Disambiguation of Portuguese Clitic Pronouns

Jorge Baptista[a,b], Joana Pinto[c,b], Nuno J. Mamede[c,b]

[a] *University of Algarve, Faro, Portugal*

[b] *INESC-ID Lisboa/Spoken Language Laboratory, Lisboa, Portugal*

[c] *Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Abstract

Part-of-speech (POS) tagging is usually considered a solved problem in NLP, with state-of-the-art precision usually above 97% precision for most languages. However, these figures are often based on small and/or coarse POS tag sets. With highly granular tag sets, the number of ambiguous forms may increase significantly, as thiner distinctions may apply to the same word form. Portuguese is a highly inflected language, and several state-of-the-art POS taggers are already available. Nevertheless, some POS categories, which are key to later NLP tasks, remain a resilient POS-tagging problem, with suboptimal results. This is due to their level of ambiguity, particularly when a fully-fledged POS tag set is used, and also for their high frequency in texts. In Portuguese, clitic pronouns are one such categories. Correct disambiguation of personal pronouns is particularly relevant in this language, not only because they constitute a very frequent category in texts, but especially because of their prominent role in the discourse structure by their anaphoric value, and their essential contribution to the semantic interpretation of the verbs they depend on and/or of the sentences they appear in. This paper presents a solution that combines statistical POS tagging techniques with rich lexical information in order to improve the POS disambiguation of clitic personal pronouns in Portuguese. The solution is fully integrated in a NLP processing chain and is directly exploited in subsequent NLP tasks. Results on the POS tagging problem significantly outperform baseline. Consequently, improvements in several NLP tasks were also assessed.

Keywords: Portuguese, Disambiguation, Statistical POS tagging, Clitic Pronoun

Email addresses: jbaptis@ualg.pt (Jorge Baptista), joana.lapas.pinto@tecnico.ulisboa.pt (Joana Pinto), Nuno.Mamede@inesc-id.pt

1. Introduction

Part-of-speech (POS) tagging is an important task in Natural Language Processing (NLP). Over the years different approaches have been solving this problem with a precision above 97%. Generally, a POS tagger combines lexical and contextual information with a technique to choose the best POS to a given word. Among these techniques, two main approaches have been applied: rule-based and statistical approaches.

Despite rule-based systems requiring less computational resources, they are ussually difficult to build. Brill [1] proposed a rule-based POS tagger, where the rules are automatically acquired, overcoming the mentioned disadvantage of the rule-based systems.

Regarding statistical techniques, Hidden Markov Models (HMMs) are very explored in POS tagging [2] [3], which assign the best sequence of tags to an input sequence of words, considering the contextual information codified in n-grams models. Usually, these models considers only previous tags in the sequence, however the use of both preceding and following tags has also been exploit, using dependency networks [4]. Nevertheless, other statistical approaches have been experimented, for example Maximum Entropy models [5], where it is possible to define more complex statistics, not limited to n-grams, that explores linguistic sophisticated features [6] [7]. Besides, as the tagset varies from language to language, a POS tagger should consider the amount of information utilized and the context shape. A method based on Support Vector Machine models addressed the concerns mentioned above, yielding good results [8]. Other approaches combined rule-based with statistical techniques, as STRING chain [9] for European Portuguese. Besides the previous system, there are other research works also focused in Portuguese [10] [11] [12].

This paper presents a solution that combines statistical POS tagging techniques with rich lexical information in order to improve the POS disambiguation of personal pronouns in Portuguese. The paper is organized as follows: The remainder of this section presents a brief description of the European Portuguese personal pronouns' case system (1.1), in order to introduce the ambiguity problem they pose to parsing (1.2); next, the NLP processing chain STRING (2.1) is presented, for which the pronoun's disambiguation module was developed, and describe succintly the lexical resource with syntactic and semantic information on verbs, ViPEr (2.2).

1.1. Personal pronouns in Portuguese

This section presents a brief overview of the European Portuguese personal pronouns system and their main syntactic properties, which is necessary to understand the nature of the problem.

Portuguese personal pronouns partially conserve Latin case distinction, and it is possible to organize this category into a five case system: nominative, accusative, dative, reflexive and oblique cases, as illustrated in Table 1.

Table 1: Personal pronouns system in European Portuguese

Pers-num	Nominative	Accusative	Dative	Reflexive	Oblique
1^{st}-sing	*eu*	*me*	*me*	*me*	*mim*, *comigo*
2^{nd}-sing	*tu*	*te*	*te*	*te*	*ti*, *contigo*
3^{rd}-sing	*ele*, *ela*	*o*, *a*	*lhe*	*se*	(*ele*, *ela*) *si*, *consigo*
1^{st}-plur	*nós*	*nos*	*nos*	*nos*	*nós*, *conosco*
2^{nd}-plur	*vós*	*vos*	*vos*	*vos*	*vós*, *convosco*
3^{rd}-plur	*eles*, *elas*	*os*[1], *as*	*lhes*	*se*	(*eles*, *elas*) *si*, *consigo*

In the course of the language's history, the grammatical distinctions expressed by this case system have been progressively eroded (see [13] for an overview). This process is still underway, being one one the major grammatical distinctions between the Portuguese (European) and Brazilian variants of the language.

(Nuno J. Mamede)

Nominative (or 'tonic') forms are not clitic. Nominative pronouns (except 1^{st}- and 2^{nd}-person singular) are also regularly used as oblique pronouns: *Ele gosta de* ***mim/ti/dela*** 'He likes/is fond of me/you/she'.

The use of the 3^{rd}-person singular oblique case *si*, *v.g. Ele gosta de* ***si*** 'He likes you', is restricted to the 2^{nd}-person singular formal way addressing (corresponding to pronoun *você*; see below).

However, these nominative pronouns are increasingly being used instead of the (normative) accusative and the dative case forms:

- as direct (accusative) complements, e.g. *Ele viu ela a cantar o fado* 'He saw she [at] singing the fado' instead of the "correct" accusative form, e.g. *o* 'him'.
- and indirect (dative) complements, in the later case with prepositions *a* 'to' or *para* 'for' *O Pedro disse isso* ***a/para ele*** 'Peter said that to/for him', instead of the "correct" dative form, e.g. *lhe* 'to_him'.

Most pronouns kept the Latin number distinction, except in the 3^{rd}-person, reflexive (*se*) and oblique (*si* 'him/herself, themselves', *consigo* 'with him/her/them') case forms. Number distinction is also lost when 3^{rd}-person dative forms are contracted with 3^{rd}-person accusative: *lho* 'to_him/her/them_it-masc.sg.', *lha* 'to_him/her/them_it-fem.sg.', *lhos* 'to_him/her/them_it-masc.pl.' and *lhas* 'to_him/her/them_it-fem.pl.').

Nowadays, only 3^{rd}-person nominative and accusative pronouns can be used to distinguish (partially) gender opposition, since dative, reflexive, and oblique pronouns have lost that distinction.

Some pronouns involve contractions with the prepositions introducing them. The dative forms incorporate the preposition *a* 'to' introducing the indirect object, which can be seen as a form of contraction, *e.g. O Pedro disse isso* ***a***_*o João* = *O Pedro disse-****lhe*** *isso* 'Peter said this to_the John' = 'Peter said-to_him that' (the contraction is noted with an underscore '_') . A special case of contraction also occurs with preposition *com* 'with' and the oblique forms, as vestige of Latin preposition duplication, v.g. *comigo* < CUM+MECUM (<CUM+ME). As in the reflexive pronouns, there is no distinction between the singular and plural forms in the 3^{rd} person.

The so-called possessive 'pronouns' are treated separately, as they function as determinative or predicative adjectives, *e.g. o meu livro* 'my book'. The full syntactic justification of this option is out of the scope of the current paper, as possessives raise less POS-tagging problems.

The case-invariable, $2^{n}d$-person-singular, simple personal pronoun *voçê* 'you_sg.' (plural: *vocês*) and the case-invariable, 1st-person-plural compound pronoun *a gente* 'we' (literally, 'the people') were also ignored in this paper.

Some pronouns are ambiguous with other *lexical* POS, e.g. *ele*/N(oun) and V(erb), *nós*/N and *consigo*/V, but their resolution is relatively easy, with generic statistical techniques. Other pronouns, are ambiguous with other *grammatical* POS. This is the case of the contraction *mas*=*me*/dative+*as*/accusative 'to_me+it_fem.pl', which is ambiguous with the coordinative conjunction *mas* 'but'; and, more problematic, the 3rd-person accusative forms (*o*, *a*, *os*, *as* 'it/him/her/them/those'), which are homographs of the definite article forms ('the'), as well as the particularly resilient form *a* 'to', ambiguous with a preposition.

As accusative, dative and reflex pronouns can undergo cliticization, that is, they can be attached to a verb form by an hyphen (enclisis), many of their instances can be disambiguated when used in enclitic position, e.g. *O Pedro leu o livro*=*O Pedro leu-o* 'Peter read the book/Peter read-it'. However, pronouns can also be moved to the front of the verb (proclisis), e.g. *O Pedro não leu o livro*=*O Pedro não o leu* 'Peter did-not read the book/Peter did-not it read', in which case 3rd-person accusative are ambiguous with definite articles.

The most complex case of ambiguity arises between accusative, dative and reflexive forms for 1^{st}- and 2^{nd}-person, singular and plural forms, *v.g. me*, *te*, *nos*, *vos*. This will be called the *ADR* ambiguity class and it constitutes the main focus of this paper.

The second most important case of ambiguity is the use of the nominative 3^{rd}-person (both singular and plural) forms,

and the 1^{st}- and 2^{nd}-person (only plural) forms as oblique pronouns, *v.g. ele* 'he', *ela* 'she', *eles* 'they$_{masc}$', *elas* 'they$_{fem}$', *nós* 'we' and *vós* 'you$_{plur}$'. This will be our *NO* ambiguity class.

1.2. Ambiguity and Parsing

The case distinction indicates the syntactic function of the pronominalized constituent. Hence, the lack of formal distinction constitutes a form of information degenerescence [14], that renders more difficult the correct parsing of the sentence, namely the attribution of the correct syntactic function.

As many verbs are also ambiguous, i.e. they may have more than one sense, which are often expressed by different syntactic-semantic constructions, identifying the syntactic function of the pronouns is necessary for a precise disambiguation of verb senses [15].

The syntactic function expressed by the case distinction is also directly related to the semantic role the pronoun performs in relation to the verb on which it depends. Therefore, its resolution is key to the correct assignment of semantic roles, especially in systems that perform that labeling based on the syntactic layer produced by a parser [16].

On the other hand, the anaphoric value of the pronoun can only be entirely sorted out if its case is solved as well, which precludes or at least makes it more difficult the anaphora resolution task and all other subsequent NLP tasks that depend on it [17].

Besides the morphologic ambiguity, a syntactic issue complicates this matter further: with many Portuguese verbs, a passive-like transformation moves the direct object noun phrase to the subject position, and a reflex pronoun is inserted, as the initial subject is either zeroed or kept as a prepositional phrase, *e.g. Isso*$_{\texttt{Cause}}$ *irritou o Pedro*$_{\texttt{Experiencer}}$ (active) = *O Pedro*$_{\texttt{Experiencer}}$ *irritou-**se** com isso*$_{\texttt{Cause}}$ (passive-reflexive) 'That infuriated Peter'='Peter irritated himself (with that)'. As transformations do not change the semantic role of the predicate's arguments (both `Cause` and `Experiencer` remain the same) in spite of their different syntactic functions in the sentence, this passive-reflexive construction must be distinguished from ordinary active sentences, where the same pronoun *me* 'myself' is used in the accusative and in the reflexive cases, e.g. *Isso irritou-**me*** (accusative, active) = *Eu irritei-**me*** (reflexive, passive) 'That infuriated-me' ='I infuriated-myself' (= 'I got furious'). Naturally, this implies disambiguating the pronoun's case.

2. Methods

2.1. STRING processing chain

STRING [9][2] is a fully-fledged NPL processing chain, that produces all the major steps of text processing, including sentence-splitting, tokenization, morphological analysis and lemmatization [18], part-of-speech (POS) rule-based and statistical disambiguation [19–21], and parsing [9, 22].

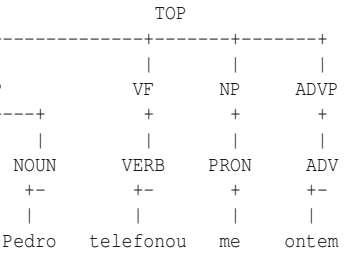
STRING has a well developed module for Named Entities Recognition (NER) [23–28], including time expressions recognition and normalisation [29, 30].

Recently, external modules for Anaphora Resolution (AR) [17, 31], Slot-Filling (SF) [32], Events and Relations Extraction (REvE) [33, 34], and Word Sense Disambiguation (verbs) [15], have been added.

New modules for Semantic Role Labelling (SRL) [35] and event relations' structuring and temporal ordering [36] are currently being developed. Large-scale linguistic resources have been developed for integration in STRING.

[2]https://string.l2f.inesc-id.pt/; a demo is also available at: https://string.l2f.inesc-id.pt/demo/ [2013-07-15]

The parsing module of STRING uses XIP (Xerox Incremental Parser) [37][3], a finite-state technology, rule-based parser, whose grammar for (European) Portuguese is being development under a collaboration between INESC-ID Lisboa/L2F and Xerox Research Center Europe. Acting on the POS-disambiguated text, the XIP parser performs a *chunking* of the sentences, that is, it produces a shallow parsing of the text, identifying the sentences' elementary constituents (NP, PP, AP, ADVP, etc.). Next, it extracts the syntactic dependencies between these constituents, such as SUBJect, CDIR (direct object), MODifier, etc. In order to obtain these major constituents (deep parsing), several auxiliary dependencies are previously extracted. For example, a determinant dependency DETD links the determiner to the head of the NP, or a PREPD dependency links the preposition of a PP to the head noun of this constituent. Fig. 1 illustrates the output of the system, with the chunking tree, the main dependencies extracted and the chunking of the sentence *O Pedro telefonou-me ontem* 'Peter phoned me yesterday'.




```
MAIN(telefonou)
DETD(Pedro,O)
MOD_POST(telefonou,ontem)
SUBJ_PRE(telefonou,Pedro)
CINDIR_POST(telefonou,me)
NE_PEOPLE_INDIVIDUAL(Pedro)
NE_T-REF-SIMULT_T-REF-ENUNC_TEMPO_T-DATE(ontem)
EVENT_OTHER(telefonou)
EVENT_AGENT-GENERIC(telefonou,Pedro)
EVENT_ADDRESSEE(telefonou,me)
EVENT_TIME-CALENDAR(telefonou,ontem)
0>TOP{NP{O Pedro} VF{telefonou} NP{me} ADVP{ontem}}
```

Figure 1: Chunking tree of sentence with dative pronoun.

In Fig. 1 one can see the main dependencies extracted by the parser: MAIN indicates the main verb; DETD links the determiner to the head of the NP; the syntactic functions SUBJ[ject], CINDIR (indirect object) and MOD[ifier] capture the main arguments and complements of the verb; the _PRE and _POST features just indicate if the constituent is before or after the governor, respectively. Next, the NER module extracts the named entity *Pedro* 'Peter' and the time expression *ontem* 'yesterday'. Each entity is given a set of features, corresponding to the entity type it belongs to. Finally, the event extraction module retrieves the EVENT *telefonou* 'phoned' (the _OTHER feature indicates this is a non-normalized event [4]), while the SRL module assigns the semantic roles of AGENT, ADDRESSEE and TIME-CALENDAR to the arguments or complements of the event.

2.2. ViPEr

As mentioned before, the choice of the case for the pronoun depends mostly on the verb it depends on. Therefore, syntactic and semantic information on verbs was deemed essential for solving the pronoun case disambiguation problem.

[3] http://open.xerox.com/Services/XIPParser/

[4] A non-normalized event is any event that does not fall into any of the previously defined classes of events already extracted by the Event Extraction module. These include relevant concepts for many Information Extraction tasks, such as date and place of birth/death, family relation, profession, employer, academic degree, place of residence, etc. For a fully detailed description, see [34].

To this end, the ViPEr [38, 39] lexicon-grammar database for European Portuguese verbs was used, which is briefly present now.

ViPEr consists of a detailed description of the main structural, distributional and transformational properties of the most frequent verbs in European Portuguese. Frequency information was obtained from CETEMPúblico, a publicly available, 189 million words corpus of journalistic text [40], after being processed by STRING [9].

ViPEr was build under the Lexicon-Grammar theoretical and methodological framework [41], based on the Transformational Operator Grammar of [14]. For each verb, the main senses, clear-cut distinct and reproducible, were identified, and to each word-sense, or construction, the most relevant syntactic and semantic properties were encoded, namely: (a) the number of essential complements, (b) the preposition introducing them, (c) the distributional constraints on the lexical fulfillment of those argument slots, (d) some transformational properties, and (e) several semantic features, including argumental semantic roles.

The verbs were organized in syntactic (formal) classes, based of these properties and directly inspired in [42] syntesis. In its current version, ViPEr contains 5,050 verbs (different lemmas), representing approximately 6,500 verb senses, distributed by 71 formal classes and encoded with over 130 features. For example, a verb like *contar* will consist of three distinct entries, or verb senses, each with a different set of syntactic and semantic features[5]:

- *contar*[1] [06]: $Nhum_0$ *contar* $QueF_1$ *a* $Nhum_2$
  O Pedro contou isso ao João 'Peter told that to John'
- *contar*[2] [32PL]: $Nhum_0$ *contar* Npl_1
  O Pedro contou os livros 'Peter counted the books'
- *contar*[3] [35R]: $Nhum_0$ *contar com* N_1
  O Pedro contava com o João 'Peter counted/relyed on John'

Most verbs (4,160; 64%) present only one entry (or verb sense). Ambiguous, polysemic verbs usually have only two (576; 9%) or three (180; 3%) senses; highly ambiguous verbs, with 4 or more verb senses are relatively few (134; 2%).

Initial experiments in verb sense disambiguation [15] have been conducted on a 250,000 words, manually annotated corpus. From the 38,291 verb forms (corresponding to 2,665 lemmas and 10,314 different inflected forms), 21,141 were manually tagged for their ViPEr class. Preliminary results show that for the WSD task the baseline (most frequent sense) is relatively high (63,64% precision) [43]. These results are detailed per verbal lemma in the Table 2.

2.3. Corpora

Evaluation corpus

For evaluation, a fragment of the LE-PAROLE corpus [44, 45] was used. This corpus consists of a large variety of texts, from multiple genres, with approximately 250,000 words. The corpus was processed by STRING [9] and automatically disambiguated [18, 20]. Afterwards, POS tags were manually corrected. Recently, the corpus has been enriched by other syntactic and semantic information, namely, the syntactic classes of ViPEr [39, 46]. For this paper, all personal pronouns were again manually revised. This corpus was intended to be used as a reference corpus for evaluation.

The corpus contains 853 personal pronouns of the ADR ambiguity class {*me-te-nos-vos*} and 964 pronouns of the NO ambiguity class {*ele-ela-nós-vós-eles-elas*}. Tables 3 and 4 show the breakdown of these pronouns and their cases in the corpus.

[5] The classification codes are conventional. `N0`, `N1` and `N2` represent the subject, and the first or second complement, respectively. The `hum`, `pl` and `QueF` indices represent distributional or structural features: human, plural, and sentential arguments, respectively.

Table 2: Results of most frequent sense on the WSD task, detailed per verbal lemma.

Verbal lemma	Number of instances	Number of Classes	Accuracy	Class
abandonar	22	4	54.17%	38L1
aceitar	64	2	51.47%	38TD
acreditar	59	3	71.19%	08
aprender	56	4	66.07%	06
assinalar	13	3	50.00%	06
atirar	12	7	41.67%	38LD
avançar	43	6	46.51%	35LD
chamar	91	4	76.19%	39
comprometer	14	3	71.43%	07
concordar	40	4	72.50%	42S
confrontar	13	5	69.23%	36R
contornar	2	2	50.00%	32C
convencer	24	3	52.00%	12
destacar	16	4	50.00%	36R
esconder	19	3	60.87%	38LD
explicar	134	3	94.81%	09
falar	206	3	96.17%	41
ler	82	4	94.38%	32C
mostrar	63	3	55.56%	36DT
pensar	166	4	63.31%	06
preparar	48	4	43.14%	32C
resolver	95	2	52.63%	32C
saber	480	2	95.87%	06
ver	450	2	48.16%	32C

Table 3: Evaluation corpus: ADR ambiguity class distribution

	me	*te*	*nos*	*vos*	total
acc	104	65	33	6	208
dat	227	105	57	11	**400**
ref	114	82	49	0	245
total	445	252	139	17	853

For the accusative-dative-reflexive pronouns (ADR ambiguity class), one can see that the 2^{nd}-person plural *vos* is seldom used in this corpus, while approximately half of the pronouns correspond to the 1^{st}-person singular *me*. On the other hand, a little less than half of them are in the dative case, whereas accusative and reflexive forms have almost the same frequency. The second most used pronoun is *te*, mostly in the dative and reflexive case. No reflexive *vos* instance was found.

Table 4: Evaluation corpus: NO ambiguity class distribution

	ele	*ela*	*nós*	*vós*	*eles*	*elas*	Total
nom	274	191	90	0	54	16	625
obl	140	84	5	2	84	24	339
Total	414	275	95	2	138	40	964

In the case of nominative-oblique pronouns (NO ambiguity class), the nominative forms are almost the double of the oblique; the 2^{nd}-person plural *vós* is even more rarely used in this corpus than the in the ADR class; on the other hand, there is a stronger asymmetry for the 1^{st}-person plural *nós* in this class; the 3^{rd}-person pronouns are the largest subset of the NO ambiguity class, and the nominative forms are, in average three times more numerous than the oblique, though the difference is larger in the singular that in the plural forms[6].

Training corpus

A Maximum Entropy-based model was initially trained on the evaluation corpus, using the MegaM tool [47], in order to help build the training corpus. The MegaM tool uses an efficient implementation of conjugate gradient (for binary problems) and limited memory BFGS (for multiclass problems).

To build a training corpus for the ADR ambiguity class, an initial random sample of 1,367 sentences, containing the target pronouns, was collected from a news corpus retrieved from the on-line edition of the daily newspaper *Público*[7]. This corpus is composed of two sets: one set contains news published from January 1999 to December 2004, totaling 189 million words; the second set contains only shorter breaking news, collected from January 2005 and June 2009, and has about 41 million words.

The sample was pre-annotated using the classifier previously built, in order to simplify the manual annotation process.

Though the annotation task may seem trivial, this is not the case, even for trained linguists. The sample was initially given to an annotator (Annotator 1), with an MA degree in Linguistics and a lot of experience in corpus annotation. The classifiers built from this initial training corpus produced very poor results, which hinted at inconsistency in the annotated data. The same sample was thus attributed a second annotator (Annotator 2), who is a Syntax specialist. This annotator corrected almost 448 tags from the first annotator.

[6]Two forms of *consigo* were correctly POS-tagged as the oblique pronoun contracted with preposition *com* (with) but with the wrong number (singular, instead of plural). These should be counted as correct, for the person distinction was not exactly the aim of the paper and only these two cases

Table 5: Training corpus: Confusion matrix for class ADR (1st phase).

		Ann 1		
		acc	dat	ref
Ann 2	acc	380	212	97
	dat	130	513	8
	ref	0	1	26

Table 5 presents the confusion matrix of this initial annotation step.

One can see that the two annotators only marked 67% of the pronouns in the same class (or case). This was due mostly to confusion between accusative and dative pronouns (25%), while the differences in the reflex pronoun were less significative (<8%). After this first annotation effort, a new classifier was built using only the sentences in this sample.

Next, a new sample of 2,087 different, randomly selected sentences was collected from the same corpus as the first sample, and the new classifier used to help the annotation process. The tags of this new sample were then manually revised by Annotator 2, and the two samples constitute then our training corpus, with 3,454 sentences, with the pronouns POS-tags manually corrected. The breakdown of the pronouns and their cases in the training corpus is shown in Table 6.[8]

Table 6: Training Corpus: ADR ambiguity class

	me	*te*	*nos*	*vos*	total
acc	360	303	101	433	1,197
dat	559	269	110	735	1,673
ref	316	186	54	28	584
total	1,235	758	265	1,196	3,454

Comparing Tables 3 and 6, it should be noted that the proportions of pronouns and their cases in each corpus – the evaluation and the training corpus – though similar, is not exactly the same.

The most important difference is the much larger proportion of the 2nd-person plural *vos*, that now represents 35%, while in the evaluation corpus, this pronoun represented only 2% of the instances; next, the 1st-person singular *me* is still the most frequent form of this ambiguity class, but now it represents just 36%, while in the evaluation corpus it constituted 52%; along with these differences, 2nd-person singular *te* and 1st-person plural *nos*, show a drop of 8% and 9% in the training corpus, respectively, when compared with their frequency in the evaluation set.

The relative rank of the different cases is the same, but there are more nominative forms, against a similar reduction in the number of reflexive pronouns. Again, the dative pronouns constitute almost half of the tags (47%), but they are immediately followed by the accusative pronouns, with little less than 35% (10% more than in the evaluation corpus), and then the reflex, with 17% (less 12%).

Since the problem of the nominative-oblique ambiguity seemed relatively simple to model, and, in fact, the models built for the NO ambiguity class yielded very satisfactory results from the onset (see Section 3), it was considered unnecessary to produce more data for training and we used 80% of the initial corpus for training and the remaining 20% for evaluation.

---

were found.

[7] www.publico.pt

[8] One form of *nos* has been parsed as a contraction of *em+o*.

2.4. Features

To investigate the problem of correct case assignment to the two sets of ambiguous personal pronouns, a set of features was automatically extracted from the training corpus, after it was processed by the STRING processing chain.

Tests were made using the full POS-tags of the words produced by STRING. The STRING tagset [48] has 11 fields: CATegory, subcategory (SCT), MODality, TENse, PERson, NUMber, GENder, DEGree, CASe, SYNtactic features, and SEMantic features. Verb tags have 8 fields, as SCT, DEG and CAS are not marked; pronouns have all the fields except MOD, TEN and DEG; and so on. In the STRING's tag set, 12 main POS categories are considered; nouns, pronouns, articles, numerals and conjunctions also have subcategories, *e.g.* nouns are distinguished between proper an common, conjunctions between subordinate and coordinate, etc.

Experiments showed that results improved if the verb tag was divided into different subsets. After this division, the fields DEG, SYN and SEM were shown to have no impact in the models.

Different solutions were found for each ambiguity class.

For the **ADR** ambiguity class (accusative-dative-reflexive pronouns: *me-te-nos-vos*), after several experiments, the best features were achieved, which are based on the target pronoun itself and the two words before and after the pronoun:

- WORD : the form of the pronoun itself;
- LEMMA : the lemma of the verb; in Portuguese, this is the non-inflected/impersonal infinitive form;
- CLITIC? : the clitic use of the pronoun, that is, whether it is attached to a verb or not (values: *yes*/*no*);
- NOTREFLEX? : a feature based on the person-number inflectional values of both the pronoun and the verb, indicating that they are not equal; for example, in *falam-me* 'speak_3rd.pl-Pron_1st.sg', the verb and the pronoun have different person-number values, while in *falamo-nos* 'speak_2nd.pl-Pron_2nd.pl', those values are the same (values: *yes*/*no*);
- PNG : the person-number and gender inflectional features of the verb; Portuguese has 6 person-number combinations per tense and the past participle shows 4 gender-number values;
- VIPER : the syntactic-semantic class of the verb, as encoded in ViPEr (see Section 2.2), each class corresponding to one of the verb word-senses (values: 71 verb syntactic-semantic classes).

For the ambiguity class of **NO** (nominative-oblique) pronouns (*ele-ela-eles-elas-nós-vós*), the best solution involved using the following features, which are based on the target pronoun itself and the words before and after the pronoun:

- WORD : the form of the pronoun itself;
- LEMMAPROX : the lemma of the neighbouring (left and right) words;
- CAT+SUBCAT : the part-of-speech (POS), both the main category and the subcategory, of the neighbouring words;
- PNG : the person-number and gender inflectional features of the verb; Portuguese has 6 person-number combinations per tense and the past participle shows 4 gender-number values;

Besides these, other features were tested and eventually discarded, in view of their negative results, namely the features on the two (for NO ambiguity class) or the three words (for ADR ambiguity class) appearing before and after the target pronoun.

	acc	dat	ref	tot
acc	178	69	61	308
dat	24	329	12	365
ref	6	2	172	180
ref	208	400	245	853

Table 7: ADR: `All` features (P=0.7960)

	acc	dat	ref	tot
acc	150	161	28	339
dat	46	231	27	304
ref	12	8	190	210
tot	208	400	245	853

Table 8: ADR: w/o feature `VIPER` (P=0.6694)

3. Results

Several experiments were set up, which were different for each ambiguity class. Firstly, the experiments with the ADR ambiguity class (*me, te, nos, vos*) are addressed, and then the NO ambiguity class (*ele, ela, nós, vós, eles, elas*)

3.1. ADR ambiguity class

The first scenario, presented in Table 7, tested the model using all features presented in the previous section. In this and in the following Tables, the lines indicate the result of the system, and the columns the values in the evaluation corpus. Precision reached 0.796. While this was not the best result, it is nonetheless close to it.

In the next experiments, each feature was discarded in turn to assess its impact on the performance of the classifier against this initial result.

Since the case of the pronouns is related to the syntactic structure determined by the verb they depend on, in the next scenario, shown in Table 8, the `VIPER` features, indicating the verb syntactic class was discarded.

As expected, a substantial 12.66% precision drop was observed. This confirms that the syntactic class of verb is determinant for the case assignment to pronouns.

As the syntactic-semantic ViPEr class tags of the verb constitute a relatively large set (around 71 classes), different groupings of these classes were also tested, deemed relevant to the classification task at hand. These groupings were based on:

(i) the existence (or not) of complements (purely intransitive constructions);
(ii) the existence (or not) of direct object (accusative) complement; and
(iii) the existence (or not) of lexical, *i.e.* essential, indirect object (dative) complement.

Each grouping constituted a subset of ViPEr classes. The different class subsets never improved the results, neither in the ADR nor in the NO ambiguity classes. The ViPEr feature alone did not improve the classifier in the NO ambiguity class so it was dropped from the solution.

In the next experiment (Table 9), the `NOTREFLEX` feature was ignored. This feature indicates if the person-number value of the verb and the pronoun are the same. In the case of reflexive pronouns, this should always be the case, although many verb forms are ambiguous.

	acc	dat	ref	tot
acc	176	141	131	448
dat	4	219	8	231
ref	28	40	106	174
tot	208	400	245	853

Table 9: ADR: w/o feature NOTREFLEX (P=0.5873)

A precision drop of 20.87% was observed, the largest difference against the All-features scenario. This result confirms the relevance of this feature for the disambiguation task of the ADR class of pronouns.

In the next experiment (Table 10), the word form of the pronoun, given by the feature WORD, was ignored. Notice that the verbs kept the information regarding their lemma, which is another feature that is separately tested below (Table 12). The precision showed a small impact of 0.35% bellow the All-features experiment. Therefore, this feature only marginally contribute to the classifier.

	acc	dat	ref	tot
acc	178	68	53	299
dat	26	331	25	382
ref	4	1	167	172
tot	208	400	245	853

Table 10: ADR: w/o feature WORD (P=0.7925)

Next, we consider the CLITIC feature, which indicates that the pronoun is attached to the verb (eventually fronted, due to certain, reasonably well-known, syntactic contexts). This allows to distinguish pronouns from other homograph words, though not necessarily their case. By removing this feature, one expects to confirm that the similar properties of clitic pronouns would not produce much difference in the classifier. Results in this experiment (Table 11) suggest this feature is more relevant than WORD, as it presents a higher precision drop (2.81%).

	acc	dat	ref	tot
acc	141	66	48	255
dat	49	326	9	384
ref	18	8	188	214
tot	208	400	245	853

Table 11: ADR: w/o feature CLITIC (P=0.7679)

By removing the feature LEMMA of the verb POS, one intends to ascertain whether lemmatization of verb form influences the disambiguation task. In Portuguese, a considerable number of verbal inflected forms are ambiguous and may be ascribed to different verb lemmas [49]. As these different verbs may also have different syntactic constructions, this has a relation to the type of clitic pronouns that co-occur with each verb. Results are shown in Tabletable:adr-wo-lemma.

Results show that the LEMMA feature does have a small impact (3.40%) on the classifier performance.

Finally, results from removing the feature PNG are shown in Table 13. This feature represents a subset of the morphosyntactic features of the entire verb POS-tag, namely the 6 person-number inflection values and, in the case of past-participle, the 4 number-gender combinations.

The PNG feature is the second most significant one for the disambiguation task of the ADR class of pronouns, since its removal caused a drop of 18.64% when compared with the All-features experiment.

	acc	dat	ref	tot
acc	170	93	42	305
dat	31	301	24	356
ref	7	6	179	192
tot	208	400	245	853

Table 12: ADR: w/o feature `LEMMA` (P=0.7620)

	acc	dat	ref	tot
acc	187	92	192	471
dat	19	304	24	347
ref	2	4	29	35
tot	208	400	245	853

Table 13: ADR: w/o feature `PNG` (P=0.6096)

To conclude, Table 14 resumes the results of these experiments. The table also compares the `All-features` experiment against a baseline, which consists in selecting the most frequent case for each pronoun (in this case, the dative).

me-te-nos-vos	**Precision**	**Precision–Baseline**
Baseline	0.4689	–
`All-features`	0.7960	32.71
	Precision	**Precision–`All`**
without `NotReflex`	0.6873	20.87
without `PNG`	0.6096	18.64
without `ViPEr`	0.6694	12.66
without `Lemma`	0.7620	3.40
without `Clitic`	0.7679	2.81
without `Word`	0.7925	0.35

Table 14: Results for ADR ambiguity class *me-te-nos-vos*: synopsis

First of all, results show that the most important features are the `NotReflex` and the `PNG`. These features involve the verb-pronoun person-number agreement, on one hand; and the verb person-number and number-gender, on the other hand. Thus, the inflection values of verb and pronoun alone are responsible for a significant part of the classifier's performance.

Secondly, the syntactic-semantic information on the verb class provided by the `ViPEr` feature, though still important for selecting the adequate case for these ambiguous pronouns, is surprisingly less significative that the former features. Two reasons for this result can be advanced: on the one hand, the performance of the word-sense-disambiguation module [15, 43], at this stage, still does not allows the use of this feature consistently and it is still less than sufficiently effective; and, on the other hand, the previous, best performing features may overlap the effect of the `ViPEr` feature on the data.

Thirdly, by discarding the information on the lemma of the verb, one can test if the case of the pronouns is highly dependent of the verb they appear with. However, the `Lemma` feature seems to contribute little to the performance of the classifier. This justifies the use, for the disambiguation task, of word forms' data, rather than information on lemmatized verbs.

The `Clitic` feature indicating that the pronoun is being used in a clitic position also contributes, but little, to the classifier's performance, while the effect of the information on the `Word` form of the pronoun itself is negligible.

3.2. NO ambiguity class

Concerning the ambiguity class involving nominative and oblique pronouns, similar experiments were carried out, though with a slightly different set of features.

Notice that the total number of elements in the following tables (962) is less than in Table 4 (964). This difference is due to an error in the processing chain when solving the contraction of the number-ambiguous form *consigo* (with_him/her or with_them), as found in the following sentence:

Os representantes dos estudantes levaram ***consigo*** *as suas propostas*
(The representatives of the students took **with_them** their proposals).

In this sentence, the underlying pronoun *eles* (them) had been incorrectly reconstructed by the system as a singular form *ele* (him) instead of the plural. Solving this problem involves resolving long-distance anaphora, namely knowing the gender-number of the subject `NP`, which the processing chain cannot solve at this time. Only two instances of this preposition+pronoun contraction occurred, and they were ignored for the purpose of this paper.

For the experiments with this ambiguity class, we defined two different baselines:

(i) *Baseline 1* consists in selecting the most frequent case from the total number of pronouns in the reference; in this case, three are 625 pronouns in the nominative case, against 339 oblique forms.
(ii) *Baseline 2* consists in selecting for each individual form of the pronouns in this class the most frequent case; in this way, the frequency of nominative forms *ele* (he/him), *ela* (she/her) and nós (we/us) supersedes that of the corresponding oblique forms, while for pronouns *vós* (you_pl), *eles* (they/them_masc) and *elas* (they/them_fem)

Notice that only one baseline was devised for the ADR class, since the dative forms are systematically more frequent that the other two cases.

The first results, shown in Table 15, correspond to the model built using all features of the best solution achieved.

	nom	obl	tot
nom	623	6	629
obl	2	331	333
tot	625	337	962

Table 15: NO:`All` features (P=0.9917)

This results are 34.33% and 30.19% above *Baseline 1* and *Baseline 2*, respectively. This improved performance of the classifier falls within the same range as for the ADR class (32.71%), though the set of features and the baseline value are different.

As in the preceding Section, we proceed to remove one feature at a time, in order to assess its impact in the performance of the classifier. Table 16 shows the confusion matrix and the precision of the model when the `CAT+SUBCAT` feature is removed from the data. As explained before, this feature takes into consideration the main POS category and subcategory of the neighbouring words.

	nom	obl	tot
nom	621	6	627
obl	4	331	335
tot	625	337	962

Table 16: NO: w/o feature `CAT+SUBCAT` (P=0.9896)

Only a small drop of 0.21% is observed. This indicates that the using the POS of the words in the immediate vicinity of the pronoun is not very relevant information for the classifier.

Next, in Table 17, we show the results obtained when the WORD feature is removed. This feature consists in the form of the pronoun itself.

	nom	obl	tot
nom	589	8	606
obl	27	329	356
tot	625	337	962

Table 17: NO: w/o feature WORD (P=0.9636)

A drop of 2.81% is observed, which is a higher value than the one found for the previous feature. These values suggest that feature WORD is more relevant for the classifier, than the feature CAT+SUBCAT. Notice that the same feature was used for modelling the disambiguation of class ADR pronouns, but its removal form the classifier only yield a 0.35% drop in its performance.

Next, in Table 18, we show the results obtained when the LEMMAPROX feature is removed. This feature takes into account the lemma of the pronoun's neighbouring words.

	nom	obl	tot
nom	620	6	626
obl	5	331	336
tot	625	337	962

Table 18: NO: w/o feature LEMMAPROX (P=0.9886)

A drop of 0.31% was observed, which is a higher but still close value to CAT+SUBCAT experiments, as shown in Table 16. This means that both of information on context words is similarly important for the disambiguation task hand. Nevertheless, the information provided by the form of the target pronoun, as illustrated in Table 17, seems much more important (a drop of 2.81%).

Finally, Table 19 shows the effect on the classifier's performance with the removal of the PNG feature. The reader will remember that this feature represents the person-number values of the verb (or the gender-number of the past participles).

	nom	obl	tot
nom	623	7	630
obl	2	330	332
tot	625	337	962

Table 19: NO: w/o feature PNG (P=0.9906)

An even smaller drop of 0.11% is observed. Notice that this had been the second most important feature in the ADR experiments, as shown by a drop of 18.52% in that classifier's performance. Thus, for different disambiguation tasks, as represented by the different classes here considered, the weight of the features may vary significantly. In this case, and once again, separating the morphosyntactic features of the verb POS-tag may yield a more accurate result.

To conclude this section, Table 20 resumes the results of these experiments.

This table also compares the All-features experiment against the two baselines (see Table 4).

Comparing Tables 14 and 20, on the one hand, we notice that the base line for each class of ambiguous pronouns is

ele-ela-nós-vós-eles-elas	**Precision**	**Precision–Baseline**
Baseline 1	0.6483	–
Baseline 2	0.6898	–
`All-features`	0.9917	30.19
	Precision	**Precision–`All`**
without `CAT-SUBCAT`	0.9896	0.21
without `WORD`	0.9636	2.81
without `LEMMAPROX`	0.9886	0.31
without `PNG`	0.9906	0.11

Table 20: Results for NO ambiguity class *ele-ela-nós-vós-eles-elas*: synopsis

significantly higher for the NO class. This, however, could be only a fortuitous feature of the data. Still, it seems to correspond, in fact, to a higher complexity of the problem of modelling the classification task for the ADR class, when compared with the linguistically-motivated rules that can be devised to disambiguate the NO class pronouns.

On the other hand, irrespective of the way we define the two baselines, their values are similar. This is mostly due to the fact that, for Baseline 2, only a small number of cases exist in the data where the oblique case is more frequent than the nominative, thus only marginally affecting the baseline values.

4. Conclusion and future work

While POS tagging is already fairly well solved, the biggest contribution of this work is the disambiguation of personal pronouns, which still cause many POS tagging problems, mainly in highly inflected languages. We have presented a solution that combines Maximum Entropy-based models with rich lexical information in order to improve the POS disambiguation of personal pronouns in Portuguese. It proved to be a suitable solution, yielding almost 100% for the NO ambiguity class (99.17%) and 80% for the ADR ambiguity class (79.60%).

For the ambiguity class of ADR pronouns, some features seem to be very relevant in the disambiguation task. Perhaps the study of others linguistic sophisticated features could improve the final accuracy. Besides, when fully integrated in a NLP processing chain, the personal pronouns disambiguation is influenced by the preceding modules in the chain. A very interesting topic to investigate in further improvements is the order of the personal pronouns disambiguation, among other disambiguation tasks as verbal inflection disambiguation [49].

5. Acknowledgements

We thank Cláudio Diniz for the first experiments. This work was supported by national funds through Fundação para a Ciência e a Tecnologia, ref. UID/CEC/50021/2013.

[1] E. Brill, A simple rule-based part of speech tagger, in: Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics, Trento, Italy, 1992, pp. 152–155. doi:10.3115/974499.974526.
URL http://www.aclweb.org/anthology/A92-1021

[2] T. Brants, Tnt: A statistical part-of-speech tagger, in: Proceedings of the Sixth Conference on Applied Natural Language Processing, ANLC '00, Association for Computational Linguistics, Stroudsburg, USA, 2000, pp. 224–231. doi:10.3115/974147.974178.
URL http://dx.doi.org/10.3115/974147.974178

[3] S. M. Thede, M. P. Harper, A second-order hidden markov model for part-of-speech tagging, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, Association for Computational Linguistics, Stroudsburg, USA, 1999, pp. 175–182. doi:10.3115/1034678.1034712.
URL http://dx.doi.org/10.3115/1034678.1034712

[4] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1 of NAACL '03, Association for Computational Linguistics, Stroudsburg, USA, 2003, pp. 173–180. doi:10.3115/1073445.1073478.
URL http://dx.doi.org/10.3115/1073445.1073478

[5] A. Ratnaparkhi, et al., A maximum entropy model for part-of-speech tagging, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol. 1, Philadelphia, USA, 1996, pp. 133–142.

[6] K. Toutanova, C. D. Manning, Enriching the knowledge sources used in a maximum entropy part-of-speech tagger, in: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Vol. 13, Association for Computational Linguistics, Hong Kong, China, 2000, pp. 63–70.

[7] E. Charniak, A maximum-entropy-inspired parser, in: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000, Association for Computational Linguistics, Stroudsburg, USA, 2000, pp. 132–139.
URL http://dl.acm.org/citation.cfm?id=974305.974323

[8] J. Giménez, L. Màrquez, Svmtool: A general pos tagger generator based on support vector machines, in: In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004, pp. 43–46.

[9] N. Mamede, J. Baptista, C. Diniz, String - an hybrid statistical and rule-based natural language processing chain for portuguese, in: P. . Demos (Ed.), PROPOR 2012, PROPOR, PROPOR, Coimbra, Portugal, 2012.
URL http://www.propor2012.org/demos/DemoSTRING.pdf

[10] N. C. Marques, G. P. Lopes, Using neural nets for portuguese part-of-speech tagging, in: In Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese, 1996, pp. 1–9.

[11] S. Aluísio, J. Pelizzoni, A. R. Marchi, L. de Oliveira, R. Manenti, V. Marquiafável, An account of the challenge of tagging a reference corpus for brazilian portuguese, in: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language, PROPOR'03, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 110–117.
URL http://dl.acm.org/citation.cfm?id=1758748.1758769

[12] C. Nogueira Dos Santos, R. L. Milidiú, R. P. Rentería, Portuguese part-of-speech tagging using entropy guided transformation learning, in: Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language, PROPOR '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 143–152.
URL http://dx.doi.org/10.1007/978-3-540-85980-2_15

[13] E. B. Williams, From Latin to Portuguese: Historical Phonology and Morphology of Portuguese Language, University of Pennsylvania Press, Oxford, 1938.

[14] Z. S. Harris, A Theory of Language and Information. A Mathematical Approach, Clarendon Press, Oxford, 1991.

[15] T. Travanca, Verb sense disambiguation, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa (2013).

[16] R. Talhadas, Semantic role labelling in european portuguese, Master's thesis, Universidade do Algarve/FCHS, Faro, Portugal (2013).

[17] J. S. Marques, Anaphora resolution, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa., Lisboa (2013).

[18] C. Diniz, N. Mamede, Lexman - lexical morphological analyser, Tech. rep., L2F / INESC ID Lisboa, Lisboa (2011).

[19] C. Diniz, Rudrico2 - um conversor baseado em regras de transformação declarativas., Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa (2010).

[20] C. Diniz, N. Mamede, J. D. Pereira, RuDriCo2 - a faster disambiguator and segmentation modifier, in: Simpósio de Informática - INForum, Universidade do Minho, Portugal, 2010, pp. 573–584.

[21] A. M. F. Vicente, Lexman: um segmentador e analisador morfológico com transdutores, Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisboa (June 2013).

[22] J. Baptista, N. Mamede, F. Gomes, Auxiliary verbs and verbal chains in european portuguese, in: T. Pardo, A. Branco, A. Klautau, R. Vieira, V. L. S. de Lima Tiago Pardo, A. Branco, A. Klautau, R. Vieira, V. L. S. de Lima (Eds.), Computational Processing of the Portuguese Language, no. 6001 in Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence, PROPOR 2010, Springer, Berlin, 2010.

[23] J. Loureiro, Reconhecimento de entidades mencionadas (obra, valor, relações de parentesco e tempo) e normalização de expressões temporais, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, mSc Dissertation (2007).

[24] C. Hagège, J. Baptista, N. J. Mamede, Reconhecimento de entidades mencionadas com o xip: Uma colaboração entre o inesc-l2f e a xerox, in: C. Mota, D. Santos (Eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM (Aveiro, 11 de Setembro de 2008), Linguateca, 2009.

[25] C. Hagège, J. Baptista, N. J. Mamede, Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo harem e o futuro, in: C. Mota, D. Santos (Eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM (Aveiro, 11 de Setembro de 2008), Linguateca, 2009.

[26] J. Hagège, Caroline; Baptista, N. J. Mamede, Caracterização e processamento de expressões temporais em português, Linguamática 2 (1) (2010) 63–76.

[27] D. Oliveira, Extraction and classification of named entities, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, mSc Dissertation (2010).

[28] J. Baptista, D. Oliveira, D. Santos, N. J. Mamede, Classification directives for named entities in portuguese texts, Tech. rep., L2F-Spoken Language Laboratory (2011).

[29] C. Hagège, J. Baptista, N. Mamede, Portuguese temporal expressions recognition: from te characterization to an effective ter module implementation, in: STIL'2009 - 7th Brazilian Symposium in Information and Human Language Technology, São Carlos, Brazil, 2009.

[30] A. Maurício, Identificação, classificação e normalização de expressões temporais, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa (November 2011).

[31] N. Nobre, Resolução de expressões anafóricas, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, mSc Dissertation (2011).

[32] F. Carapinha, Extração automática de conteúdos documentais, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, mSc Dissertation (2010).

[33] D. Santos, N. J. Mamede, J. Baptista, Extraction of family relations between entities, in: L. Barbosa, M. Correia (Eds.), INForum 2010 - II

Simpósio de Informática, Universidade do Minho, Braga, Portugal, 2010, pp. 549—560.
[34] N. J. M. Baptista, Jorge; Vera Cabarrão, Classification directives for events and relations extraction between named entities in portuguese texts, Tech. rep., L2F-Spoken Language Laboratory (2012).
[35] R. Talhadas, N. Mamede, J. Baptista, Semantic roles portuguese verbs, in: J. Baptista, M. Monteleone (Eds.), Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013), CLG'2103, Universidade do Algarve – FCHS, Faro, Portugal, 2013, pp. 127–132.
[36] V. Cabrita, Identificar, ordenar e relacionar eventos, dissertation project, MSc Electronic Engenneering, IST and INESC-ID Lisboa/L2F (September 2012).
[37] S. Ait-Mokhtar, J. Chanod, C. Roux, Robustness beyond shallowness: incremental dependency parsing, Natural Language Engineering 8 (2/3) (2002) 121–144.
[38] J. Baptista, Viper: A lexicon-grammar of european portuguese verbs, in: J. Radimsky (Ed.), Proceedings of the 31st International Conference on Lexis and Grammar, Università degli Studi di Salerno (Italy)/University of South Bohemia in Nové Hrady (Czech Republic), Università degli Studi di Salerno (Italy)/University of South Bohemia in Nové Hrady (Czech Republic), Nové Hrady (Czech Republic), 2012, pp. 10–16.
[39] J. Baptista, Viper: uma base de dados de construções léxico-sintáticas de verbos do português europeu, in: Actas do XXVIII Encontro da APL, APL, Faro, Portugal, 2012.
[40] P. Rocha, D. Santos, Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa, in: Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000, Atibaia, são Paulo, Brasil, 2000, pp. 131–140.
[41] M. Gross, Lexicon-grammar, in: K. . Brown, J. Miller (Eds.), Concise Encyclopedia of Syntactic Theories, Pergamon, Cambridge, 1996, pp. 244–259.
[42] C. Leclère, Organization of the lexicon-grammar of french verbs, Linguisticae Investigationes 25-1 (2002) 29–48.
[43] G. Suissas, Verb sense disambiguation, Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, Lisboa (2014).
[44] M. Nascimento, J. Bettencourt, P. Marrafa, R. Ribeiro, R. Veloso, L. Wittmann, Le- parole – do corpus à modelização da informação lexical num sistema multifunção., in: Actas do XIII Encontro da APL, APL, Lisboa, Portugal., 1997.
[45] R. Ribeiro, Anotação morfossintáctica desambiguada em português, Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, mSc Dissertation (2003).
[46] J. Baptista, A lexicon-grammar of european portuguese verbs, in: 31st International Conference on Lexis and Grammar, Institut Gaspard-Monge, Marne-la-Vallée,Paris (France)/Institut de Langues Romanes, University of South Bohemia in Nové Hrady (Czech Republic), Nové Hrady, Czech Republic, 2012.
[47] H. I. Daumé. Notes on cg and lm-bfgs optimization of logistic regression [online] (2004).
[48] N. J. Mamede, J. Baptista, C. Hagège, Nomenclature of chunks and dependencies in portuguese xip grammar 4.0, Technical report, L2F-Spoken Language Laboratory, INESC-ID Lisboa, Lisboa (May 2013).
[49] J. C. L. Pinto, Fine-grained POS-tagging: Full Disambiguation of Verbal Morphosyntactic Tags, Dissertation project, Universidade de Lisboa – Instituto Superior Técnico/INESC-ID Lisboa – Spoken Language Laboratory, Lisboa (December 2015).