Elisabete Ranchhod & Paula Carvalho

Unidades lexicais complexas
Problemas de análise e etiquetagem

In Actas del VIII Simposio Internacional
de
Comunicación Social

pp. 212-217

Santiago de Cuba, 2003

ELISABETE MARQUES RANCHHOD
Faculdade de Letras de Lisboa & LabEL (CAUTL-IST)
Lisboa - Portugal
elisabet@label.ist.utl.pt

PAULA CARVALHO
Faculdade de Letras de Lisboa & LabEL (CAUTL-IST)
Lisboa - Portugal
paula@label2.ist.utl.pt

Unidades lexicais complexas. Problemas de análise e etiquetagem[*]

Abstract. Compound words form the bulk of the lexicon of languages, and they are numerous in all types of texts. They are usually built from the vocabulary of simple words, but their meaning is not always compositional. From a syntactic point of view, there are combinations of words that are found to be ambiguous, with one analysis as multi-word lexical unit and the other as a free sequence of simple words. However, when they occur in an adequate syntactic context such ambiguity disappears.
We discuss some of the problems that that kind of ambiguity poses to natural language processing. The discussion is based on the results of applying a set of disambiguation grammars that were conceived to eliminate lexical ambiguities and to tag the disambiguated words.

1. Apresentação

A maioria das operações de processamento automático das línguas naturais tem como objectivo último extrair dos textos unidades de significado. É assim em tradução automática, pesquisa e extracção de informação, elaboração de resumos e sumários, etc.

Se, em geral, as unidades de significado correspondem a unidades frásicas, muitas vezes, uma boa parte do conteúdo informativo dos textos está contida em unidades lexicais complexas (palavras constituídas por mais do que uma palavra simples). Estas unidades são muito frequentes em textos de caracter técnico e científico, designando-se, neste caso, por termos ou termos técnicos, mas são igualmente numerosas em todo o tipo de textos, nomeadamente nos de carácter jornalístico.

Como ilustração, veja-se o seguinte extracto de um texto publicado na edição on-line do jornal «Expresso»[i] de 31 de Agosto de 2002, onde estão destacadas as unidades lexicais compostas:

> Mas o tema não é pacífico. É complicado romper a curto prazo com um modelo tecnológico, dominante desde o século XIX, assente no uso maciço de carvão, petróleo e gás. Estes representam actualmente 80% do consumo total de energia e são responsáveis pela libertação de seis mil milhões de toneladas de dióxido de carbono para a atmosfera, conduzindo aos problemas na camada de ozono e consequentes alterações climáticas. Aposta nas energias alternativas. As outras fontes de energia actualmente utilizadas no mundo são a nuclear (16% da electricidade gerada) e as modernas energias renováveis, como a hidroeléctrica, eólica, solar e geotérmica, que ocupam apenas 4,5% da produção total de energia. Daí que um dos objectivos estabelecidos para o desenvolvimento sustentável seja a diversificação das fontes de energia mais limpas e o aumento da quota de fontes de energia renováveis para, pelo menos, 5% em todos os países, até 2010. Neste caso, a meta da União Europeia é mais ambiciosa. A UE pretende passar da actual média de 6% de consumo proveniente das energias renováveis para 15%. Para Portugal não será difícil, uma vez que já conta com uma fatia de 13% graças à energia hidroeléctrica, de acordo com o Eurostat. Porém, mais uma vez, os Estados Unidos são os principais opositores à definição de objectivos quantificáveis para o uso das energias renováveis, tentando esquecer-se de que são responsáveis por um quarto da poluição mundial.

Na sua maioria trata-se de nomes compostos: *dióxido de carbono*, *camada de ozono*, *alterações climáticas*, *energias alternativas*, *desenvolvimento sustentável*, etc., mas existem igualmente advérbios: *a curto prazo*, *mais uma vez*; conjunções: *uma vez que*, preposições: *graças a* [ii], etc.

2. Identificação e tratamento de nomes compostos

Como se vê por esta pequena amostra, dado, por um lado, o importante número de unidades lexicais compostas, e, por outro, a sua importância na recuperação do conteúdo informativo dos textos, é conveniente que os sistemas de processamento automático de textos as possam identificar e tratar adequadamente.

Os nomes compostos que se grafam com hífen são normalmente identificados pelos lexicógrafos e registados nos dicionários. Contudo, muitos deles não têm tal grafia, nem esta seria um critério linguístico

adequado para caracterizar uma sequência de palavras como unidade lexical. De facto, na identificação de compostos há que utilizar um conjunto de critérios linguísticos que vão desde a verificação do comportamento morfológico dos seus constituintes até à análise das propriedades sintácticas e semânticas. Vejamos, por exemplo, o comportamento linguístico do nome *desenvolvimento sustentável* e do advérbio *a curto prazo*. O nome *desenvolvimento*, quando ocorre em combinações livres, varia em número: *haverá certamente um novo desenvolvimento*; *haverá certamente novos desenvolvimentos*. Como constituinte do nome composto, perde estas propriedades morfológicas: *estabeleceram-se objectivos para o desenvolvimento sustentável* vs **estabeleceram-se objectivos para os desenvolvimentos sustentáveis*. O adjectivo *sustentável* pode ser quantificado e modificado: *uma posição facilmente sustentável*, mas isso não é possível no composto: **estabeleceram-se objectivos para o desenvolvimento facilmente sustentável*.

Quanto a *a curto prazo*, embora seja constituído por um adjectivo que varia em género e número (*curto*) e por um nome pluralizável (*prazo*), quando estes elementos se combinam para formar o advérbio, perdem todas essas propriedades. Além disso, o adjectivo *curto* é, em construções livres, um adjectivo predicativo que ocorre à direita dos nomes: *cabelo curto* (**curto cabelo*); *vestido curto* (**curto vestido*). No advérbio, ocorre obrigatoriamente à esquerda: **a prazo curto*. De facto, integrado nesta unidade lexical, não tem valor de adjectivo predicativo, quando muito pode ser aproximado de um quantificador.

Se um sistema de processamento de texto não utilizar a noção de composto, cada uma das palavras que o constituem será analisada e etiquetada como uma palavra simples, o que não é de modo algum adequado. Retomando o advérbio *a curto prazo*, ele deve ser analisado como:

a curto prazo	*Advérbio*

e não:

a	*Determinante*
a	*Preposição*
a	*Pronome*
curto	*Adjectivo*
prazo	*Nome*

2. 1. O Problema da Ambiguidade

Pelas razões brevemente apontadas antes, os compostos devem ser adequadamente representados e utilizados na análise automática de texto. Um dos problemas que se coloca é o da sua eventual ambiguidade. Se bem que a ambiguidade nos compostos seja menor do que a que se observa nas palavras simples (veja-se acima o caso de *a*, que, como palavra simples, pode ser um determinante, uma preposição ou uma forma pronominal, mas no advérbio só tem valor preposicional), há unidades lexicais que tanto podem ser analisadas como compostas, e receberem um traço categorial único, como podem constituir sequências livres de palavras simples.

No texto acima, os compostos ambíguos são poucos e, por isso, a maioria pode ser imediatamente etiquetada, com os atributos adequados, na fase de pré-processamento de texto (preparação do texto para a análise sintáctica). O tratamento daqueles que são ambíguos necessitará, contudo, de uma análise sintáctica do texto que esclareça qual o seu estatuto linguístico no contexto em que ocorrem. É o caso de:

Ele está sempre a dizer graças à Maria

Ele alugou um quarto perto da faculdade

graças a não é, neste exemplo, uma preposição, nem sequer constitui uma unidade sintáctica: *graças* é o complemento directo de *dizer* e *a* uma preposição que introduz o complemento dativo desse verbo; no segundo exemplo, *um quarto* não é, como no texto, um número fraccionário, mas um sintagma nominal, complemento directo de *alugar*.

2.2. Análise de grupos nominais que contêm, ou não, nomes compostos

A fim de analisar correctamente grupos nominais ambíguos, isto é, os GN que contêm palavras que, descontextualizadas, podem pertencer a mais do que uma categoria gramatical, construímos gramáticas que restringem as possibilidades de co-ocorrência das diversas categorias gramaticais. Estas gramáticas estão formalizadas em transdutores de estados finitos e são aplicadas aos textos, em combinação com os dicionários de palavras simples e compostas[iii], pelo sistema INTEX[iv]. A Fig. 1 representa parcialmente

uma dessas gramáticas.

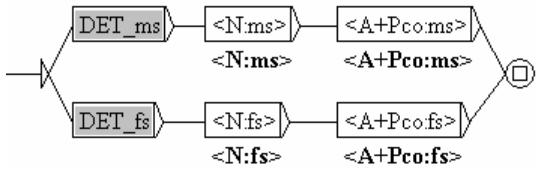

Fig. 1

Descreve as restrições - morfológicas, sintácticas e combinatórias - que se devem observar entre nomes e adjectivos (as duas categorias gramaticais que maior número de formas homógrafas têm em comum), no interior de grupos nominais. Depois de aplicado aos textos o conjunto de gramáticas construído, as análises que violam as restrições especificadas nas gramáticas são eliminadas.

A gramática da Fig. 1 é bastante simples: aplica-se a grupos nominais que contenham um nome, especificado por um ou mais determinantes, representados na subgramática *DET* (nó sombreado), e modificado por um adjectivo predicativo de cor, *A+Pco*. Dado que, em português, os constituintes do grupo nominal têm de concordar em género e número, estão ainda representadas na gramática duas estruturas paralelas, uma em que os constituintes se encontram no masculino do singular e outra no feminino do singular. Na versão integral da gramática estão, naturalmente, especificadas as estruturas cujos constituintes são formas do masculino e feminino plural.

Como a utilização das gramáticas não está dependente dos textos a processar, elas são aplicadas pelo sistema na análise e etiquetagem de grupos nominais, contenham eles ou não nomes compostos. Assim, é suposto que as gramáticas resolvam correctamente as ambiguidades existentes em grupos nominais livres, como por exemplo: *um vestido preto*, *um lápis vermelho*, *uma pasta amarela*, mas que sejam igualmente adequadas para processar grupos nominais que contenham palavras compostas, como é o caso de: *uma caneta de feltro vermelha, uma camisa de forças branca*, *uma guitarra eléctrica preta*, etc.

De modo a ilustrar o desempenho da gramática representada na Fig. 1, tomemos como exemplo os grupos nominais: *um vestido preto* e *uma caneta de feltro vermelha*. A análise lexical destes grupos nominais, antes da aplicação da gramática, está representada nas Fig. 2 e Fig. 3 sob a forma de um transdutor, automaticamente construído pelo sistema.

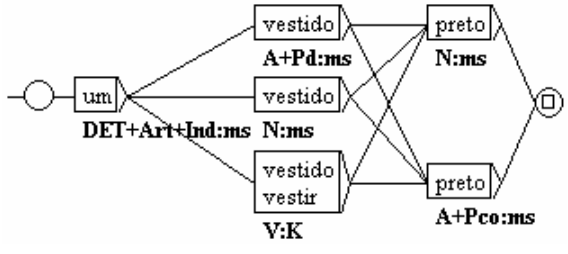

Fig. 2

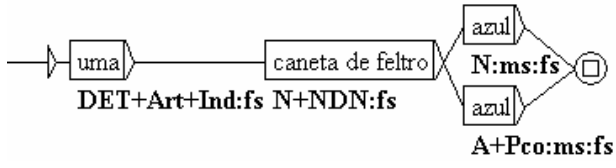

Fig.3

A maioria das unidades lexicais que compõem o primeiro grupo nominal são ambíguas, conforme ilustra a Fig. 2. *Vestido* pode corresponder a um nome (N), à forma participial do verbo *vestir* (V:K) ou ainda a um adjectivo de tipo predicativo (A+Pd); *preto* é ambíguo entre nome e adjectivo, neste caso um adjectivo de cor (A+Pco).

Relativamente ao transdutor da Fig. 3, que representa o grupo nominal *uma caneta de feltro azul*, verifica-se que o número de análises ambíguas é substancialmente menor do que no anterior. A razão pela qual isso acontece deve-se ao facto de a combinação *caneta de feltro* ter sido previamente reconhecida como um nome composto não ambíguo, impedindo, desse modo, a atribuição de análises lexicais individuais a cada um dos seus constituintes.

Depois da aplicação da gramática da Fig. 1, as análises incorrectas foram eliminadas, e as unidades lexicais etiquetadas com os atributos linguísticos adequados. Os resultados deste processamento podem novamente ser visualizados sob a forma de transdutores (Fig. 4 e Fig. 5).

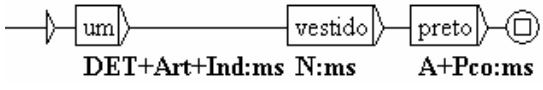

Fig. 4

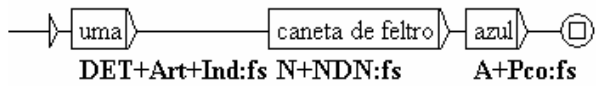

Fig. 5

2.3. Grupos nominais com nomes compostos ambíguos

Se, diferentemente dos casos anteriores, um dado texto contiver compostos ambíguos (isto é, susceptíveis de, dependendo do contexto, constituírem, ou não, um composto), como é o caso de *saco azul*, cerveja *preta*, etc., em que o adjectivo pode ser analisado como um modificador livre do nome que o precede ou, pelo contrário, formar com ele uma unidade lexical complexa, nem sempre este tipo de gramática dá resultados satisfatórios.

A análise lexical de *saco azul* está representada no transdutor da Fig. 6, automaticamente construído pelo sistema.

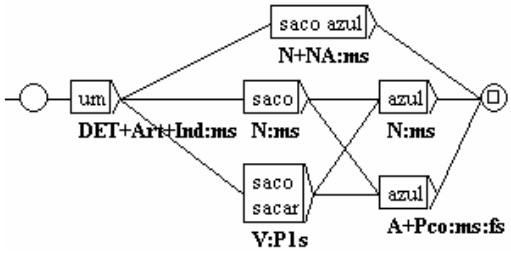

Fig. 6

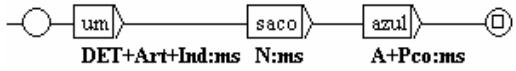

Fig. 7

O transdutor contempla a análise de *saco azul* como nome composto, mas fornece, além desta, outras possibilidades de análise: *saco* pode corresponder a uma forma do verbo *sacar* e *azul* pode corresponder a um nome ou a adjectivo. A forma adjectival é ainda ambígua no que diz respeito aos seus traços morfológicos. Por não ter morfemas de género, *azul* deverá, em função do nome a que esteja associado, ser analisado como uma forma masculina ou feminina.

Embora a aplicação da gramática de resolução de ambiguidades permita eliminar algumas análises incorrectas, como ilustra o transdutor da Fig. 7, elimina também, inadequadamente, a análise de *saco azul* como composto. Esse resultado é, de certa forma, esperado, uma vez que a análise da sequência em questão como grupo nominal livre é a única que satisfaz todos os requisitos impostos pela gramática. Porém, do ponto de vista linguístico, não é adequado, uma vez que o estatuto de *saco azul* depende da estrutura sintáctica em que esteja integrado. Se se encontrar na posição de complemento de um verbo como, por exemplo, *comprar*, o resultado fornecido pela gramática é correcto (*saco azul* não pode ser analisado como um nome composto):

A Maria comprou um saco azul

Mas a mesma sequência pode ocorrer na posição de complemento de um verbo como *constituir:*

Essa empresa constituiu um saco azul, ilegalmente.

Neste caso, o valor (lexical e semântico) de *saco azul* é diferente do do primeiro exemplo, e apenas a análise da sequência como nome composto, incorrectamente eliminada pela gramática, é adequada.

Mas nem sempre se observam estes maus resultados. Há situações em que as mesmas gramáticas fornecem uma análise correcta dos grupos nominais que contêm nomes compostos ambíguos. É, por exemplo, o caso do grupo nominal: *uma colher de açúcar amarela*, cuja ambiguidade está explícita no transdutor da Fig. 8.

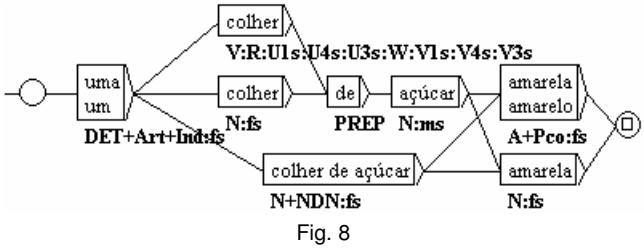

Fig. 8

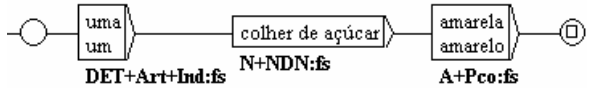

Fig. 9

O nome composto é ambíguo com uma estrutura sintáctica livre do tipo: *Nome Preposição Nome*. Contudo, a gramática que temos vindo a utilizar obriga à verificação dos traços morfológicos das várias categorias gramaticais, e, por isso, analisa correctamente *colher de açúcar* como nome composto, a única análise que satisfaz as condições de concordância impostas pela gramática.

2.3.1. Sobreposição de nomes compostos

As condições representadas na mesma gramática são, porém, insuficientes para analisar uma sequência como: *uma caneca de cerveja preta*, que possui uma estrutura aparentemente idêntica à anterior. Neste caso, todos os constituintes do grupo nominal estão no feminino singular, sendo, por isso, difícil calcular qual dos nomes (*caneca* ou *cerveja*) são modificados pelo adjectivo, que é adequado aos dois.

Este exemplo introduz um novo problema, que se prende com a sobreposição de nomes compostos: o nome *cerveja* tanto pode ser analisado como sendo um dos constituintes do composto *caneca de cerveja* como do composto *cerveja preta* (Cf. o transdutor fornecido pelo sistema após a aplicação dos dicionários).

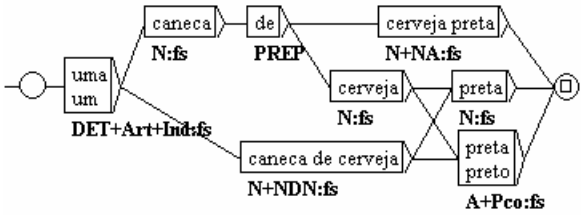

Fig. 10

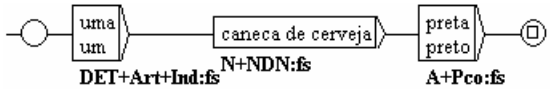

Fig. 11

A aplicação ao texto da gramática de resolução de ambiguidades (Fig. 1) fornece apenas a análise de *caneca de cerveja* como nome composto, eliminando todas as restantes leituras. No grupo nominal em questão, o problema não reside em optar por um grupo nominal livre ou um nome composto, mas em decidir de que nome composto se trata. Essa decisão não pode ser tomada no âmbito restrito da análise da estrutura de grupos nominais. Ela obriga a uma análise sintáctica da frase que contém a sequência. Tal como no exemplo anterior, é necessário verificar os traços argumentais dos verbos de que depende a sequência antes de poder decidir se se trata do composto *caneca de cerveja* se do composto *cerveja preta*. Numa construção com o verbo *partir* :

O Zé partiu uma caneca de cerveja preta

apenas é adequada a análise de *caneca de cerveja* como nome composto, sendo *preta* um modificador livre desse nome. Essa mesma análise é incorrecta se, em vez de *partir*, tivermos um verbo como *beber* :

O Zé bebeu uma caneca de cerveja preta

Agora, a etiqueta nome composto (N+NA) deve ser atribuída à expressão *cerveja preta*, constituindo *uma caneca de* um especificador daquele nome composto.

3. Conclusão

As palavras compostas, isto é as unidades lexicais formadas por sequências de palavras simples, constituem uma parte significativa do léxico de qualquer língua e são muito frequentes em qualquer tipo de texto, em particular nos de natureza técnica e científica. Em geral, são formadas por utilização das regras gerais de combinação de palavras e categorias gramaticais. No entanto, apresentam restrições a essas combinações e o seu significado é também muitas vezes não composicional. As operações de processamento automático de texto não podem ignorar estas unidades, tanto mais que grande parte do conteúdo dos textos é recuperável através da sua correcta identificação e análise.

As palavras simples são muito ambíguas (uma mesma forma pode pertencer a várias categorias gramaticais, ter variadíssimos significados, etc.), o que causa grandes problemas em processamento automático de texto. A ambiguidade dos compostos é muito menor. Contudo, alguns compostos são ambíguos, uma vez que também podem ser analisados como uma combinação livre de palavras. A resolução desta ambiguidade passa pela elaboração de gramáticas específicas. No que diz respeito aos nomes compostos (cujo número se estima ser várias vezes superior ao dos nomes simples) é natural pressupor que, quando são ambíguos, possam ser adequadamente tratados por gramáticas que analisem grupos nominais. Demostrámos que essa solução é, de facto, adequada para um número

significativo de casos, mas ficou igualmente claro que muitas vezes a resolução da ambiguidade dos grupos nominais só pode feita a um nível sintáctico superior, que tenha em conta as restrições argumentais que os verbos de que dependem impõem.

Referências

Baptista, Jorge. (1995). *Estabelecimento e formalização de classes de nomes compostos*. Tese de mestrado, Faculdade de Letras da Universidade de Lisboa.

Carvalho, Paula. (2001). *Gramáticas de Resolução de Ambiguidades Resultantes da Homografia de Nomes e Adjectivos*. Tese de mestrado, Faculdade de Letras da Universidade de Lisboa.

Gross, Maurice. (1988). Methods and Tactics in the Construction of a Lexicon-Grammar. In: *Linguistics in the Morning Calm*, selected papers from SICOL-86. Seoul: The Linguistic Society of Korea .

Joshi, Aravindi. (1999). A parser from antiquity: An early application of finite state transducers to natural language parsing. In: Kornai, A. (ed.): *Extended Finite State Models of Language. Studies in Natural Language Processing*. Cambridge University Press.

Laporte, Éric; Monceaux, A. (1999). Elimination of lexical ambiguities by grammars: The ELAG system. In: Fairon, C. (ed.): «Analyse lexicale et syntaxique: Le système INTEX». *Lingvisticae Investigationes*, XXII (Volume Spécial). Jonh Benjamins, Amsterdam Philadelphia.

Laporte, Éric. (2001). Resolução de ambiguidades. In: Ranchhod, E. (ed.): *Tratamento das Línguas por Computador*. Caminho, Lisboa.

Ranchhod, E., Mota, C., Baptista, J. (1999). A Computational Lexicon of Portuguese for Automatic Text Parsing. In *Proceedings of SIGLEX' 99: Standardizing Lexical Resources*, 37th Annual Meeting of the ACL, Maryland.

Ranchhod, Elisabete. (1999). Ressources linguistiques du portugais implémentées sous INTEX. In: Fairon, C. (ed.): «Analyse lexicale et syntaxique: Le système INTEX». *Lingvisticae Investigationes*, XXII (Volume Spécial). Jonh Benjamins, Amsterdam Philadelphia.

Silberztein, Max. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Masson, Paris.

[*] Este trabalho foi, em parte, financiado pela FCT no âmbito do Projecto ENLEX – *Enhancement of Large-scale Lexicons*. Ref. POSI/PLP/34729/99.

[i] http://www.expresso.pt/

[ii] Em Português as preposições aparecem frequentemente contraídas com os determinantes. É o caso de *à*, que corresponde à contracção da preposição *a* e do artigo *a*. Esta é uma questão que não será tratada no âmbito deste trabalho.

[iii] Ver neste volume *Dicionários Electrónicos do Português. Características e Aplicações.*

[iv] Ver http://www.nyu.edu/pages/linguistics/intex/