Building a Dictionary of Anthroponyms

J. Baptista [1,2], F. Batista [1,3], N. Mamede [1,4]

[1] L2F – Laboratório de Sistemas de Língua Falada - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
jbaptis@ualg.pt, {Fernando.Batista, Nuno.Mamede}@inesc-id.pt
http://www.l2f.inesc-id.pt/
[2] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
Campus de Gambelas, P – 8005-139 Faro, Portugal
[3] ISCTE – Instituto de Ciências do Trabalho e da Empresa
Av. Forças Armadas, 1649-026 Lisboa, Portugal
[4] Instituto Superior Técnico - Universidade técnica de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa

Abstract. This paper presents a methodology for building an electronic dictionary of anthroponyms of European Portuguese (DicPRO), which constitutes a useful resource for computational processing, due to the importance of names in the structuring of information in texts. The dictionary has been enriched with morphosyntactic and semantic information. It was then used in the specific task of capitalizing anthroponyms and other proper names on a corpus automatically produced by a broadcast news speech recognition system and manually corrected. The output of this system does not offer clues, such as capitalized words or punctuation. This task expects to contribute in rendering more readable the output of such system. The paper shows that, by combining lexical, contextual (positional) and statistical information, instead of only one of these strategies, better results can be achieved in this task.

1 Introduction

The recognition of proper names in texts is a recurring problem in different domains of Natural Language Processing (NLP) such as Information Retrieval, Information Extraction, Machine Translation, Syntactic Analysis, Named Entity Recognition and, for example, in actor's identification in the discourse for dialogue systems [1,2,3,4]. Conventionally, the identification of proper names in an untagged text is carried out by using a linguistic [1] or a probabilistic approach [3,5,6]. In each case, a dictionary containing information about this type of words may constitute an important NLP resource for automatic corpora processing [7].

It may strike one as unusual to create a dictionary of proper names because, as opposed to words of the common lexicon, they are regarded as constituting a (potentially) infinite set. However, certain subclasses can be closed. For example, most *first names* (or *given* or *Christian names*)[1] can be intuitively identified whereas *last names* (or *surnames*, or *family names*) are more often ambiguous with words of the common lexicon. Several kinds of proper names can be considered: anthroponyms (names of persons) and institutions' names [8]; toponyms (names of locations), hidronyms (names of rivers, lakes, seas and oceans) and other geographical accidents (mountains, chains, isles, and so on) [1]; ergonyms (manufactured objects and products), commercial names; siglae and acronyms [8]; and several others. Each one of these subclasses poses specific representation and recognition problems, such as their internal structure, multilexical forms and lexical ambiguity.

This paper presents an electronic dictionary of anthroponyms (DicPRO), which constitutes a resource for computational processing of texts in European Portuguese. In the future, we aim at treating and including other proper names' subclasses, allowing for a broader use of the resource.

The paper is structured as follows: the next section resumes the main linguistic features that can be associated to anthroponyms. Section 3 presents the methodology applied in the building of the DicPRO. Section 4 describes an experiment to assess the usefulness of this resource, in the specific task of capitalization of proper names. This was done by applying the dictionary to a corpus of broadcast news speech recognition system [9], automatically produced and then manually corrected. Usually, the output of the system does not offer formal clues, such as capitalized words and punctuation, which are otherwise used to identify proper names. We then compare this strategy with the alternative approaches of using only probabilistic methods for capitalizing proper names, using contextual (positional) rules based on lexical information made available by DicPRO and, finally, by combining together these strategies. The paper ends with some conclusions and perspectives for future work.

2 Anthroponyms

Anthroponomy is a subdomain of onomastics that deals mainly with the formation of personal names. Anthroponyms are of particular syntactic and semantic interest in linguistics, since they show special referential values and determination/modification constraints, and particular discourse values [10, 11, 12, 13]. This paper, however, will only focus on the most relevant linguistic features that should be encoded in an electronic dictionary of anthroponyms in view of their automatic recognition in texts. Furthermore, our approach will mainly consider the recognition of anthroponyms based on internal evidence [3], that is, without reference to contextual clues. We will see, however, that contextual (positional) and orthographic/probabilistic information can be used *in combination* to improve results in the specific task of capitalization that we have in mind.

As far as (European) Portuguese is concerned, it is possible to structure this class of names in two major subsets: *first names* and *last names*, each having different

[1] This terminology is somewhat inadequate, when dealing with non-Christian cultures (e.g. Muslim, Hebraic) or with cultures where surnames precede given names (e.g. Chinese, Korean). Naming conventions vary according to culture but, for the purpose of this paper, we center on Portuguese naming conventions.

morphosyntactic properties. However, many names may be used both as first and last name: *Rosa*, even if one of the uses is often more frequent than the other.

First names can be marked with the gender property (*João*, masc./ *Sara*, fem.). In some cases they can present a diminutive form, which may be derived from the deletion of syllables: *Tó* (=*António*) or from affixation of diminutive suffixes: *Carlinhos* (=*Carlos*) or even by combining both processes together: *Nelinha* (=*Manuela*). Last names do not have gender and usually do not admit the formation of diminutives. In this paper, however, diminutives were not considered.

The information about gender is relevant for the syntactic processing, for example, anaphora resolution, and should be taken into consideration when building a resource such as DicPRO. The information about diminutives may also be of pragmatic importance, since they may be used in several ways in discourse, to express affection, irony, familiarity, or other values.

Anthroponyms in texts only seldom show plural inflection. First names also differ from last names in this respect since they can often take a plural morpheme (usually the *–s* ending: *António*, sing./ *Antónios*, pl., cases like: *João*, sing./ ?*Joões*, pl., being rather rare and barely acceptable) while last names, even if they may (rarely) take plural endings: *o Silva* / *os Silvas* (the_sing. Silva / the_pl. Silva_*s*), they may also refer to a group of persons (usually a family) without any explicit marking (*os Silva*, the_pl. Silva). In this case, number of last name can only be ascertained from the preceding article. In the current state of the dictionary, plural was not considered.

The naming of a person is frequently made by combining several anthroponyms, eventually using connectors such as *e* (and) or *de* (of); preposition *de* can also be contracted with definite articles: *do*, *da*, *dos*, *das* (of+the) or, in special cases, be abbreviated as *d'* (of). These combinations are governed by culturally determined rules. Several aspects of the automatic syntactic analysis require the full identification of these sequences of proper names, in order to assign them the status of a named entity. However, this analysis may be error prone; for example, the conjunction may be ambiguous between a compound proper name (*Vale e Azevedo*, former president of a football club) or two coordinated noun phrases (*Soares e Cavaco*, two candidates for Presidency).

Finally, a significant percentage of DicPRO proper names (about 43%, mainly last names) are also ambiguous with words of the common lexicon, e.g. *Figo*. This ambiguity leads to an additional difficulty in recognizing the sequence used in naming a person, but it can be (at least partially) solved by means of disambiguating local grammars [6, 14].

3 Building the dictionary

The anthroponyms dictionary was constructed semi-automatically, by selecting word forms from institutional official lists of proper names. This approach consisted of selecting words from two different types of sources: (a) a **List1** of isolated words collected from the phone directory of a Portuguese telephone company [15], and (b) a **List2** of complete person names, collected from lists of university students.

Each word form in List1 included some additional information such as: (i) its frequency in the entire phone directory and the indication of its occurrence as a first (F) or as a last (L) name. From the approximately 10,000 different forms having frequency higher than 10, a manual selection was made, resulting in about 4,500 anthroponyms. The classification of each name as first or last was then manually verified and (eventually) corrected, and first names were given their gender values.

List2, containing approximately 8,100 complete person names, was first processed manually, by establishing the border between first (= given) names and last names (= surnames). This first step allowed us to unequivocally determine the frequency of use of each word as first or/and last name. To all new names, that is names not yet present in the first list, gender, as well as other information was added, in particular, indication of foreign names: *Yuri*, and orthographic variants: *Melo* vs. *Mello*.

The resulting dictionary integrates, in its current state, approximately 6,200 different entries. In addition to the previously described process, each word in the dictionary was checked against a lexical resource of common lexicon entries [16], and all ambiguous words were marked (Amb). Table 1 characterizes the current content of DicPRO.

Table 1. DicPro content

Total Entries:	6,173	
from List1:	4,533	73.5 %
from List2: (not in List1)	1,640	26.5 %
(already in List1)	1,756	
F: first names (only):	1,870	30.3 %
L: last names (only):	4,200	68.0 %
names both F and L:	103	1.7 %
ambiguous :	2,629	42.6 %
ambiguous F	468	7.0 %
ambiguous L	2,196	35.0 %
ambiguous both F and L	35	0.6 %

The entries of DicPRO were formatted using the DELA formalism for INTEX [14]. Some entries are shown below:

```
Alegre,Alegre.N+Hum+Npr+CAP+L+Amb+L1
Gil,Gil.N+Hum+Npr+CAP+L+F+L1:ms
Tó,António.N+Hum+Npr+Dim+CAP+F+Amb:ms
```

From the information encoded in the dictionary, it is possible to remark that last names constitute the majority of DicPRO entries, and thus this set is likely to be expanded. On the other hand, first names are likely to constitute a more close set, not easily subject to further expansion. The number of proper names functioning simultaneously as first and last names may be considered residual. Secondly, ambiguous anthroponyms constitute an important part of DicPRO, especially last names, half of which are ambiguous with common lexicon entries.

To deal with orthographic variation[2] of proper names (for example, the use of voiceless consonants: *Ba**p**tista/Batista*, *Vi**c**tor/Vítor* or double consonants: *Estre**ll**a/Estrela*, *Ma**tt**os/Matos*; use of digraphs instead of simple consonants: *So**ph**ia/Sofia*, ***Ath**aíde/Ataíde*; use of *y* instead of *i*: *Heytor/Heitor*), a set of enhanced morphological finite-transducers (FSTs) were built using INTEX linguistic development platform. These FSTs take an input string (an orthographic variant of a proper name) and validate them against the DicPRO entries (lemmas), while undoing the formal variation that gave rise to the variant and thus producing the lemma. Fig. 1 illustrates some variations described by these FSTs.

At this stage, only simple words, i.e., one-word names have been included in the dictionary, thus ignoring several compound names (e.g. *Corte-Real*, *Mil-Homens*, *Sim–Sim*, *Vila-Real*), seldom appearing in texts.

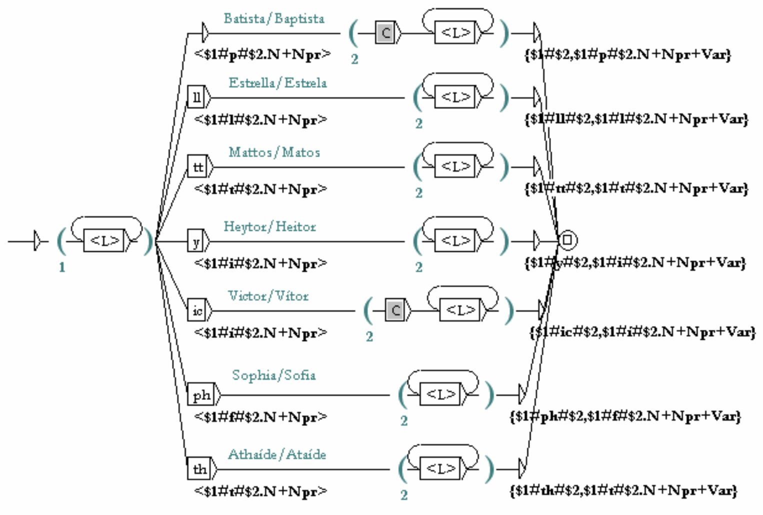

Fig. 1. A finite-state transducer to recognize and lemmatize orthographic variants of proper names.

4 An Experiment on Capitalization

In order to assess the usefulness of the DicPRO resource, several experiments were carried out on the specific task of capitalization of proper names. Two subtasks were

[2] In some cases, these are not exactly variants, but correspond to former (often etymologically derived) spellings of the name, in previous orthographic conventions of Portuguese, which may coexist with the new spellings. Some of these spellings, however, should not be considered ‘orthographic’ variants, but rather a new and somehow increasingly fashionable spelling of proper names.

considered: *subtask 1* – only evaluates the capitalization of anthroponyms; *subtask 2* – evaluates the capitalization of all proper names, regardless of their anthroponymic status.

4.1 Methods

Corpus. The experience was carried out by applying the dictionary to a corpus of broadcast news speech recognition system [9], automatically produced and then manually corrected. Each proper name (anthroponyms and other) in the corpus has been capitalized and it was preceded by the sign '^'. Usually, the system's output does not offer any formal clues, such as capitalized words and punctuation, which are otherwise used to identify proper names. This fact results in a less then optimal reading quality of the output, that the capitalization task is intended to improve.

The corpus contains approximately half million words, and was divided in two subcorpora: (i) a *training corpus* with about 498,000 (27,513 different) words; and (ii) an *evaluation corpus* with 49,306 (7,513 different) words. Anthroponyms were then distinguished from other proper names in the evaluation corpus, by manually replacing the sign '^' by '#'. The following is a small sample of the evaluation corpus (anthroponyms and other proper names have been highlighted in bold)[3]:

Jornal Dois, a informação com **#Manuel #Menezes**.
Boa noite.
A Comissão Europeia decidiu pedir a **^Portugal** que explique alguns aspectos do traçado da auto-estrada do **^Algarve**. Em causa está o projectado troco da ~A dois, que atravessa a zona de protecção especial de **^Castro ^Verde**, e que poderá constituir uma violação da directiva comunitária sobre protecção das aves selvagens.

The evaluation corpus contains 3,001 proper names, of which 1,101 are anthroponyms.

P(robabilistic)-Dic. In order to determine how much the DicPRO might improve the capitalization task of proper names, as compared with orthographic probability of a given word to be written in upper case, we produced a probabilistic dictionary from the training corpus. Each word of this corpus was assigned a capitalization probability depending on how many times it appeared in upper case (CAP) or in lower case (MIN) form. A word is given the CAP tag if it had >50% number of occurrences in capitals; else, the MIN tag was accorded. The list of word forms of the training corpus with this probabilistic information constitutes the P(robabilistic)-Dic. 15% of P-Dic forms were tagged with CAP. P-Dic covers about 83% of the word forms of the evaluation corpus. CAP or MIN information was also added to the DicPRO entries in order to enrich the resource. Table 2 resumes the content of P-Dic.

[3] Notice that the proper name *Castro* in the compound toponym *Castro Verde* has not been given the anthroponyms tag (#), even if it can also be used as such.

Table 2. P-Dic content

Total Entries:	25,528 [4]	
CAP:	3,822	15.0 %
MIN	21,706	85.0 %
also in DicPRO	1,373	
CAP	899	65.5 %
MIN	474	34.5 %

Local Grammars. A set of local grammars were built combining positional and lexical information[5]. These automata may be viewed as rules to recognize sequences of words (and eventual connectors), that are likely to be proper names and should therefore be spelled in upper case. In the scope of this paper, a set of 10 rules were considered. Every rule can be used in a standalone mode, however, the idea is to build an integrated grammar, combining all the separated rules, which will provide the best result. Fig. 2 illustrates two examples of these rules.

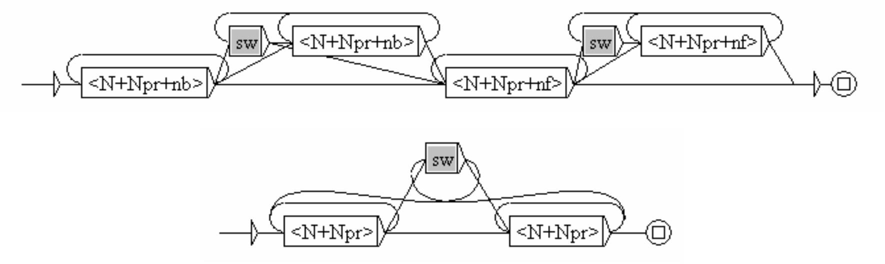

Fig. 2. Rules 8 and 9. Two finite-state automata to recognize sequences of words, candidates for proper names. The first rule (rule 8) identifies sequences of at least one first name: `<N+Npr+F>` followed by at least one last name: `<N+Npr+F>`; eventual connectors – but not conjunction e (and) – are represented by the auxiliary graph (grey box) `sw` (=stopword). The second rule (rule 9) is less specified: it identifies any sequence of proper names (and any eventual connectors), regardless of their F or L tags

4.2 Results and Discussion

Several experiments were conducted in order to find the best way of identifying: anthroponyms and proper names in general. In these experiments, different methods of capitalization were compared, namely: a) using only the DicPRO information (experiment 1); b) using only probability information regarding the use of upper case in a training corpus (experiment 2); c) using the DicPRO with contextual (posi-

[4] Difference between number of entries of P-Dic and the number of different word forms of the training corpus is due to different tokenization criteria, namely, forms with hyphen were kept together in P-Dic.

[5] These local grammar were adapted from [17].

tional) information (experiments 3, 4 and 5); d) combining the different methods (experiments 6 and 7). Results are shown in Table 3.

Table 3. Results obtained from 7 different experiments on identifying anthroponyms and proper names. F = F-Measure, MaxF = F-measure for the two subtasks' F-measures

	subtask 1 (anthroponyms)			subtask 2 (proper names)			
Experiment	Prec.	Recall	F	Prec.	Recall	F	Max F
1. <N+Npr>	32,6%	79,3%	0,462	60,1%	53,6%	0,566	0,509
2. <WORD+CAP>	30,3%	79,9%	0,439	72,7%	70,4%	0,715	0,544
3. Rule 8	86,6%	51,0%	0,641	97,3%	20,5%	0,338	0,443
4. Rule 9	70,5%	65,5%	0,679	92,2%	31,4%	0,468	0,554
5. Rules 1-9	63,6%	74,8%	0,687	93,6%	40,4%	0,564	0,619
6. Rule 10	58,4%	69,5%	0,634	89,2%	39,0%	0,542	0,585
7. Rules 1-10	30,5%	87,0%	0,451	71,8%	75,2%	0,734	0,559

The first two experiments help define precision and recall baseline values for the two main methods of capitalization, namely the separate use of DicPRO against the separate use of P-Dic. For subtask 1, their results are approximately equivalent, even if the P-Dic shows a slightly better F-measure. Overall, a similar baseline precision of ±30% and ±80% recall can be expected for both methods in this subtask. However, for subtask 2, as expected, P-Dic achieves much better results since the lexical coverage of DicPRO is limited to anthroponyms.

Experiments 3, 4 and 5 illustrate the combined use of DicPRO with several contextual (positional) rules; here the separate results of the best performing rules (rules 8 and 9, shown in Fig. 2) are given, while experiment 5 shows the result of the combination of all rules. It is clear that, for subtask 1, the combined use of contextual rules and the DicPRO achieves much better results, compared to the two baseline experiments. Previous rules (rules 1 to 7), not shown here, were highly specific having very high precision but very low recall. Each rule seems to capture different combinatorial phenomena, but experiment 5, which is the conjunction of all rules, achieves a better F-measure. For subtask 2, these experiments achieved the best precision results, at the cost of lowering recall. Nevertheless, F-measure of experiment 5 is only slightly less that that of experiment 1. Still, one should bear in mind that DicPRO information only regards anthroponyms, thus a low recall on subtask 2 should be expected.

Experiment 6 combines the three methods, by using rule 10, shown in Fig. 3.

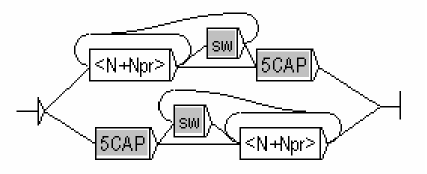

Fig. 3. Rule 10 - Combining DicPRO, P-Dic and contextual rules. A finite-state automata to recognize sequences of (up to 5) words not included in DicPRO but otherwise marked as CAP on the P-Dic, followed or preceded by anthroponyms

In both subtasks 1 and 2, adding probabilistic information does seem to improve results significantly, as compared to experiments 3 to 5. However, it outperforms precision of both experiments 1 and 2 in subtask 1, while showing a lower recall.

Experiment 7, combines experiments 2, 5 and 6. In both subtasks 1 and 2, this experiment attains the best performance in recall (87% and 75%, respectively), even if precision is equivalent to the baselines defined in experiments 1 and 2. This experiment also gives the best F-measure for subtask 2.

The rightmost column calculates the F-measure for both subtasks' F-measures. From these values we can conclude that using DicPRO together with contextual rules is a good choice for identifying anthroponyms, while giving a good precision in detecting proper names. Furthermore, the recall value obtained on the subtask 2 constitutes a base value, which may be subsequently improved with other methods. The introduction of statistical information of P-Dic made possible to obtain better results in subtask 2, but at the cost of impoverishing results of subtask 1, which was the main purpose for building DicPRO and the motivation for this experiment.

5 Conclusions and Future Work

Building lexical databases for proper names has not been the main approach in NLP to deal with this kind of linguistic elements. One of the reasons for this is the general assumption that the set of proper names is potentially infinite. This may not be exact for all classes of proper names, and most probably is not as far as anthroponyms (especially first names) are concerned.

This paper described a methodology employed in the building of an electronic dictionary of proper names (DicPRO). The paper evaluated the usefulness of this resource in a specific task: the capitalization of anthroponyms and proper names in general. The main purpose of this task was to improve the reading quality of the output of a speech recognition system, where every word, including proper names appears in lower case. Of course, the usefulness of this new tool could be extended to other scenarios where information regarding proper names is lacking or has been removed and needs to be restored (automatic spellchecking, named entities recognition, just to cite a few).

We compared and combined several approaches, namely, the use of a probabilistic dictionary (P-Dic) based on the use of upper case in a training corpus, and the use of contextual rules based on the information of DicPRO. Results consistently show improved results benefiting from the use of DicPRO.

Nevertheless, we expect to get better results applying automatic learning techniques like decision lists, such as described in [5,6], which are reported to achieve much better results for problems of similar nature. Hopefully, this will also help to enlarge the dictionary.

Furthermore, we intend to expand DicPRO, possibly by making use of extant lists of names and of lists of recognized named entities, the latter already made available for Portuguese in recent Named Entity Recognition (NER) evaluation contest, HAREM [6]. DicPRO should also evolve in order to encompass other types of proper names (toponyms, hidronyms and the like), and to integrate both simple and compound forms. By the expansion of the DicPRO coverage, we expect to apply this tool to the NER task in the near future.

References

1. Fourour, N., Morin, E, Daille, B.: Incremental recognition and referential categorization of French proper names. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), vol. III, pp. 1068-1074, (2002)

2.Traboulsi, H.: A Local Grammar for Proper Names. MPhil Thesis. Surrey University (2004)

3. McDonald, D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In: Boguraev, B., Putejovsky, J. (eds.): Corpus Processing for Lexical Acquisition. pp. 61-76. MIT Press, Cambridge, Mass.(1993)

4. Friburger, N., Maurel, D.: Finite-state transducer cascades to extract named entities in texts. Theoretical Computer Science, Volume 313(1): 93-104 (2004)

5. Yarowsky, D.: Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. Proceedings of ACL'94, pp. 88-95 (1994).

6. Yarowsky, D. Hierarchical Decision Lists for Word Sense Disambiguation. Computers and the Humanities, 34(1-2): 179-186. 2000

7. Piton, O., Maurel, D. : Les noms propres géographiques et le dictionnaire PROLINTEX. C. Muller, J. Royauté e M. Silberztein (eds.) INTEX Pour la linguistique et le traitement automatique des langues. Cahiers MSH Ledoux 1, pp. 53-76. Presses Universitaires de Franche-Comté: Besançon (2004).

8. Moura, P.: Dicionário electrónico de siglas e acrónimos. MSc Thesis, Faculdade de Letras da Universidade de Lisboa (2000, unpublished).

9. D. Caseiro, I. Trancoso: "Using dynamic wfst composition for recognizing broadcast news", in proc. ICSLP '2002, Denver, Colorado, EUA (2002).

10. Marie-Noël Gary-Prieur (ed.) Syntaxe et sémantique des noms propres. Langue Française 92. Larousse : Paris (data).

11. S. Leroy. Le nom propre en français. Paris: Ophrys (2004).

12. Jean Molino (ed.) Le nom propre. Langue Française 66.: Paris: Larousse (data)

13. Anderson, J.: On the Grammar of names. (to appear in Language 2004/05)

14. Silberztein, M. Dictionnaires électroniques et analyse automatique de texts. Le système INTEX. Masson, Paris (1993)

15. Trancoso, I.: "The ONOMASTICA Inter-Language Pronunciation Lexicon" Proceedings of EUROSPEECH'95 - 4th European Conference on Speech Communication and Technology - Madrid, Spain, September 1995.

16. Ranchhod, E., Mota, C., Baptista, J.: A Computational Lexicon of Portuguese for Automatic Text Parsing. SIGLEX-99: Standardizing Lexical Resources, pp. 74-80. ACL/Maryland Univ., Maryland (1999)

17. Baptista, J.: A Local Grammar of Proper Nouns. Seminários de Linguística 2: pp. 21-37. Faro: Universidade do Algarve (1998).

[6] http://www.linguateca.pt/HAREM/