A Construção de um Analisador Morfossintáctico do Português: a Implementação de Verbos e Clíticos

CAROLINE HAGÈGE
ANTÓNIO MEIRELES BRÍGIDA TRINDADE
CARLA DIOGO FERNANDO LEITE
(ILTEC)

1. Introdução

O Analisador desenvolvido no ILTEC é uma ferramenta que, partindo de duas componentes, uma linguística e uma informática, tem como objectivo atribuir uma classificação morfossintáctica às palavras do português.

A componente linguística ocupou-se da definição do tipo de informação que devia estar registada no dicionário e da construção de uma bateria de regras que permitisse o reconhecimento das palavras flexionadas e de outros tipos de palavras morfologicamente complexas (como por exemplo, diminutivos, advérbios em *-mente*, etc.), cujos lemas não estão registados em qualquer ficheiro. Por sua vez, a componente informática consistiu na criação de uma infraestrutura que, partindo dos dados fornecidos pela componente linguística, constrói uma análise para o fornecimento de etiquetas a cada uma das palavras encontradas num determinado texto.

Optou-se sempre que possível pela construção de regras em detrimento do registo de lemas flexionados ou derivados, o que permitiu reduzir o número de entradas registadas sem diminuir o número de palavras reconhecidas. Por exemplo, o reconhecimento dos advérbios em *-mente* é feito a partir do lema do adjectivo, que está na base do advérbio, registado no dicionário e de um conjunto de regras (a título de exemplo, ao adjectivo *óbvio* vai ser aplicada uma regra de derivação que permitirá o reconhecimento do advérbio correspondente—*obviamente*—sem que este esteja dicionarizado).

2. Descrição do Analisador Morfossintáctico

A componente linguística do Analisador está dividida em ficheiros de lemas e ficheiros de regras. Neste momento, o dicionário contém 43 000 lemas (ou seja, formas não flexionadas de cada entrada lexical) e a respectiva informação morfossintáctica, exceptuando os verbos e, por razões idiossincráticas à sua implementação, os clíticos, as palavras compostas, as contracções, as locuções, as siglas, as abreviaturas, os acrónimos, os numerais e a pontuação, que se encontram descritos em ficheiros à parte.

O modo de implementação da componente informática permitiu aplicar sucessivamente várias regras a um determinado lema. Consequantemente, um advérbio é reconhecido quer este seja formado a partir do superlativo quer a partir do lema, visto que o programa «procurará» o maior «sufixo» possível e caso não reconheça a palavra aplica as regras até chegar ao lema:

bonito > bonit**íssimo** > bonitissima**mente**

Por sua vez, cada uma das regras tem também a sua especificidade, de forma a impedir o reconhecimento de formas agramaticais, como por exemplo **cãomente*.

Neste momento a partir de um lema como *lindo*, adjectivo, reconhecem-se trinta e três formas diferentes. No caso de um radical verbal como *am-* reconhecem-se setenta e sete formas, para além das formas cliticizadas.

Informaticamente, esta ferramenta foi escrita tendo em vista uma maior portabilidade e visando conseguir a maior independência possível do código-fonte em relação ao suporte máquina, possibilitando a alteração dos dados linguísticos sem recorrer a modificações na componente do programa. Desta forma, a plataforma primária de desenvolvimento é de base *UNIX* (*SunOS* 4.1.3 e *Linux*), o que não impede, visto que é usada a linguagem *C* e respeitado o padrão *standard*, que, com relativa facilidade, se possa compilar e executar o suporte lógico em plataformas não *UNIX*, como o *Windows NT* ou o *Windows* 95. Relativamente à velocidade do programa, este reconhece 1200 palavras/segundo num *Pentium* 166 - 64 *Mb* de *RAM*.

O programa recebe e trata todo o tipo de *input*, que pode ser introduzido directamente pelo utilizador na linha de comando ou pode provir de um ficheiro. Depois de separadas as palavras todos os elementos delimitados por hífenes, tabulações, sinais de pontuação ou espaços são colocados elemento a elemento numa estrutura, já que seguidamente o programa irá analisar parágrafo a parágrafo.

Passada esta fase, o Analisador está apto a «pronunciar-se» sobre cada uma das unidades significativas (uma palavra, um sinal de pontuação, um número...), consultando o dicionário. Caso não encontre a informação de imediato, o programa vai chegar a uma entrada no dicionário através da aplicação das regras, se o lema estiver registado.

3. Implementação da morfologia verbal

Num trabalho desta natureza surgem frequentemente problemáticas, sobretudo na selecção dos critérios que devem prevalecer no tratamento de algumas categorias, se os linguísticos, se os informáticos.

Imagine-se uma máquina. Qual será o primeiro problema com que se depara ao tentar reconhecer uma forma verbal como *amamo-los*? O entrave mais imediato é o facto de o programa não ter registado nem a forma *amamo* nem a forma *los*, problema que será ultrapassado através da construção de regras.

Assim, de forma a resolver esta questão, implementou-se, em primeiro lugar, o reconhecimento dos verbos, seguidamente o reconhecimento dos clíticos e só depois as formas verbais cliticizadas.

Os objectivos subjacentes à construção do motor de reconhecimento verbal foram, por um lado, tentar manter a maior simplicidade possível para o linguista que manipula e introduz os dados, e por outro «regularizar» (de um ponto de vista informático) o maior número de verbos possível.

Tendo em conta a estratégia utilizada, dividiram-se os verbos em três grupos: regulares, irregulares e aqueles que optámos por designar como «semi-regulares»[1].

Consideraram-se verbos regulares os «verbos cuja flexão não exibe qualquer característica estranha ao paradigma de flexão a que esse verbo pertence» (cf. MATEUS & XAVIER, 1992), bastando, para concretizar a sua implementação, registar o radical num ficheiro e noutro os morfemas de tempo/modo e pessoa correspondentes a cada conjugação, que estão indexados uns aos outros. A cada radical aglutinar-se-á a unidade correspondente aos morfemas de tempo/modo e pessoa relativos à conjugação a que pertence o verbo.

Definidos os verbos irregulares como aqueles que apresentam assistematicidades na flexão ou no radical, optou-se por diferenciá-los conforme o tipo de irregularidade. Nos casos dos verbos que apresentam irregularidades na flexão decidiu-se registar todas as suas formas nos ficheiros.

No tratamento dos verbos designados por «semi-regulares», ou seja, aqueles cuja irregularidade consiste na apresentação de alternâncias de radical, e também os que apresentam alternâncias gráficas, optou-se, assim, por constituir paradigmas de verbos. Por exemplo, o verbo medir apresenta o radical *med-* nas 2.ª e 3.ª pessoas do singular e nas três do plural do presente do indicativo. Cada uma das pessoas foi indexada:

med {-;+;+;+;+;+}.

Desta forma, todos os verbos que apresentam esta alternância foram incluídos mesmo paradigma. Assim, foi possível reduzir significativamente a informação.

Vejamos, passo a passo, qual o comportamento do programa perante uma forma verbal como *amarei*. Como se sabe, esta é uma forma regular do verbo *amar*. Para chegar a esta conclusão, o Analisador começa por tentar identificar (da esquerda para a direita) os radicais que estão definidos no primeiro campo de um determinado ficheiro:

Extracto do ficheiro:

1.º campo	2.º campo	3.º campo	4.º campo
am	**amar**	**a**	**full**
↓	↓	↓	↓
Radical verbal	Infinitivo	Vogal temática	Índice que indica que o programa não necessita de ler mais informação

1) Encontrado o radical ***am***-, o programa sabe que este pertence à «pseudo-conjugação» ***a*** (informação relativa aos morfemas flexionais e à vogal temática a aplicar, que está registada em outro ficheiro), seguidamente verificará se ***-arei*** é uma componente válida dessa conjugação.
2) Caso a terminação seja válida, o Analisador passa a verificar a informação presente no 4.º campo. Neste exemplo, ***full*** significa que ao verbo amar, amar, só se aplica um radical. No caso dos verbos com alternância de radical, o índice é outro.
3) Assim, o programa reconhece este verbo como 100% regular, visto ter encontrado uma forma «legal» do verbo, neste caso a do futuro imperfeito, 1.ª pessoa do singular.

Por outro lado, com esta estratégia, foi possível reduzir significativamente a informação linguística, uma vez que no caso dos verbos derivados por prefixação como por exemplo *redizer*, *transpor* e *repor* foi suficiente indexar o prefixo ao índice correspondente ao verbo «primitivo» sem ter sido necessário rescrever todo o verbo, independentemente do facto de o verbo ser regular ou irregular.

4. As formas verbais cliticizadas: como reconhecê-las informaticamente?

Em português, os clíticos podem ocorrer à esquerda do verbo, posição proclítica, à sua direita, posição enclítica ou, ainda, no «interior» do verbo de que dependem, posição mesoclítica.

Em termos informáticos, uma palavra é concebida como uma sequência entre dois espaços, sendo o hífen considerado um carácter como qualquer outro. Sendo assim,

revelam-se problemáticos fenómenos como a ênclise e a mesóclise, já que a próclise não coloca qualquer entrave pois o clítico é lido como uma unidade independente da forma verbal.

Quando o programa se depara com uma forma verbal cliticizada, esteja o clítico em posição mesoclítica ou enclítica, vai reconhecê-la como duas unidades ou três (fê-lo, dar-te-ei), no entanto, nalguns casos, terá, para que o reconhecimento seja profícuo, de reconstituir as formas verbal e clítica.

Para contemplar o reconhecimento das formas verbais mesocliticizadas e encliticizadas foi necessário ter disponível informação específica:

Regras de modificação da forma verbal	Regras de variantes de clíticos

Lista de terminações verbais válidas nas formas cliticizadas

Ao pretender-se que o Analisador reconheça uma forma «composta», vai-se tentar, em função do contexto em que o clítico é encontrado (isto é, se se encontra isolado, precedido de um hífen, ou se lhe sucede outro hífen), chegar à informação mais correcta sobre o mesmo. A título de exemplo, se o programa encontrar um clítico entre hífenes e caso suceda que o segundo hífen seja uma terminação válida de mesóclise, o Analisador vai verificar, depois de aplicadas determinadas regras, se a fusão do que o precede com o que lhe sucede é uma forma válida do futuro ou do condicional—únicos tempos em que o clítico surge em posição mesoclítica:

dar-lhe-ei → darei-lhe

É também neste momento que, em função do contexto, algumas informações encontradas em fases anteriores são consideradas irrelevantes (ex. *no* é clítico e contracção, encontra-se amam-no e *no* é só clítico).

5. Considerações finais

Na construção de um analisador, deve-se, sempre que possível, implementar a informação de modo a torná-la o menos ambígua possível. Se nalguns casos, e em determinados contextos, algumas formas lexicais são ambíguas e não é possível «desfazer» essa ambiguidade aquando da «construção» do dicionário, no caso dos pronomes pessoais átonos esse problema foi ultrapassado através da construção de regras.

A fase seguinte do projecto centra-se no desfazer da ambiguidade categorial das restantes categorias. Para tal está-se a desenvolver um módulo de desambiguação.

NOTAS:

[1] Atente-se no facto de estas designações terem um objectivo exclusivamente confinado ao âmbito deste projecto.

BIBLIOGRAFIA:

AÏT-MOKHTAR, Salah (1995), *SMORPH: Guide d'utilisacion*. Rapport tecnhique, GRIL, Université Blaise Pascal, Clermont-Ferrand.

——————— (1997), «Du texte ASCII au texte lemmatisé: la présyntaxe en une seule étape», in *Actas de TALN'97 (Traitement Automatique du Langage Naturel)*, Grenoble.

ANDRADE, Ernesto d' (1993), *Dicionário Inverso do Português*, Lisboa: Edições Cosmos.

CHANOD, J.-P. e Pasi TAPANAINEN (1995), «Tagging French – comparing a statistical and a constraint-based method», in *Proceedings of the EACL-95*, Dublin.

COSTA, J. Almeida e A. Sampaio e MELO (1990), *Dicionário da Língua Portuguesa*, 6.ª edição revista e aumentada, Porto: Porto Editora.

CUESTA, Pilar Vázquez e Maria Albertina Mendes da LUZ (1971), *Gramática da Língua Portuguesa*, Lisboa: Edições 70.

CUNHA, Celso e Lindley CINTRA (1985), *Nova Gramática do Português Contemporâneo*, Rio de Janeiro: Editora Nova Fronteira.

FERREIRA, Aurélio Buarque de Holanda (1986), *Novo Dicionário da Língua Portuguesa*, 2.ª edição revista e aumentada, 2.ª impressão, Rio de Janeiro: Editora Nova Fronteira.

GREFENSTETTE, Gregory e Pasi TAPANAINEN (1994), «What is a word. What is a sentence? Problems of Tokenization», in *The Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)*, Budapeste.

HAGEGE, Caroline, António MEIRELES, Brígida TRINDADE, Carla DIOGO e Fernando LEITE (1997), *Analisador Morfossintáctico*, Relatório técnico, Instituto de Linguística Teórica e Computacional, Lisboa.

MATEUS, Maria Helena Mira, Ana Maria BRITO, Inês DUARTE e Isabel Hub FARIA (1989), *Gramática da Língua Portuguesa*, 2.ª edição revista e aumentada, Série Linguística, Lisboa: Editorial Caminho.

SÁ NOGUEIRA, Rodrigo de (1991), *Dicionário de verbos portugueses conjugados*, 9.ª edição, Lisboa: Clássica Editora.

SANCHEZ LEON, Fernando (1995), «Development of a Spanish Version of the Xerox Tagger», Facultad de Filosofía y Letras, Universidade Autónoma de Madrid, Madrid.

SILVA, Emídeo e António TAVARES (1989), *Dicionário dos Verbos Portugueses: conjugação e referências*, Porto: Porto Editora.

TEYSSIER, Paul (1984), *Manuel de Langue Portugaise – Portugal-Brésil*, 12ème édition revue et corrigée, Paris: Editions Klincksieck.

VILLALVA, Alina (1994), *Estruturas Morfológicas: Unidades e Hierarquias nas Palavras do Português*, Dissertação de Doutoramento apresentada à FLUL.

VILELA, Mário (1990), *Dicionário do Português Básico*, Porto: Edições Asa.

WILKENS, Mike e Julian KUPIEC (1995), «Training Hidden Markov Models for Part of Speech Tagging». Xerox Corporation, Palo Alto.