AUTOMATIC TERM EXTRACTION AND THESAURUS CONSTRUCTION

Lucelene Lopes
lucelene.lopes@pucrs.br
Roger Leitzke Granada
roger.granada@acad.pucrs.br

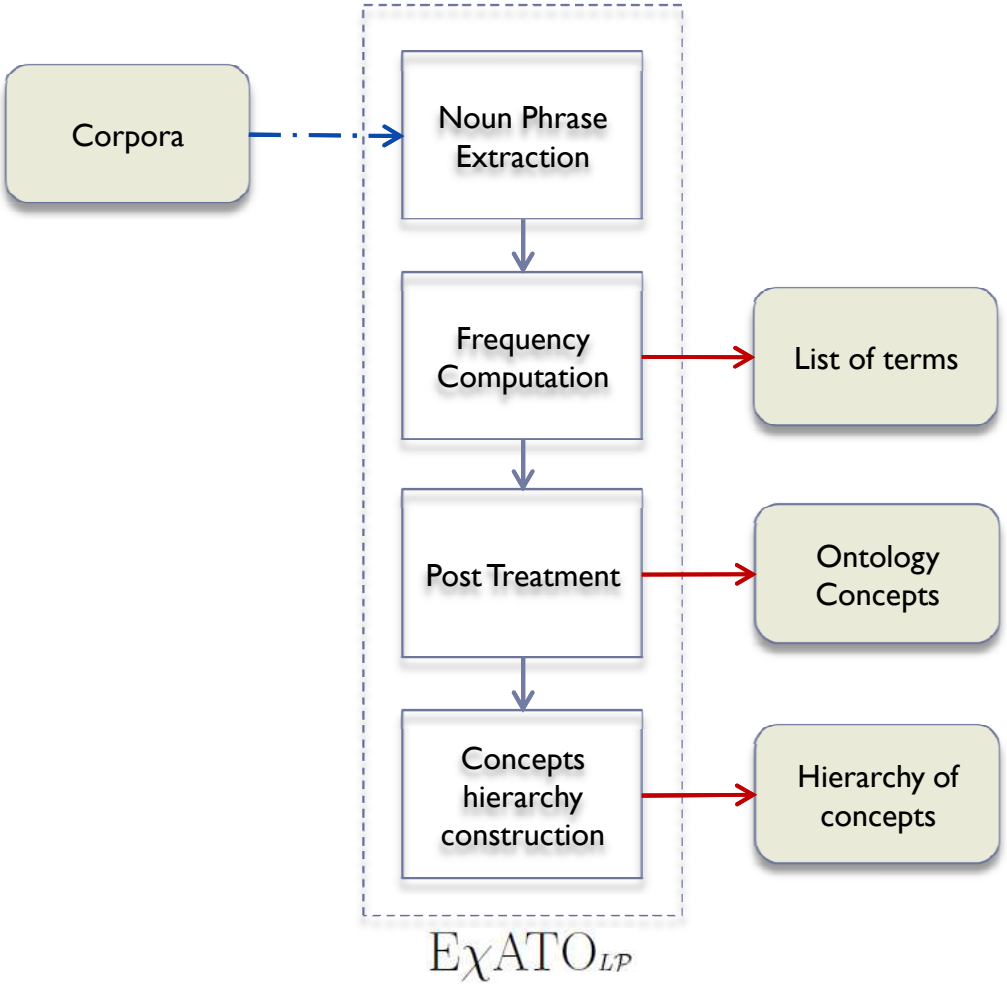
Renata Vieira
renata.vieira@pucrs.br

Who we are

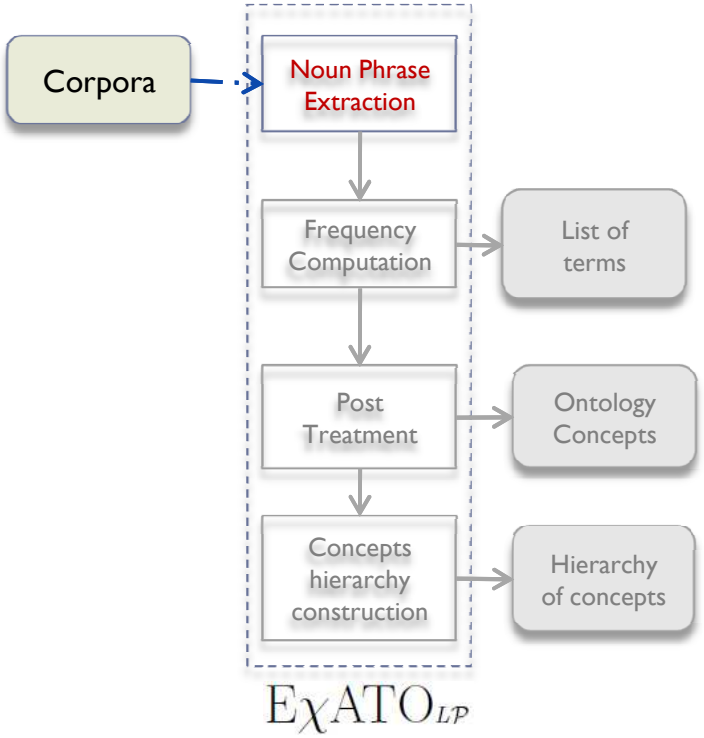
- Lucelene Lopes
  - Phd Student
  - Automatic Term Extraction

- Roger Leitzke Granada
  - Phd Student
  - Automatic Thesaurus Construction

Automatic Term Extraction (EχATO$_{LP}$)

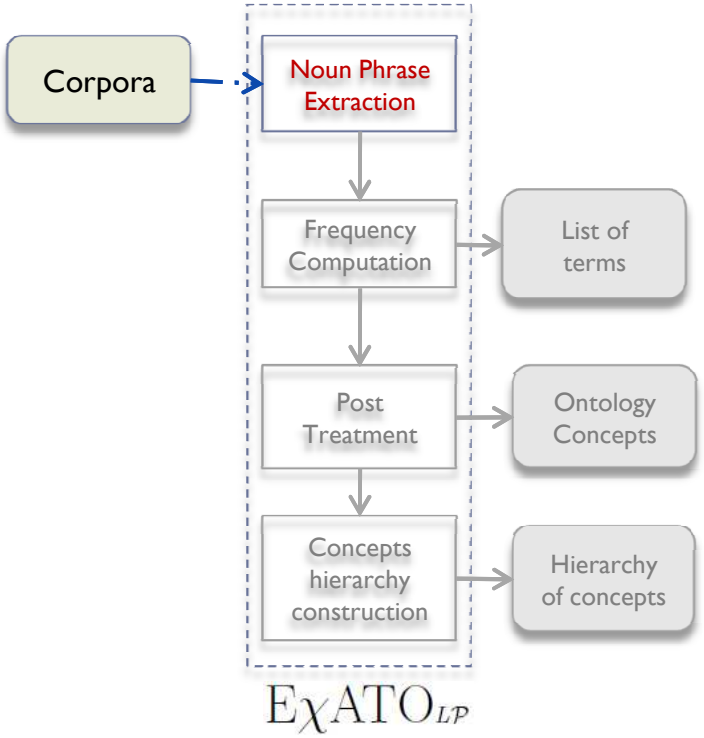


EχATO$_{LP}$ - Noun Phrase Extraction



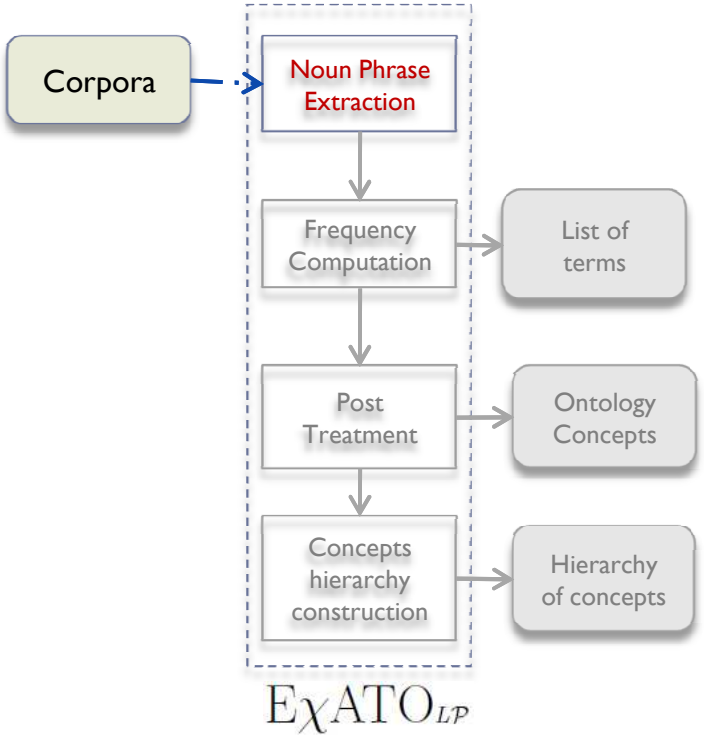
- Input: Tiger XML format (PALAVRAS)
  - Detection of Noun Phrases
  - Application of discard rules
  - Application of transformation rules

EχATO$_{LP}$ - Noun Phrase Extraction



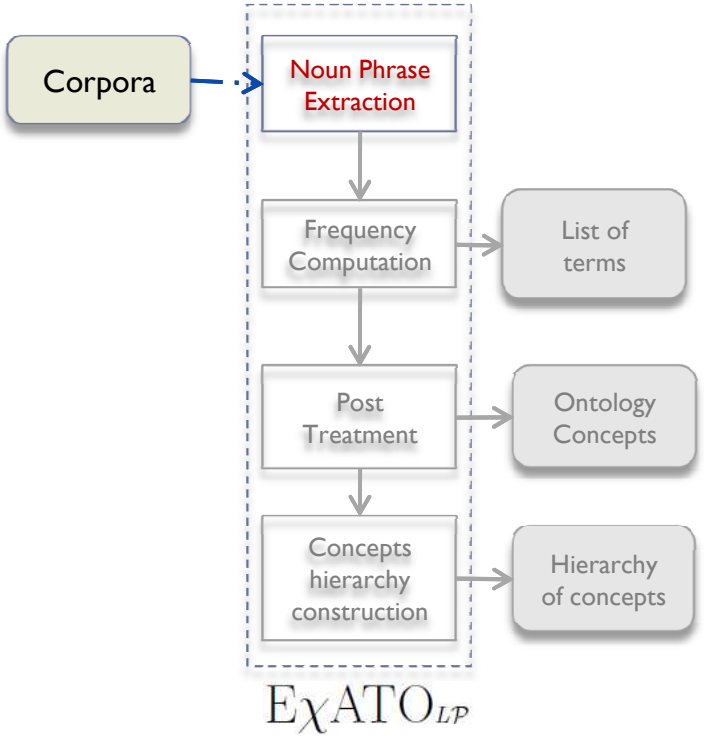
- Input: Tiger XML format (PALAVRAS)
  - Detection of Noun Phrases
    - Original form, canonical form, head, syntactic tag and semantic tag
  - Application of discard rules
  - Application of transformation rules

EχATO$_{LP}$ - Noun Phrase Extraction



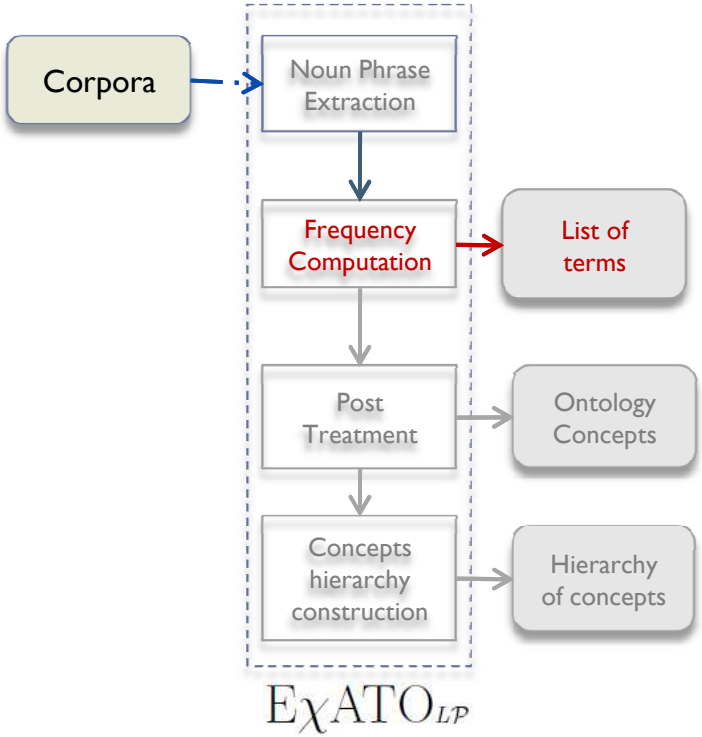
- Input: Tiger XML format (PALAVRAS)
  - Detection of Noun Phrases
  - Application of discard rules
    - Discard numerals
    - Discard special characters
    - Discard annotation errors
  - Application of transformation rules

EχATO$_{LP}$ - Noun Phrase Extraction



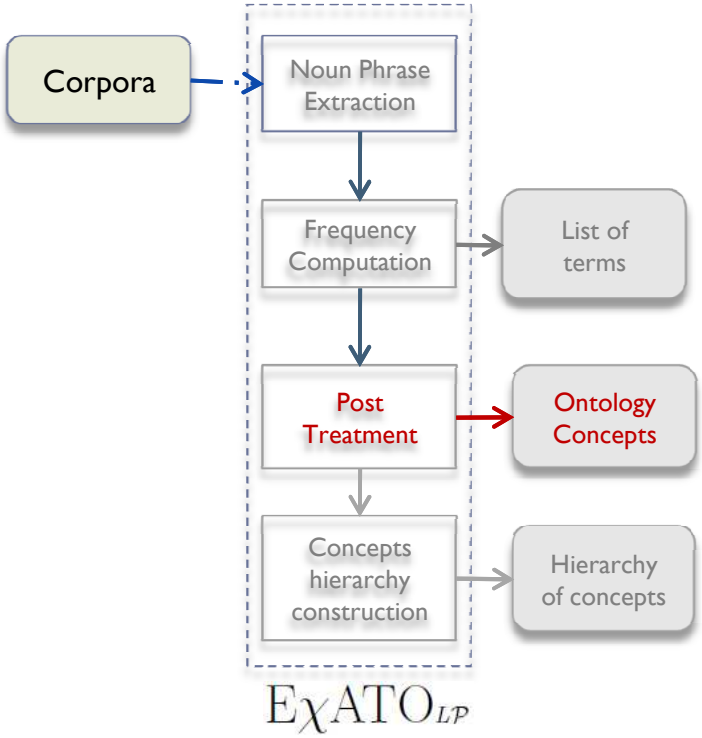
- Input: Tiger XML format (PALAVRAS)
  - Detection of Noun Phrases
  - Application of discard rules
  - Application of transformation rules
    - Conjunction at the end of a NP removal
    - Pronoun at the beginning of a NP removal
    - Article removal
      - From preposition + article
      - Ex: "*Deriva dos continentes*" = *de* + *os*

EχATO$_{LP}$ - Frequency computation



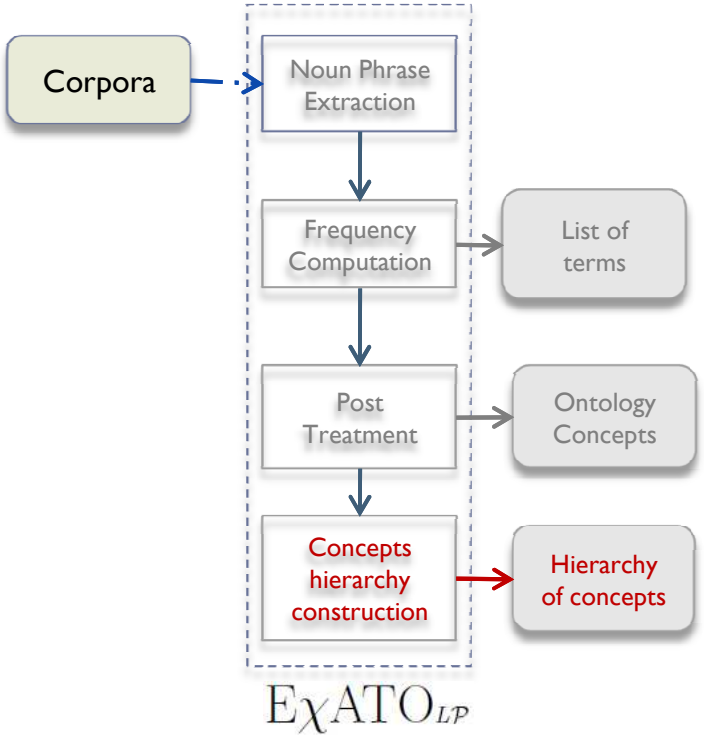
- Computes absolute and relative frequency
  - Unigrams, bigrams…
- Applies a cut-off threshold
  - Discarding less frequent terms
- Output: List of the most frequent terms

EχATO$_{LP}$ - Post-treatment



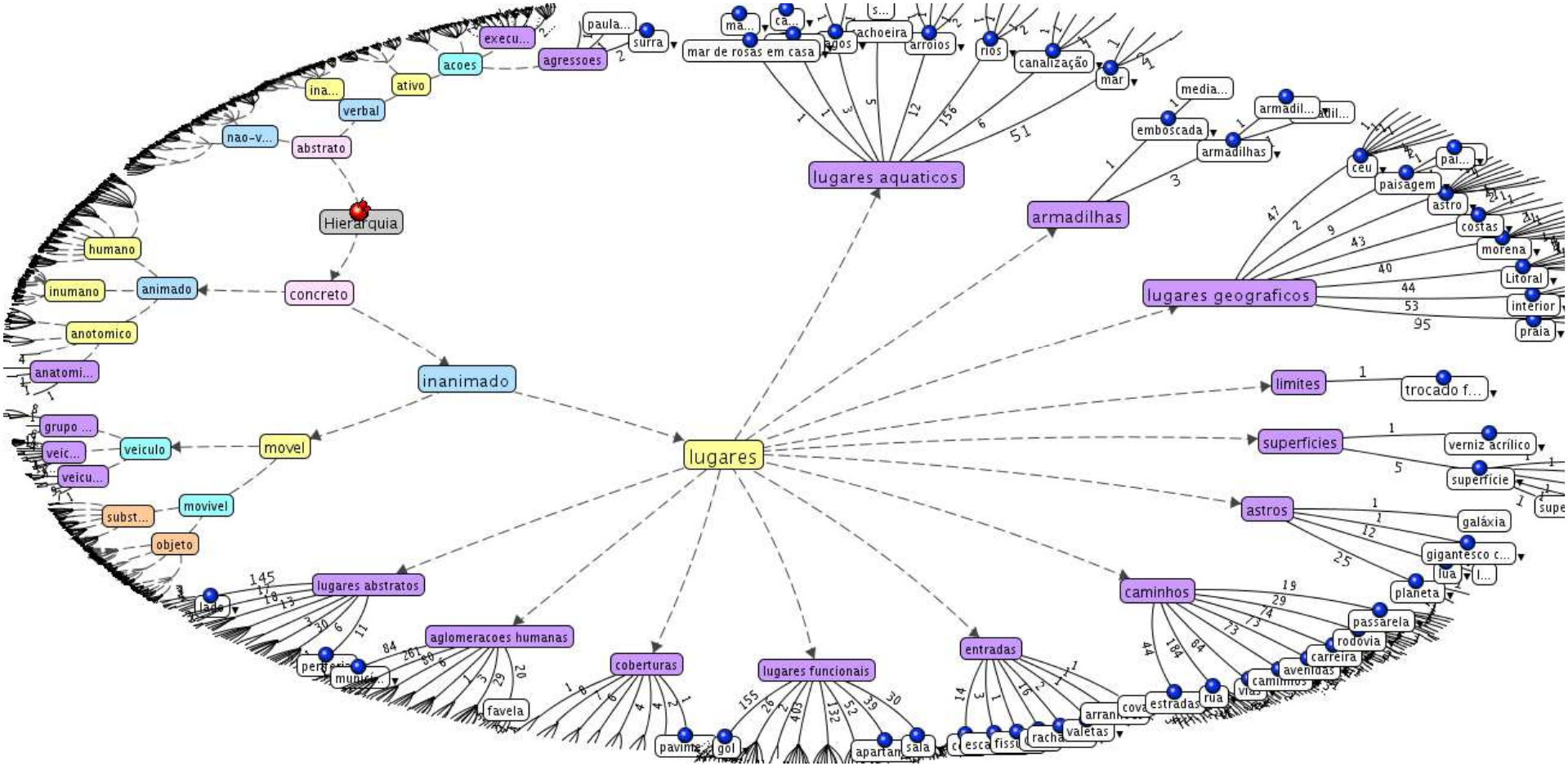
- Refinement of the terms using corpora
  - Better results
- Comparison with a reference list of terms
  - Precision
  - Recall
  - F-measure

EχATO$_{LP}$ - Post-treatment



- Building of a hierarchical structure
  - Based on semantic tags of PALAVRAS

$E\chi ATO_{LP}$ - Semantic hierarchy



- http://www6.ufrgs.br/textecc/porlexbras/porpopular/Treebolic/HVerbos.html

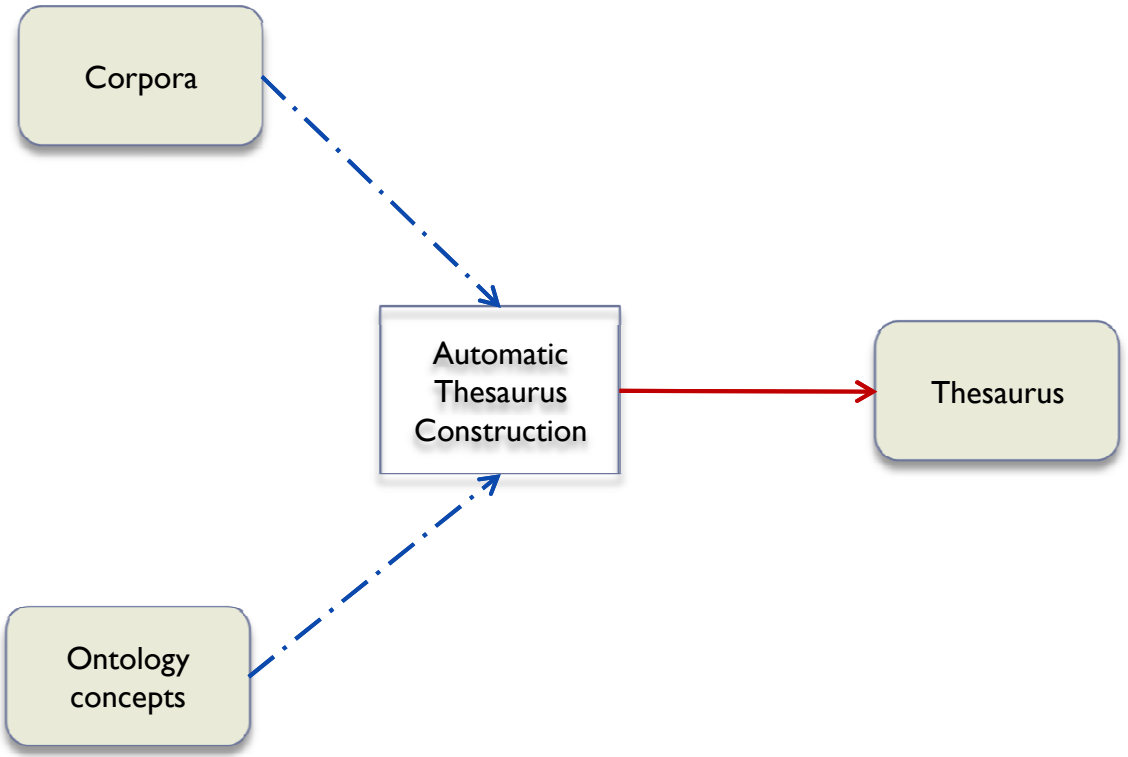
ExATO$_{LP}$ - User's interface

ExATO$_{LP}$ - More about...

- LOPES, L. ; OLIVEIRA, L. H. M ; VIEIRA, R. . Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches. In: International Conference on Computational Processing of Portuguese Language - PROPOR, 2010, 2010. p. 1-6.

- LOPES, L. ; FERNANDES, P. ; VIEIRA, R. ; FEDRIZZI, G. . ExATO lp - An Automatic Tool for Term Extraction from Portuguese Language Corpora.. In: Proceedings of the Fourth Language and Technology Conference LTC'09, 2009. p. 427-431.

Automatic Thesaurus Construction



What is a thesaurus?

- Thesaurus for "*tezgüino*":

  A bottle of tezgüino is on the table.
  Everyone likes tezgüino.
  Tezgüino makes you drunk.
  We make tezgüino out of corn.

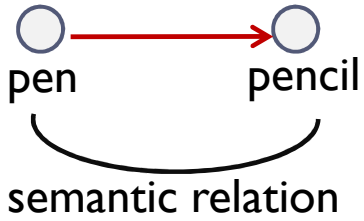
  - Seed: *tezgüino*
    - Related: beer
    - Related: wine

D. Lin. "Automatic retrieval and clustering of similar words". In: Proceedings of the 17th international conference on Computational linguistics. 1998, pp. 768-774.

Methods to build thesaurus

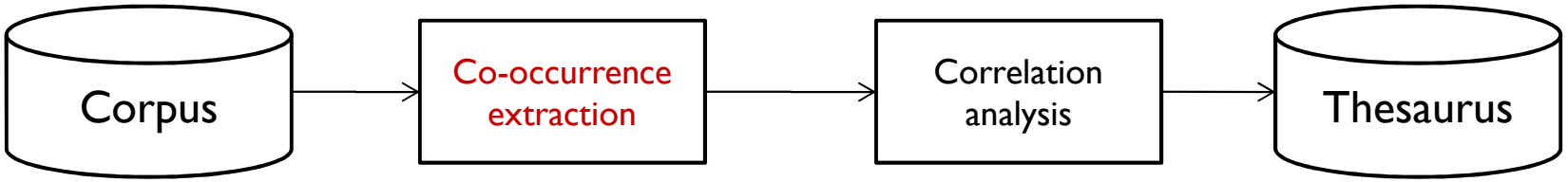
- Identifying **first order co-occurrence**
  - Extract terms in a window
  - First order relation between terms

» Ex: "The man bought a pen and a pencil."
  » "Man bought pen pencil"
Size of the window = 3
Relations: "<man, bought>",
"<man, pen>"
"<bought, pen>",
"<bought, pencil>"
"<pen, pencil>"



How to build?

- Identifying **first order co-occurrence**



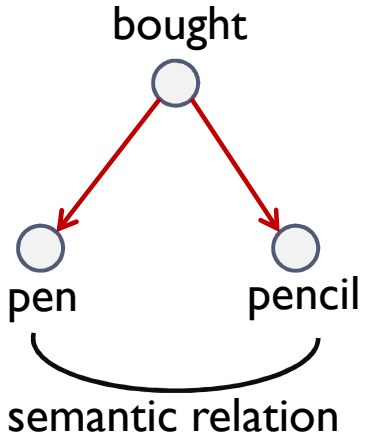
Extraction in a Window

Using Mutual Information

Methods to build thesaurus

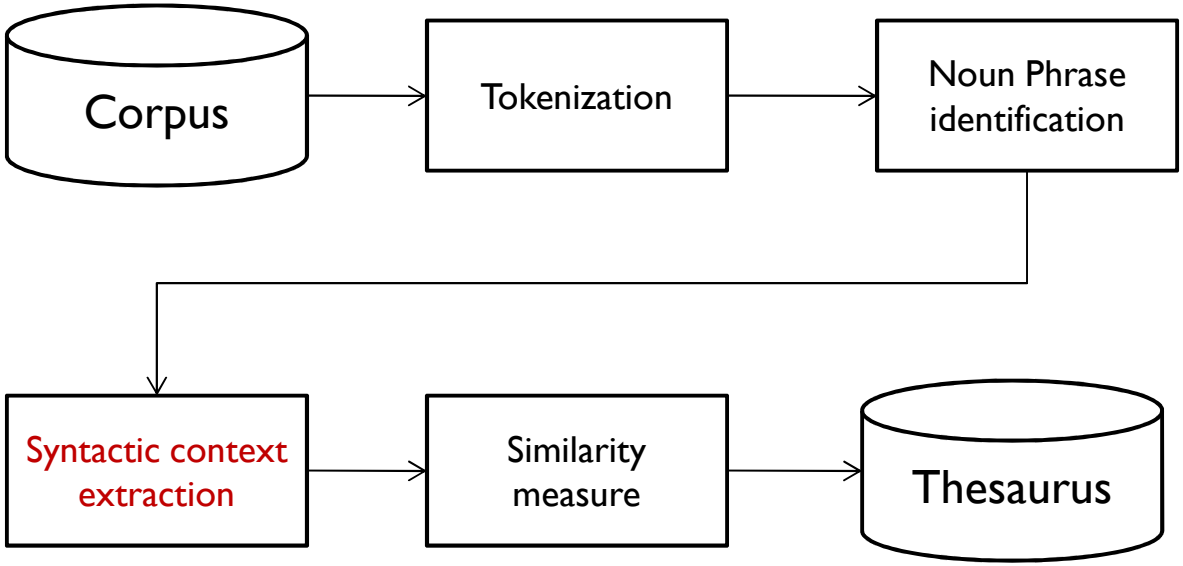
- Identifying **second order co-occurrence**
  - Syntactic contexts extraction
  - Terms are related because they share the same modifiers

» Ex: "Yesterday I bought a pen. Today I bought a pencil."
Relations: "<bought, pen>",
"<bought pencil>"



How to build?

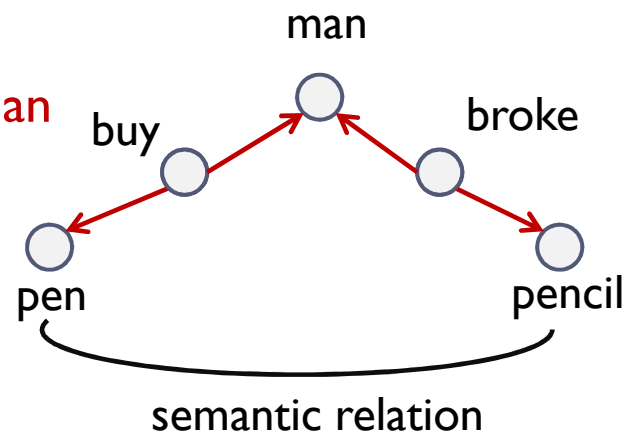
- Identifying **second order co-occurrence**



Methods to build thesaurus

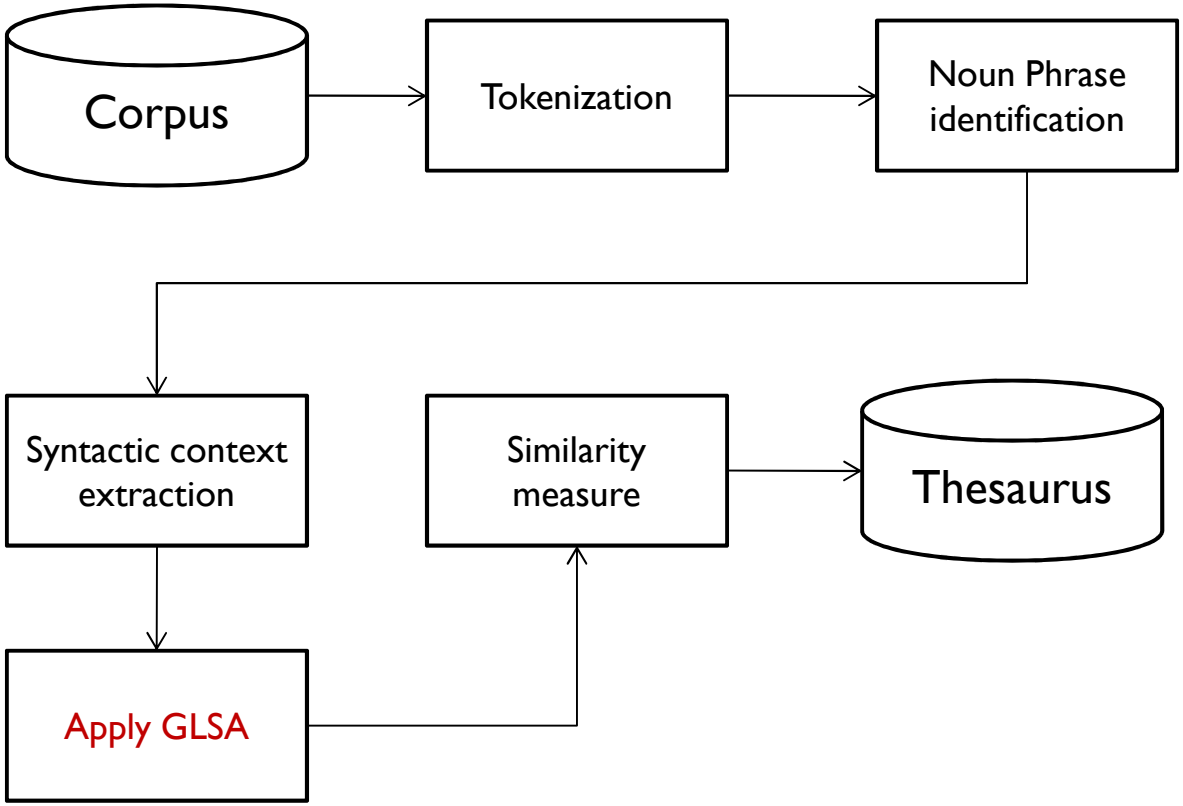
- Identifying **third or more order co-occurrence**
  - Application of General Latent Semantic Analysis
  - Terms are related because they share terms that share the same modifiers

» Ex:"A man buy a pen. This man broke a pencil."
Relations: "<pen, man>" through buy
"<pencil, man>" through broke
Latent relations: "<pen, pencil>" through man



How to build?

- Identifying **third or more order co-occurrence**



Automatic Thesaurus Construction

- Output of the automatic thesaurus construction:
  - Seed: car
    - Related: bus
    - Related: sleeping car
    - Related: …

- Related terms give the meaning to the seed

AUTOMATIC TERM EXTRACTION AND THESAURUS CONSTRUCTION

Lucelene Lopes
lucelene.lopes@pucrs.br
Roger Leitzke Granada
roger.granada@acad.pucrs.br

Renata Vieira
renata.vieira@pucrs.br