Automatic detection of spelling variation in historical corpus

An application to build a Brazilian Portuguese spelling variants dictionary

Rafael Giusti, Arnaldo Candido Jr, Marcelo Muniz, Lívia Cucatto, Sandra Aluísio

Corpus Linguistics 2007

Agenda

- Introduction to the DHPB project
- Spelling variation in historical *corpora*
- Related Works
- Our approach: an iterative process for detecting spelling variants
  - Transformation rules
- Experiments and evaluation
- Brazilian Portuguese dictionary of spelling variants
- Conclusion

DHPB Project

- Historical Dictionary of Brazilian Portuguese (DHPB)
  - **XVI-XVIII** centuries (beginning of Brazil's history)
  - First dictionary of this kind
- It's a three-year project (2006-2008)
  - Sponsorship of the funding agency CNPq

DHPB *Corpus*

- Texts from **1500-1808**
  - Written by Brazilians or Portugueses who have lived in Brazil for a long time
- Corpus size: more than **3,000** texts and **7.5 million** words
  - Working Corpus Size: **1,733** texts, **4.9 million** words and **57.1** MB (UTF-16LE)

DHPB *Corpus* (2)

- Text types: Letters of Jesuit missionaries, Inquisition's documents, reports of Brazilian explorers, etc
- Text sources:
  - **Manuscripts** (manually keyboarded)
  - **Original printed documents** (OCR)
  - **PDF files composed of images** (OCR)

DHPB *Corpus* (3)

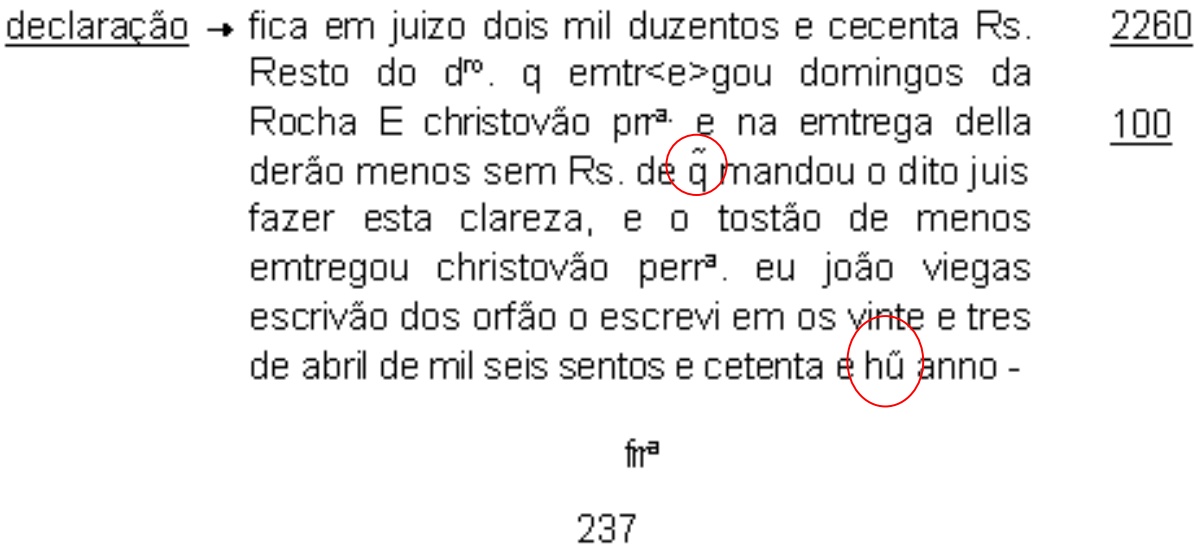
Data	Centuries			
	16th	17th	18th	19th
Texts	11.16%	27.64%	52.06%	9.13 %
Sentences	28.99%	15.94%	43.17%	11.90 %
Words (approx.)	18.68%	20.67%	47.68%	12.98 %

Distribution of texts by century

Challenges in dealing with historical corpora

- 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
  - Great **spelling variation** (even within the same text)

Use of non-standard typographical symbols



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

Spelling variants problems

- Distorts **frequency counts**
- Difficulties indexing techniques for **Information Retrieval** (Hauser et al., 2007)
- Hinders corpus **annotation tools** trained on contemporary language (Crane and Jones, 2006)
- Difficulties **NLP tasks** such as named entity extraction (Rayson et al., 2007)

Related works

- **VARD** (VARiant Detector): spelling variation detection and normalization (Rayson et al., 2007) (focus on English language)
- **RSNSR**: German spelling variation (Archer et al., 2006)
- **Tycho Brahe** spelling variant normalizer (Hirohashi, 2005) (focus on Brazilian Portuguese (BP) language)
- **AGREP** in **Philologic**: spelling variation detection (language independent)

VARD

- Trained on sixteenth to nineteenth-century texts
- Focus on **precision** rather than recall
  - since it was developed to detect and normalise spelling variants to their modern equivalents in running text
- Use XML to normalize and preserve original variant form
- SoundEx and edit distance algorithms

RSNSR

- Rule-based fuzzy search engine
  - Created by statistical analyses, historical material and linguistic principles
- Focus on **recall** rather than precision,
  - since it is a web-based system focuses on finding and highlighting historical spellings

Tycho Brahe spelling variant normalizer

- Supervised machine learning
- Modules based
- Indirect effectiveness evaluation through Tycho Brahe POS Tagger

AGREP

- Fuzzy string searching
- Variety of well-known fastest string searching algorithms
  - Manber and Wu's bitap algorithm, mgrep, amonkey, mmonkey, etc
  - Best-suited algorithm used

Our objectives

- To present
  - an approach based on **transformation rules** to cluster distinct spelling variations around a common form
    - our aim is that the groupings reduce the impact of spelling variation on the frequency count
    - the choices made to build a **dictionary of spelling variants of BP** based on these clusters
  - a **system** to support both
    - the detection of spelling variants and
    - the development of **new rules**

Our approach

- Transformation Rules (TR)
  - Letter and string replacement rules
  - Same format as those in Hirohashi (2005)
  - Grouping spelling variations around a common form
    - Not always the orthographic (or modern) form

Transformation Rules

- It's a triplet (**C1 C2 S**) applied over strings, where:
  - **C1**: a regular expression that determines if a string is covered by the rule
  - **C2**: a regular expression that determines the substring that will be replaced
  - **S** is a the replacement substring

Transformation Rule Example

- (**e[ao]** **e** **ei**)
  - "**e[ao]**" will cover forms like “ald**ea**”, “m**eo**”, “ch**ea**s”, etc
  - "e" define the substring will be replaced (ald**e**a, m**e**o, ch**e**as, etc)
  - "**ei**" define the replacement (the normalized forms ald**ei**a, m**ei**o, ch**ei**as, etc)

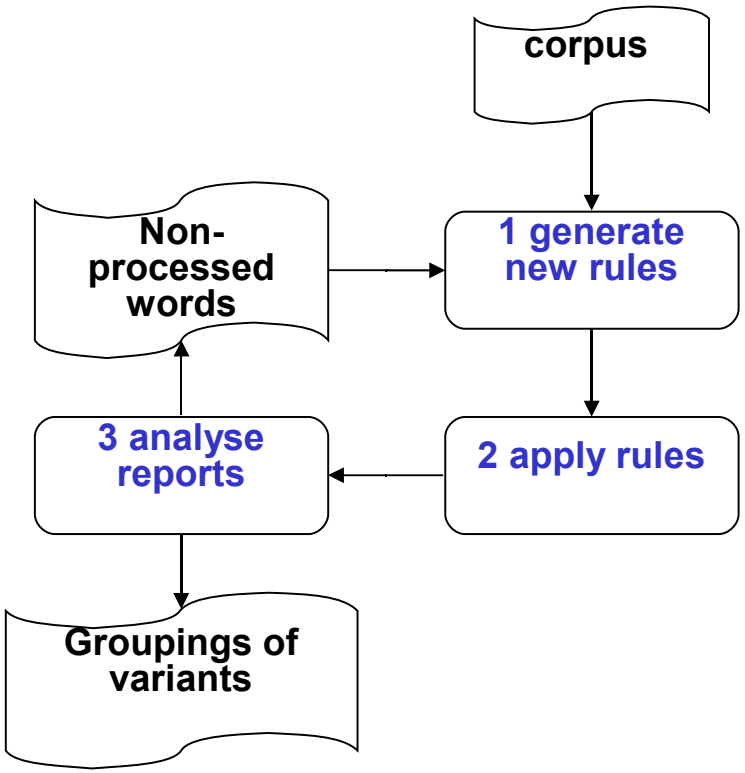
Siaconf system

- Support System for Frequency Counting in Corpus
  - Based on **TRs**
- Freely available
  - http://moodle.icmc.usp.br/dhpb/siaconf.tar.gz
  - Currently, documentation is only in Portuguese
- Generates several reports…

Siaconf reports

- Groupings/clusters including **spelling variants** of the same word
- Information on the **rules applied**
- List of **non-processed words**

Using Siaconf



Iterative process for detecting spelling variants in a given historical corpus

Using Siaconf (2)

- A start set of **rules** that are applied to the corpus
- Reports are generated and **variants are grouped**
- The **reports** are analysed and rules validated (with the report of rules applied)
- New rules are created and applied to the corpus with the aid of the list of **non-processed words**
- Go back to step 2 until dictionary of spelling variants be satisfactory

TRs created

- Six classes of rules created:

1. Rules to deal with spellings that fell in disuse (4 rules)
   - Example: all "ph" are replaced to "f", because in "ph" is no longer used
     - **phármacia -> fármacia**

TRs created (2)

2. Rules to deal with double consonants (13 rules)
   - Example: **ffoy** -> **foi**, **edittou** -> **editou**
3. rules according orthographic norm (6 rules)
   - Example: "n" must be replaced by "m" before "b" or "p“
   - **tenpo** -> **tempo**

TRs created (3)

4. Rules based on frequency analysis (14 rules)
   - Example: replace "ch" by "x"
   - **Cham -> xam**
5. Rules used in Tycho Brahe (5 rules)
   - Example: "z" by "s" in the infix "preciz"
   - **preciza -> precisa**

TRs created (4)

6. Lexicalised rules (1 rule): specific rules to cover spellings which are not grouped by general rules
   - Example: replace "o" by "u" to forms ending in "deos"
   - deos -> deus, judeos -> judeus

Experiments

- **43 rules** applied in **4.9 millions** word *corpus*
  - **12,189** clusters
  - **27,199** variants

Grouping variations of “floor” through several rules

Words	Rules applied	Spellings generated
CHAÕ	ch ch x aõ aõ ão [^r][aã]o$ [aã]o am	"xaõ" "xão" "xam"
CHAÃO	ch ch x aã aã ã [^r][aã]o$ [aã]o am	"xão" "xaão" "xam"

“**Chaõ**” and “**chaão**” (floor) are grouped under “**xam**”, witch doesn't exist in Portuguese

Sample groupings

apelido (90) appellido (48) apelido (30) appelido (7) apellido (5)	**nam** (37,100) não (33,684) naõ (2,652) nam (439) nao (325)
mais (23053) mais (22,918) majs (67) maes (38) mays (30)	vila (5,218) villa (4,073) vila (1,113) vyla (13) vjlla (9) vylla (9) vjla (1)

Evaluation

- **Transformation Rules** (Siaconf) was compared with **Edition Distance** (Philologic with Agrep)
- Experiments divided in two parts
  - 23 random words for each letter of the Portuguese alphabet (except for “X”, plus “k”)
  - 5 most frequent words
    - “Que” (that), “com” (with), “não” (not), “mais” (more), “seu” (your)

23 random words

Technique	True positive	False positive	Precision	Comparative recall
Transformation rules	36	0	100%	72%
Edit distance (AGREP)	41	196	20.92%	84%

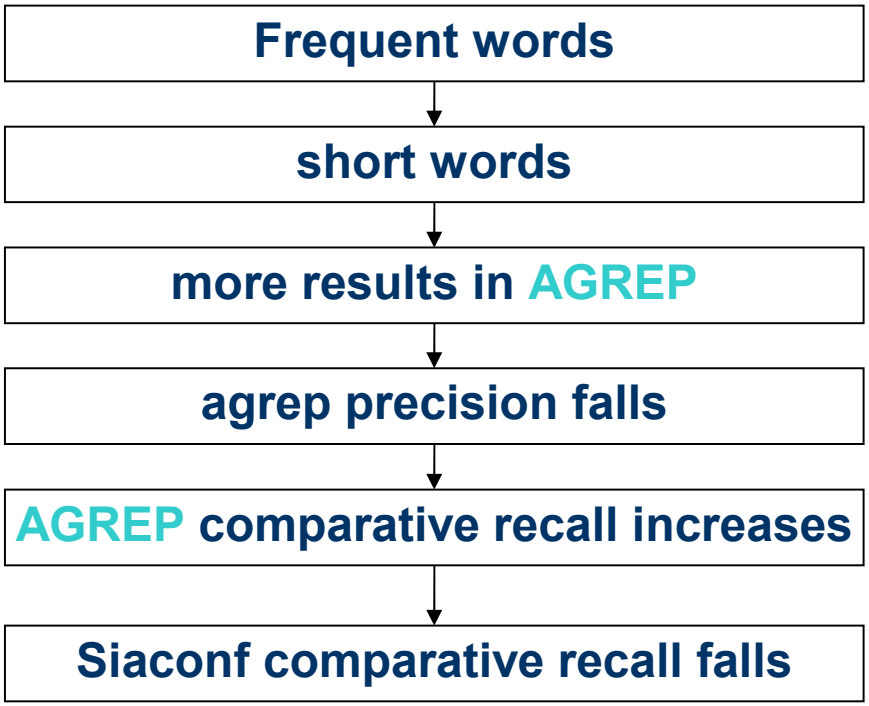
TR -> better precision

ED -> better recall

5 very frequent words

Technique	True Positives	False Positives	Precision	Comparative Recall
Transformation rules (Siaconf)	7	2	77.77%	23.33%
Edition distance (AGREP)	27	217	11.06%	90%

5 very frequent words (2)



Evaluation by lexicographers

- Some variants not covered by the transformation rules was reported (Siaconf focus on precision)
- To solve this problem:
  - Develop more transformation rules
  - Include the results from AGREP

DELA Dictionary Created – Entry sample

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%

- All entries were masculine-singular (MS) nouns (N) because the process was automatic
- Can be useful also insert lemmatised form to in Dela entry (as semantic atribute or replacing normalized form)

The lexical entries in DELAF have the following general structure:
(*Inflected word*),(*canonical form*).(*part of speech*)[+(subcategory)]:*morphological behaviour*

How to build the dictionary entries

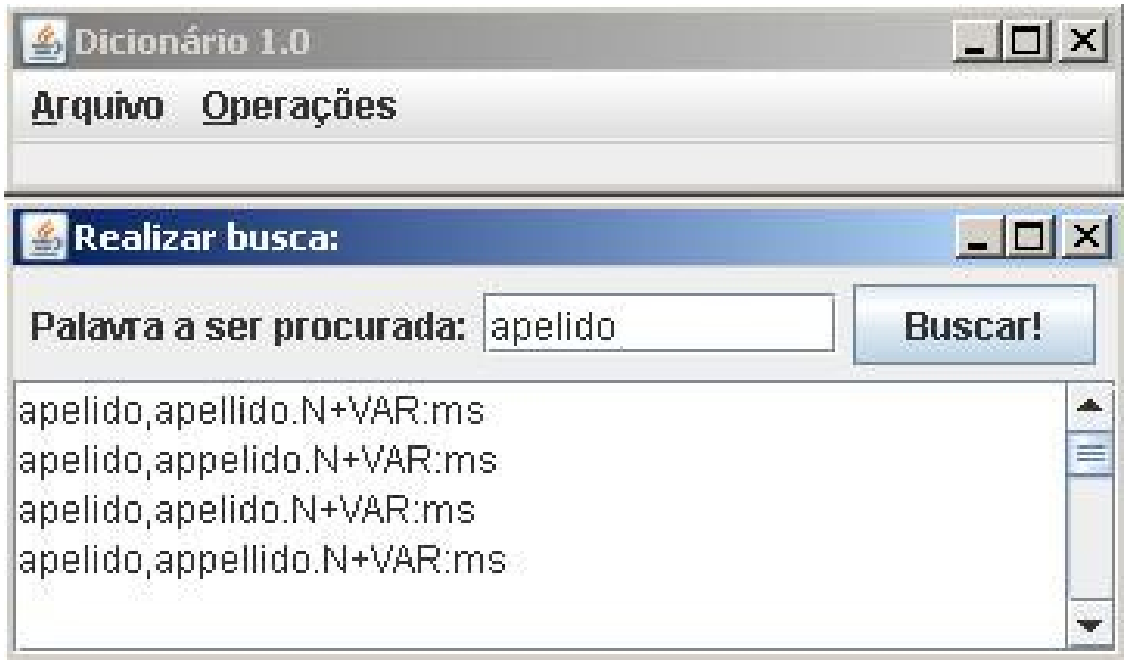
- A possible change is to insert the lemmatised form of the spelling in the proposed structure.
- Searches based on the lemmatised form are particularly useful for verbs in Portuguese, since they have a great number of inflections.
- The lemmatised form can be inserted in the place of the spelling generated by Siaconf:

  appellidos,apelido.N+VAR:ms/50.0%

- An alternative is to insert the normalised form as a semantic attribute:

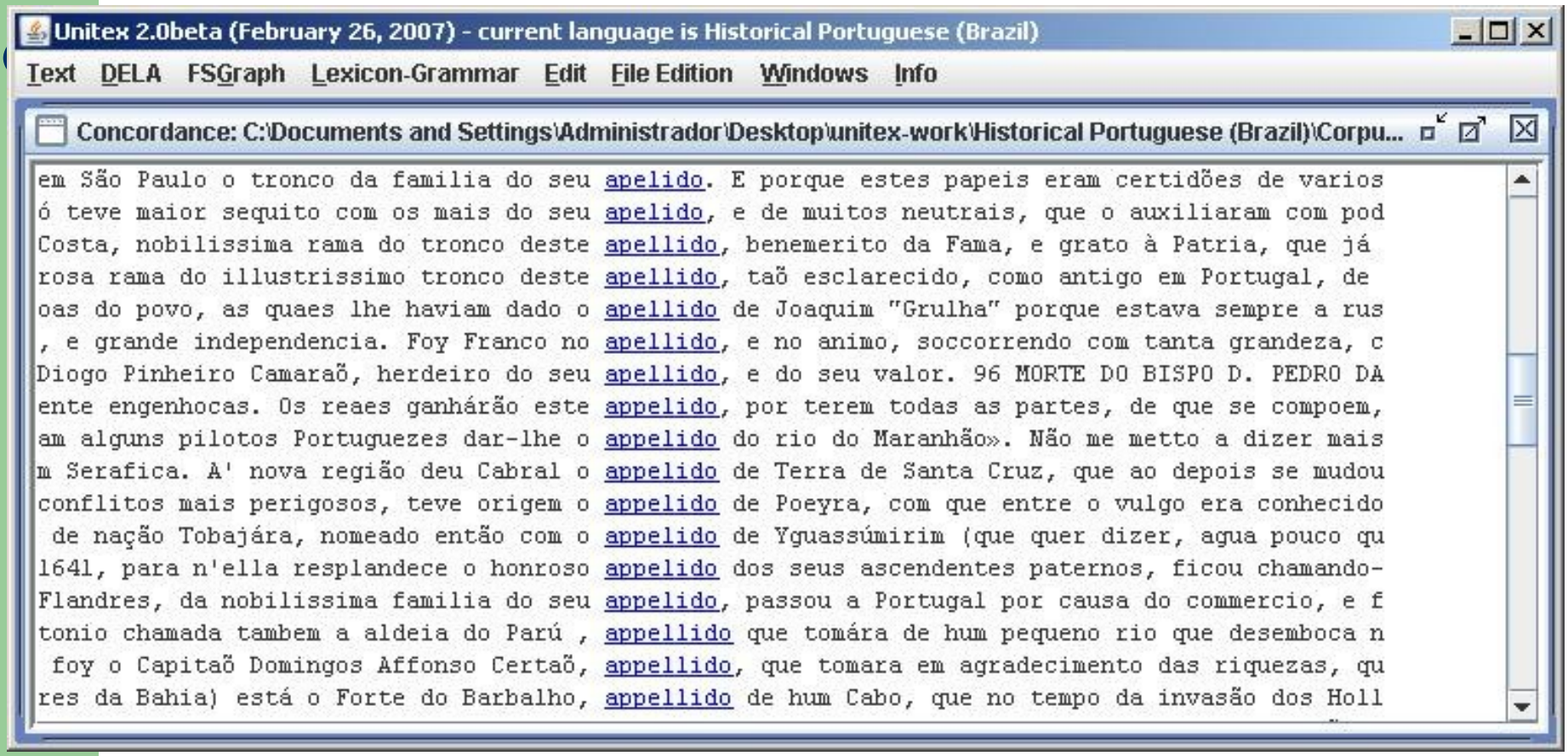
  appellidos,apelidos.N+VAR+apelido:ms/50.0%

Searching entries in Dicionário



Search for variants using *Dicionário system*

Using the dictionary to search the *corpus*



Search in the corpus with the aid of the dictionary of spelling variants

Conclusions and Future Work

- In this work was presented: a methodology and a system to dealing with spelling variants in Portuguese historical texts
- The dictionary of spelling variants is freely available
  - http://moodle.icmc.usp.br/dhpb/spelling-variants.gz

Conclusions and Future Work (2)

- Transformation rules can be an efficient way to detect spelling variations in historical corpora
  - Just forty-three rules can detected almost 30,000 variants in a corpus of 4.9 million words with high precision

- Develop more transformation rules, including phonetic rules
- Include the results from AGREP

References

- Archer, D., A. Ernst-Gerlach, S. Kempken, T. Pilz and P. Rayson (2006) The identification of spelling variants in English and German historical texts: manual or automatic? In E. Vanhoutte et al. (eds.) Proceedings abstracts of Digital Humanities 2006, 3–5. Paris: Sorbonne.
- Crane, G. and A. Jones (2006) The challenge of Virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection, In G. Marchionini et al. (ed.) Proceedings of 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 31-40. Chapel Hill, USA: ACM Press.
- Hauser, A., M. Heller, E. Leiss, K. U. Schulz and C. Wanzeck (2007) Information Access to Historical Documents from the Early New High German Period, In C. Knoblock et al. (eds.) Proceedings of IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data (AND-07), pp. 147-154. Hyderabad, India. Available on-line at http://research.ihost.com/and2007/cd/Proceedings_files/p147.pdf (accessed: 22 june 2007).
- Hirohashi, A. (2005) Aprendizado de regras de substituição para normatização de textos históricos. Master's thesis. IME: Universidade de São Paulo, Brasil. (In Portuguese)

References (2)

- Rayson, P., D. Archer, A. Baron and N. Smith (2006) Tagging historical corpora: the problem of spelling variation, In L. Burnard et al. (eds.) Proceedings of Digital Historical Corpora - Architecture, Annotation, and Retrieval, no. 6491. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI).
- Rydberg-Cox, J. A. (2003) Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries, In G. Henry et al. (eds.) Joint Conference on Digital Libraries (JCDL 2003), 372-373. Houston, USA: ACM Press.
- Sanderson, R. (2006) "Historical Text Mining", Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities, Talk presented at Historical Text Mining Workshop, Lancaster University, UK. Available on-line at http://ucrel.lancs.ac.uk/events/htm06/RobSandersonHTM06.pdf (accessed 22 june 2007).

Comparative Recall

The recall is the fraction of relevant document terms which has been found by a method. If we define R as the set of document terms relevant to a specific query term q, A as the set of document terms found by the method in response to query term q and Ra as the intersection of R and A, the recall is given by

$$\text{Recall} = \frac{|Ra|}{|R|} \tag{4.3}$$

When computing the comparative recall the set R is not defined as the set of all document terms relevant to the query term q, but as the union of the sets of relevant document terms found by any of the methods tested. So

$$R = \cup_{i=1}^{n} R_n \tag{4.4}$$

where n is the number of methods tested and R_n is the set of relevant documents terms found by method n.