UNIVERSIDADE FEDERAL DE SÃO CARLOS

USP ICMC

unesp UNIVERSIDADE ESTADUAL PAULISTA

Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora

Oto Vale, Arnaldo Candido Jr., Marcelo Muniz, Clarissa Bengtson, Lívia Cucatto, Gladis Almeida, Abner Batista, Maria C. Parreira, Maria Tereza Biderman, Sandra Aluísio

Federal University of São Carlos (UFSCar), University of São Paulo (USP), State University of São Paulo (UNESP)

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

This work is supported by the Brazilian National Research Council

Agenda

- Introduction to the Historical Dictionary of Brazilian Portuguese (HDBP) project
  - The HDBP corpus
  - Challenges to process our historical corpus
  - Corpus processing tools: Philologic & Unitex and DELA formalism
- **Our work**: the hybrid process to build a dictionary of abbreviations with semantic information (predefined set of named entities – NEs)
  - Printed resources: Flexor (1991) dictionary
  - Generic patterns to extract **different categories of NEs** for an abbreviation
  - Specific patterns to extract **new NEs of a given category** from HDPB corpus
- Building the dictionary: difficulties
  - To indentify and classify the NEs
  - To make the dictionary/gazetteer publicly available: solutions

Hidden Agenda of our work

- To motivate students (mainly undergrads) to work with Portuguese historical corpora

Oto Vale, Arnaldo Candido Jr., Marcelo Muniz, Clarissa Bengtson, Lívia Cucatto, Gladis Almeida,
Abner Batista, Maria C. Parreira, Maria Tereza Biderman,
Sandra Aluísio (CL coordinator)

- They are exposed to a large project – HDPB. Our team comprises 18 senior researchers
- They have access to the resources, methods and corpus processing tools being developed
- They tackle a challenging problem: abbreviations in historical corpora
- They are required to create **resources** for important NLP tasks - **Named Entity* Delimitation and Classification** according to a predefined set

*entities that may be indentified by a proper name plus numeric and time references

HDBP project

- It's a three-year project (2006-2008)

- Integrated network of researchers from various regions of Brazil and Portugal:
  - 11 universities, 18 senior researchers, 23 students

- **Main purpose**: To fill a gap in Brazilian culture,
  - for it is developing a dictionary that describes the **vocabulary of Brazilian Portuguese** in the beginning of the country's history.

- The *Historical Dictionary of Brazilian Portuguese* (*HDBP*):
  - Is based on corpus from the 16 through the 18 centuries (some texts from the beginning of the 19 century).

HDPB corpus

- Texts from 1500 – 1808:
  - Portuguese monarchy transferred the government of the Portuguese empire to Brazil; no press agencies in colonial Brazil

- The texts in the *HDBP* corpus:
  - written by Brazilian authors and Portuguese authors who have lived in Brazil for a long time.
  - collected in public archives and libraries all over Brazil and in Portugal.
  - published texts with minor editing, since the emphasis is on word meaning.

- Text types:
  - letters of Jesuit missionaries,
  - reports of Brazilian explorers,
  - documents of the Inquisition, etc.

HDPB corpus

- Corpus size: 2,458 texts; 287,570 sentences;
  - ~ 16 million tokens (of which 368,850 are different from each other);
  - ~ 7,5 million simple forms (letters of the Historical Portuguese alphabet)

Data	XVIth	XVIIth	XVIIIth	XIXth
Texts (%)	6.24	26.39	59.78	7.59
Sentences (%)	6.30	18.32	64.34	11.04
Simple Forms (%)	7.60	20.18	62.57	9.65
Megabytes (%)	7.23	19.95	63.09	9.73

Table 1: Distribution of texts by century

Challenges to process the HDPB corpus

- Frequent problems (Rydberg-Cox, 2003; Sanderson, 2006):
  - common words and word-endings are **abbreviated with non-standard typographical symbols**
  - **Broken words** at the end of lines are not always hyphenated
  - **Word breaks** are not always used
  - **Uncommon typographical symbols** also in non-abbreviated words

Footnotes and spelling variations; capitalized common words

ções que lhe insinamos, e nom parece honesto estarem nuas
entre os christãos na igreja, e quando as insinamos. E disto
peço ao P.e M. João [21] tome cuidado, por elle ser parte na
conversão destes gentios, e nom fique senhora nem pessoa
a que nom importune [5r] para cousa tam sancta; e a isto se
avião de applicar todas as restituições que lá se ouvessem
de fazer, e isto agora soomente no começo que elles farão
algodões para se vestirem [illegible]o diante.

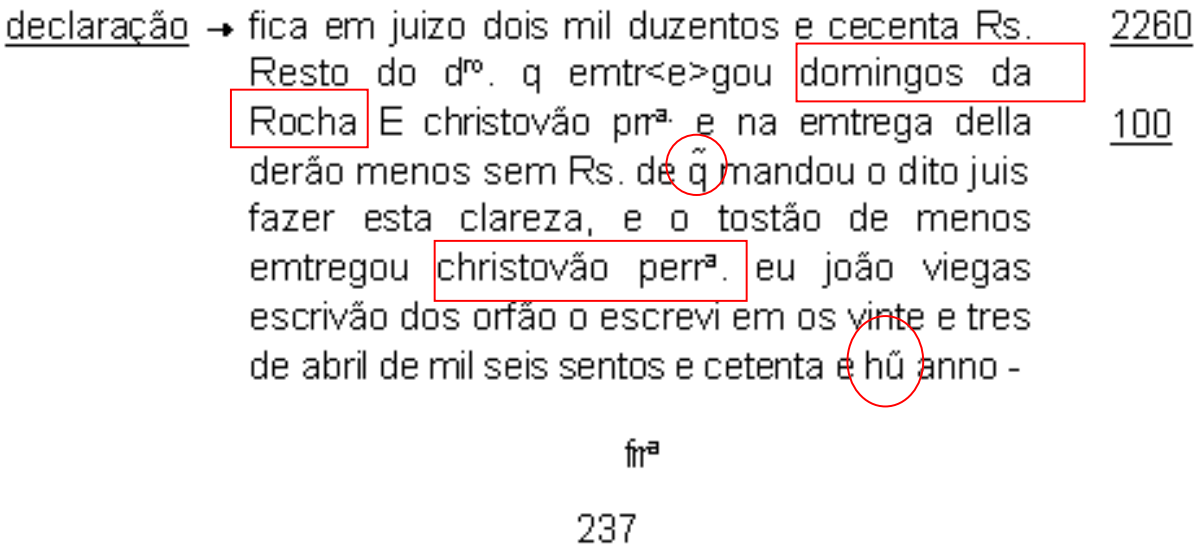
14. Os Irmãos todos estão de saude e fazem o officio a
que forão enviados: somente Antonio Pirez se acha mal das
pernas, que lhe arebentarão depois das maleitas [22] que teve,
e nom acaba de ser bem são.

Leonardo Nunez mandei aos Ilheos, huma povoação
daqui perto, onde dá muito exemplo de si e faz muito fruito,
e todos se spantão de sua vida e doctrina. Foi com elle
Diogo Jácome, que faz muito fruito em insinar os moços e
escravos.

15. Agora pouco há vierão aqui a consultar-me algu-
mas duvidas, e esteverão aqui por dia do Anjo [23], onde

CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO 1549, SERAFIM LEITE S. J (ed.)

Non-standard typographical symbols and proper names both in lower case and abbreviated



declaração → fica em juizo dois mil duzentos e cecenta Rs. 2260
Resto do dº. q emtr<e>gou domingos da
Rocha E christovão prrª. e na emtrega della 100
derão menos sem Rs. de q̃ mandou o dito juis
fazer esta clareza, e o tostão de menos
emtregou christovão perrª. eu joão viegas
escrivão dos orfão o escrevi em os vinte e tres
de abril de mil seis sentos e cetenta e hũ anno -

frª

237

PEDRO CARAÇA, INVENTÁRIO E TESTAMENTO, 1653 - VILA DE SÃO PAULO. APENSO: INVENTÁRIO E TESTAMENTO DE MARGARIDA RODRIGUES 1634 - VILA DE SÃO PAULO, SÍLNIA NUNES MARTINS, EDITORA RESPONSÁVEL PELA DIVISÃO DE ARQUIVOS DO ESTADO DE SÃO PAULO

Graphic forms for the abbreviations

Nos partimos do Rio de Janr.º a 18 do passado e chegamos qui a 4 do corr.te agora não se sabe q.do partiremos daqui porq. hums dizem a 25 do prez.te outros a 10 do q. vem e ha pessoas q. dizem não hirão se não p.ª janr.º q. vem; esta inçerteza tem esta cid.e com g.de comfuzão e a cauza de haver esta incerteza he por não ter chegado a nau de guerra e não haver mais q. hua, que lhe aseguro a VM. que o cazo depende de m.to g.de cuidado que ainda depois que o Brasil he Brazil não foi frotta tão emportante; pois so a nau de guerra dizem leva mais de outo milhões agora o q. pesso a D.s he escolha o milhor p.ª bem de todos e a VM. o g.de m.s a.s VM. me perdoe me fara m.ce dar minhas lembranças a toda a familia, e ao s.r M.el Pretto.

Menor capp.to de VM.
João Deniz de Az.do

CARTAS REMETIDAS PARA LISBOA EM 12 DE JULHO DE 1715 IN: NEGÓCIOS COLONIAIS, L. LISANTI (ED.)

To process the abbreviations forms computationally we used the character circumflex to denote superscript:

(a) **dot followed by a superscript piece of text**: Janr.^o (January) and corre.^te (current)

(b) **when the abbreviation does not possess a dot, but has a supers**

(The Lord Jesus Christ)

Problems to expand abbreviations and recognize named entities

Ambiguity

alteza (highness)	Rio de Jan.ro
alvará (warrant)	Rio de Jarº
Amaro (proper name)	Rio de Janr.º
Ana (proper name)	Rio de Jan.º
anima (cheers up)	Rio de Jnrº
ano (year)	Rio de janrº
anos (years)	Rio de janrº
Antônio (proper name)	Rº de janº
arroba (measure of weight, singular)	Rº de Jarº
arrobas (measure of weight, plural)	Rº de janerº
Assembléia (assembly)	Rº de Janrº
assinado (signed)	Rº de Jnrº
Atual (current)	Rio de Janrº

Spelling Variation

Figure 1. 13 different expansions for the abbreviation “A” in the first column. The second column illustrates 13 different forms of abbreviating the name of the famous Brazilian city “Rio de Janeiro”

- Inherent ambiguity of proper names abbreviations
- Capitalized common words, as if they were proper names
- Proper names in lower case
- Many proper nouns are spelt both with and without initial

Related Projects: Brazilian Portuguese historical corpora

- Language Studies projects expand abbreviations manually:
  - "**Para uma História do Português do Brasil**". For a History of Brazilian Portuguese": Diachronic Studies of Portuguese Language (texts from 17 through the 20 centurie
  - www.letras.ufrj.br/phpb-rj/
  - "**Projeto Programa para a História da Língua Portuguesa**" (PROHPOR). Project Program for a History of Portuguese Language: Syntactic and morphosyntactic changes of Portuguese Language (texts from 13 through the 17 centuries)

    www.prohpor.ufba.br/projetos.html
  - **Tycho Brahe Project** contains tagged and parsed texts written by Portuguese authors born between the 16 and 19 centuries.These texts had their abbreviations expanded manually to facilitate tagging and parsing.

    www.ime.usp.br/~tycho/
- Their **development contexts**, **purpose** and **corpus size** differ

Why we are using Unitex in HDPB project

- Besides using **Philologic** (http://philologic.uchicago.edu/)
  - since it is Web-based and includes several corpus-processing tools, as for example AGREP, used to check for similar or alternative spellings

- We use **Unitex** (http://www-igm.univ-mlv.fr/~unitex/):
  - It allows the use of DELA standard for developing **electronic lexicons**
  - It supports several languages, including *Contemporary Brazilian Portuguese*
  - It allows the use of idioms: we created the *Historical Portuguese*

- Both types of DELA are used:
  - DELAF (simple words associated with grammatical – morphological - and inflectional information)
  - DELACF (compound words associated with grammatical – morphological – and inflectional information)

- DELA also allows to add syntactic and semantic information

Disappear in the binaries

(Inflected word),(canonical form).(POS) [*+(subcategory)*]*:morphological features/comments*

dogs,dog.N+Animal:mp/mammal

DELA dictionaries used in HDPB project (1/3)

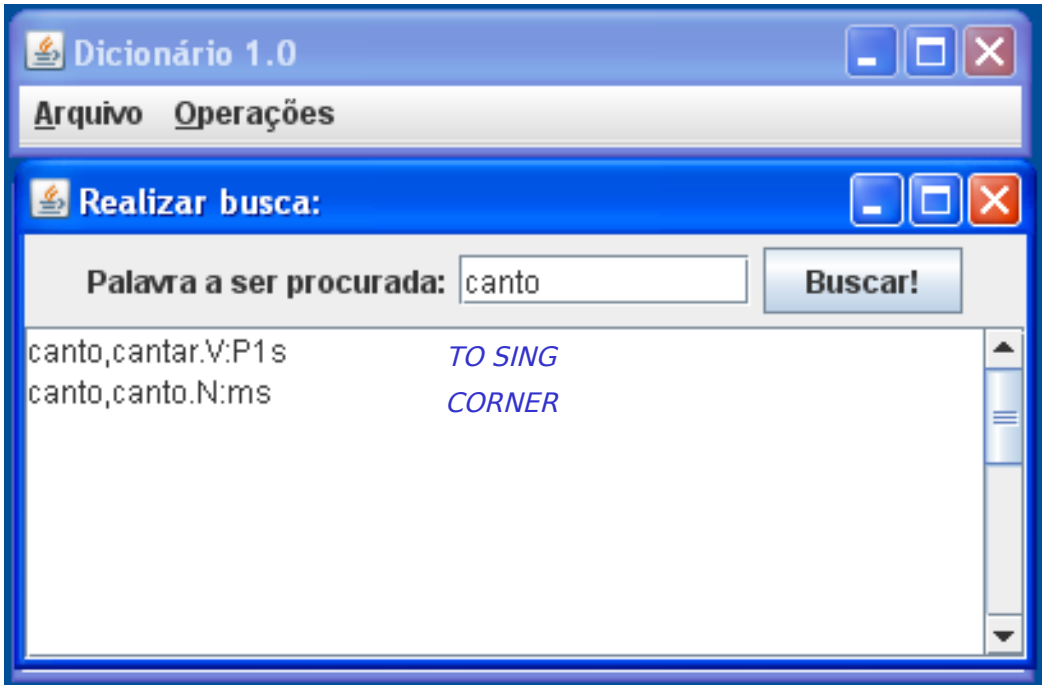
Unitex-PB

Contemporary Brazilian Portuguese (Muniz et al., 2005)

DELAF_PB (~880.000 entries)

DELACF_PB (~4.000 entries)

Dicionário is a Java application that handles any dictionary compacted in the DELA format



www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/

DELA dictionaries used in HDPB project (2/3)

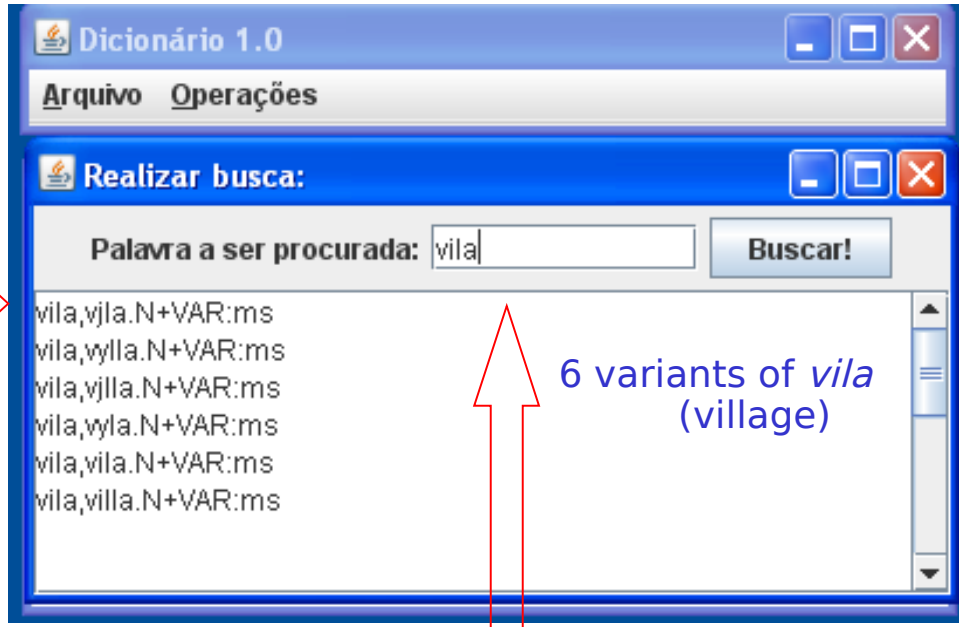
Spellings variants dictionary (SVD)

- **SVD** was created with SIACONF, a system based on 43 transformation rules (Giusti et al., 2007)
- SIACONF clusters variants of a corpus under a common form, sometimes corresponding to the modern form
- **HDPB corpus**: 18,082 clusters; 41,710 variants

Disappears in the binaries

appellidos,apelidos.N+**VAR**:ms/50.0%
apelidos,apelidos.N+**VAR**:ms/36.36%
appelidos,apelidos.N+**VAR**:ms/9.09%
apellidos,apelidos.N+**VAR**:ms/4.54%

inversion



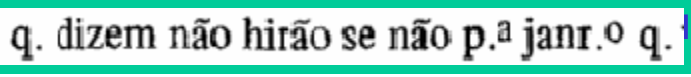
6 variants of *vila* (village)

Search for the modern form of a word

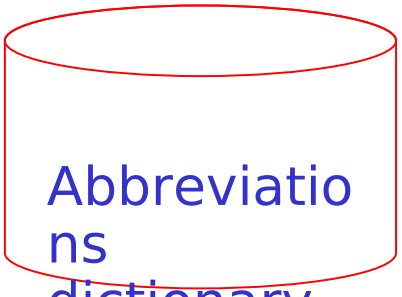
- **To use SVD with Dicionário**: variant and modern form have been inverted to facilitate searching

- Morphological information is the default *ms,* since SVD was automatically generated

moodle.icmc.usp.br/dhpb/



DELA dictionaries used in HDPB project (3/3)



- Flexor (1991) is a large, alphabetically organized, dictionary of abbreviations from the 16 through the 19 centuries.

  3 types of information: abbreviation, expansion and centuries

- Thus far, Flexor (1991) – letters A, B, C, D – and some of the authoritative lists of abbreviations to be used in the UNITEX system to create AbbDic.

Disappears in the binaries

B^s,bastardo.N+INIT+PARENTE+**ABREV**+**SEC18**:mp/**bastardos**

Bas^tos,bastardo.N+INIT+PARENTE+**ABREV**+SEC19:mp/bastardos

Bastrd^os,bastardo.N+INIT+PARENTE+**ABREV**+SEC19:mp/bastardos

3 abbreviations of *bastardo* (the illegitimate offspring of unmarried parents)

Expansion of abbreviation is in the comments

Flexor in Numbers

Simple and Multi-word Abbreviations by Century					
Types	XVIth	XVIIth	XVIIIth	XIXth	Total
Flexor	2,050	4,091	14,376	9,939	21,869
Flexor (%)	9.37	18.70	65.74	45.45	139.26
Intersection of Flexor and Corpus	754	1,323	2,447	1,710	3,529
Intersection of Flexor and Corpus (%)	21.37	37.49	69.34	48.46	176.65
Coverage (%)			16.13		

- Flexor has a large number of abbreviations but **only 16% appear in the *HDBP* corpus**.
- **Experiment (how many abbs is out there?):** Using heuristic rules to recover a set of abbreviations of HDPB we found 7.045 **simple** abbreviations; **only 35% of them are in Flexor.**
- **Flexor is still worth using** as it has abbreviations expansion and the centuries they have been used.

Simple and Multi-word Abbreviations by n-grams							
Types	1	2	3	4	5	6 or +	Total
Flexor	17,872	1,624	833	527	302	711	21,869
Flexor (%)	81.73	7.42	3.81	2.41	1.38	3.25	100.00
Intersection of Flexor and Corpus	3,237	234	33	18	5	2	3,529
Intersection of Flexor and Corpus (%)	91.75	6.60	0.94	0.51	0.14	0.06	100.00

Dictionary of Abbreviations: semantic attributes

<ENT>: named entity (NE) – 10 top-categories of Linguateca´s Evaluation Contest HAREM:

Person, Organization, Artifact, Location, Object, Event, Abstract Entity, Quantity, Time, Miscellaneous

<INIT>: a trigger expression to extract subtypes of NE

- 10 HAREM top-categories plus three subcategories:
- <TITULO> : for jobs/professions and titles/positions,
- <PARENTE>: for family relations, and
- <TRATAMENTO>: for forms of address, since they are very pervasive in Flexor's dictionary.

Brg^es,Borges.N+ENT+PESSOA+ABREV+SEC19:ms/Borges

Brag.,Braga.N+ENT+PESSOA+LOCAL+ABREV+SEC18:ms/Braga

Br^ça,Braça.N+ENT+VALOR+ABREV+SEC19:fs/Braça

7^bro,setembro.N+ENT+TEMPO+ABREV:ms/setembro

B^eis,bacharel.N+INIT+TITULO+ABREV:mp/bacharéis

B.,beco.N+INIT+LOCAL+ABREV+SEC18:ms/beco

Bat^am,batalhão.N+INIT+ORGANIZAÇÃO+ABREV+SEC16:ms/batalhão

Bas^tos,bastardo.N+INIT+PARENTE+ABREV+SEC19:mp/bastardos

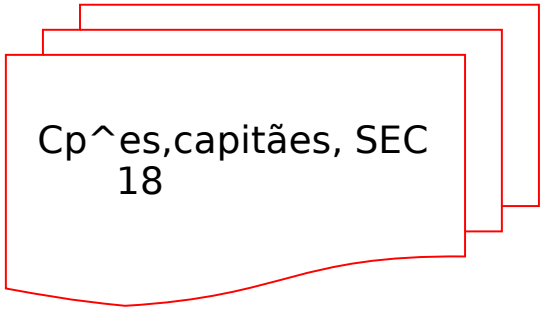
previation),(canonical form).(POS)[+ENT/INIT+subcategories+ABREV+centuries]*:morphological features/expans*

Using the canonical form for information retrieval

FLEXOR, Maria H. *Abreviaturas, Manuscritos do século XVI ao XIX*.

DHPB corpus

Cp^es,**capitão**.N+INIT+TITULO+ABREV+SEC18:mp/capitães



Cap^aens,capitão.N+INIT+TITULO+ABREV+SEC19:mp/capitães
Cap^ams,capitão.N+INIT+TITULO+ABREV+SEC19:mp/capitães
Cap^ans,capitão.N+INIT+TITULO+ABREV+SEC18:mp/capitães
Cap^ens,capitão.N+INIT+TITULO+ABREV+SEC18+SEC19:mp/capitães
Cap^es,capitão.N+INIT+TITULO+ABREV+SEC16+SEC17+SEC18+SEC19:mp/capitães
...
Cp^es,capitão.N+INIT+TITULO+ABREV+SEC18:mp/capitães

Search for the canonical form "capitão":

capitães,capitaães.N+VAR:ms
capitães,capitães.N+VAR:ms
If we include the canonical form

Spelling variantes:

Capitão capitam Capitaõ Cappitam ...

and the abbreviated forms:

Cap^{aens}	Cap^{ams}	Cap^{ans}	Cap^{ens}
Cap^{es}	Cap^{ms}	Cap^{ns}	Cap^{s}
Cap^{tens}	Cap^{tes}	Capa^{ens}	Capitt^{es}
Cap.m^{s}	Capn^{es}	Capn^{s}	Capns
Capp^{ães}	Capp^{es}	Capp^{tes}	Capt^{es}
Capt^{s}	Captaẽs	Cp^{es}	C^{m} ...
C^{tam}	C^{tan}	Ca^{m}	Camp^{m}
Cap.	Cap^{am}	Cap^{an}	Cap^{ão}
Cap^{ão}	Cap^{m}	Cap^{n}	Cap^{on}
Cap^{t}	Cap^{tam}	Cap^{tan}	Cap^{tão}

Capitaães,**capitão**.N+VAR+capitães:ms

Hybrid process to build the dictionary of abbreviations

- **To enrich the Flexor dictionary with NE information**

- **To enlarge Flexor dictionary with new and different entries of a given category and its sub-specifications**

Hybrid process to build the dictionary of abbreviations (1/2)

- **To enrich the Flexor dictionary with NE information:**

We first classified the entries with the categories ENT/INIT using common-sense knowledge.

- They allow the extraction of new NEs using an **iterative process**:
  - searching for the INIT r^o (river) we can find the NE "r^o de s. fran^co" (São Francisco river).
  - searching for the NE "Fran.^co" (Francisco), we can find "Mosteiro de Sam Fran.^co" (São Francisco monastery).

INIT attribute is given to trigger words such as those of Fig 3. INIT has also subcategories:

- LOCATION for names of places (**fazenda/farm**, **arrayal/hamlet)**
- ORGANIZATION for organizations (**fortaleza/fortress**).

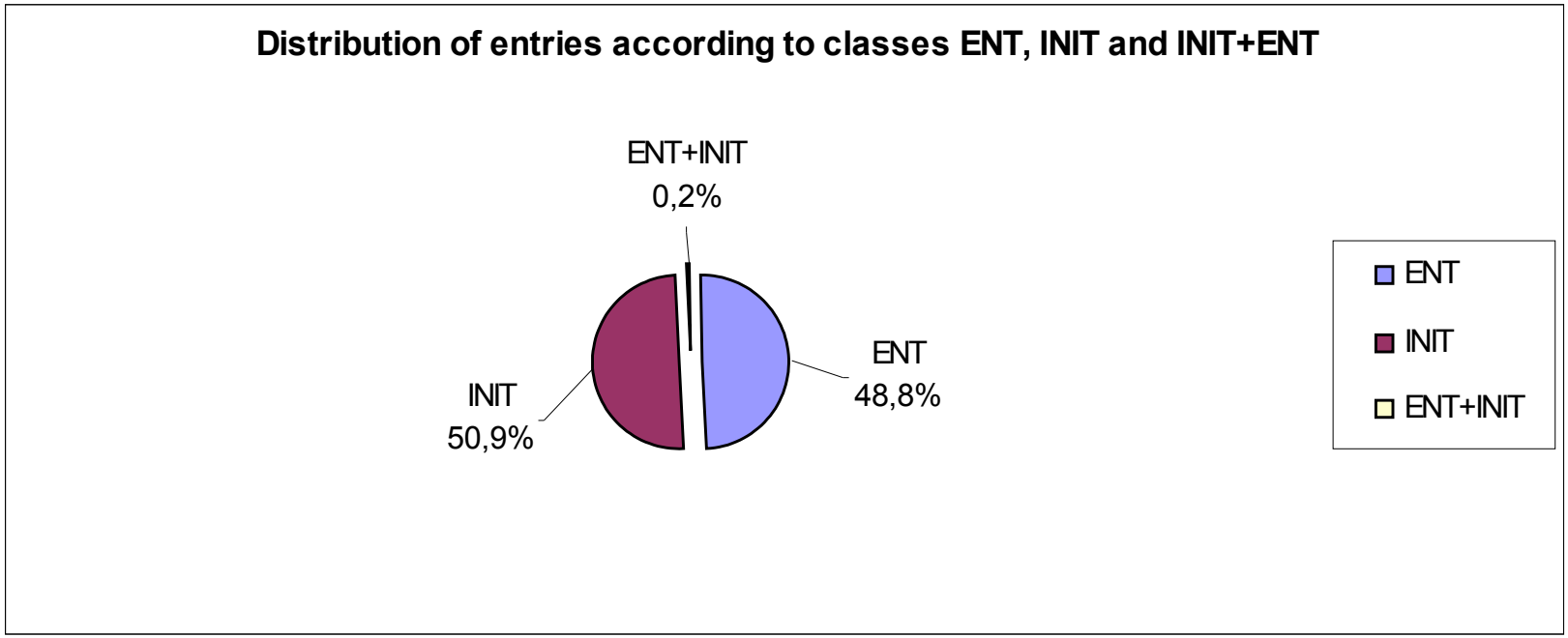
(<INIT>+((o+ao+do+ho).p\.)).
(<MOT>+<MOT><MOT>+
<MOT><MOT><MOT>+<MOT><MOT><MOT>
<MOT>)

até à fazenda de S. Bento, tambem notavel n
cupa o arrayal de S. Bento, e poem em aperta
dezagoa o rio de S. Bento; Duas legoas pouc
ina o mosteiro de S. Bento, para cujo reparo
ssos na aldeia de S. Barnabé, se serviu Deus
nço, no bairro de S. Bento. Veio depois o co
com o engenho de S. Bernardo, situado sobre
sse no arraial de S. Caetano 3 léguas antes
' 58". Arrayal de S. Caetano, 20^o 12' 58".
o para o forte de S. Carlos do Rio Negro cor
15' 45". Corgo de S. Domingos, 16^o 48' 42"
z tambem o rio de S. Domingos, ao rio de Jaq
essem no campo de S. Domingos, aonde tinham
o da bocca do rio S. Domingos, de pequeno cu
lro do Rio Fundo e S. Domingos da Saubara. Vi
vindo do pôrto de S. Filipe, vizinho a Nossa
ler a Fortaleza de S. Filippe, que na Bertioq

Figure 3: Concordances retrieved from sentences that have the abbreviated form "S."

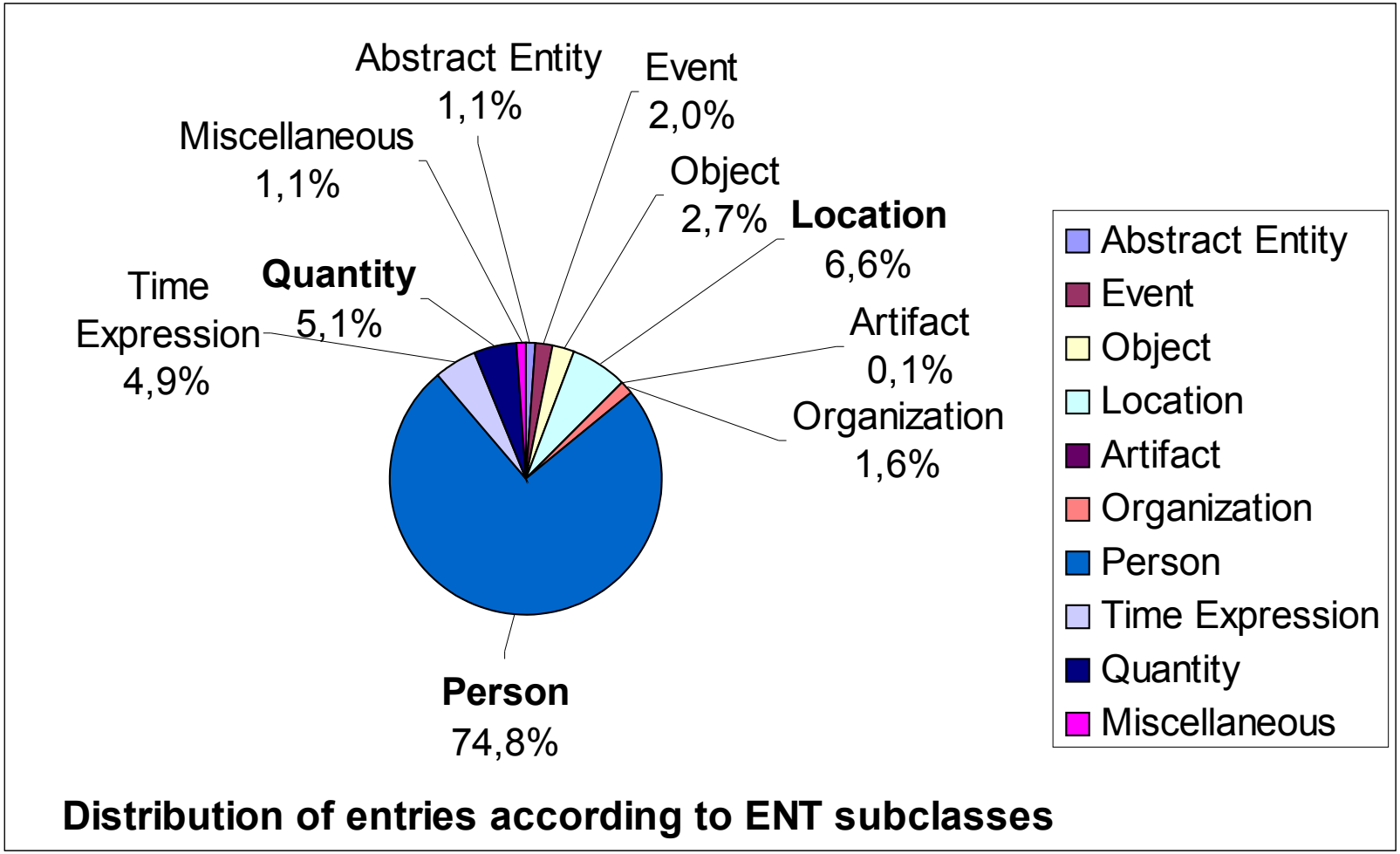
Results of the semantic classification (1/4)

Letters A, B, C, D – and some of the authoritative lists of abbreviations:
7314 **simple** abbreviations - 3604 classifications ENT,INIT, ENT+INIT

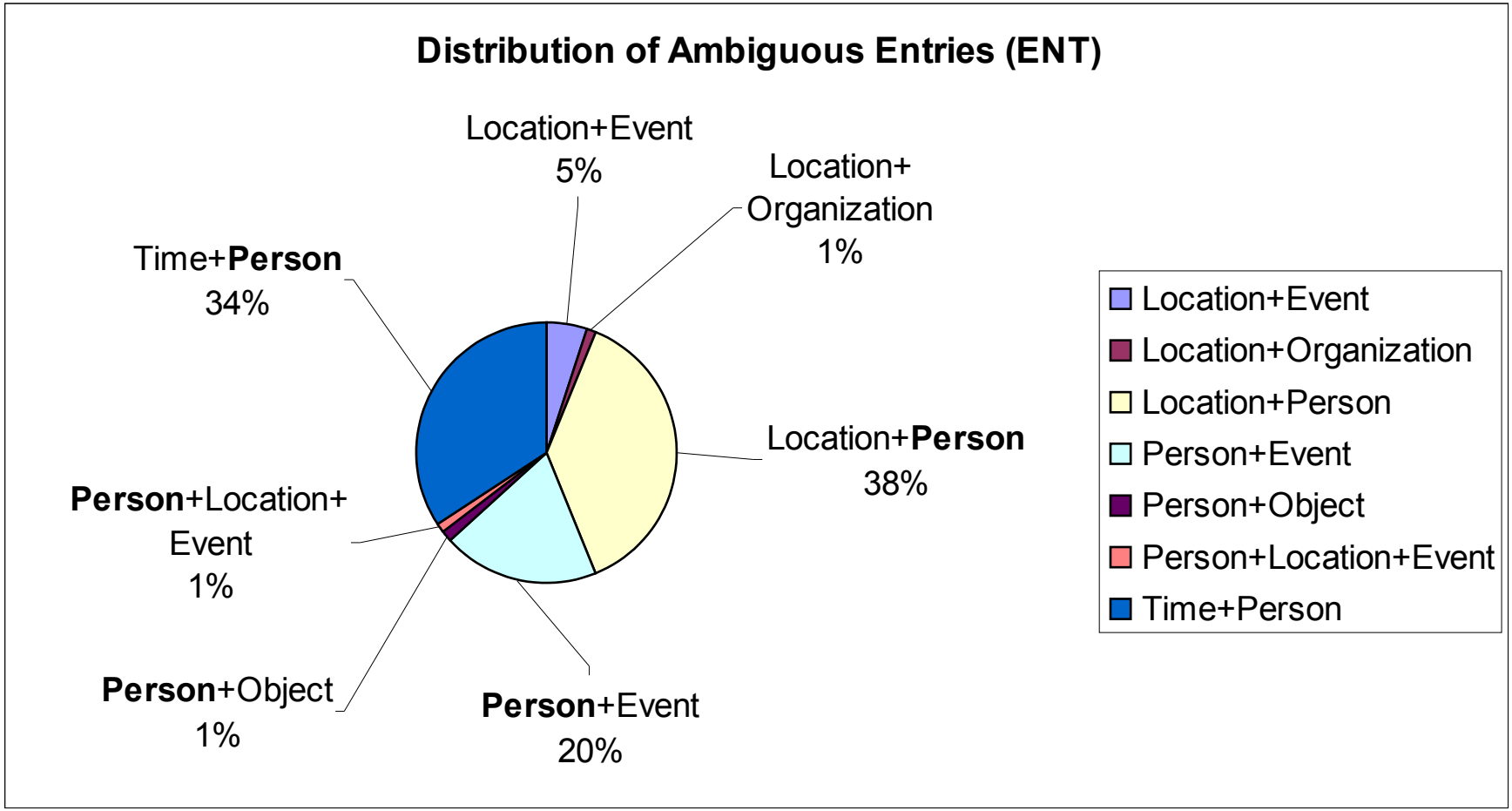


Some entries classified as <ENT> are <INIT> as well, such as “Barb^ro” (barber), a family name and a pattern used to introduce this profession.

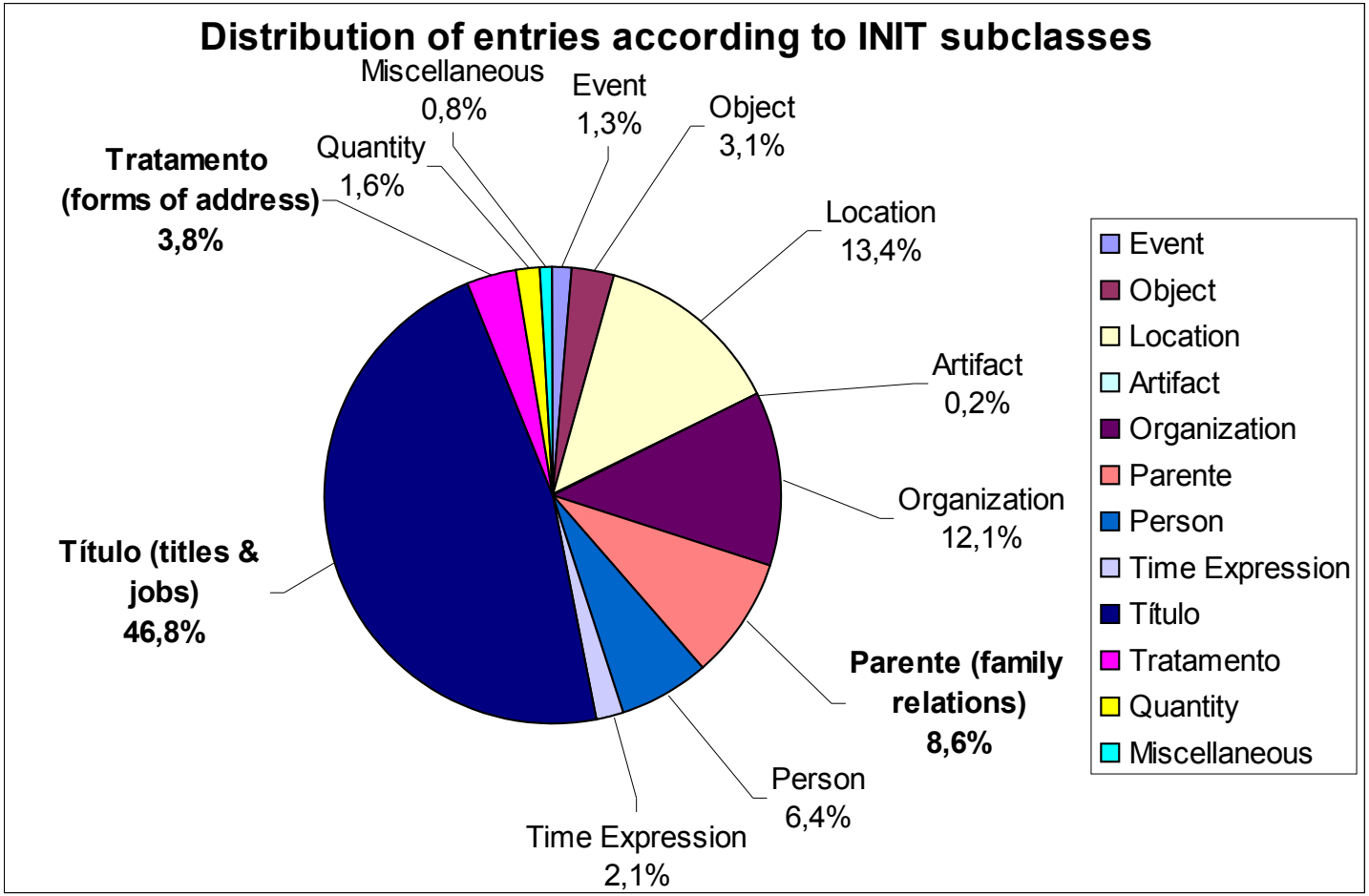
Results of the semantic classification (2/4)



Results of the semantic classification (3/4)



Results of the semantic classification (4/4)



Hybrid process to build the dictionary of abbreviations (2/2)

- **To enlarge Flexor dictionary with new and different entries of a given category and its sub-specifications:**

Same process defined by REPENTINO, a repository of NEs from modern Portuguese without expanding the classification:

1) choose a category for which you intend to search examples of entities;
2) decide which is the most appropriate strategy to search for the examples:
   a) by tag <INIT>, such as in Rio S. Francisco;
   b) by context, such as in "localizado na XXX" (located at XXX), which strongly suggests that "XXX" is a place; or
   c) by discriminating suffixes (modern organizations have in their names characteristic particles such as "Ltda."/Ltd. or "S.A"/Co.);
3) construct the respective pattern to be searched in a given corpus processor or to act as an independent program, and conduct the search;
4) validate manually the obtained candidates, considering the intended category;
5) include positive candidates in the repository;

BUT we had to adapt it to work with historical corpora:

in step 2, we included spelling variants;

in step 5:

no capitalization requirement;

at least one of the components should be abbreviated

Case study about hydronyms – names of rivers, streams, creeks, and brooks found in the *HDBP* (1/2)

- Flexor (1991) contains 18 entries with the pattern Rio XXX (River XXX)
  - but eight of them refer to the city/state of Rio de Janeiro
  - Others are: R^o da Ribr^a, R^o de Reg^o, R^o de S. Fran^co, R^o dos Alm^das, R^o G^de, R^o G^re, R^o Gdr^e, R^o Gr^de, R^o G^re e R^o P^do
  - there is nothing about Creek XXX (or its variants, brooks, streams),

- We began with ten entries

Case study about hydronyms (2/2)

Searching patterns (63)	Sources (5)	Right Occurrences (112)
rio	river	79
arroio, córrego, corrente, regato, regueira, regueiro, ribeirão, ribeiro, riacho, rio, veia, veio	synonyms	13
arroyo, corrego, corego, corgo regueyro, ribeiraõ, ribeyrão, ribeyraõ, rybeirão, rybeyrão, rebeirão, rebeyrão, ribeirao, ribeiro, ribeyro, rybeiro, ribejro, rjbeyro, rybeyro, riaxo, ryo, rjo, rrio, veya, veyo	spelling variants	7
c^te, cor, cor^e, cor^te, corr^e, corr^te, cort^e, crr^e, curr^te, r^bro, r^o, r^ro, reb^o, rib.^ro, rib^o, rib^ro, riber^o, ribr, ribr^o, ryb^o, ryb^ro, rybr^o, r^bro, r , r.^o	abbreviations from Flexor (1991)	11
naveg*	context	2

Table 5: Searching patterns for hydronyms

- The manual validation is the slowest step:

  we checked 27,808 occurrences in 160 minutes – 1,100 checking per hour).

- Results:

  122 abbreviations under category LOCAL, specifically rivers and words related to watercourses, displaying their morphology in this semantic group in the *HDBP* corpus.

- Examples:

  Ribeyrão de N. Sr.^a do Carmo (Ribeirão de Nossa Senhora do Carmo);

  Corgo de S. Gonçalo (Córrego de São Gonçalo);

  rib.^o do Tombadouro, (ribeirão do Tombadouro);

  coRego Ant.^o da Silua (Córrego Antonio da Silva);

Problems & Solutions

Problems related to semantic classification:

- **Has an abbreviation used in previous centuries the same current meaning?**

  Yes if there is at least one example.

  "companheiro" (companion, partner) is classified as PARENTE (family relation) + TRATAMENTO (address form)

  Foi um dia fóra desta cidade a confessar uma mulher, que estava muito no cabo, cujo marido, por nome Domingos Saraiva, muito triste e choroso saiu fóra de casa a receber o Padre, o qual vendo-o tão sentido lhe disse: Não vos desconsoleis, bom velho, que não vos hade morrer desta vossa **companheira**; e isto foi antes de chegar á casa aonde a doente estava.

- **Same spelling, different meaning due to capitalization**:

  Since proper names can appear in lower case, certain words such as:

  - "dias" (days) **will be classified as** ENT = Person and INIT = Quantity;

Problems & Solutions

Problems related to variants:

- **How to deal with spelling variants in Flexor that are not abbreviations?**
  - We will eliminate entries such as:
  - (Bẽs, bens)
  - (Bẽz, bens)
  - (Bãda, banda)

  since the tilde was part of the writing system of historical Portuguese.

General Problem:

- **Making the dictionary/gazetteer publicly available:**
  - AbbDic will be available under request, (due to copyright issues) by October, 2008
  - All the abbreviations found using the Repentino methodology

Applications

1) Resource for NE recognition systems

2) Creation of a golden corpus to evaluate NER systems (after manual disambiguation of some abbreviations)

3) Abbreviation lookup and expansion (used in the HDBP project)

- When lexicographers are creating the entries for the main HDBP dictionary
  - She/he may find examples of abbreviations in the corpus, but may not be aware of the possible expanded forms for a given abbreviation
- Using UNITEX and the software Dicionário
  - can assist lexicographers in manually identifying possible canonical forms for an abbreviation
  - can assist historical linguists to expand an abbreviation
- Reduces the effort required to expand all abbreviations in the corpus

Dicionário is a Java application that handles any dictionary compacted in the DELA format

Search for pattern “b” in UNITEX; Search program
onário disambiguates the abbreviation “b” (in the top right corne

Concluding Remarks

- We have used the HAREM categories since
  - This ontology is devoted to Portuguese language,
    - it used a collection of documents from several genres
    - besides being well documented (guidelines are publicaly available)

- AbbDic has been created semi-automatically to ensure the high quality of the resource and to give us clues on the problems of NE recognition

- The development of AbbDic is strategic and valuable for both
  - the **linguistic commnity** working with Historical Portuguese corpora since to have an online dictionary of abreviations will facilitate their studies
  - the **Natural Language Processing** community working with the Portuguese language

- AbbDic will be released by October 2008
  - and will continuosly be enlarged with abbreviated NE from other Brazilian Portuguese corpora by students who love historical corpora.

Thank you!

NILC 1993 - 2008

An Interinstitutional Center for Research and Development in Computational Linguistics

http://www.nilc.icmc.usp.br/nilc/

Unitex-BR:

http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/

Spelling Variants and Abbreviations Dictionaries:

http://moodle.icmc.usp.br/dhpb/

References

- Rydberg-Cox, J. A. (2003). Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries. In *Proceedings of JCDL*, 03, pp. 372--373.

- Sanderson, R. (2006). "Historical Text Mining", Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities. Talk presented at the Historical Text Mining Workshop, July 2006, Lancaster University, UK.

- Flexor, M. H. (1991). Abreviaturas - Manuscritos dos Séculos XVI ao XIX. 2nd ed. São Paulo: UNESP. 468 p.

- Muniz, M., Nunes, M. G. V., Laporte, E. (2005). UNITEX-PB, a set of flexible language resources for Brazilian. In III Workshop em Tecnologia da Informação e da Linguagem Humana, pp. 2059--2068.

- Giusti, R., Candido Jr, A., Muniz, M., Cucatto, L., Aluísio, S. (2007). Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary. In Proceedings of the Corpus Linguistics 2007 Conference, Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.).