Margarita Correia
FLUL / ILTEC / SILEX, URA 382 du CNRS

As palavras derivadas: objectivos e modos de tratamento em bases de dados lexicais*

0. Introdução

Esta comunicação tem como principal objectivo dar a conhecer a descrição atribuída às palavras derivadas, pela equipa do ILTEC, no âmbito da sua participação no projecto europeu GENELEX (EU 524). A participação portuguesa consistiu basicamente na adaptação ao português deste modelo de base de dados lexicais e na sua validação por meio do tratamento de uma lista de cerca de 5000 unidades do léxico comum.

Num primeiro momento, apresentar-se-ão sumariamente os conceitos de *dicionário legível com máquina* e *tratável com máquina* e de *base de dados lexicais*. Num segundo momento, expor-se-á muito sumariamente o tratamento dado aos derivados no GENELEX português, sendo apresentados alguns exemplos particulares.

1 Dicionários legíveis com máquina, tratáveis por máquina e bases de dados lexicais

Um dos problemas que se colocam a quem inicia a sua investigação no domínio das descrições computacionais do léxico é o da indeterminação terminológica. As

* Agradeço à Paula Guerreiro (ILTEC) pelas frutuosas discussões de conceitos, bem como pela leitura de versões prévias desta comunicação

denominações usadas na bibliografia desta área não correspondem, em geral, aos mesmos conceitos (cf., por exemplo, BRISCOE, T. (1991), CALZOLARI, N. (1990), MARTIN W. & M. WOLTERING (1989), WILKS, Y et al. (1993)). A necessidade de estabelecimento de uma terminologia na área da lexicografia computacional é já referida em CORREIA, M. & P. GUERREIRO (1993c), onde, de resto, são apresentadas tipologias de dicionários em suporte informático, com base no *critério processual*, no *critério utilizador final* e, ainda, *no critério possibilidade de alteração de dados*. Porém, tal como acontece com os dicionários impressos, estas diferentes tipologias tendem a sobrepor-se em função da concepção e carcterísticas de cada objecto em particular. Em CORREIA, M. (a publicar) desenvolvem-se os conceitos a seguir apresentados, acrescentando-se o de base de conhecimento lexical (do inglês *lexical knowledge base*).

Porém, dado a discussão da terminologia não se enquadrar nos objectivos da presente comunicação, utilizarei como denominações básicas as de dicionário legível com máquina (do inglês *machine-readable dictionary*), dicionário tratável por máquina (de *machine-tractable dictionary*) e base de dados lexicais (de *lexical database*), cujas principais características passo a apresentar.

1.1 Dicionários legíveis com máquina

Os dicionários legíveis com máquina (de agora em diante DLMs) são dicionários coligidos por lexicógrafos e concebidos para uso humano, sendo apresentados em suporte informático. Destes dicionários são publicadas versões impressas e versões em

suporte informático armazenadas em bases de dados (seja em CD-ROM, seja em disquetes)[1]

A microstrutura destes dicionários é, no essencial, semelhante à dos dicionários impressos, embora o facto de serem armazenados em bases de dados contribua para potenciar toda uma rede de relações morfológicas, sintagmáticas, semânticas e paradigmáticas entre diferentes unidades lexicais, possibilitando o acesso à informação por outras vias que não apenas a entrada, perdendo a ordenação alfabética das entradas (único meio para a sua localização nos dicionários impressos) parcialmente a sua validade nos DLMs.

Estes dicionários não são susceptíveis de ser utilizados directamente em sistemas de processamento de linguagem natural (PLN), devido fundamentalmente a serem concebidos para uso humano, ao índice de formalização da informação ser bastante baixo e à descrição das unidades ser sempre feita em linguagem natural.[2]

1.2. Dicionários tratáveis por máquina

WILKs, Y *et al.* (1993) definem um dicionário tratável por máquina (DTM) como um DLM transformado, apresentando um formato que o torne apto a ser usado em sistemas de PLN (cf. p. 341). Esta aptidão resulta basicamente na descrição do conhecimento lexical num formalismo que o sistema possa facilmente reconhecer, traduzindo a informação que nos dicionários humanos surge em linguagem natural, bem como na explicitação de todo o conhecimento que nos dicionários para uso humano permanece

implícito na descrição (como sejam a informação sintáctica, a colocacional, a estilística, ou outras). Os DTMs são, à partida, apenas utilizáveis em sistemas de PLN (cf. CORREIA, M. (a publicar)).

1.3 Bases de dados lexicais

Uma base de dados lexicais (BDL) é uma estrutura computacional concebida de modo a ser capaz de suportar os mais variados tipos de conhecimento sobre cada unidade lexical, permitindo estabelecer conexões quer entre unidades lexicais distintas, quer entre características pertencentes a unidades lexicais distintas. Por outras palavras, a estruturação de uma BDL permitir-nos-á observar as unidades lexicais sob os mais variados prismas e aceder a elas das mais variadas formas possíveis. Numa BDL o léxico é entendido como uma complexa rede de relações (morfológicas, sintagmáticas, semânticas e paradigmáticas), onde o conhecimento sobre uma unidade lexical é composto a vários níveis ou camadas.

A informação carregada numa BDL é primeiramente destinada a ser utilizada em sistemas de PLN. Porém, devido ao modo como essa informação é codificada, ela pode também ser utilizada por humanos, quer através de produtos impressos, quer através de produtos em suporte informático.

As BDLs são entendidas em Lexicografia Computacional como grandes repositórios de informação lexical, estruturada de tal forma que as torne passíveis de servirem de base a diferentes sistemas de PLN. Assim, um dos princípios fundamentais no momento da sua

concepção é o princípio da 'reutilizabilidade' (do inglês *reusability*) dos objectos conseguidos, preocupação que é determinada sobretudo pelos elevados custos materiais e temporais que a criação de qualquer destes objectos implica. A coerência e a sistematicidade das descrições obtidas são factores fundamentais para assegurar a reutilizabilidade das BDLs (cf. CORREIA, M. (a publicar)).

2. O tratamento dado aos derivados no GENELEX português

O objectivo do projecto GENELEX é a criação de BDLs monolingues de várias línguas europeias, informatizadas segundo uma modelização comum, que poderá permitir, no futuro, a sua utilização como base para sistemas de PLN envolvendo uma, duas ou mais línguas.[3]

A BDL GENELEX tem uma estrutura tripartida, que por falta de espaço não poderei descrever, comportando uma camada morfológica, uma camada sintáctica e uma camada semântica.[4] Foi ao nível da camada morfológica que se realizou a descrição das palavras derivadas.[5]

2.1 Breve descrição da camada morfológica do GENELEX

No GENELEX, cada lema corresponde a uma UM (unidade morfológica). As UMs são caracterizadas em função da sua autonomia prevendo o modelo os seguintes tipos de unidades:

- as simples (UM_S), que podem ser autónomas ou não autónomas (ex. **cavalitas** que apenas ocorre na locução **às cavalitas**); as palavras derivadas são classificadas como UM_Ss;
- as compostas (UM_C), autónomas, permitindo descrever compostos sintagmáticos e compostos por temas;
- as aglutinadas (UM_AGG), autónomas, permitindo descrever as contracções de preposição e determinante, frequentes em português;
- as unidades morfológicas afixais (UM-AFF), obviamente não-autónomas, cuja descrição pode ser observada na figura 4.[6]

2 2. A descrição dos derivados no GENELEX português

O tratamento informático de palavras derivadas pode ter basicamente dois objectivos distintos:

i. uma descrição passiva: da estrutura interna das palavras que fazem parte do dicionário de base, tal como é feita frequentemente nos dicionários impressos;
ii. uma descrição activa: de regras de formação de palavras (RFPs), visando tornar o sistema apto a gerar e/ou reconhecer palavras que não constem do dicionário de base.

A descrição atribuida aos derivados no GENELEX é feita de acordo com i., de modo que, para cada palavra derivada, é apresentada uma estrutura interna, dando conta da base e do(s) afixo(s) envolvido(s) no processo derivacional, como pode verificar-se nas figuras 3 e 5. Porém, na medida em que os afixos são alvo de uma descrição pormenorizada dando conta, designadamente, da categoria das bases seleccionadas, bem

como da categoria dos derivados, afigura-se possível adaptar o modelo de modo a permitir-lhe o comportamento apresentado em ii. (a figura 1 apresenta esquematicamente o tratamento da alomorfia do prefixo *in-*, enquanto que a figura 4 apresenta o tipo de informação associada a um afixo).

No quadro do GENELEX, dois princípios básicos são considerados: por um lado, apenas é possível descrever derivados cuja base seja uma palavra autónoma[7] e, por outro, todas as unidades apresentadas como bases devem ser tratadas como entradas do dicionário. Este segundo princípio levou a um alargamento do número de entradas inicialmente previstas para o GENELEX português.

No GENELEX português, a descrição dos derivados foi feita, sempre que possível, seguindo de perto o modelo morfológico associativo e estratificado proposto para o francês por CORBIN, D (1987 e 1991) e a investigação desenvolvida para a nossa língua por RIO-TORTO, G. M. (1993). Porém, tal não foi sempre possível, dada, por exemplo, a limitação imposta de apenas atribuir descrição a palavras cujas bases fossem autónomas e atestadas.

A coerência na descrição foi um dos princípios considerados ao longo do trabalho: assim, procurou-se descrever os afixos de modo a que a descrição prevista pudesse dar conta do maior número possível de casos, evitando sempre que possível soluções 'ad hoc' para casos pontuais. Essa coerência tem a ver não só com a conformidade ao modelo de descrição escolhido, mas foi também entendida como um aspecto indispensável para conseguir uma descrição válida para eventuais reutilizações da

informação armazenada. A coerência é também condição 'sine qua non' para uma possível implementação automática ou semi-automática de informação lexical com base em critérios de índole derivacional. Porém, por vezes a coerência teve que ser sacrificada aos dados em causa, uma vez que o trabalho foi fundamentalmente lexicográfico (dando conta apenas de formas atestadas, isto é, de um vocabulário específico) e não lexicológico (no sentido de dar conta das possibilidades do léxico da língua).

Os derivados foram ainda descritos assumindo o princípio geral de que em cada processo derivacional se junta à base apenas um afixo de cada vez. Este princípio não foi respeitado apenas nos casos de parassíntese[8] e no de derivados que apresentam dois sufixos em sequência, como é o caso de muitos avaliativos, nos quais a forma correspondente à base mais o primeiro afixo corresponde a uma forma não atestada em português (é o caso adjectivo *brincalhão* cuja descrição é apresentada na figura 5)[9]

A descrição feita é basicamente sincrónica. Como é sabido, existem casos em que a forma do derivado é diferente da que seria previsível pela aplicação das RFPs no momento actual, devido por exemplo ao facto de terem sido construídas ainda em latim - ex.. **conceptual**$_{Adj}$ (do latim *conceptuale-*) por oposição a **conceitual**$_{Adj}$ (de **conceito**$_{N}$). As palavras do tipo de **conceptual**$_{Adj}$ foram descritas atribuindo um radical combinatório à base correspondente à forma que esta assume no seio do derivado:

conceptual|mfgn034|1|base|**conceito**|conceito|**conceptu**|2|suff|al1|al|al|||||[10]

Descrição semelhante foi atribuída aos derivados que são construídos sobre variantes da forma considerada como lema. É o caso da palavra **bonecreiro**$_N$ (único exemplo deste tipo de palavras no dicionário de base do GENELEX português), cuja descrição é apresentada esquematicamente na figura 3. É ainda o caso dos derivados que apresentam fenómenos de haplologia, como **feminismo**$_N$:

feminismo|mfgn040|1|base|**feminino**|feminino|**femin**|2|suff|ismo|ismo|ismo|||||

Os afixos são descritos no GENELEX de acordo com os parâmetros apresentados na figura 4. A sua alomorfia é também passível de ser descrita neste modelo sob a forma de radicais combinatórios que são indexados à forma canónica do afixo, o que é visível na figura 1, a propósito do prefixo *in-*. A descrição dos afixos foi feita basicamente considerando o princípio de que a cada afixo corresponde a selecção de uma categoria de bases e a produção de uma categoria de derivados, como pode ser verificado a seguir:

ada1|suff|N|N|fem|ada,056;zada,056;lada,056|
ada2|suff|V|N|fem|ada,056|

em que **-ada1** permite formar nomes denominais (do tipo **ninhada**, **pazada** e **paulada**) e **-ada2** permite formar nomes deverbais (do tipo **entrada**).

Porém, em alguns casos este princípio não foi mantido, designadamente quando uma das operações categoriais se revelou pouco produtiva (pelo menos no seio do dicionário de base considerado), como acontece com **-eiro** que intervém prioritariamente na formação

de adjectivos denominais (como **bacalhoeiro**) e menos frequentemente na de nomes deverbais (como **herdeiro**)

```
eiro:suff|N ADJ||eiro,001, eir,
            V ADJ:eiro,001,
```

É o caso também de derivados cuja categoria previsível não é uma forma atestada (por razões extralinguísticas como a história, a organização do mundo, etc.), mas apenas a já resultante de um processo de conversão (ex. °**pedinte**$_{Adj}$) [11]

Na generalidade, o modelo GENELEX revelou-se bastante eficiente para a descrição dos diferentes tipos de derivação existentes em português: a prefixação, a sufixação, a parassíntese, a conversão e a derivação regressiva. Se em relação aos três primeiros tipos não se colocaram problemas de maior, os dois últimos merecem uma menção especial. A conversão foi tratada como um tipo de derivação no qual apenas intervém um elemento, ao qual é atribuído o estatuto de base (na figura 2 pode observar-se o modo de tratamento de uma palavra resultante de conversão, **impermeável**$_N$); porém, no caso das conversões em que base e derivado apresentam o mesmo padrão flexional (como **dentista**$_{ADJ/N}$), o sistema não tem meios para distinguir qual a base e qual o derivado, pelo que, se o alargamento do dicionário GENELEX for empreendido, este problema terá que ser ultrapassado pela indicação da categoria de cada uma das palavras. A derivação regressiva foi tratada como um tipo de derivação onde também apenas intervém um componente, a essa base foi atribuído um radical combinatório de forma correspondente à do derivado regressivo, como pode observar-se no exemplo abaixo:

fuga|mfgn056i1|base|fugir|fugir|fuga|||||||||

As palavras construidas com afixos avaliativos iniciados por **z** (como **-zinh(o/a)** em **cãozinho**) foram aquelas cuja estrutura interna não foi possível descrever. Três hipóteses se afiguraram: tratá-las como derivados (mas o modelo apenas admite a marcação da flexão na periferia do derivado e não no seu interior); tratá-las como compostos (mas se descrevêssemos o primeiro componente como invariável a forma flexionada gerada seria ***cãozinhos**, enquanto que se o descrevêssemos como variável obteríamos ***cãeszinhos**); tratá-las como palavras sem estrutura interna. Embora esta solução fosse a menos adequada, ela foi a única viável no quadro deste modelo. Esta é, portanto, uma das questões a resolver se for empreendida uma a descrição de uma larga porção do léxico da língua portuguesa no quadro do modelo GENELEX.

Bibliografia

BOGURAEV, Bran & Ted BRISCOE (1989), "Introduction", in BOGURAEV, Bran & BRISCOE, Ted (eds.), *Computational Lexicography for Natural Language Processing*, Londres/N. Iorque, pp. 1-40.

BRISCOE, Ted (1991), "Lexical Issues in Natural Language Processing", in KLEIN, E. & F. VELTMAN (eds.), *Natural Language and Speech*, Spinger-Verlag, pp. 39-68

CALZOLARI, Nicoletta (1990), "Structure and access in an automated lexicon and related issues", in *Linguistica Computazionale vol. VI - Computational Lexicology*

and Lexicography: Special Issue dedicated to Bernard Quemada, Pisa, Giardini Editori e Stampatori, pp 139-161

CALZOLARI, Nicoletta (1991), "Acquiring and representing semantic information in a Lexical Knowledge Base", in PUSTEJOVSKY, James & Sabine BERGLER (ed.), *Lexical Semantics and Knowledge Representation*, Association for Computational Linguistics, pp 188-197

CAMEIRA, Célia, M. CORREIA & Paula GUERREIRO (1994), *Final Report on the Morphological Specifications for Portuguese Computational Lexica*, Lisboa, ILTEC (disponível)

Consortium GENELEX (1993), *Couche morphologique*, Version 3.0, Paris (disponível)

CORBIN, Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vols., Tubinga, Max Niemeyer Verlag (2ª ed., Villeneuve d'Ascq, Presses Universitaires de Lille, 1991)

CORBIN, Danielle (1991), «Introduction - La formation des mots: structures et interpretations», in *Lexique 10*, Villeneuve d'Ascq, Presses Universitaires de Lille

CORREIA, Margarita & Paula GUERREIRO (1993a), "GENELEX: um modelo para a construção de bases de dados lexicais do português", *Actas do 1º Encontro de Processamento da Língua Portuguesa - Escrita e Falada*, Lisboa, INESC / UNINOVA / CLUL, pp. 147-150

CORREIA, Margarita & Paula GUERREIRO (1993b), "Modèle morphologique du portugais", comunicação apresentada ao Club Utilisateurs GENELEX, Paris, IBM-France, Abril (disponível).

CORREIA, Margarita & Paula GUERREIRO (1993c), "Bases de dados lexicais", *in Actas do Seminario, «Engenharia Linguistica»*, Lisboa, Working Papers do ILTEC (a publicar pelas Edições Colibri).

CORREIA, Margarita (a publicar), «Bases digitais lexicais na União Europeia», *in Actas do Simpósio de Lexicologia, Lexicografia e Terminologia* (UNESP, Araraquara, 25 a 27 de Outubro de 1994).

COSTA, J. Almeida & A. Sampaio e MELO (1994), *Dicionário da Língua Portuguesa*, 7ª ed. revista e ampliada, Porto, Porto Editora.

Dicionário Aurélio Eletrônico (1993) (baseado em FERREIRA, Aurélio Buarque da Holanda (1986), Rio de Janeiro, Ed. Nova Fronteira.

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

MARTIN, Willy & Marc WOLTERING (1989), *Basic Issues in Computational Linguistics* (relatório redigido sob a responsabilidade de Bernard AL), Utreque, Van Dale Lexicografie.

WILKS, Yorick, Dan FASS , Cheng-Ming GUO, James McDONALD, Tony PLATE & Brian SLATOR (1993), "Providing Machine Tractable Dictionary Tools", in PUSTEJOVSKY, James (ed.), *Semantics and the Lexicon*, Dordrecht/Boston/Londres, Kluwer Academic Publishers, pp. 341-401.

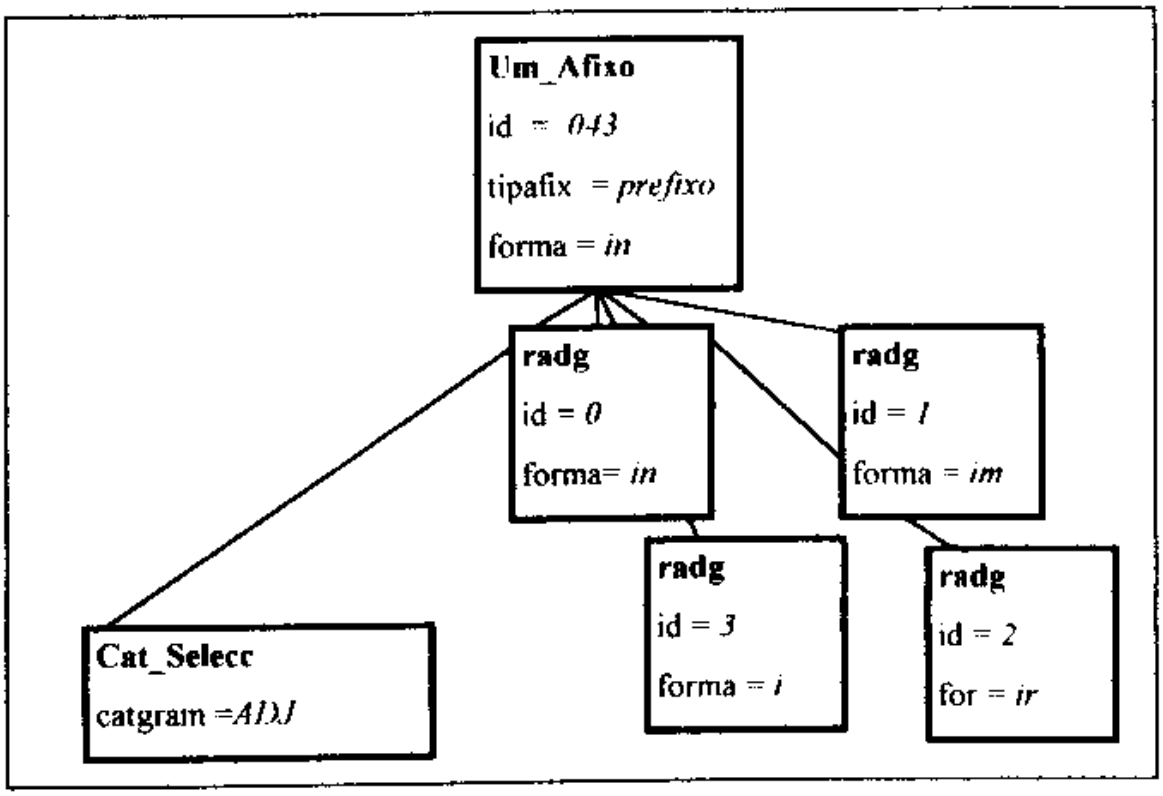


Figura 1: (em cima): representação do prefixo *in-*.

Figura 2: (em baixo): a representação da conversão de *impermeável*.

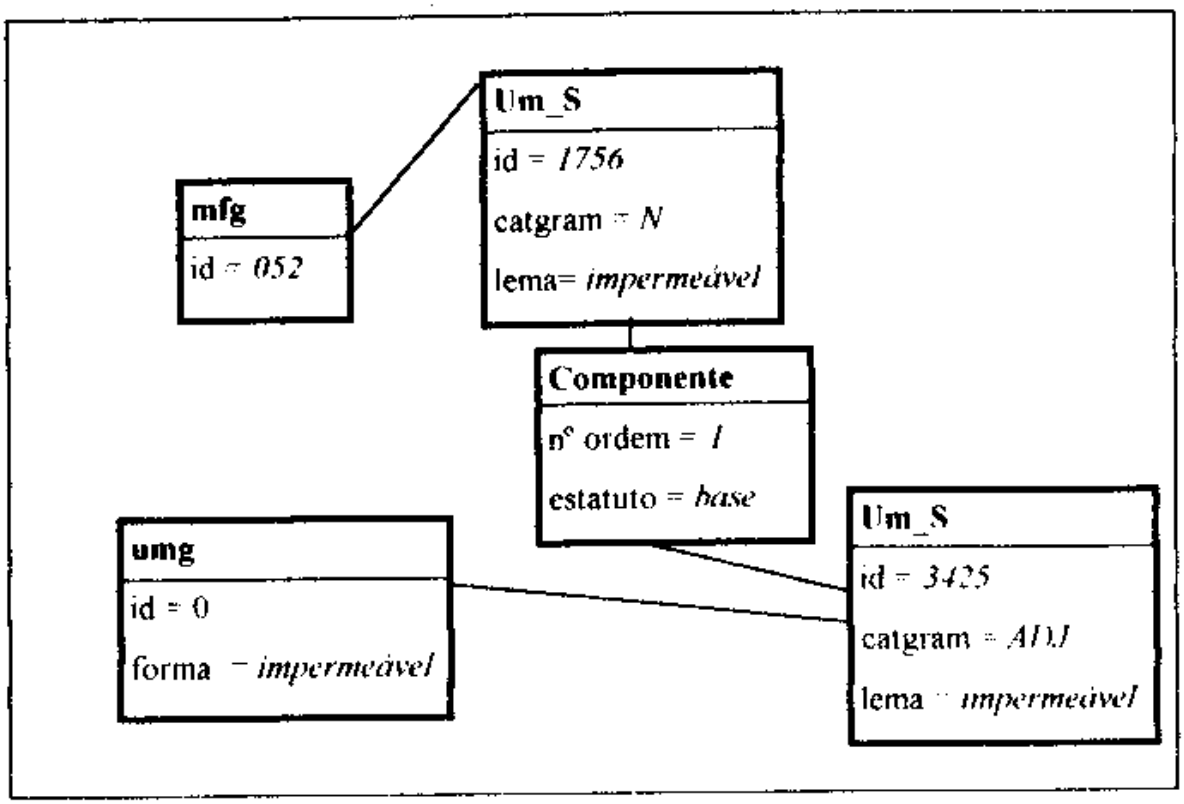
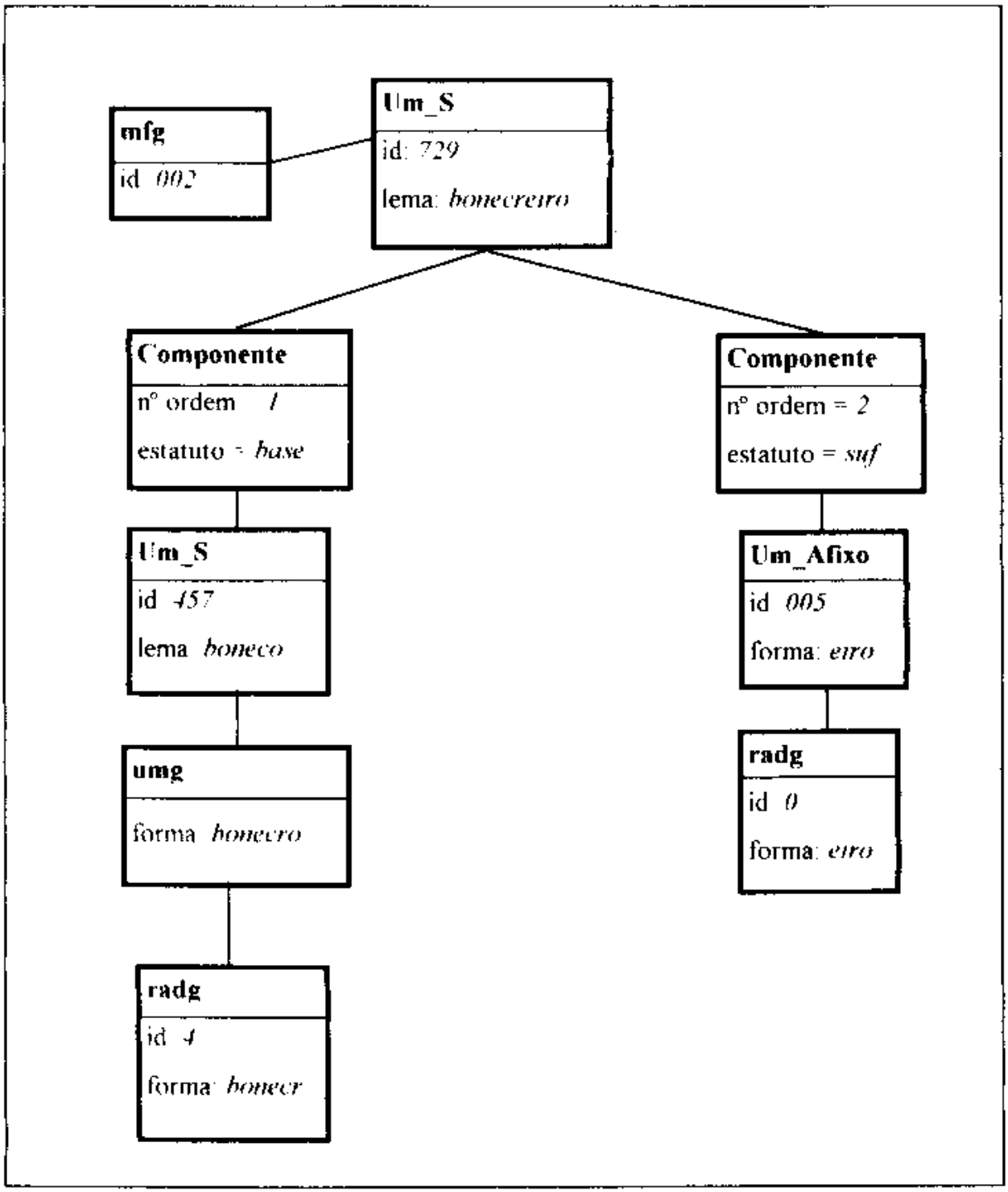


Figura 3: a derivação de *bonecreiro*.



Umaff: a forma representativa do afixo Ex *-eiro*

Typaff: o tipo de afixo (prefixo ou sufixo). ex *sufixo*

CatGram_Select: a categoria de palvras seleccionadas como bases de derivação Ex.. *N (pastel / barba) / V (herdar)*

CatGram_Result: a categoria resultante do processo derivacional envolvendo o afixo.ex *ADJ (pasteleiro), N (barbeiro herdeiro / alheira)*.

GenreN_Result: se o derivado for um substantivo, o seu género na forma lematizada Ex . *masc / fem*.

umgaff: as várias formas que o afixo pode assumir nos vários derivados (alomorfes ou variantes condicionadas historicamente) Ex : *-eiro / -eir- (toleirão)*

mfg: o modo de flexão gráfica, traduzido num índice, que os derivados apresentando dado afixo podem apresentar. Ex.: *-eiro,001 040; -eira,056;-eir-,Ø.*

Figura 4: Tipos de informação associados a cada Um_Afixo: exemplificação com o sufixo *-eiro*

Informações directamente relacionadas com o derivado:

Um: o lema a ser descrito. Ex.. *brincalhão*

mfg: o modo de flexão gráfica (expresso através de um índice. Ex.: *014*

Informações directamente relacionadas com cada um dos seus componentes

ordre_linéaire: a posição que o componente ocupa na estrutura do derivado (1, 2 ou 3).

statut: o tipo de componente descrito Ex.: *1 - base ; 2 -sufixo ; 3 - sufixo*

um: o lema da base ou a forma canónica do afixo. Ex.: *brincar -alho -ão.*

umg: a forma gráfica seleccionada no processo derivacional (geralmente igual à de um).

radg: o radical combinatório ou a forma truncada assumida pelo componente no processo derivacional. Ex.: *brinc -alh- -ão*

Figura 5: Tipos de informação associados a cada Um_S (derivada) exemplificação com *brincalhão*$_{ADJ}$

[1] Para o português do Brasil, foi publicada a versão informatizada do 2ª edição *do Novo Dicionário da Língua Portuguesa*, o *Aurélio Eletrônico* (em diquetes para ambiente DOS ou Windows). Embora a base da sua constituição tenha sido a versão impressa do dicionário, a forma como foi concebida a base de dados que o suporta permitiu um considerável aumento dos meios de acesso à informação, tornando explícita muita informação que na versão impressa se encontrava apenas implícita. Para o português europeu, está anunciada a publicação da versão informatizada da 7ª edição do *Dicionário da Língua Portuguesa*, da Porto Editora.

[2] Cf. CORREIA, M. & GUERREIRO, P. (1993c) para uma exposição mais detalhada das virtudes da apresentação dos dicionários em suporte informático, bem como das consequências para a descrição lexical que derivam do facto de o produto ser concebido para utilizadores humanos.

[3] Fazem do Consórcio GENELEX, além do ILTEC, a GSI-Erli, a IBM-France, a SEMA-Group, o ASSTRIL-LADL (França), o Consorzio Lexicon Ricerche e a SERV EDI e SOGESS, srl (Itália).

[4] Para uma descrição sucinta do modelo, remeto o leitor para CORREIA, M. & P. GUERREIRO (1993a). Para uma descrição mais detalhada da estrutura modelar GENELEX, remeto o leitor para os relatórios disponíveis do Consórcio, bem como para os Relatórios Finais de Especificações (morfológicas, sintácticas e semânticas) da equipa do ILTEC.

[5] A estrutura modelar GENELEX permite que a descrição de derivados e compostos possa ser feita ao nível da camada sintáctica, como se se tratasse de unidades sintácticas.

[6] Para descrições mais pormenorizadas da estrutura da camada morfológica e das especificações definidas para o português, remeto o leitor para CORREIA, M. & P. GUERREIRO (1993b) e para CAMEIRA, C., M. CORREIA & P. GUERREIRO (1994).

[7] Esta prerrogativa acaba por limitar o modelo, dado que impossibilita a descrição de unidades cuja base é não-autónoma (ex.: **eléctrico**$_{Adj}$); porém, ela é facilmente ultrapassável.

[8] Ex.: a estrutura do verbo **amaciar** foi descrita do seguinte modo: **||a| |maci(o)|**$_{Adj}$ **|ar||**.

[9] A forma "**brincalho**$_{Adj}$ é uma palavra possível não atestada. O símbolo "?" precede uma palavra possível não atestada.
Aos sufixos que, como **-alh(o)**, podem ocorrer numa posição interna ao derivado, foi atribuído um radical combinatório sem marca flexional (no caso, **-alh-**). Cf. figura 5.

[10] A descrição é apresentada em formato *delimited text file*, o utilizado nos ficheiros com que foi carregada a base de dados. O valor das convenções aqui incluídas pode encontrar-se na figura 5.

[11] É, porém, minha convicção que, se pudesse ter tido acesso a um extenso corpus de dados textuais informatizado, teria certamente encontrado atestações que permitissem confirmar as intuições relativas aos processos derivacionais.