

AVALON'2003

Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português

Faro, 28 de Junho de 2003

Vantagens da integração de dicionários de palavras compostas em sistemas de PLN

Elisabete Ranchhod

FLUL & LabEL (CAUTL/IST)

<http://label.ist.utl.pt>

Estruturação de dados lexicais e gramaticais

Palavras compostas

Dicionários

Palavras Simples

(120 000)

partir, V
partido, Adj
partido, N

Palavras Compostas

bom partido, N+AN
a partir de, Prep
à partida, Adv

Formas Flexionadas

(1 250 000)

partir, parto, parti, ...
partido, partidos

Formas Flexionadas

(50 000)

bom partido, N+AN:ms
bons partidos, N+AN:mp
a partir de, Prep

FST Lexicais

I, II, III, IV, ...
vinte e um, ...
vinte e uma, ...

Gramáticas

Resolução de Ambiguidades

Pregar
{uma, DET+Art+Def:fs}
{boa, A+Pde:fs}
{partida, N:fs}

Gramáticas Locais

Sábado, 28 de Junho
de 2003

Léxico-Gramática

Construções Livres

partir, V+Vmt
 N_0 partir de NLoc para NLoc

partir, V+Vt
 N_0 partir N_1

partir, V+Vtpc
 N_0 partir Npc (=: braço, nariz ...)

Expressões Fixas

Nhum partir a (cara+ fronha + ...) a Nhum

PCAN

Nhum partir desta para melhor

PCPP

Expressões linguísticas não composicionais

Unidades lexicais compostas

(São constituídas por mais do que uma palavra, mas comportam-se como uma unidade lexical)

- (i) **Expressões nominais** (nomes compostos);
- (ii) **Expressões adverbiais** (advérbios compostos);
- (iii) **Expressões adjectivais** (adjectivos compostos);
- (iv) **Conjunções** (conjunções compostas);
- (v) **Preposições** (preposições compostas);
- (vi) **Determinantes** (determinantes compostos).

Expressões linguísticas não composicionais

Expressões oracionais

- (i) **Frases fixas**, em que se observam restrições linguísticas muito fortes entre um verbo e as várias posições sintáticas envolvidas, em geral, o(s) complemento(s);
- (ii) **Provérbios**, em que, a par de uma interpretação específica, se podem observar certos 'desvios' sintáticos.

Nomes compostos

Os **nomes compostos** constituem a parte mais numerosa do léxico nominal das línguas. Muitos pertencem ao vocabulário corrente, outros são termos técnicos e científicos, criados continuamente, ao ritmo do desenvolvimento tecnológico.

São geralmente formados a partir de palavras simples por meio de regras gerais de combinação de palavras. O seu significado é, na maior parte dos casos, não composicional.

Exemplos de estruturas produtivas:

N Adj: cordão umbilical, fibra óptica, via verde, lugar-comum, obra-prima, .

N de N: efeito de estufa, fibra de vidro, golpe de teatro, braço-de-ferro, ...

Exemplos de estruturas heterogéneas:

quem de direito, maria-vai-com-as-outras, vitamina C, ...

Nomes compostos

A tradição gramatical e lexicográfica considera que os nomes compostos deveriam ser ortografados com hífen (caso de **braço-de-ferro**, **café-concerto**, **cara-metade**, **bem-me-quer**). Assim, só os compostos que têm esta grafia constituem entradas de dicionário.

Contudo, na identificação de nomes compostos, há que utilizar um conjunto de critérios linguísticos, que vão desde a verificação do **comportamento morfológico** dos seus constituintes (restrições sobre a flexão) até à verificação da sua, total ou parcial, **perda de composicionalidade**, lexical, sintáctica e semântica.

Ocorrem em distribuições tipicamente nominais, como nos exemplos:

- (1) O tronco é fabricado em **betão** e **fibra de vidro**, enquanto os ramos e as folhas são de plástico.
- (2) É uma questão que **quem de direito** deve resolver. A **administração** deve pronunciar-se sobre isto.

Nomes compostos

As classes mais representativas dos nomes compostos binários (de acordo com o recenseamento do LabEL) são:

Classes	Estruturas	Exemplos
NA	Nome Adjectivo	via verde; batata-doce
NDN	Nome de Nome	efeito de estufa; braço-de-ferro
AN	Adjectivo Nome	falsa modéstia; mau-olhado
NPN	Nome Prep Nome	barco a remos; voto em branco
NPV	Nome Prep Verbo	canção de embalar; ferro de engomar
VN	Verbo Nome	coca-bichinhos; ganha-pão
PN	Preposição Nome	sem-abrigo; sob-roda
NN	Nome Nome	cara-metade; raio laser
NCN	Nome Conj Nome	saia e casaco; prós e contras
XX	---	habeas corpus; modus vivendi

Fig. 1: *Classes Formais de Nomes Compostos*

Advérbios compostos

Os **advérbios compostos** (ou expressões adverbiais fixas ou advérbios idiomáticos) são o caso mais simples de formas compostas. Ocupam as posições sintáticas características dos advérbios e complementos circunstanciais, e, em geral, não são interpretáveis composicionalmente.

Nas frases:

- (1) O Zé fez isso **a contragosto**
- (2) O Zé expôs a questão **de viva voz**
- (3) O Zé contou isso à Ana **tintim-por-tintim**

a contragosto, **de viva voz** e **tintim-por-tintim** são adjunções facultativas às frases, **comutam com (e têm o valor de) advérbios**:

- (1') O Zé fez isso (**a contragosto** + constrangidamente)
- (2') O Zé expôs a questão (**de viva voz** + pessoalmente)
- (3') O Zé contou isso à Ana (**tintim-por-tintim**, pormenorizadamente)

Advérbios compostos

Não permitem:

Inserções

*O Zé fez isso **a esse contragosto**

*O Zé expôs a questão **de muita viva voz**

Reduções

* O Zé contou isso à Ana **tintim**

*O Zé expôs a questão **de voz**

Comutações

O Zé expôs a questão **de viva** (**voz** + ^{*}**presença**)

Alterações morfológicas

*O Zé expôs a questão **de vivas vozes**

Advérbios compostos

Nos dicionários do LabEL, os advérbios compostos foram classificados tendo em conta a sua constituição categorial, embora a noção de categoria gramatical perca, nestes casos, grande parte da sua pertinência.

Classe	Estrutura	Exemplos
P-PADV	Adv	tão-somente
P-PC	Prep C	de rompante
P-PDETC	Prep Det C	à pressa
P-PAC	Prep Adj C	de bom grado
P-PCA	Prep C Adj	a olhos vistos
P-PCDC	Prep C de C	com pezinhos de lã
P-PCPC	Prep C Prep C	de alto a baixo
P-PCDN	Prep C de N	em matéria de N
P-PCPN	Prep C Prep N	no tocante a N
P-PCONJ	Prep C Conj C	contra ventos e marés
P-PV	Prep V W	a bem dizer
PACO	(Adj) como C	como uma porta
P-PVCO	(V) como C	como sopa no mel
P-PPCO	(V) como Prep C	como do dia para a noite
P-PJC	Conj C	e assim por diante
P-PF	F	sem tugar nem mugir

Fig. 2: *Classes Formais de Advérbios compostos*

Adjectivos compostos

Adjectivos compostos com função predicativa

Entram em frases com a seguinte forma geral:

NO (ser + estar) Adj W

A estrutura interna destes adjectivos é variada, e pode ser complexa. Como acontece com todas as expressões não composicionais, observa-se uma fixidez total ou parcial entre os elementos constituintes. Os tipos mais frequentes são os seguintes:

Adj Prep C =: **duro de ouvido, novo em folha, baço para espelho, cheio de nove horas**

Adv Adj =: **bem-parecido, mal-agrado, meio-doido, muito visto**

Adj Conj Adj =: **certo e sabido, impávido e sereno, velho e relho, pobre e mal-agrado**

Adjectivos compostos

Os adjectivos compostos foram integrados em classes sintácticas, de acordo com os seguintes critérios principais:

- (i) Construção dos adjectivos com **ser** e **estar** ou apenas com um deles;
- (ii) Existência ou não de complementos livres;
- (iii) Aceitação de uma completiva na posição de sujeito e/ou complemento.

Classes	Estruturas	Exemplos
SA	N_0 <i>ser</i> Adj	O Zé é maior e vacinado
EA	N_0 <i>estar</i> Adj	O bife está mal-passado
SEA	N_0 (<i>ser</i> + <i>estar</i>) Adj	O Zé (é + está) doido varrido
QSA	(Que F) $_0$ <i>ser</i> Adj	É certo e sabido que vai haver problemas
SAPN	N_0 <i>ser</i> Adj Prep N	A Ana é mal-empregada para o Zé
EAPN	N_0 <i>estar</i> Adj Prep N	O Zé está bem-visto junto do eleitorado

Fig. 3: *Classes Sintácticas de Adjectivos Compostos*

Frases fixas

Trata-se de **frases** que, pertencendo a registos variados, possuem uma característica comum: contêm combinações *verbo-nome* que **não são distribucionalmente produtivas** nem são interpretáveis composicionalmente:

- (1) O Zé esticou o pernil
- (2) O Zé salvou a honra do convento
- (3) O Zé meteu a mão na consciência
- (4) O Zé chamou a Ana à pedra

As frases (1) a (4) têm uma estrutura sintáctica idêntica à das frases ‘livres’:

- (5) O Zé esticou (a perna + o braço)
- (6) O Zé salvou o património da empresa
- (7) O Zé meteu a mão no bolso
- (8) O Zé chamou a Ana ao gabinete

Frases fixas

No recenseamento das frases fixas, tiveram-se fundamentalmente em conta os seguintes factores:

- (i) as fortes restrições distribucionais que se observam entre os verbos e os grupos nominais que se encontram formalmente na posição de complemento (mais raramente na posição de sujeito);
- (ii) o facto de essas restrições bloquearem a aplicação às frases de algumas operações sintácticas que envolvem verbos e grupos nominais;
- (iii) a interpretação não composicional das construções.

Frases fixas

As frases foram incluídas em classes sintáticas, de acordo com os princípios gerais utilizados na classificação dos verbos.

Classes	Estruturas	Exemplos
P VC0	C ₀ V W	O Senhor chamou o Zé à Sua presença
P VC1	N ₀ V C ₁	O Zé perdeu a cabeça
	N ₀ V (C de C) ₁	O Zé salvou a honra do convento
P VC2	N ₀ V (C de N) ₁	O Zé exige a cabeça dos culpados
P VC3	N ₀ V C ₁ a N ₁	O Zé deu carta-branca à Ana
P VC4	N ₀ V C ₁ Prep N ₂	O Zé passou uma esponja sobre o assunto
P VC5	N ₀ V Prep C ₁	O Zé rema contra a maré
P VC6	N ₀ V Prep (C de N) ₁	O Zé tocou na corda sensível da Ana
P VC7	N ₀ V N ₁ Prep C ₂	O Zé meteu a Ana num chinelo
P VC8	N ₀ V Prep C ₁ C ₂	O Zé fez das tripas coração
P VC9	N ₀ V C ₁ Prep C ₂	O Zé entregou a alma ao Criador
P VC10	N ₀ V Prep C ₁ Prep C ₂	A novidade voou de boca em boca
P VC11	N ₀ V que F Prep C ₂	Ele soube isso de fonte segura
P VC12	N ₀ V C ₁ Prep que F	O Zé reflecte duas vezes antes de falar
P VC13	N ₀ V C ₁ Prep C ₂ Prep N ₃	O Zé tirou as palavras da boca à Ana

Fig. 4: *Classes de Frases Fixas*

A detenção de [Fátima Felgueiras] **na semana passada** acabou por levar a uma declaração de [Ferro Rodrigues] onde este retirou a "**confiança política**" à autarca e realizou aquele difícil **número de malabarismo** que consiste em **meter os pés pelas mãos**: o **PS** acha que [Fátima Felgueiras] não tem condições para continuar como autarca, já não tem "**confiança política**" na autarca, acha que ela se deve demitir se for acusada formalmente, mas se se provar que houve uma cabala, lá estará para a ajudar a **subir ao pódio** e desmascarar os cabalistas.

A posição é incompreensível porque o que os dirigentes do **PS** sempre disseram foi que [Fátima Felgueiras] era inocente e seria considerada como tal pelo partido **até prova em contrário**. Como os tribunais ainda não concluíram nada, não se percebe a mudança de posição.

Para além de todo o circo que envolve a questão e do evidente **mal-estar** dos dirigentes do **PS** (e do **PSD**: veja-se [Durão Barroso], com a prudência de um vidraceiro, a dizer que [Felgueiras] é uma questão interna do **PS**...), há uma questão que **vale a pena** reter: a facilidade com que os políticos **em geral** invocam a independência da justiça e a **presunção de inocência** para evitar comprometer-se no julgamento desta ou daquela figura.

Que a independência da justiça e a **presunção de inocência** são princípios fundamentais já se sabe. Mas isso não pode servir de cobertura **para que** os partidos **fujam à responsabilidade** de avaliar os homens e mulheres que propõem ao país nas suas listas e muito menos à responsabilidade de os sancionar (e afastar) quando a sua prática é inaceitável.

O que gostaríamos de saber, no caso de [Fátima Felgueiras] e dos outros autarcas suspeitos, é se os partidos que os fizeram eleger continuam a **pôr a mão no fogo** e a **cabeça no cepo** por eles ou se, **pelo contrário**, lhes retiraram realmente a sua confiança - o que é, em política, uma posição **sem retorno**. É importante saber isto porque, se um dado autarca pode vir a reconquistar a confiança do seu partido se se safar das **teias da lei**, isso pode significar que esse partido deve deixar de merecer a nossa.

Dizer que só aos tribunais cabe julgar está muito bem, mas isso não pode significar que os partidos se desvinculam dos actos dos seus eleitos **na hora H**, em que a porcaria chega à ventoinha - mas não antes.

Observações finais

Por razões de natureza estritamente linguística, mas também por razões que se prendem com as aplicações da linguística computacional (tradução automática, pesquisa e extracção de informação, elaboração de resumos, etc.), parece desejável que, aquando do processamento de um texto, os analisadores lexicais identifiquem adequadamente as palavras compostas.

Assim, sequências de palavras como:

teias da lei

até prova em contrário

devem ser tratadas como unidades lexicais, respectivamente um nome e um advérbio, e incluídas nos módulos de dicionários:

teias da lei, teias da lei, N+NDN:fp

até prova em contrário, ADV+PCPC

Observações finais

Se, pelo contrário, essas combinações forem consideradas sequências livres de palavras simples, obtêm-se resultados como :

teias,teia.N:fp

da,do.PREPXDET+Art+Def:fs

da,do.PREPXPRO+Dem:fs

lei,lei.N:fs

[4 unidades lexicais]

até,até.ADV

até,até.PREP

prova,prova.N:fs

prova,provar.V:P2s:P2's:P3s:Y2s

em,em.PREP

contrário,contrário.A:ms

contrário,contrário.N:ms

[7 unidades lexicais]

Estes resultados, além de serem desadequados, do ponto de vista da análise linguística dos textos, manteriam ambiguidades que podem ser imediatamente resolvidas nesta fase do processamento.