC. Durante o seu desenvolvimento, a principal dificuldade encontrada foi a falta de modelos lingüísticos do português que pudessem servir como base, indicando quais opções de valores poderiam ser utilizadas para os diferentes traços identificados e quais os valores desses traços.

Após a investigação de diferentes abordagens, adotamos como modelo a Teoria da Valência de Borba (1990), que forneceu, direta ou indiretamente (juntamente com a análise do *corpus*), todas as informações semânticas necessárias.

Certamente, a existência de outros recursos semânticos para a língua portuguesa, como léxicos, poderia beneficiar muitas pesquisas em PLN, pois, como podemos verificar com o desenvolvimento do Léxico Enriquecido, o esforço e o tempo despendidos para essa tarefa são grandes, e podem, portanto, acabar inviabilizando tais pesquisas.

Nesse sentido, o Léxico Enriquecido, apesar de ser de abrangência bastante limitada, por ser restrito a um número pequeno de palavras, apresenta uma base teórica bem fundamentada e poderia, portanto, ser estendido para cobrir um número maior de palavras. Uma possível forma de extensão seria o enriquecimento, com as suas informações semânticas, do Lex-Port ou da base lexical DIADORIM. Essa extensão exigiria um tempo relativamente grande e, possivelmente, algumas adaptações do modelo semântico definido para o Léxico Enriquecido, mas, certamente, esse modelo serviria

como ponto de partida para o desenvolvimento de um recurso que poderia ser utilizado em outros sistemas de PLN.

Referências Bibliográficas

Abrahão, P.R.C. (1997). *Modelagem e Implementação de um Léxico Semântico para o Português*. Dissertação de Mestrado, PUC/RS, Porto Alegre.

Borba, F.S. (1990). *Dicionário gramatical de verbos do português contemporâneo do Brasil*. Fundação Editora Unesp, São Paulo.

Briscoe, T. (1991). *Lexical Issues in Natural Language Processing*. University of Cambridge Computer Laboratory, Cambridge.

Chafe, W. (1970). *Meaning and the Structure of the Language.* University of Chicago Press, Chicago.

Copestake, A. (1993). *The Compleat LKB*. University of Cambridge Computer Laboratory, Cambridge.

Dias, M.C.P. (1994). *O Léxico em Sistemas de Análise e Geração Automática de Textos em Língua Portuguesa*. Tese de Doutorado, PUC/RJ, Rio de Janeiro.

Dias-da-Silva, B.C.; Sossolote, C.; Zavaglia, C.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N.; Aluísio, S.M. (1998). The design of the Brazilian Portuguese machine tractable dictionary for an interlíngua sentence generator. In *III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada.* PUCRS, Porto Alegre.

Dik, S.C. (1997). *The theory of functional grammar part 1*: *the structure of the clause*. Mouton de Gruyter, Berlim.

Dorr, B.J. (1992). The Use of Lexical Semantics in Interlingual Machine Translation. *Journal of Machine Translation*, 7:3, pp. 135-193.

Dorr, B.J. (1993). *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge.

Evans, R.; Gazdar, G. (1996). *DATR – A Language for Lexical Knowledge Representation*. School of Cognitive and Computing Sciences, The University of Sussex, Brighton.

García, L.S. (1995) *Linx: Um Ambiente Integrado de Interface para Sistemas de Informação Baseados em Conhecimento*. Tese de Doutorado, Departamento de Informática, PUC-Rio, Rio de Janeiro.

Greghi, J.G; Martins, R.T.; Nunes, M.G.V. (2001). *O Processo de Desenvolvimento da BDL-NILC*. Série de Relatórios do NILC, NILC-TR-01-7. São Carlos, Outubro, 57p.

Greghi, J.G. (2002). *Projeto e desenvolvimento de uma base de dados lexicais do português*. Dissertação de Mestrado, ICMC/USP, São Carlos.

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Lima, V.L.S.; Silva, J.L.T; Oliveira, F.M. (1996). O Enfoque Paralelo e Distribuído do Projeto Nálamas. In *II Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada.* CEFET-PR, Curitiba.

Lobato, L.M.P. (1986). *Sintaxe Gerativa do Português: da Teoria Padrão à Teoria da Regência e Ligação*. Vigília, Belo Horizonte.

Miller, G. (1990). Wordnet: An On-line Lexical Database. In *International Journal of Lexicography*, 3, pp. 235-312.

Moraes, H.R. (2002). *Categorias Aspectuais: Vendler, Dik, Chafe e Pustejovsky*. 50º Seminário do Grupo de Estudos Lingüísticos do Estado de São Paulo. Faculdade de Filosofia, Letras e Ciências da USP.

Nunes, M.G.V.; Vieira, F.M.C.; Zavaglia, C.; Sossolote, C.R.C.; Hernandez, J. (1996). *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Série de Relatórios Técnicos do ICMC, (Tech. Rep. 42), Universidade de São Paulo, São Carlos.

Nunes, M.G.V.; Aluísio, S.M.; Bonfante, A.G.; Dias-da-Silva, B.C.; Hasegawa, R.; Jesus, M.A.C.; Martins, R.T.; Montilha, G.; Oliveira Jr., O.N.; Rino, L.H.M.; Sossolote, C.R.; Zavaglia, C. (1997). Developing a UNL decodifier for Brazilian Portuguese. In *Proceedings of the II Workshop of UNL/Brazil Project.* NCE/UFRJ, Rio de Janeiro.

Oliveira Jr., O. N.; Martins, R. T.; Rino, L. H. M.; Nunes, M. G. V. (2001). *O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil.* Série de Relatórios do NILC. NILC-TR-01-3. São Carlos, Julho, 14p.

Onyshkevych, B.; Nirenburg, S. (1994). *The Lexicon in the Scheme of KBMT Things*. Technical Report MCCS-94-277, Computing Research Laboratory, New Mexico State University , Las Cruces.

Pereira, F.C.N.; Warren, D.H.D. (1980). Definite Clause Grammars for Language Analysis – A Survey of the Formalism and a Comparison with Augmented Transition Networks. In *Artificial Intelligence*, 13, pp. 231-278.

Pustejovsky, J. (1991). The syntax of the event structure. In *Cognition*, Vol. 41, pp. 47-81.

Pustejovsky, J. (1995). *The generative lexicon*. The MIT Press, London.

Rino, L. H. M.; Martins, R. T.; Marchi, A. R.; Kuhn, D. C. S.; Pinheiro, G. M.; Pardo, T. A. S.; Di Felippo, A.; Nunes, M. G. V. (2001). *Projeto TraSem: A investigação teórica sobre o problema da ambigüidade categorial.* Série de Relatórios do NILC, NILC-TR-01-1. São Carlos, Abril, 42p.

Schank, R. (1975). *Conceptual Information Processing*. North-Holland Publishing Company, New York.

Specia, L. (2002). *Um gerador conceitual para o português visando à produção de códigos UNL*. Monografia de Qualificação ao Mestrado. DC-UFSCar, São Carlos.

Specia, L.; Rino, L.H.M. (2002). *Representação Semântica: Alguns Modelos Ilustrativos*. Serie de Relatórios Técnicos do NILC, NILC-TR-02-12. São Carlos, Julho, 29p.

Uchida, H.; Zhu, M.; Senta, T.D. (1999). *The UNL, a Gift for a Millennium*. UNU/IAS/UNL Center, Tokyo.

UNL (2001). *The Universal Networking Language (UNL) Specifications*. UNU/IAS/UNL Center, Tokyo.

Vendler, Z. (1967). *Linguistics in philosophy.* Cornell University Press, New York.

Viegas, E.; Raskin, V. (1998). *Computational Semantic Lexicon Acquisition: Methodology and Guidelines*. Computing Research Laboratory, New Mexico State University , Las Cruces.