Documentação dos recursos linguísticos subjacentes ao AnELL

Isabel Marcelino
27 de Setembro de 2005

O AnELL (**An**otador **E**lectrónico **L**abEL-**L**inguateca) é um serviço online de anotação linguística de corpora cujo principal objectivo é produzir corpora anotados com informação gramatical. Foi desenvolvido em parceria entre o **LabEL**, que elaborou os recursos linguísticos, e a **Linguateca**, que implementou a interface web. Foi apresentado em (Mota&Moura, 2003). Encontra-se neste momento instalado na máquina acdc da Linguateca, acessível através do endereço www.linguateca.pt/ANELL.

O AnELL utiliza léxicos computacionais e gramáticas desenvolvidas pelo LabEL. Estes recursos são formalizados e aplicados aos textos usando técnicas de estados finitos (finite state transduces, FST), cada vez mais reconhecidas no processamento da linguagem natural.
Estes recursos são utilizados pelo INTEX (Silberztein, 1993), e é com estes recursos que a análise linguística das palavras de um texto se torna possível com o AnELL.
Antes, os dicionários eram de forma DIGRAMA (Eleutério et al., 1995), e foram transformados para forma INTEX (Mota, 2004).

Os recursos linguísticos subjacentes ao AnELL dividam-se em três grupos: (i) dicionários (do tipo DELACF) e transdutores de pré-processamento, (ii) dicionários (do tipo DELAF e DELACF) e transdutores de análise de unidades lexicais simples e compostas, e (iii) transdutores do módulo de desambiguação (DISAMB).

Recursos linguísticos utilizados para o pré-processamento de textos:

(Estes recursos encontram-se em: /usr/local/INTEX/Prv/Portuguese)

Recursos utilizados:

Alphabet (147 bytes)
(Ficheiro que contém todos os caracteres que fazem parte do alfabeto (maiúsculas e minúsculas com e sem acentos). É com base neste ficheiro que a atomização é feita.)

Norm.fst (10,4 KB)
(Transdutor que faz a descontracção e análise linguística das contracções)

Replace.fst (50,2 KB)
(Transdutor que analisa e etiqueta números e endereços electrónicos.)

Sentence.fst (6 KB)
(Transdutor que segmenta o texto em unidades menores que o parágrafo, as quais correspondem, de uma forma geral, a frases. É inserida a etiqueta **{S}** no início de cada frase identificada no texto)

Tag-All.fst (81,15 KB)
(Chama os transdutores Tag-ELLE.fst e Tag-Verbo-Pro.fst)

Tag-ELLE.fst (78 KB)
(Transdutor que analisa e etiqueta as entidades mencionadas)

Tag-Verbo-Pro.fst (2,33 KB)
(Transdutor que delimita as combinações verbo-clítico, que serão mais tarde analisadas pelo grafo de resolução de ambiguidades)

UCompounds.bin (165 KB)
UCompounds.inf (16,2 KB)
(Dicionário de compostos não ambíguos. Para já é constituído por nomes compostos que incluem hífens e por advérbios compostos)

Descrição do pré-processamento:

O pré-processamento do texto consiste em transformá-lo no formato INTEX. Divide-se em três fases:

(1) Aplicação de um transdutor em modo de inserção ("Merge") que insere a etiqueta de delimitação de unidades de texto '{S}'. Este transdutor (***Sentence.fst***) identifica frases em textos normais (jornalísticos, romances, etc., em oposição a poemas, por exemplo) com base em regras tipográficas.

(2) Aplicação de um dicionário de palavras compostas não ambíguas (***UCompounds.bin***) para identificar e etiquetar palavras compostas que de outra forma seriam analisadas como combinações livres de palavras simples.

Compostos que contenham palavras simples que não existem isoladamente (apesar de, etc.) e hífens (chapéu-de-chuva, a bel-prazer) devem fazer parte deste dicionário.

Neste momento o ficheiro ***UCompounds.bin*** contém 6500 palavras compostas que integram hífens.

(3) Aplicação de um transdutor em modo de substituição ("***Replace***") para normalizar sequências desviantes. Poderá ser usado para analisar contracções não ambíguas (nele → em ele). No entanto, para tratamento das contracções aconselha-se a aplicação do fst ***Norm.fst*** no momento da construção do FST do texto ou, em alternativa, a aplicação do FST ***Contracções.fst*** no momento da aplicação dos recursos lexicais (mas actualmente, o AnELL não está a usar esse último transdutor, apesar de poder vir a fazê-lo.).

Está a ser usado para etiquetar números e endereços URL.

Os dicionários DELACF e DELAF:

Os dicionários electrónicos são ficheiros no formato DELAF (palavras simples) ou DELACF (palavras compostas). Os dicionários descritos em ficheiros «.dic» são

editáveis (pode-se visualizar o seu conteúdo e modificá-lo); os dicionários descritos em ficheiros «.bin» são compactados e o seu conteúdo não pode ser modificado.

Os dicionários DELACF:

(Estes recursos encontram-se em: /usr/local/INTEX/Prv/Portuguese/Delacf)

Os dicionários de palavras compostas flexionadas (DELACF) são análogos aos dicionários DELAF de palavras simples; a única diferença é que as entradas lexicais (o texto que aparece ao princípio da linha e antes da vírgula no dicionário) e/ou os lemas (o texto entre a virgula e o ponto) podem conter separadores.

Recursos presentes no ficheiro DELACF:

DCard-.fst (3,18 KB)
(Identifica números cardinais compostos de "vinte e um" a "novecentos e noventa e nove mil novecentos e noventa nove", associando-lhes o valor numérico.
Garante a concordância em género e número.
Exemplos após a aplicação do transdutor:
vinte e uma,vinte e uma.DET+Num+VAL=21:Cfp
vinte e três,vinte e três.DET+Num+VAL=23:Cmp:Cfp
A versão utilizada pelo AnELL foi revista em Março de 2004)

DelacAdv_V2.1.bin (51,1 KB)
DelacAdv_V2.1.inf (304 bytes)
(Dicionário de advérbios compostos (2 194 entradas)
Recolhidos e formalizados por Elisabete Ranchhod.
Exemplos:
a par e passo,a par e passo.ADV+PCONJ
a prazo,a prazo.ADV+PC
a toda a hora,a toda a hora.ADV+PDETC
com falinhas mansas,com falinhas mansas.ADV+PCA
de bom grado,de bom grado.ADV+PAC
A versão utilizada pelo AnELL foi revista em Março de 2004)

DelacConj.dic (952 bytes)
(Dicionário de conjunções compostas (54 entradas))
Exemplos:
à medida que,.CONJ
bem como,.CONJ
cada vez que,.CONJ
como se,.CONJ)

DelacfAdj_V1.bin (8,64 KB)
DelacfAdj_V1.inf (1,25 KB)
(Dicionário de adjectivos compostos com 361 entradas, flexionadas a partir de um DELAC de adjectivos compostos com 96 entradas)
Exemplos:
amiga da onça,amigo da onça.A+ADN+Pd:fs

amigas da onça,amigo da onça.A+ADN+Pd:fp
amigo da onça,amigo da onça.A+ADN+Pd:ms
amigos da onça,amigo da onça.A+ADN+Pd:mp
azuis-bebé,azul-bebé.A+AN+Pco:mp:fp
azul-bebé,azul-bebé.A+AN+Pco:ms:fs
bem falante,bem falante.A+ADVA+Pd:ms:fs
bem falantes,bem falante.A+ADVA+Pd:mp:fp

DelacfEmp_V1.dic (88 bytes)
(Dicionário de palavras estrangeiras compostas que são usadas em português (5 entradas))

Exemplos:

e-mail,.N:ms
on-line,.ADV

DelacfNomes_V3.bin (1,25 MB)
DelacfNomes_V3.inf (54,7 KB)
(Dicionário de nomes compostos – inclui os não ambíguos, ou seja, os que têm hífen) (13 465 + 25 259/76 400)
13 465 compostos foram formalizados e flexionados no sistema DIGRAMA
25 259 compostos foram formalizados e flexionados por um programa compatível com o sistema INTEX implementado por Cristina Mota
Recolhidos e formalizados conjuntamente por:
Pedro Cordeiro, Ana Oliveira, Jorge Baptista e Paula Carvalho

Exemplos:

amarelo(N291) vivo(A291),N+NA+Cor
amargo(N201) de boca,N+NDN
ambição(N308) desmedida(A301),N+NA+Pred+Vsup=ter
abaixo-assinados,abaixo-assinado.N+ADVA:mp
abalo de terra,abalo de terra.N+NDN:ms
abalo nervoso,abalo nervoso.N+NA:ms
abalo sísmico,abalo sísmico.N+NA:ms

Incluem também:
Os termos de seguros (nd/741), marcados com +Seg, que foram recolhidos e formalizados por Vítor Franca.

Exemplo:

abono das fazendas seguradas,abono das fazendas seguradas.N+Seg:ms

Os nomes de cargos e funções (714/2 410), marcados com +Func, que foram recolhidos e formalizados por Tânia Veríssimo.

Exemplo:

administrador-adjunto,administrador-adjunto.N+NA+Func:ms

A versão utilizada pelo AnELL foi revista em Março de 2004)

DelacfNpr_V3.bin (257 KB)
DelacfNpr_V3.inf (5 KB)
(Dicionário de nomes próprios compostos, Versão 3, com 4.858 entradas
Foram acrescentados alguns traços semânticos.

Exemplos:

Combatentes da Grande Guerra,.N+Hum+Grupo

Feira Internacional de Lisboa,.N+Evento+Org:fs
Os Rouxinóis,.N+Grupo+Mus
Última revisão (Isabel Marcelino): 20 de Setembro de 2005)

DelacPrep.dic (661 bytes)
(Dicionário de preposições compostas (39 entradas))
Exemplos:
em vez de,.PREP
graças a,.PREP
junto a,.PREP)
Os dicionários de preposições e o grafo das contracções são necessários para que as entidades mencionadas que comportam preposições e contracções sejam correctamente anotadas.

DOrd-.fst (3,21 KB)
(Identifica números ordinais compostos de "décimo primeiro" a "milésimo nongentésimo nonagésimo nono", associando-lhes o valor numérico.
Garante a concordância em género e número.
Exemplos após a aplicação do transdutor:
décima primeira,décima primeira.DET+Num+VAL=11:Ofs
centésimos vígésimos segundos,centésimos vigésimos segundos.DET+Num+VAL=122:Omp
A versão utilizada pelo AnELL foi revista em Março de 2004)

linguateca_adicC.dic (112 bytes)
(Dicionário que conterá novas entradas compostas identificadas pelo revisor e que eventualmente serão adicionadas ao respectivo dicionário de palavras composta.)
(com 3 entradas)

linguateca_naoambiguos.dic (30 bytes)
(Dicionário de novos compostos não ambíguos.)
(com 1 entrada)

linguateca_NprC.dic (48 bytes)
(Dicionário que conterá os novos nomes próprios compostos.)
(com 2 entradas)

linguateca_SiglasC.dic (153 bytes)
(Dicionário que conterá novas siglas compostas e também as bases lexicais compostas de siglas simples.)
(com 3 entradas)

linguateca_substC.dic (63 bytes)
(Dicionário que conterá entradas preferenciais compostas que substituirão as que já existem.)
(com 2 entradas)

Siglacf-d_V1.bin (586 bytes)
Siglacf-d_V1.inf (103 KB)
(Dicionário de desenvolvimentos compostos de siglas e acrónimos.

Recolhidos e formalizados por Paulo Moura
Exemplo:
A-Adenosil Metionina,sam.N+DSig:ms
Associação Moçambicana de Bancos,amb.N+DSig+HumCol:fs
Associação Portuguesa de Astrónomos Amadores,apaa.N+DSig+HumCol:fs
Conjuntamente com os dicionários:
Siglaf.bin
Siglaf-d.dic
Siglacf.dic
Siglacf-dE.bin
Forma o módulo de siglas e acrónimos.
A versão utilizada pelo AnELL foi revista em Março de 2004)

Siglacf-dE_V1.bin (63,8 KB)
Siglacf-dE_V1.inf (9,55 KB)
(Dicionário de desenvolvimentos compostos de siglas e acrónimos estrangeiros.
Recolhidos e formalizados por Paulo Moura
Exemplo:
Aides Nationales - Échanges de Données,ain-ed.N+DSigE:fp
Conjuntamente com os dicionários:
Siglaf.bin
Siglaf-d.dic
Siglacf.dic
Siglacf-d.bin
Forma o módulo de siglas e acrónimos.
A versão utilizada pelo AnELL foi revista em Março de 2004)

Siglacf_V1.dic (1,52 KB)
(Dicionário de siglas e acrónimos compostos (e siglas estrangeiras) com 56 entradas)
Recolhidos e formalizados por Paulo Moura
Exemplo:
ACA-M,aca-m.N+Sig+HumCol:fs
Conjuntamente com os dicionários:
Siglaf.bin
Siglaf-d.dic
Siglacf-d.bin
Siglacf-dE.bin
Forma o módulo de siglas e acrónimos
A versão utilizada pelo AnELL foi revista em Março de 2004)

Os dicionários DELAF:

(Estes recursos encontram-se em: /usr/local/INTEX/Prv/Portuguese/Delaf)

Os dicionários DELAF são listas de palavras simples flexionadas às quais estão associadas o seu próprio lema e também a sua categoria morfo-sintáctica.
As informações codificadas nos dicionários DELAF são de dois tipos: informações sintatico-semânticas e informações de flexão.

Recursos presentes no ficheiro DELAF:

Delaf_V3.bin (0,98 MB)
Delaf_V3.inf (397 KB)
(Dicionário geral de palavras simples, contém 940 380 entradas.
Corresponde a uma versão do dicionário em formato DIGRAMA, revista e aumentada por Cristina Mota aquando da sua conversão para o formato Intex.
Este dicionário é gerado pelo sistema INTEX a partir do dicionário de lemas DELAS (105 770 entradas) e com base em grafos de flexão.
A versão utilizada pelo AnELL foi revista em Março de 2004)

Exemplos:

abade,abade.N:ms
abafar,abafar.V+t+z1:R[cq]:U1s[i]:U2"s[i]:U3s[i]:W[icq]:V1s[icq]:V2"s[icq]:V3s[icq]
abafem,abafar.V+t+z1:S2"p[i]:S3p[i]:Y2"p[icqn]
abaixo,abaixo.ADV+z1
abaixo,abaixo.INTERJ+z1
abalada,abalado.A+Pd+z1:fs
abalado,abalado.A+Pd+z1:ms
abalado,abalar.V+t+z1:K[i]
abalam,abalar.V+t+z1:P2"p[icqn]:P3p[icqn]
abalei,abalar.V+t+z1:J1s[icqo]
abalizado,abalizado.A+Pd:ms
abalizado,abalizar.V+t+z1:K[i]
abalo,abalar.V+t+z1:P1s[icqo]
abalo,abalo.N+z1:ms
abalos,abalo.N+z1:mp
abalou,abalar.V+t+z1:J2"s[icqo]:J3s[icqo]

DelafEmp_V1.dic (756 bytes)
(Dicionário de palavras estrangeiras que são usadas em português (41 entradas))

Exemplos:

alors,alors.ADV
assez,assez.ADV
baby,baby.XC)

DelafNpr_V3-.bin (70 KB)
DelafNpr_V3-.inf (4 KB)
(Dicionário de nomes próprios simples, Versão 3, com 6.582 entradas
Foram acrescentados alguns traços semânticos.

Exemplos:

Amazonas,.N+Geo:ms
Árabes,.N+Hum+Grupo:mp
Congresso,.N+Evento+Org:ms
Doors,Doors.N+Grupo+Mus:ms

Última revisão (Isabel Marcelino): 20 de Setembro de 2005)

DelafTmp_V1.dic (1,36 KB)
(Dicionário de neologismos que talvez venham a ser integrados no dicionário geral (42 entradas)).

Exemplos:

nogueiristas,nogueirista.N:mp:fp
portunhol,portunhol.N:ms
semimilitarizada,semimilitarizado.A:fs)

Filtro-.dic (4,49 KB)
(Dicionário que é aplicado antes de todos os outros. As análises que contém são preferenciais, ou seja, se uma palavra for analisada por este dicionário já não é analisada por nenhum outro (de palavras simples).
(168 entradas))

Exemplos:

a,a.PREP
a,me.PRO+Pes+A+z1:2"fs[o]:3fs[o]
a,o.DET+Art+Def+z1:fs
a,o.PRO+Dem+z1:fs
ainda,ainda.ADV+z1)

linguateca_adic.dic (101 bytes)
(Dicionário que conterá novas entradas identificadas pelo revisor e que eventualmente serão adicionadas ao dicionário de palavras simples geral.)
(com 4 entradas)

linguateca_Npr-.dic (93 bytes)
(Dicionário que conterá os novos nomes próprios.)
(com 5 entradas)

linguateca_Siglas-.dic (58 bytes)
(Dicionário que conterá novas siglas.)
(com 3 entradas)

linguateca_subst-.dic (24 bytes)
(Dicionário que conterá entradas preferenciais que substituirão as que já existem.)
(com 1 entrada)

NomesProprios+.fst (95 bytes)
(Identifica candidatos a siglas e nomes próprios:
- sequências de letras maiúsculas são marcadas como sendo candidatas a siglas (N+CSig)
- sequências começadas em maiúsculas, sendo as seguintes letras minúsculas, são marcadas como sendo candidatas a nome próprio genérico (N+CNpr)
Usado em combinação com \/Graphs\/Siglas\/siglacp.fst permite fazer o recenseamento de siglas associadas às suas descrições
Última revisão: Março, 2004)

Romanos.fst (693 bytes)
(É o grafo que permite analisar números romanos.)
(útil para os nomes de pessoas e de ruas, exemplos: D. João VI, avenida Afonso III)

Siglaf-d_V1.dic (1,04 bytes)
(Dicionário de desenvolvimentos simples de siglas e acrónimos simples (33 entradas)
Conjuntamente com os dicionários:

Siglaf.bin
Siglacf.dic
Siglacf-d.bin
Siglacf-dE.bin
Forma o módulo de siglas e acrónimos
Recolhidos e formalizados por Paulo Moura
Última revisão: Março, 2004)

Exemplos:

Azidothymidine,azt.N+DSig:ms
Citomegalovirus,cmv.N+DSig:ms
Clorofluorocarbonos,cfc.N+DSig:mp

Siglaf_V1-.bin (50,9 KB)
Siglaf_V1-.inf (100 KB)
(Dicionário de siglas e acrónimos simples (4803 entradas)
Conjuntamente com os dicionários:
Siglaf-d.dic
Siglacf.dic
Siglacf-d.bin
Siglacf-dE.bin
Forma o módulo de siglas e acrónimos
Recolhidos e formalizados por Paulo Moura
Última revisão: Março, 2004)

Exemplos:

AZT,azt.N+Sig:ms
Acam,acam.N+Acr+HumCol:fs

Os transdutores integrados no módulo DISAMB:

(Estes recursos encontram-se em: /usr/local/INTEX/Prv/Portuguese/Disamb)

No ficheiro DISAMB, encontram-se transdutores de resoluções de ambiguidades.

disamb.fst (205 KB)
(Trasndutor que resulta da união entre os transdutores: 1-Adjectivos.fst, A-Vaux.fst e Disamb-Verbo-Pro.fst.

1-Adjectivos.fst (141 KB)
(Analisa certos grupos nominais constituídos por pelo menos um nome e um adjectivo.
Elaborado por Paula Carvalho
Última revisão: Março, 2004)

A-Vaux.fst (110 KB)
(Analisa sequências de verbos auxiliares seguidos de um predicado verbal ou adjectival.
Elaborado por Elisabete Ranchhod
Última revisão: Março, 2004)

Disamb_Verbo_Pro.fst (2,20 KB)
(Analisa combinações verbo-clítico.

Elaborado por Cristina Mota
Última revisão: Março, 2004)

Agradecimento
O AnELL foi (parcialmente) financiado pela Fundação para a Ciência e Tecnologia, co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001 (Linguateca).

Ordem pela qual os recursos são chamados:

fst2txt	**Sentence.fst**

etiqc	**UCompounds.bin:** guarda-chuva **linguateca_naoambiguos.dic:** guarda-fatos **Siglacf_V1.dic:** ACA-M **Siglacf-d_V1.bin:** Associação Moçambicana de Bancos **Siglacf-dE_V1.bin:** Aides Nationales - Échanges de Données **linguateca_SiglasC.dic:** Instituto Nacional de Estatística **DelacfNpr_V2.bin:** Combatentes da Grande Guerra **linguateca_NprC.dic:** América Central **DelacConj.dic:** à medida que **DelacPrep.dic:** em vez de **DelacAdv_V2.1.bin:** a bem ou a mal **DelacfAdj_V1.bin:** bem falante **DelacfNomes_V3.bin:** abalo de terra **DelacfEmp_V1.dic:** e-mail **linguateca_substC-.dic:** chapéus de chuva **linguateca_adicC.dic:** entidade mencionada **DOrd-.fst:** três mil e quarenta e três **DCard-.fst:** décimo terceiro

fst2txt	**Replace.fst:**	URL:	{http://www.nyu.edu/pages/linguistics/intex/,.X+url}
		Números:	{123,.X+num}

indexer

dicos	**Delaf_V3.bin:** abalou **DelafEmp_V1.dic:** assez **DelafTmp_V1.dic:** portunhol **Romanos.fst:** **linguateca_subst-.dic:** desmiserabilismo **linguateca_adic-.dic:** inaceitabilidade **filtro-.dic:** ainda **Siglaf_V1.bin:** AZT **linguateca_Siglas-.dic:** INE **DelafNpr_V2-.bin:** Amazonas **NomesProprios+.fst:** **linguateca_Npr-.dic:** Linguateca

dicoc	**DelacConj.dic:** à medida que **DelacPrep.dic:** em vez de **DelacAdv_V2.1.bin:** a bem ou a mal **DelacfAdj_V1.bin:** bem falante **DelacfNomes_V3.bin:** abalo de terra **DelacfEmp_V1.dic:** e-mail **linguateca_substC-.dic:** chapéus de chuva **linguateca_adicC.dic:** entidade mencionada **DOrd-.fst:** **DCard-.fst:**

Siglacf_V1.dic:	ACA-M
Siglacf-d_V1.bin:	Associação Moçambicana de Bancos
Siglacf-dE_V1.bin:	Aides Nationales - Échanges de Données
linguateca_SiglasC.dic:	Instituto Nacional de Estatística
DelacfNpr_V2.bin:	Combatentes da Grande Guerra
linguateca_NprC.dic:	América Central

fst2txt

Tag-All:	**Tag-Verbo-Pro.fst:**	<V+Clit>livro-o</V+Clit>
	Tag-ELLE:	<EM>João</EM>

indexer

dicos

Delaf_V3.bin:	abalou
DelafEmp_V1.dic:	assez
DelafTmp_V1.dic:	portunhol
Romanos.fst:	
linguateca_subst-.dic:	desmiserabilismo
linguateca_adic-.dic:	inaceitabilidade
filtro-.dic:	ainda
Siglaf_V1.bin:	AZT
linguateca_Siglas-.dic:	INE
DelafNpr_V2-.bin:	Amazonas
NomesProprios+.fst:	
linguateca_Npr-.dic:	Linguateca

dicoc

DelacConj.dic:	à medida que
DelacPrep.dic:	em vez de
DelacAdv_V2.1.bin:	a bem ou a mal
DelacfAdj_V1.bin:	bem falante
DelacfNomes_V3.bin:	abalo de terra
DelacfEmp_V1.dic:	e-mail
linguateca_substC-.dic:	chapéus de chuva
linguateca_adicC.dic:	entidade mencionada
DOrd-.fst:	
DCard-.fst:	
Siglacf_V1.dic:	ACA-M
Siglacf-d_V1.bin:	Associação Moçambicana de Bancos
Siglacf-dE_V1.bin:	Aides Nationales - Échanges de Données
linguateca_SiglasC.dic:	Instituto Nacional de Estatística
DelacfNpr_V2.bin:	Combatentes da Grande Guerra
linguateca_NprC.dic:	América Central

dicoe

Norm.fst:	Contracções:	do → de o

etiqa

interg

Disamb.fst:	NP que contêm adjectivos:	a flor que era amarela
	Predicados complexos:	tenho estado a andar
	Verbo-clítico:	livro-o

tokenslist

Definições das etiquetas a cinzento:

fst2txt → Etapa do pré-processamento.

etiqc → Etiqueta as palavras compostas dum texto.

indexer → Indexa o texto.

dicos → Aplica os recursos lexicais para identificação de palavras simples.

dicoc → Aplica os recursos lexicais para identificação de palavras compostas

dicoe → Aplica o transdutor para identificação das contracções.

etiqa → Constrói um transdutor por cada frase, pondo em evidência as ambiguidades.

interg → Faz a intersecção entre os transdutores do texto e as gramáticas de resolução de ambiguidades

tokenslist → Cria a lista de todos os átomos de um texto, associando-lhes a sua frequência no texto.

REFERÊNCIAS

- Carvalho, Paula; «Gramáticas de Resolução de Ambiguidades Resultantes da Homografia de Nomes e Adjectivos.» Tese de Mestrado. Faculdade de Letras da Universidade de Lisboa, 2001. http://label.ist.utl.pt/label/download/TesePaulaCarvalho2001.pdf

- Eleutério, Samuel; E. Ranchhod; H. Freire; J. Baptista, «A System of Electronic Dictionaries of Portuguese». Lingvisticae Investigationes, XIX:1, pp.57-82, Amsterdam/Philadelphia: John Benjamins, 1995.

- Mota, Cristina, «Inflection of the Portuguese DELAS using FST». In C. Muller, J. Royaute, M. Silberztein, INTEX Pour la Linguistique et le Traitement Automatique des Langues, Série "Archive, Bases, Corpus", Nº1, pp 35-51, Presses Universitaires de Franche-Comté, France, 2004.

- Mota, Cristina; P. Moura, «ANELL: A Web System for Portuguese Corpora Annotation». In Mamede, Nuno; J. Baptista; I. Trancoso; G. Volpes Nunes (eds.), Computational Processing of the Portuguese Language, LNAI 2721, pp. 184-188, Berlin: Springer, 2003. http://www.linguateca.pt/documentos/CMotaPMoura-Propor2003.pdf

- Ranchhod, Elisabete, «O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais». In Elisabete Marques Ranchhod (org.), *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações*, pp. 13-47, Lisboa: Caminho, 2001.

- Ranchhod, Elisabete; P. Carvalho; C. Mota; A. Barreiro, «Portuguese Large-scale Language Resources for NLP Applications». In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th LREC*, pp. 1755-1758, Lisbon, 2004.

- Silberztein, Max, «Dictionnaires électroniques et analyse lexicale du français. Le système INTEX». Paris, Masson, 1993.

- Silberztein, Max, INTEX 4.12 french manual (http://msh.univ-fcomte.fr/intex/downloads/Manuel.pdf - último acesso: 27 de Setembro de 2005).