technology
from seed

Jorge Baptista, Nuno Mamede and Ilya Markov

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

L² F - Spoken Language Systems Laboratory

technology
from seed

Integrating a Lexicon-Grammar of Verbal Idioms in STRING Portuguese NLP system

Jorge Baptista, Nuno Mamede and Ilya Markov

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

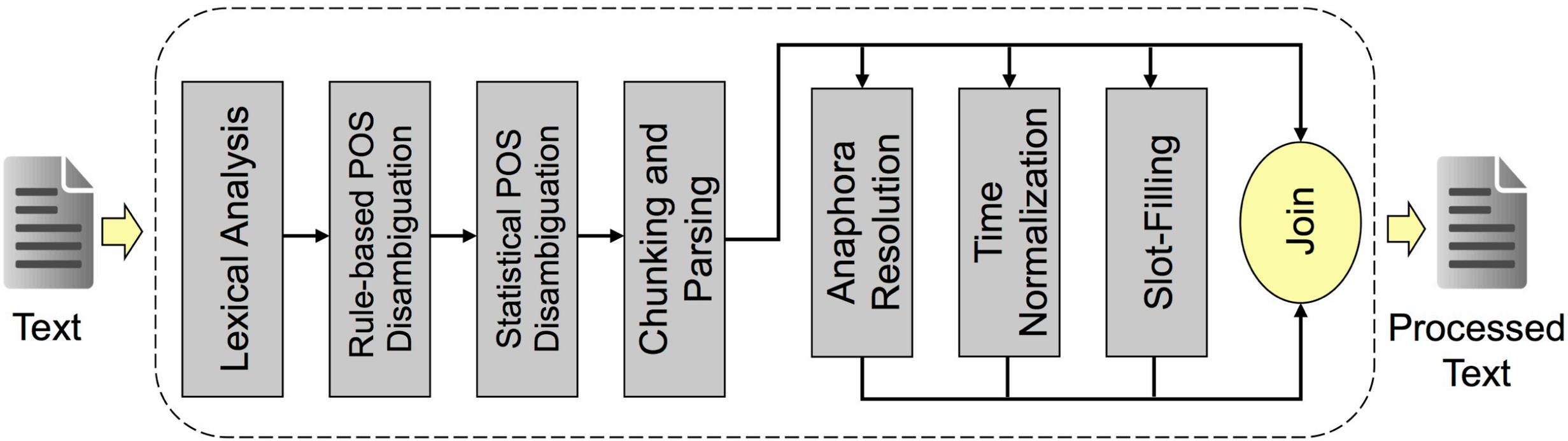
L² F - Spoken Language Systems Laboratory

System Architecture

technology
from seed

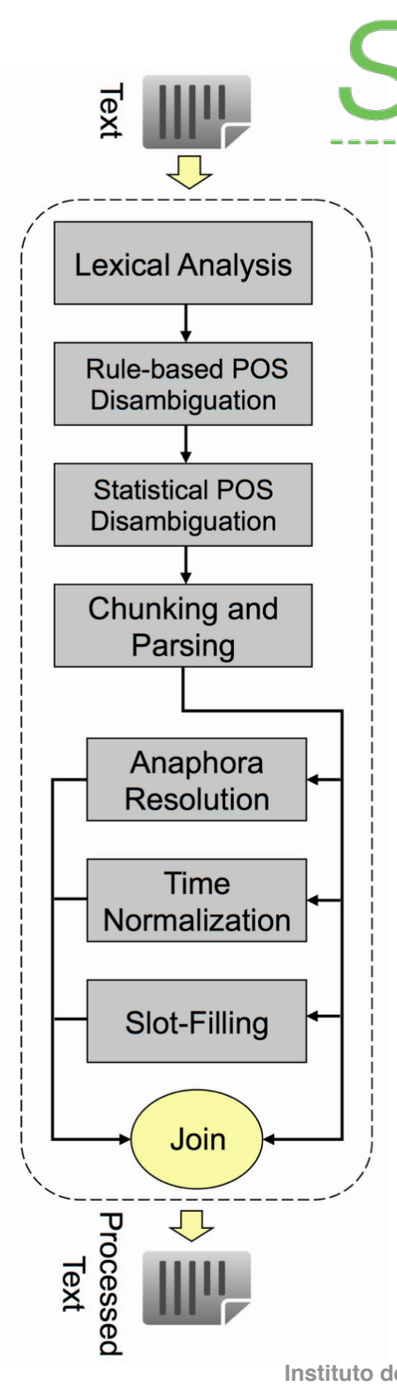
inesc id lisboa

String *statistical and rule-based natural language processing chain*



string.l2f.inesc-id.pt

MWEs in STRING



← unambiguous compounds — *lua-de-mel* ‘honeymoon’

← ambiguous/context dependent compounds — *, a saber,* ‘namely’

← Lexicon-Grammars

perder a cabeça
‘lose one’s mind’

Lexicon-Grammar of Verbal Idioms

Class	Structure	Example	Size
C1	N_0 V C_1	O Pedro matou a galinha dos ovos de ouro	800
CAN	N_0 V (C de N)$_1$ = C_1 a N_2	O Pedro arrefeceu os ânimos (da Maria = à Maria)	200
CDN	N_0 V (C de N)$_1$	O Pedro queria a cabeça da Maria	100
CP1	N_0 V Prep C_1	O Pedro bateu com a porta	900
CPN	N_0 V Prep (C de N)$_1$	O Pedro foi aos cornos do João	100
C1PN	N_0 V C_1 Prep N_2	O Pedro arrastou a asa à Maria	400
CNP2	N_0 V N_1 Prep C_2	O Pedro tirou o relógio do prego	350
C1P2	N_0 V C_1 Prep C_2	O Pedro deitou mãos à obra	400
CPP	N_0 V Prep C_1 Prep C_2	O Pedro foi de cavalo para burro	200
CPPN	N_0 V C_1 Prep C_2 Prep C_3	O Pedro deitou o bebé fora com a água do banho	50
		Total	3,500

Table 1. Lexicon-Grammar of Frozen Sentences of European Portuguese
N and *C* stand for free or frozen noun phrases, respectively; N_0 is the subject, N_1 , N_2 and N_3 the first, second and third complement; *V* is the verb and *Prep* a preposition. Figures are approximate.

Baptista *et al*. 2004, 2005

Lexicon-Grammar of Verbal Idioms

N0=:Nhum	*N0=:N-hum*	V	*Vse*	*NegObrig*	*Prep*	*Det*	C	*N1=: Nhum*	*N1=: N-hum*	*de N = a N*	*de N = Poss*	Example
+	-	<acabar>	-	-	com	a	raça	+	-	+	+	*O Pedro acabou com a raça da Maria*
+	-	<atirar>	+	-	a	os	pés	+	-	+	+	*O Pedro atirou-se aos pés da Maria*
+	-	<chegar>	-	+	a	os	calcanhares	+	-	+	+	*O Pedro não chega aos calcanhares da Maria*
+	-	<cortar>	-	-	em	a	casaca	+	-	+	-	*O Pedro cortava na casaca da Maria*
+	-	<ir>	-	-	a	as	trombas	+	-	+	-	*O Pedro foi às trombas do João*
+	-	<ir>	-	-	em	a	cantiga	+	-	-	+	*O Pedro foi na cantiga da Maria*
+	-	<ir>	-	-	a	a	cara	+	-	+	-	*A Maria foi à cara do Pedro*
+	-	<pegar>	-	-	em	a	deixa	+	-	+	+	*O Pedro pegou na deixa da Maria*
+	-	<rir>	-	-	em	a	cara	+	-	+	+	*O Pedro riu na cara da Maria*
+	-	<rir>	+	-	em	a	cara	+	-	+	+	*O Pedro riu-se na cara da Maria*
-	+	<sair>	-	-	de	o	pelo	+	-	+	-	*O salário sai-lhe do pelo*
-	+	<subir>	-	-	a	a	cabeça	+	-	+	+	*A fama subiu à cabeça do Pedro*
+	-	<viver>	-	-	em	a	sombra	+	+	-	+	*O Pedro vive na sombra da Maria*

Table 2. Class CPN : *N0 V Prep (C de N)1*

Lexical elements presented in basic order; verbs are indicated by <lemma> and other elements by surface form; contractions of preposition and determiner are undone; distributional and transformational properties are represented by binary features: '+/-'. *Vse* : intrinsically reflexive constructions, *NegObrig* : obligatory negation, *deN=aN* : dative restructuring, *deN=Poss* : possessive pronouning.

STRING's Strategy for verbal idioms

- parse sentences as usual
- use syntactic dependencies built until then

- to extract a new dependency FIXED linking the core elements of the verbal idiom
- use this dependency to:
  - by-pass full, ordinary verbs' constructions (WSD)
  - extract the idioms as EVENTs (predicates)
  - extract the semantic roles of the idiom' free argument slots
  - avoid extracting part-whole relations
  - ...

Semi-automatically produced rules

At the last stages of parsing:

O Pedro não chega aos calcanhares do João
'Peter does not reach the heels of John'

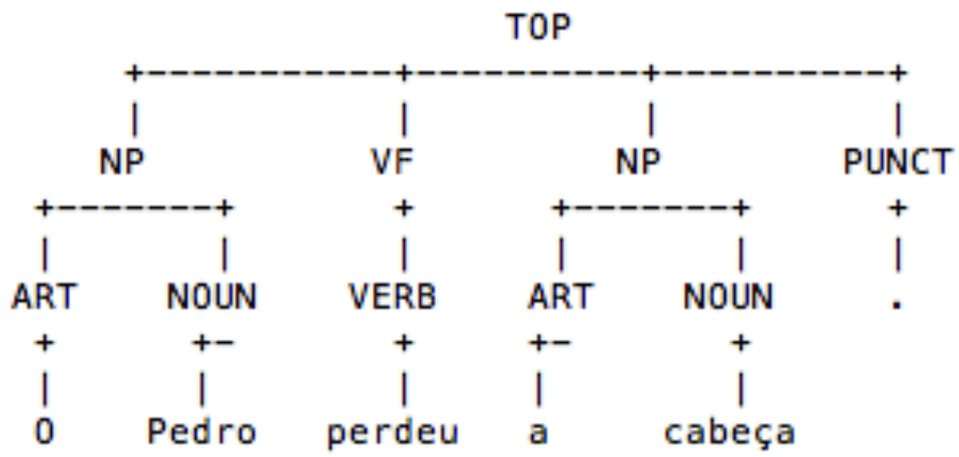
N=:Nhum	N=:N-hum	V	Vse	NegObrig	Prep	Det	C	N≠:Nhum	N≠:N-hum	de N = a N	de N = Poss	Example
+	-	<acabar>	-	-	com	a	raça	+	-	+	+	*O Pedro acabou com a raça da Maria*
+	-	<atirar>	+	-	a	os	pés	+	-	+	+	*O Pedro atirou-se aos pés da Maria*
+	-	<chegar>	-	+	a	os	calcanhares	+	-	+	+	*O Pedro não chega aos calcanhares da Maria*
+	-	<cortar>	-	-	em	a	casaca	+	-	+	-	*O Pedro cortava na casaca da Maria*
+	-	<ir>	-	-	a	as	trombas	+	-	+	-	*O Pedro foi às trombas do João*
+	-	<ir>	-	-	em	a	cantiga	+	-	-	+	*O Pedro foi na cantiga da Maria*
+	-	<ir>	-	-	a	a	cara	+	-	+	-	*A Maria foi à cara do Pedro*
+	-	<pegar>	-	-	em	a	deixa	+	-	+	+	*O Pedro pegou na deixa da Maria*
+	-	<rir>	-	-	em	a	cara	+	-	+	+	*O Pedro riu na cara da Maria*
+	-	<rir>	+	-	em	a	cara	+	-	+	+	*O Pedro riu-se na cara da Maria*
-	+	<sair>	-	-	de	o	pelo	+	-	+	-	*O salário sai-lhe do pelo*
-	+	<subir>	-	-	a	a	cabeça	+	-	+	+	*A fama subiu à cabeça do Pedro*
+	-	<viver>	-	-	em	a	sombra	+	+	-	+	*O Pedro vive na sombra da Maria*

```
if (VDOMAIN(#1,#2[lemma:chegar]) &
PREPD(#3[surface:calcanhares],#4[lemma:a])
& MOD[post](#2,#3) & MOD[pre,neg](#2,#5) )
FIXED(#5,#2,#4,#3)
```

Chunking and dependency parsing
Example: *O Pedro perdeu a cabeça*
Peter lost the head ‘Peter lost his mind’

Terminal — ssh — 104×28

```
x05.l2f.inesc-id.ptjbaptis[8:44] echo "O Pedro perdeu a cabeça." | xip/string.sh -t -tr -f -indent
```




```
MAIN(perdeu)
DETD(Pedro,O)
DETD(cabeça,a)
VDOMAIN(perdeu,perdeu)
SUBJ_PRE(perdeu,Pedro)
CDIR_POST(perdeu,cabeça)
FIXED(perdeu,cabeça)
NE_PEOPLE_INDIVIDUAL(Pedro)
0>TOP{NP{O Pedro} VF{perdeu} NP{a cabeça} .}

x05.l2f.inesc-id.ptjbaptis[8:45]
```

```
if (VDOMAIN(?,#2[lemma:perder])
& CDIR[post](#2,#3[surface:cabeça]))
FIXED(#2,#3)
```

Example: Peter lost the head=Peter lost his mind

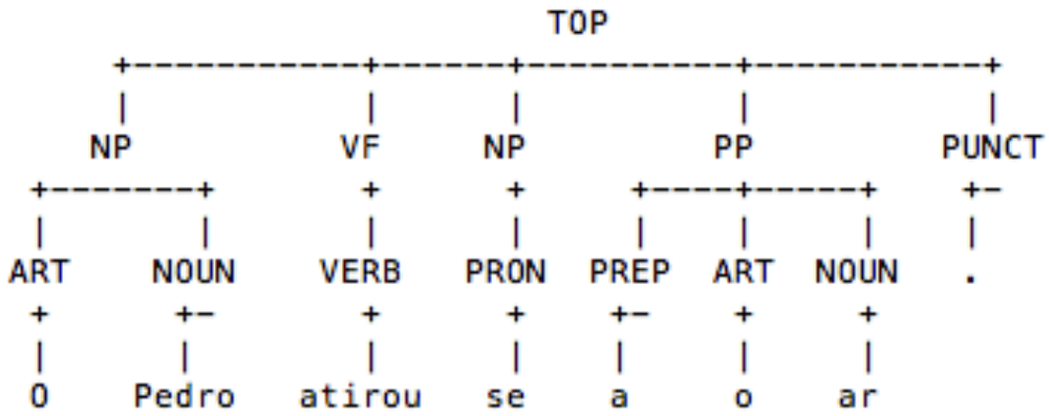
Terminal — ssh — 121×28

```
x05.l2f.inesc-id.ptjbaptis[8:45] more trees.out

TOP(0-23)+[cat:0]
 NP(0-6)-[np:+,noun:+,start:+,first:+]
  ART(0-0)+[def:+,toutmaj:+,maj:+,masc:+,sg:+,art:+,hmmselection:+,!start:+,first:+]
    O(0-0)+[def:+,toutmaj:+,maj:+,masc:+,sg:+,art:+,hmmselection:+,!start:+,first:+]
  NOUN(2-6)-[enp3:+,sem-hpeople:+,sem-hindividual:+,firstname:+,human:+,start_people:+,maj:+,proper:+,ma
sc:+,sg:+,noun:+,hmmselection:+,last:+,first:+]
    Pedro(2-6)-[enp3:+,sem-hpeople:+,sem-hindividual:+,firstname:+,human:+,start_people:+,maj:+,proper:+
,masc:+,sg:+,noun:+,hmmselection:+,last:+,first:+]
 VF(8-13)-[mark_ger:+,fin:+,verb:+]
  VERB(8-13)+[sr-n1-nhum:+,sr-n1-cdir:+,sr-pass-estar:+,sr-pass-ser:+,sr-n1-object-gen:+,sr-n0-agent-gen
:+,mark_ger:+,32c:+,markviper:+,perf:+,ind:+,3p:+,sg:+,verb:+,hmmselection:+,last:+,first:+]
    perdeu(8-13)+[sr-n1-nhum:+,sr-n1-cdir:+,sr-pass-estar:+,sr-pass-ser:+,sr-n1-object-gen:+,sr-n0-agent
-gen:+,mark_ger:+,32c:+,markviper:+,perf:+,ind:+,3p:+,sg:+,verb:+,hmmselection:+,last:+,first:+]
 NP(15-22)-[np:+,noun:+]
  ART(15-15)+[def:+,fem:+,sg:+,art:+,hmmselection:+,first:+]
    a(15-15)+[def:+,fem:+,sg:+,art:+,hmmselection:+,first:+]
  NOUN(17-22)+[sem-pos-soc:+,sem-part:+,sem-cc:+,sem-anmov:+,common:+,fem:+,sg:+,noun:+,hmmselection:+,l
ast:+]
    cabeça(17-22)+[sem-pos-soc:+,sem-part:+,sem-cc:+,sem-anmov:+,common:+,fem:+,sg:+,noun:+,hmmselection
:+,last:+]
 PUNCT(23-23)-[dots:+,punct:+,hmmselection:+,!end:+,last:+]
   .(23-23)-[dots:+,punct:+,hmmselection:+,!end:+,last:+]
 NOUN(-100--10)-[noun:+]
   outro(-100--100)-[noun:+]
x05.l2f.inesc-id.ptjbaptis[8:46]
```

Chunking and dependency parsing
Example: Peter jump in the air ‘Peter got furious’

Terminal — ssh — 117×29




```
MAIN(atirou)
DETD(Pedro,0)
DETD(ar,o)
VDOMAIN(atirou,atirou)
MOD_POST(atirou,ar)
SUBJ_PRE(atirou,Pedro)
CDIR_POST(atirou,se)
CLITIC_POST(atirou,se)
FIXED(atirou,a,ar)
NE_PEOPLE_INDIVIDUAL(Pedro)
EVENT_LEX(atirou,outro)
EVENT_OTHER(atirou)
EVENT_AGENT-GENERIC(atirou,Pedro)
0>TOP{NP{0 Pedro} VF{atirou} NP{se} PP{a o ar} .}

x05.l2f.inesc-id.ptjbaptis[9:00]
```

References

Jorge Baptista; Correia, Anabela; Fernandes, Graça (2004). Frozen Sentences of Portuguese: Formal Descriptions for NLP. Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics, Barcelona (Spain), July 26, 2004. ACL: Barcelona, pp. 72-79.

Jorge Baptista; Correia, Anabela; Fernandes, Graça (2005). Léxico Gramática das Frases Fixas do Portugués Europeo, in *Cadernos de Fraseoloxía Galega* 7, pp. 41-53, Santiago de Compostela, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades.

Jorge Baptista, Structuring of cross-linguistic database of frozen sentences, 2008. in: González-Royo, Carmen; Mogorròn-Huerta, Pedro (eds.), Estudios y análisis de fraseología contrastiva: lexicografía y traducción. Alicante: Univ. Alicante, pp. 37-46.

Jorge Baptista; Mamede, Nuno; Gomes, Fernando. 2010. Auxiliary verbs and verbal chains in European Portuguese. Proceedings of PROPOR'2010. LNCS/LNAI 6001: pp. 110-119. Berlin: Springer.

Jorge Baptista; Català, Dolors. 2010. What glues idioms together may not be just statistics after all: the case for compound adverbs in Portuguese and Spanish. EUROPHRAS'2010 Cross-Linguistic and Cross-Cultural Perspectives on Phraseology and Paremiology. Granada, Spain. June 30th-July 2nd 2010

Jorge Baptista (2010), Verb Classification Guidelines. (Technical Report), Lisboa: INESC-ID Lisboa (10 pp.)

Jorge Baptista (2012). ViPEr: A Lexicon-Grammar of European Portuguese Verbs. Proceedings of the 31th International Conference on Lexis and Grammar, Nové Hrady (Czech Republic), September 19-22, 2012, pp. 10-16.

Jorge Baptista, Lucas Vieira, Cláudio Diniz, Nuno Mamede, Coordination of '-mente' ending adverbs in Portuguese: an integrated solution. 2012. in Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.) Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. pp. 24-34. Berlin: Springer.

Dolors Català, Jorge Baptista, 2007. Spanish Adverbial Frozen Expressions. Proceedings of the Workshop on A Broader Perspective on Multiword Expressions (MWE 2007, Prague, June 2007), International Conference of the European Chapter of the Association for Computational Linguistics, pp. 33–40.

Graça Fernandes, Baptista, Jorge. 2008. Frozen sentences with obligatory negation: linguistic challenges for natural language processing. in Mellado-Blanco, Carmen (ed.), Colocaciones y fraseología en los diccionarios, Frankfurt: Peter Lang, pp.85-96.

Nuno Mamede, Jorge Baptista, Cláudio Diniz Eva Cabarrão. 2012. STRING – An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. in Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.) Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. http://www.propor2012.org/demos.html.

Ricardo Portela, Nuno Mamede and Jorge Baptista, 2011. Multiword Identification. INForum, III Simpósto de Informática, Coimbra, Portugal, September 8-9, 2011.

Lucas Nunes Vieira, Cláudio Diniz, Nuno Mamede, Jorge Baptista. 2012. A Lexicon of Verb and –mente Adverb Collocations in Portuguese: Extraction from Corpora and Classification. Proceedings of the 31th International Conference on Lexis and Grammar, Nové Hrady (Czech Republic), September 19-22, 2012, pp. 155-161.

L² F - Spoken Language Systems Laboratory