Indexação assistida por computador de documentos textuais em língua portuguesa

[1]*João Sequeira*

[2]*Maria Teresa Balesteros*

É hoje uma evidência comumente aceite que foi graças ao advento da informática e sobretudo ao desenvolvimento das comunicações, que se tornou possível o modelo social e económico em que assenta a chamada Sociedade da Informação.

A necessidade de acesso rápido à informação estratégica é neste modelo um factor de vantagem comparativa. Essa constatação levou a que surgissem, a partir da década de sessenta, em face do já então previsível crescimento acelerado da quantidade de documentos produzidos (explosão documental) e sobretudo pelo desenvolvimento das formas de transmissão e acesso à informação a partir das redes informáticas, organizações (CDIs) e pessoas (Documentalistas) cuja função passou de simples *guardiões* dos documentos a agentes do processo de tratamento e disponibilização da informação.

A descrição documental tradicional assente basicamente na catalogação e classificação dos documentos revelavam-se então já insuficientes; Tornavam-se necessários sistemas de acesso rápido e fiável ao seu conteúdo. As tarefas de resumo e indexação ganharam no seio da comunidade arquivistica e bibliotecária importância crescente e ocupam ainda hoje nas bibliotecas e arquivos, um conjunto muito vasto de técnicos e consomem enormes recursos financeiros.

A complexidade e morosidade dos processos manuais de indexação e resumo, assentando fundamentalmente em mão-de-obra qualificada e portanto melhor remunerada, transforma os produtos documentais em bens de grande valor acrescentado e consequentemente em produtos caros não acessíveis à generalidade dos consumidores de informação.

Por outro lado a explosão documental facilitada pela capacidade de distribuição e acesso decorrente do acesso às redes de comunicação, exige aos Centros de Documentação e Informação capacidades de tratamento documental, cada vez menos compatíveis com o seu tratamento pelos processos tradicionais.

A verdade é que não é mais possivel travar o "caos" informativo existente na maioria das organizações e sobretudo nas actuais redes de informação (veja-se o enorme "ruído" presente na maioria dos actuais motores de pesquisa existentes na INTERNET), sem recorrer por um lado às técnicas de tratamento documental normalizadoras bem como à utilização por esses técnicos de sistemas de indexação da informação automáticos e sobretudo de sistemas "*inteligentes*", capazes eles próprios com reduzida intervenção humana de "*compreender*" o conteúdo dos documentos e proceder à sua indexação.

---

[1] Licenciado em informática, assistente universitário na Universidade Autónoma de Lisboa e Chefe de Departamento de Informação Documental da Direcção de Arquivos e Documentação da Radiotelevisão Portuguesa, SA (RTP).
Email: jsequeira@mail.rtp.pt

[2] Licenciada em Informática e Chefe de Serviço de Documentação Escrita da Direcção de Arquivos e Documentação da Radiotelevisão Portuguesa, SA (RTP).
Email: balesteros@mail.rtp.pt

Quais as vantagens reais para os CDI da automatização da análise de conteúdo dos documentos?

- A **Rapidez** com que os CDI passariam a disponibilizar a informação é talvez a vantagem mais óbvia.

- **Economia** - Sabemos por outro lado que a tarefa de indexação exige profissionais qualificados e detentores de múltiplas competências, tanto no âmbito das técnicas documentais como no que diz respeito à compreensão das matérias tratadas nos documentos. Um perito com tal perfil nem sempre é fácil de encontrar e demora tempo a formar. Ao reduzir o número de peritos necessários e o tempo de formação de cada um, estas serão, certamente, vantagens económicas não desprezíveis.

- **Homogeneidade e Coerência** - Ao permitir reduzir a multiplicidade de critérios (inevitável na indexação humana) conseguiremos sistemas documentais mais homogéneos e coerentes.

- **Especialização** - Por último, é necessário ter em conta que determinados ramos do conhecimento evoluem tão rapidamente que é cada vez mais difícil aos documentalistas manter-se permanentemente actualizados. Uma vantagem adicional seria assim a partilha por todos os técnicos dos vários contributos individuais.

Tipos de Sistemas de Indexação Automática

A constatação da necessidade de sistemas de indexação automática não é, como se disse, nova. Os primeiros sistemas surgiram na década de 70 nos EUA e na Europa, mas a sua implementação era grandemente constrangida pela capacidade dos equipamentos e pela investigação que dava então os primeiros passos. Alguns desses conceitos estão já hoje presentes nos sistemas de pesquisa em bases de dados. A grande revolução é no entanto só hoje possível, com o desenvolvimento da chamada ***Inteligência Artificial***.

No essencial podem caracterizar-se três tipos de sistemas de indexação assistida , a saber:

I. Sistemas não selectivos

A. A indexação é feita palavra a palavra, sendo tidas em conta todas as palavras não-vazias (i.e., com conteúdo semântico) presentes no documento.

B. São sistemas amplamente divulgados, relativamente simples de implementar, utilizados para possibilitar a pesquisa a bases de dados em texto integral.

C. Estes sistemas colocam alguns problemas ao nível da pesquisa documental que se relacionam com a não selectividade da indexação, problemas de sinonímia e polissemia, não-normalização das formas gramaticais, e decomposição de termos compostos representativos de conceitos, em unitermos pouco ou nada significativos.

D. Para reduzir o impacto de alguns destes problemas, as linguagens de interrogação incluem, para além da possibilidade de utilizar operadores booleanos, outras funcionalidades como a truncatura (ao permitir a pesquisa de termos a partir de um radical, minimiza-se o problema da não-normalização das formas gramaticais) e a pesquisa de referências próximas e adjacentes permitindo (em muitos casos recuperar termos compostos).

E. Para além disso, alguns incluem ainda uma normalização da linguagem natural, permitindo o tratamento de: variantes **(oiro = ouro),** erros ortográficos **(açoreano = açoriano),** erros de digitação **(dcumento = documento)** e formas flexionadas **(feminino = masculino, plural = singular)**.

II. Sistemas selectivos

A. Apenas são tidos em conta certos termos: aqueles que forem considerados pelo algoritmo do sistema como os mais representativos do conteúdo do documento. Podem utilizar os termos presentes no próprio documento (linguagem natural) ou uma linguagem controlada.

III. Indexação assistida por computador

A. A indexação é feita em duas fases:

1. uma fase de pré-indexação automática, em que o sistema analisa um texto e lhe associa um conjunto de descritores, normalmente extraídos de uma lista de autoridade ou de um tesauro;

2. e uma fase de diálogo entre o sistema e o documentalista, em que a lista de descritores proposta na fase precedente é validada.

B. De acordo com os trabalhos já desenvolvidos podemos concluir que estes sistemas se dividem em dois grandes grupos:

1. sistemas de orientação probabilística e

2. sistemas de orientação linguística

Indexação assistida de orientação probabilística - SINTEX (Système d'Indexation de Textes)

Este sistema prevê uma fase de pré-aprendizagem no decurso da qual é analisado um conjunto de cerca de 4000 documentos indexados e classificados manualmente. É durante esta fase que são criados automaticamente a maioria dos ficheiros em que o sistema se baseia. Estes ficheiros são essencialmente de dois tipos: léxicos (de palavras, de descritores e de domínios) e ficheiros de correspondência ponderada (entre descritores e palavras, entre descritores e descritores, entre descritores e domínios, etc.).

A pré-indexação automática compreende duas fases: reconhecimento das palavras do texto e classificação e indexação do documento por consulta aos ficheiros de correspondência ponderada.

Indexação assistida de orientação linguística - ALEXDOC

O sistema baseia-se em: dados (tesauro e dicionário base da língua) e regras (de derivação morfológica, de reestruturação de termos compostos, regras sintácticas - para resolver a ambiguidade criada pelas palavras homógrafas, regras de transformação semântica e regras de reconhecimento de frases.

O processo de indexação compreende essencialmente as seguintes fases:

- Identificação de frases
- Identificação de cada uma das palavras, por consulta do tesauro e do dicionário base da língua e por aplicação das regras de derivação morfológica
- Resolução de ambiguidades
- Identificação de termos compostos
- Aplicação das regras de transformação semântica
- Eliminação de não descritores
- Eliminação de descritores simples que fazem parte de descritores compostos
- Adição de termos genéricos, ao primeiro nível, de cada descritor retido
- Validação da indexação proposta

Estes sistemas ainda não se encontram muito divulgados e muitos dos que existem estão ainda em fase experimental.

Este estado de coisas tem a ver com o facto de que para indexar um texto é necessário compreendê-lo.

Entramos aqui num domínio particular da Inteligência Artificial:

*o **Processamento da Linguagem Natural ou Engenharia da Linguagem.***

Esta é uma área de investigação que constitui o tronco comum a múltiplas aplicações práticas. Um dos primeiros campos de aplicação da investigação levada a cabo nesta área foi a tradução automática. A partir dos anos 50/60 foi desenvolvido um grande trabalho nesta matéria e, neste momento, estes sistemas já têm grande utilização.

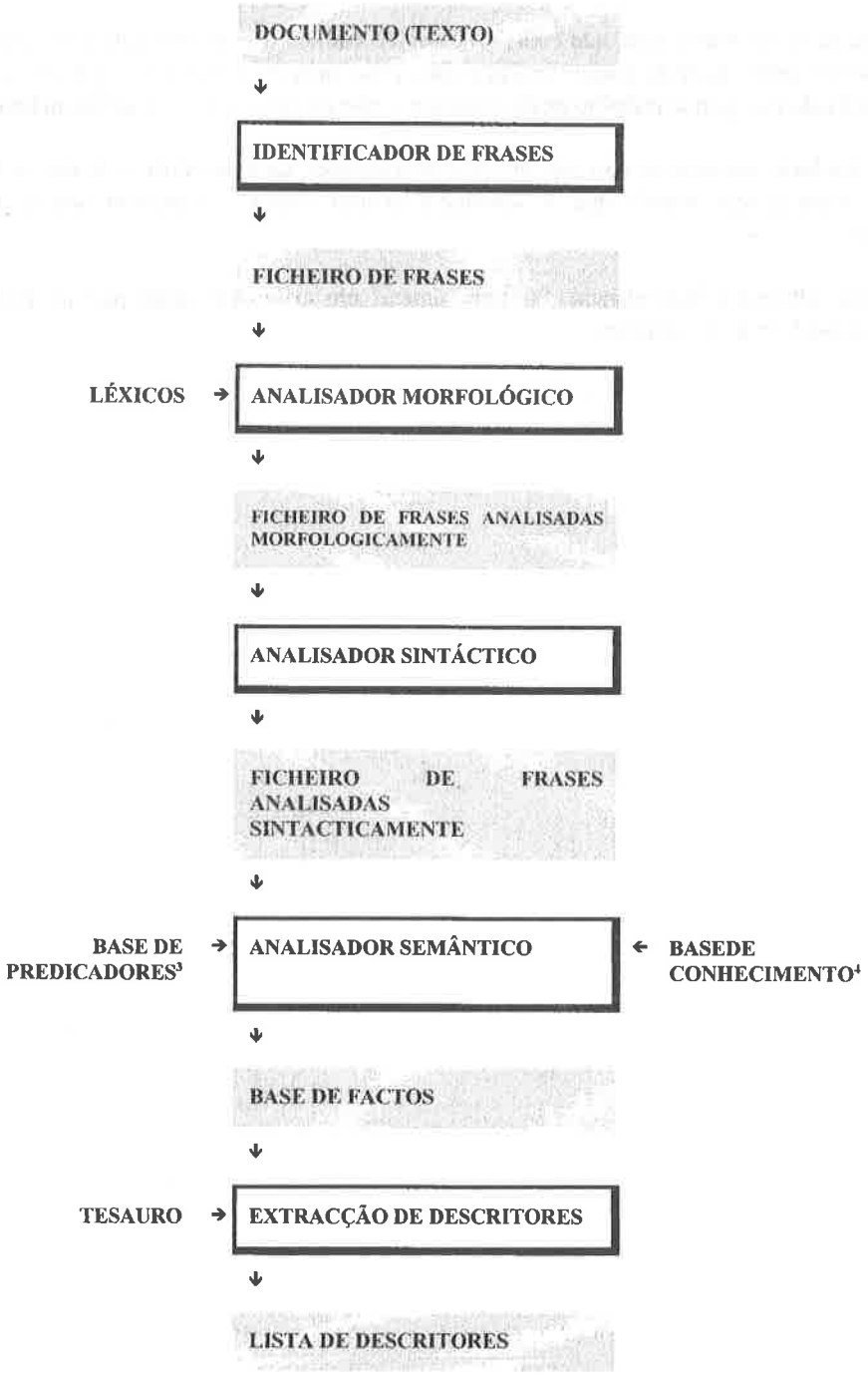
Embora a tradução automática não seja um processo simples, qualquer erro de tradução é detectável com relativa facilidade dada a redundância que a linguagem natural encerra. O mesmo não se passa com a indexação automática.

A indexação elimina a redundância, faz uma síntese: um artigo de várias páginas pode ficar reduzido a uma pequena quantidade de descritores.

Posta a questão teórica importa agora apresentar o trabalho de investigação em que nos encontramos envolvidos, tendente a incorporar alguns destes conceitos num protótipo de um **Indexador Automático para a Língua Portuguesa (IndeLiP)**, assente em critérios linguísticos e explicitar a sua arquitectura.

Estrutura do IndeLiP

(Detalhe do módulo de indexação)



[3] A sua função é permitir a resolução de algumas ambiguidades

[4] A sua função é produzir inferências

Identificador de Frases

O seu objectivo consiste em identificar as unidades mínimas de comunicação que irão constituir a base de todo o processamento posterior. A sua implementação não constitui problema computacional. Trata-se apenas de reconhecer os conjuntos de caracteres que constituem as palavras, agrupá-las em frases através do reconhecimento dos sinais de pontuação que normalmente indicam o fim de uma frase (ponto, ponto de interrogação, ponto de exclamação, etc.) e tratar as excepções relativamente à utilização desses sinais.

Em termos computacionais um texto é uma lista de listas com a seguinte representação:

((Frase 1) (Frase 2) .... (Frase n))

Analisador Morfológico

O Analisador Morfológico é na realidada o módulo inicial de processamento do texto. A sua função consiste em classificar morfológicamente as unidades lexicais básicas (palavras) .

O seu objectivo último é o de facilitar na fase de aprendizagem a construção do Léxico Básico da Língua, a partir da leitura e processamento de uma apreciável quantidade de textos num dado domínio do conhecimento.

Na prática a sua principal característica consiste em, a partir do reconhecimento de uma palavra, propor ao analisador sintáctico uma ou várias Classificações Morfológicas possíveis para a palavra. A Análise morfológica de uma palavra é assim uma operação descontextualizada da sua posição numa determinada frase. Essa operação de desambiguação é feita posteriormente, na fase ou fases seguintes de processamento (Análise Sintáctica ou até na Análise Semântica) .

O Analisador Morfológico é fundamentalmente um módulo de Consulta de um Léxico completo da Língua portuguesa, expresso numa estrutura computacional que se explicita seguidamente.

Construção de um Léxico completo da Língua Portuguesa

A filosofia subjacente a todo o projecto é a de que se trata de construir um protótipo inteligente e assistido. Existe pois uma fase de aprendizagem do sistema.

A construção do léxico básico da língua é assim um dos objectivos dessa fase.

Internamente o Analisador Morfológico é composto por dois sub-módulos:

- Módulo de Consulta e apresentação a partir do Léxico;

- Módulo de validação de palavras novas, não existentes ainda no léxico, mas que possam ser Derivadas ou Compostas a partir de morfemas lexicais básicos da língua, já existentes no léxico.

No final do processamento o léxico é actualizado com a nova palavra, que passa assim a constituir conhecimento já adquirido pelo sistema para uso posterior.

Estrutura do Léxico

O Léxico não é mais do que uma Base de Dados indexada por palavra, implementada fisicamente como um ficheiro binário de listas de estruturas próprias da maioria das linguagem de programação e de acordo com o seguinte formato:

(PALAVRA (CLASSE_1 (RADICAL+AFIXO) ((FLEXÃO_1 VALOR_1) ... (FLEXÃO_N VALOR_N)) (TRAÇO_SEMÂNTICO1) (TRAÇO_SEMÂNTICO_2))

(CLASSE_N (RADICAL+AFIXO) ((FLEXÃO_1 VALOR_1) ... (FLEXÃO_N VALOR_N)) (TRAÇO_SEMÂNTICO_1) (TRAÇO_SEMÂNTICO_2)))

Procedimentos de Composição e Derivação

O processo de formação das palavras é basicamente *" um conjunto de processos morfo-sintácticos que permitem a criação de unidades novas da língua com base em morfemas lexicais. Utilizam-se assim, para formar palavras novas, os afixos de derivação ou os procedimentos de composição*[5]*"*.

Processo de Derivação

O processo de derivação/composição faz-se através da aplicação de um conjunto de regras contidas numa base de conhecimento - BASE DE AFIXOS.

A Base de Afixos é no essencial uma base de dados de Afixos correntes da língua portuguesa (Prefixos e Sufixos) a que foram associados procedimentos de composição e regras da sua aplicação.

No essencial os Afixos são agrupados segundo a Categoria Morfológica dos Radicais a que são normalmente associados, gerando por esse processo sempre outras palavras de uma determinada Classe Morfológica.

Na implementação de um Analisador Morfológico, a grande discussão centra-se nas seguintes questões:

- Nunca é possível ter um léxico completo da língua. E será que se justifica ir tendendo para esse objectivo? O que é que custa, em termos de recursos e de tempo, manipular um léxico tendencialmente completo da língua?

- Se tivermos apenas um léxico básico da língua, teremos que recorrer amiúde ao processamento de regras de derivação e composição de palavras e de regras de reconhecimento de formas verbais. Porque não ir aproveitando esses resultados para evitar ter que repetir no futuro os mesmos processamentos?

- O léxico deve incorporar para cada entrada informação morfológica completa?

Um dado essencial nesta discussão será sempre o tipo de aplicação em que o analisador morfológico vai ser utilizado.

Analisador Sintáctico

O objectivo é, a partir do *output* do **analisador morfológico**[6], identificar, para cada frase, as várias orações e os seus termos constituintes, através da aplicação de um conjunto de regras gramaticais.

Mais uma vez o âmbito em que vai ser utilizado o analisador é uma questão essencial a ter em conta no momento de definir a sua arquitectura.

- Precisamos de uma análise sintáctica completa, ou é suficiente analisar os sintagmas nominais?
- Precisamos de validar os aspectos de concordância entre os vários elementos de uma oração, ou não?

No nosso caso, provavelmente seria suficiente o tratamento dos sintagmas nominais. Tentámos no entanto implementar um analisador sintáctico completo que poderia eventualmente ser utilizado em outras aplicações.

A primeira tentativa consistiu em seleccionar um conjunto suficientemente amplo de regras gramaticais e representá-las através de uma gramática formal (do tipo 2, independente do contexto, na classificação de Chomsky).

[5] Ver DUBOIS, Jean et alli. *Dictionnaire de Linguístique*: Paris, Larrousse, 1977.
[6] Ver Anexo 1

Primeiro problema: Se o conjunto de regras seleccionado não for suficientemente amplo, corremos o risco de fazer uma análise incorrecta ou de não conseguir reconhecer frases perfeitamente correctas do ponto de vista sintáctico. Se o conjunto de regras for mais completo, e dado que uma palavra pode pertencer a várias classes morfológicas, deparamo-nos com o problema da explosão combinatória: se os recursos da máquina permitirem, serão produzidas múltiplas análises para a mesma frase, tornando assim mais difícil escolher a apropriada.

Processamento Preliminar

Dado que o **analisador morfológico** trata palavras, não tendo em consideração o contexto concreto, e que o **analisador sintáctico** deve, para cada palavra, escolher uma, de entre uma série de categorias morfológicas possíveis, cada frase é, numa fase preliminar do processamento, desdobrada segundo as várias combinações possíveis das categorias morfológicas a que cada palavra pode pertencer. Cada combinação constitui uma possibilidade[7] de classificação morfológica das palavras da frase que será validada (ou não) pela análise sintáctica.

Implementação de uma GRAMÁTICA DA LÍNGUA PORTUGUESA

Trata-se aqui de encontrar as estruturas de dados e as técnicas de computação adequadas à representação das regras que constituem a GRAMÁTICA DA LÍNGUA PORTUGUESA e de encontrar uma forma de controlar a sua aplicação.

Numa primeira fase seleccionámos um conjunto suficientemente amplo de regras gramaticais que permitisse reconhecer o português escrito corrente e representámo-las como foi dito através de uma gramática formal (do tipo 2 na classificação de Chomsky),

$$G=<Vt, Vn, S, R>$$

em que:

Vt (vocabulário terminal) é constituído pelos símbolos representativos das categorias morfológicas e por λ, o símbolo vazio[8];

Vn (vocabulário não terminal) é constituído pelos símbolos representativos das categorias sintácticas e sintagmáticas e por outros símbolos auxiliares[9];

S (símbolo inicial) é o símbolo F representativo de uma frase;

R é o conjunto de regras de reescrita, podendo cada regra incluir condições de aplicabilidade[10].

Estas gramáticas têm as seguintes características:

- A parte esquerda de uma regra é constituída por **um e um só** elemento do vocabulário não terminal;
- A parte direita de uma regra obedece a uma única restrição - deve ser diferente da parte esquerda para evitar ciclos infinitos.

Esta última restrição levou-nos a adicionar ao vocabulário não terminal alguns símbolos auxiliares para o tratamento de estruturas de coordenação.

---

[7] Ver Anexo 2
[8] Ver Anexo 3
[9] Ver Anexo 4
[10] Ver Anexo 5

Este tipo de gramática é bastante utilizado em aplicações informáticas, tendo no entanto algumas limitações para representar os fenómenos linguísticos.

Problemas como o reconhecimento de constituintes descontínuos (não só ... mas também), a verificação das concordâncias e a resolução de anáforas e referências cruzadas não podem ser directamente tratados por este tipo de gramáticas.[11]

A tentativa de implementação deste tipo de gramática revelou-se altamente consumidora de recursos de processamento, nomeadamente de memória. Optámos então por um sistema mais pragmático de reconhecimento de padrões. O sistema faz a análise sintáctica comparando a estrutura morfológica da frase com padrões existentes *a priori*; se não encontra nenhum padrão aplicável, permite gerar um novo padrão de uma forma interactiva.

Ao nível da análise sintáctica abriram-se-nos desde logo um sem número de caminhos a explorar, nomeadamente:

- Estudar outras formas de implementação das regras gramaticais, utilizando outras ferramentas, nomeadamente ao nível do sistema operativo
- Estudar a possibilidade de integrar num mesmo sistema diversos métodos, inclusive métodos estatístico-probabilísticos.

Analisador Semântico

Trata-se aqui de validar e interpretar o sentido de cada uma das frases que formam um texto, com o objectivo de produzir uma base de factos directamente utilizável no processo de indexação.

Vejamos alguns exemplos.

Ex1.:

O que é que uma frase como "*O peixe frito comeu o gato*" pode significar?

Do ponto de vista sintáctico a frase está correcta. Do ponto de vista semântico só podemos concluir uma de duas coisas: ou houve um erro e a frase correcta seria "*O gato comeu o peixe frito*" ou encontramo-nos no universo da ficção.

Como somos levados a tirar estas conclusões?

O sujeito do predicador comer deve ser uma entidade com vida, ou seja com o traço semântico [+ANIMADO].

Ex2.:

Na frase "*A Câmara de Fafe encerrou a lixeira de Passadouro*" é necessário perceber que

A Câmara é A Câmara Municipal do concelho de Fafe

e não A câmara de vídeo do Sr. Fafe.

Como é possível escolher a interpretação correcta?

---

[11] HAGÈGE, Caroline; DUARTE, Inês. - Construção de gramáticas formais para o processamento da linguagem natural. «Engenharia da Linguagem», Lisboa, Edições Colibri, 1995, p. 71-93.

Se soubermos que existe em Portugal uma localidade - FAFE que é sede de concelho e que em cada sede de concelho o órgão executivo do poder local é a Câmara Municipal poderemos, com uma grande probabilidade de não errar, escolher a primeira hipótese.

Ex3.:

Se estivermos a analisar um texto sobre a BIBLIOTECA DA FACULDADE DE CIÊNCIAS e se o termo eleito como descritor for BIBLIOTECA UNIVERSITÁRIA, é necessário que o conhecimento sobre a relação entre FACULDADE e UNIVERSIDADE permita concluir que a Biblioteca da Faculdade de Ciências é uma Biblioteca Universitária.

Para resolver estes problemas o Analisador Semântico explora e processa um conjunto de informação que se encontra armazenada

- em cada entrada do léxico, como uma lista de traços semânticos
- na base de predicadores
- na base de conhecimento

Base de Predicadores

A base de predicadores inclui, para cada predicador, a respectiva estrutura argumental, ou seja: que argumentos podem ocorrer, que argumentos devem obrigatoriamente estar presentes, qual a função semântica de cada um dos argumentos no esquema predicativo e que restrições (ao nível dos traços semânticos) existem ao desempenho dessas funções.[12]

Veja-se a título de exemplo um esquema predicativo possível para o verbo comprar. A acção de comprar, na sua acepção mais comum, exprime a transferência de propriedade de um objecto de uma entidade para outra. De outra forma, podemos dizer que o verbo comprar tem basicamente três argumentos: X, Y, Z
X - tem a função semântica de Agente e corresponde ao componente da frase com função sintáctica de Sujeito. Para que o Sujeito possa desempenhar esta função é necessário que o núcleo do Sintagma Nominal que o constitui tenha o traço semântico [+HUMANO].
Y - tem a função semântica de Objecto e corresponde ao componente da frase com função sintáctica de Objecto Directo.
(Z) - tem a função semântica de Origem e corresponde ao componente da frase com função sintáctica de Objecto Indirecto.
O argumento Z entre parêntesis indica que pode ser omitido.

A função da base de predicadores é sobretudo a de fornecer informação que possibilite a resolução das ambiguidades que ainda subsistam após a análise sintáctica

Base de Conhecimento

A base de conhecimento, que deve fornecer conhecimento de senso comum e conhecimento específico sobre o tema em análise, foi implementada sob a forma de uma rede semântica, cuja estrutura se exemplifica de seguida.

Temos a considerar fundamentalmente dois tipos de objectos: Nós e Arcos.
Os NÓS são objectos que representam um conceito sob a forma de uma ou mais palavras
Os ARCOS são objectos que relacionam nós; têm uma direcção, que identifica o nó de origem e o nó de destino, e uma etiqueta que exprime o tipo de relação existente entre esses nós.

---

[12] Veja-se a este propósito, BUSSE, Winfried.-Dicionário Sintáctico de Verbos Portugueses e MATEUS, Maria Helena Mira; e outros.-Gramática de Língua Portuguesa.

A função da base de conhecimento é produzir todas as inferências possíveis e adicioná-las à base de factos.

Extractor de Descritores

A função do Extractor de Descritores é, a partir do Tesauro e da Base de Factos, seleccionar os descritores aplicáveis.
Cada descritor é internamente implementado como um conjunto de Regras que levam à sua escolha. A cada regra está associada uma probabilidade que indica a medida de aplicabilidade do descritor.

Uma regra é pois uma estrutura do tipo
SE <condição>
ENTÃO <descritor> com probabilidade <x>

ANEXO 1

O *output* do analisador morfológico é o *input* do analisador sintáctico.

Trata-se de um ficheiro de texto constituído por frases. Uma frase é constituída por várias palavras. Cada palavra vem seguida de uma ou várias categorias morfológicas. Apresenta-se a seguir uma frase a título de exemplo.

```
((A (PREP () () ())
   (ART_DEF () ((GEN F) (N SING)) ())
   (PRO_PES_OB_AT () ((PES 3) (N SING) (GEN F)) ())
   (PRO_DEM_VAR () ((N SING) (GEN F)) ()))
(Câmara (S_COM () ((N SING) (GEN F) (GR NOR)) (ORG)))
(de (PREP () () ()))
(Fafe (S_PRO () () (LOC)))
(encerrou (VB_TD (ENCERRAR) ((N SING) (PES 3) (MD INDIC) (TP PRET_PERF)) ())
        (VB_I (ENCERRAR) ((N SING) (PES 3) (MD INDIC) (TP PRET_PERF)) ()))
(a (PREP () () ())
   (ART_DEF () ((GEN F) (N SING)) ())
   (PRO_PES_OB_AT () ((PES 3) (N SING) (GEN F)) ())
   (PRO_DEM_VAR () ((N SING) (GEN F)) ()))
(lixeira (S_COM (LIXO + EIRA) ((N SING) (GEN F) (GR NOR)) ()))
(de (PREP () () ()))
(Passadouro (S_PRO () () (LOC))
        (S_COM (PASSAR+DOURO) ((N SING) (GEN M) (GR NOR)) (LOC))))
```

ANEXO 2

Desdobramento de uma frase de *input* num conjunto de possibilidades

A frase apresentada como exemplo no ANEXO 1 é, durante o processamento preliminar, desdobrada em 64 possibilidades segundo as várias combinações das categorias morfológicas a que cada palavra pode pertencer.

A estrutura de cada uma destas possibilidades é semelhante à estrutura de uma frase de *input* com uma diferença: a cada palavra só corresponde uma categoria morfológica.

Durante o processamento muitas destas possibilidades serão rejeitadas por serem agramaticais, sendo retidas apenas as que correspondem a padrões sintácticos possíveis.

Exemplo de uma possibilidade:

1.

((*A* (PREP () () ()))

(***Câmara*** (S_COM () ((N SING) (GEN F) (GR NOR)) (ORG)))

(***de*** (PREP () () ()))

(***Fafe*** (S_PRO () () (LOC)))

(***encerrou*** (VB_TD (ENCERRAR) ((N SING) (PES 3) (MD INDIC) (TP PRET_PERF)) ()))

(***a*** (PREP () () ()))

(***lixeira*** (S_COM (LIXO + EIRA) ((N SING) (GEN F) (GR NOR)) ()))

(***de*** (PREP () () ()))

(***Passadouro*** (S_PRO () () (LOC))))

ANEXO 3

Vt - VOCABULÁRIO TERMINAL

s_pro	-	substantivo próprio
s_com	-	substantivo comum
s_abst	-	substantivo abstracto
art_def	-	artigo definido
art_ind	-	artigo indefinido
adj	-	adjectivo
num_car	-	numeral cardinal
num_ord	-	numeral ordinal
vb_td	-	verbo transitivo directo
vb_ti	-	verbo transitivo indirecto
vb_b	-	verbo bitransitivo
vb_i	-	verbo intransitivo
vb_lig	-	verbo de ligação
vb_ip	-	verbo impessoal
vb_aux	-	verbo auxiliar
prep	-	preposição
conj_coord	-	conjunção coordenativa
conj_sub	-	conjunção subordinativa
intj	-	interjeição
cont	-	contracção
virg	-	vírgula
pt	-	ponto
pt_virg	-	ponto e vírgula
pt_pt	-	dois pontos
pt_int	-	ponto de interrogação
pt_excl	-	ponto de exclamação
ret	-	reticências
col	-	aspas
par	-	parêntesis
trv	-	travessão
λ	-	vazio

ANEXO 4

Vn - VOCABULÁRIO NÃO-TERMINAL

F	-	FRASE
OR	-	ORAÇÃO
RF	-	RESTO DA FRASE
SU	-	SUJEITO
PRED	-	PREDICADO
SU_INV	-	SUJEITO INVERTIDO
EL_COORD	-	ELEMENTO DE COORDENAÇÃO
SN_COORD	-	SINTAGMAS NOMINAIS COORDENADOS
OR_SUBST	-	ORAÇÃO SUBORDINADA SUBSTANTIVA
SV	-	SINTAGMA VERBAL
SN	-	SINTAGMA NOMINAL
RSN	-	RESTO DE SINTAGMAS NOMINAIS COORDENADOS
EXP_VB	-	EXPRESSÃO VERBAL
COMPL_SV	-	COMPLEMENTOS DO SINTAGMA VERBAL
ESP_SN	-	ESPECIFICADORES DO SINTAGMA NOMINAL
PRE_COMPL_SN	-	PRÉ-COMPLEMENTOS DO SINTAGMA NOMINAL
NUC_SN	-	NÚCLEO DO SINTAGMA NOMINAL
POS_COMPL_SN	-	PÓS-COMPLEMENTOS DO SINTAGMA NOMINAL
VB	-	VERBO
LOC_VB	-	LOCUÇÃO VERBAL
OD	-	OBJECTO DIRECTO
OI	-	OBJECTO INDIRECTO
OBL	-	OBLÍQUO
PRED_OD	-	PREDICATIVO DO OBJECTO DIRECTO
PRED_SU	-	PREDICATIVO DO SUJEITO
X	-	Argumentos opcionais do predicador
DET	-	DETERMINANTE
QUANT	-	QUANTIFICADOR
EXP_QUAL	-	EXPRESSÃO QUALITATIVA
SADJ_COORD	-	SINTAGMAS ADJECTIVAIS COORDENADOS
S_COM_COORD	-	SUBSTANTIVOS COMUNS COORDENADOS
S_PRO_COORD	-	SUBSTANTIVOS PRÓPRIOS COORDENADOS
S_ABST_COORD	-	SUBSTANTIVOS ABSTRACTOS COORDENADOS
SPREP_COORD	-	SINTAGMAS PREPOSICIONAIS COORDENADOS
SADJ	-	SINTAGMA ADJECTIVAL
SPREP	-	SINTAGMA PREPOSICIONAL
RSADJ	-	RESTO DE SINTAGMAS ADJECTIVAIS COORDENADOS
RSCOM	-	RESTO DE SUBSTANTIVOS COMUNS COORDENADOS
RSPRO	-	RESTO DE SUBSTANTIVOS PRÓPRIOS COORDENADOS

ANEXO 5

R - ALGUNS EXEMPLOS DE REGRAS DE REESCRITA[40]

			CONDIÇÕES	OBSERVAÇÕES
F	→	OR RF		Uma frase é constituída por uma ou mais orações coordenadas.
OR	→	SU PRED		
OR	→	PRED		Caso de sujeito oculto, nulo ou indeterminado.
OR	→	PRED SU_INV	VB=*vb_i* ∨ VB=*vb_lig*	Caso de inversão do sujeito
RF	→	λ		O resto da frase pode ser vazio, ou
RF	→	EL_COORD OR RF		constituído por um elemento de coordenação seguido de uma ou mais orações.
SU	→	SN_COORD		
PRED	→	SV		
SU_INV	→	SN_COORD		
SU_INV	→	OR_SUBST	VB=*vb_lig*	
EL_COORD	→	*conj_coord*		
EL_COORD	→	*virg*		
SN_COORD	→	SN RSN		
OR_SUBST	→	*conj_sub* OR	*conj_sub*=que se	
SV	→	EXP_VB COMPL_SV		
LOC_VB	→	*vb_aux* VB	MD_VB=INF ∨ TP_VB=GER ∨ TP_VB=PART	
LOC_VB	→	*vb_aux prep* VB	MD_VB=INF ∨ TP_VB=GER ∨ TP_VB=PART	
OD	→	SN_COORD		
OD	→	OR_SUBST		

[40] Em maiúsculas representam-se os símbolos não-terminais; em minúsculas e itálico representam-se os símbolos terminais.

Bibliografia

BUSSE, Winfried (coord.). - Dicionário sintáctico de verbos portugueses. Coimbra: Almedina, 1994. 449p. ISBN 972-40-0803-7.

CAMARA JR., Joaquim Mattoso - Estrutura da língua portuguesa.Petrópolis: Editora Vozes, 1982.125p.

CUNHA, Celso; CINTRA, Luís F. Lindley. - Nova Gramática do Português Contemporâneo, 10ª ed. Lisboa: Edições Sá da Costa, 1994.

DUBOIS, Jean et alli. *Dictionnaire de Linguístique*: Paris, Larrousse, 1977.

GILCHRIST, Alan - The Thesaurus in Retrieval. London: Aslib, 1971.184 p.

HAGÈGE, Caroline; DUARTE, Inês. - Construção de gramáticas formais para o processamento da linguagem natural. «Engenharia da Linguagem», Lisboa, Edições Colibri, 1995, p. 71-93.

HUMPHREY, Susanne M. - A knowledge-Based Expert System for Computer-Assisted Indexing. «IEEE Expert», Outono 1989.

JUN-TAE KIM. - Semantic knowledge acquisition for information extraction from texts on parallel marker-passing computer. - University of Southern California, 1993.

KELLER, Eric (ed.) - Fundamentals of Speech Synthesis and Speech Recognition. Basic Concepts,State of the Art and Future Challenges. Chichester: John Willey and Sons.1994.379 p. (ISBN 0 471 94449 1).

KITANO, Hiroaki - Speech-to-speech Translation. A massively Parallel Memory-Based Approach. Boston: Kluwer Academic Publishers.1994.193p.(ISBN 0 7923 9425 9).

KNIGTH, G. NORMAN - The art of indexing. A guide to the indexing of books and periodicals. London: George Allen & Unwin,1983.218 p. (ISBN 0 04 029002 6).

LANCASTER, F.W. - Vocabulary control for information retrieval. Arlington: Information Resources Press.1986. (ISBN 0 87815 053 6).

MARTINS, J. A. Legatheaux; MONTEIRO, Luís. - Linguagens formais e autómatos. Universidade Nova de Lisboa: Faculdade de Ciência e Tecnologia: Departamento de Informática, 1981.

MATEUS, Maria Helena; e outros. - Engenharia da Linguagem. Lisboa: Edições Colibri; Faculdade de Letras da Universidade de Lisboa, 1995.

MATEUS, Maria Helena; e outros. - Gramática da Língua Portuguesa, 4ª ed. Lisboa: Caminho, 1994.

NINA, Nuno. - Visual Basic 3.00 for windows. Curso completo. Lisboa: FCA Editores de Informática, 1994. 288p. ISBN 972-722-024-X

ROLE, François. - De la lettre au sens: les recherches en texte integral. «Documentaliste: Sciences de l'information». Paris: 30:3 (1993) p. 136-147.

ROWLEY, Jennifer E. - Abstracting and Indexing. London: Clive Bingley,1982.155 p.

SIBERTIN-BLANC, Martine - Nouvelles Technologies et Communication de l'information. Des besoins des utilisateurs l'ingenierie documentaires. Paris: ADBS,1994.277p. (ISBN 2 901046 71

VAN SLYPE, Georges. - Les langages d'indexation: conception, construction et utilisation dans les systèmes documentaires. Paris: Les éditions d'organisation, 1987.