



# Natural Language Processing: Dealing with unstructured data

Renata Vieira and Joaquim Neto

FGV Workshop on Data Science  
2019

# Day 2

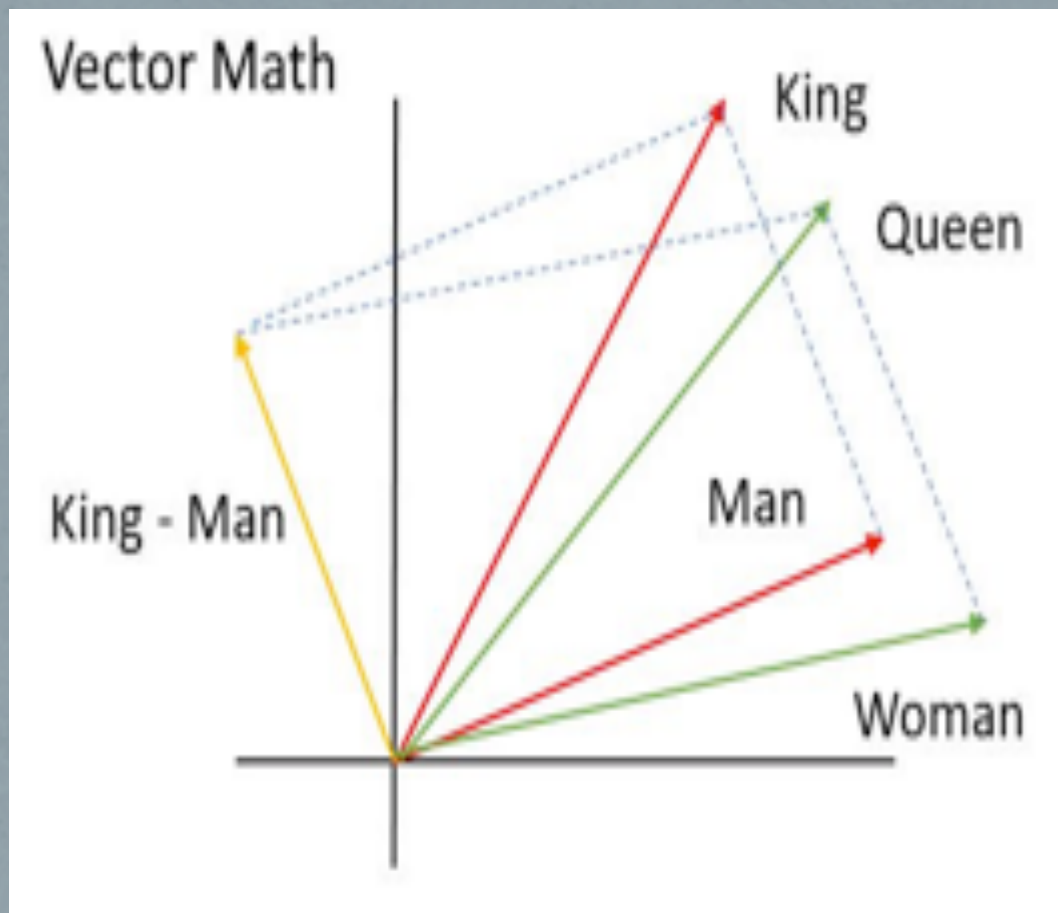
---

- Part 1
  - Ontologies and NLP
  - Research developed in these areas
- Part 2
  - Hands on
  - Task: Named entity recognition

# Word Embeddings

Words are complicated

Let's do it with numbers



# Word Embeddings

“Mathfication”  
of semantics

```
homen 0.07261448 0.011477063 -0.059029713 -0.06815444 0.000315139 0.13613994 -0.010905103 0.03647272
0.0046794554 0.012363336 0.05122622 0.06120117 -0.025945924 -0.10619497 -0.011104553 0.03546393
-0.049336046 -0.014955301 0.018162776 0.098726794 0.11303935 0.10115415 0.03336638 -0.1211666
0.087380424 0.0076276166 -0.033710524 -0.08676267 0.012442611 0.044099476 -0.039013147 -0.05161607
-0.03466344 0.05581801 0.005582102 0.06978353 -0.061807737 0.083326966 -0.002639462 0.075194046
0.06816695 -0.030302491 -0.0444801 0.07226238 -0.13232417 -0.042606737 0.00024308544 -0.11833146
0.04990067 0.008662857 0.048492823 0.0031767057 0.028667089 0.0056069755 -0.050273538 -0.0902299
0.015317945 0.015738878 -0.08506806 -0.043690708 -0.046372592 -0.08591505 -0.03909378 0.003179637
0.06132095 0.053531215 -0.0058316817 0.016137231 -0.08680149 -0.0040189293 -0.07468639 -0.07766885
0.0237779 -0.044383865 -0.036448117 0.17054246 -0.12848511 0.05719996 -0.02637475 -0.08567893
-0.10194281 -0.003074954 -0.0004163344 0.010612284 0.00090030336 0.0908042 0.067862526 -0.036803927
-0.07499598 0.013177385 -0.028992051 -0.070595205 0.034598302 0.00043570256 -0.023202796 -0.09209203
0.05096268 0.084347956 0.06024575 0.05295832 -0.0032937685 -0.061215702 -0.1137704 -0.03777014
0.06371082 -0.014589474 -0.09514905 0.0010437447 0.04241006 -0.09469389 0.08812129 -0.15705208
0.13484064 0.005074025 -0.02087597 0.0112976665 -0.06631411 0.014097601 0.031810734 0.03818184
-0.00044033732 0.0273135 0.060578868 -0.04485547 -0.19702299 0.041902576 -0.02765138 0.063520804
0.012723747 -0.113430455 0.064140104 0.040206973 0.01891557 -0.015529406 0.017346673 0.08863067
0.046590216 0.0116492845 0.09702383 -0.054204565 0.04783237 -0.06626555 -0.0042909663 -0.021025686
-0.025487859 -0.047604036 -0.028673595 0.047583032 0.003949422 0.012293308 -0.010999885 -0.055124238
0.02525984 -0.0666319 -0.017591277 -0.06522282 0.020636952 0.087872066 -0.024650618 0.0738292
-0.058370646 -0.03861952 0.027964164 -0.030344957 0.02267978 -0.019425957 0.06508563 -0.097758114
0.054441534 0.002201869 0.0012622409 -0.025332088 -0.032579288 0.10625123 0.05849593 -0.006813332
0.030056842 0.053884037 0.01207854 -0.012244217 0.014707071 -0.096495904 -0.0006245469 0.01814357
-0.010457365 0.011510279 -0.055072326 0.010400103 -0.029228477 0.092694186 0.03617669 -0.003040525
0.08626235 0.07363217 -0.09763069 -0.03824259 0.08025189 0.008204898 -0.047679786 0.03126295
0.025605323 0.010096034 -0.041474633 0.09828671 0.0006555138 -0.014995557 -0.029222693 0.0074388934
-0.09993945 0.011226459 -0.0036442268 0.022372002 -0.051770672 0.0055837794 0.035182234 -0.041797873
-0.035836212 -0.05107003 0.060495876 -0.07280102 0.04319779 0.037585646 -0.083571605 -0.036574934
0.029968044 0.024327407 -0.06921862 0.06985777 -0.058870323 -0.03389398 0.005586085 -0.0007671036
-0.009262291 0.0071850438 -0.00462656 -0.028525935 0.041870175 0.052383933 0.07833644 0.0027441343
0.022994895 0.02505833 -0.02868195 -0.038224958 0.05325009 -0.005883336 -0.023396244 0.0696365
-0.06898364 -0.0029008826 0.021996574 -0.07313164 0.029192403 0.040505964 -0.06797771 0.023516202
-0.054204125 0.0438581 -0.05967696 -0.04818756 -0.0019798842 -0.00084776996 -0.03431618 -0.045810133
0.016368818 -0.07088095 0.011881486 0.016981719 -0.026865609 -0.0007908625 -0.016643373 -0.0031557288
-0.07521733 0.03494255 0.057291813 0.11807121 -0.002152817 0.06028637 0.055090886 -0.014183729
0.04670447 0.013351276 0.016821636 0.07800454 0.055741187 0.03986755 0.038796682 0.06390219
-0.105016455 -0.12859024 -0.13362458 -0.05233001 -0.07458396 0.09157962 0.012326395 0.007187728
0.01306481 0.041843165 0.134377 0.07176255]
```



**BANCO**













**BANCO WE**

```
banco 0.181041 0.107700 -0.104667 0.243361 0.060638 0.392829 -0.333944 -0.381778 0.142200  
0.085936 -0.116615 0.395722 -0.612684 -0.076898 0.334396 0.081127 -0.051770 -0.321950  
-0.691509 -0.331210 -0.543213 0.609881 0.243700 0.037324 0.116518 0.178859 -0.378839  
0.127430 0.194497 0.000732 0.314395 -0.204550 0.534431 -0.005551 0.352343 -0.049200  
-0.138384 0.023163 -0.340013 0.500201 -0.011417 -0.129925 -0.006128 -0.180481 0.199391  
0.137645 -0.766434 -0.226784 -0.061611 0.090592
```



**Ambiguidade**



O que é um  
“banco”?

**Depende**

# Pesquisa

- Estudar polissemia em word embeddings
  - Amenizar o efeito de polissemia em vetores de palavras
- 
- Voltamos aos métodos simbólicos



# Ontologies for NLP

## NLP for Ontologies





# Overview

---

- Introduction
- NLP for Ontologies
- Ontologies for NLP
- Related research (6 PhD Thesis)

# Introduction

We think, talk, write, store and share

A lot more to think about (and much to read)

We think about the way we think and talk to build machines to help us communicating



# From the same fundamental principles



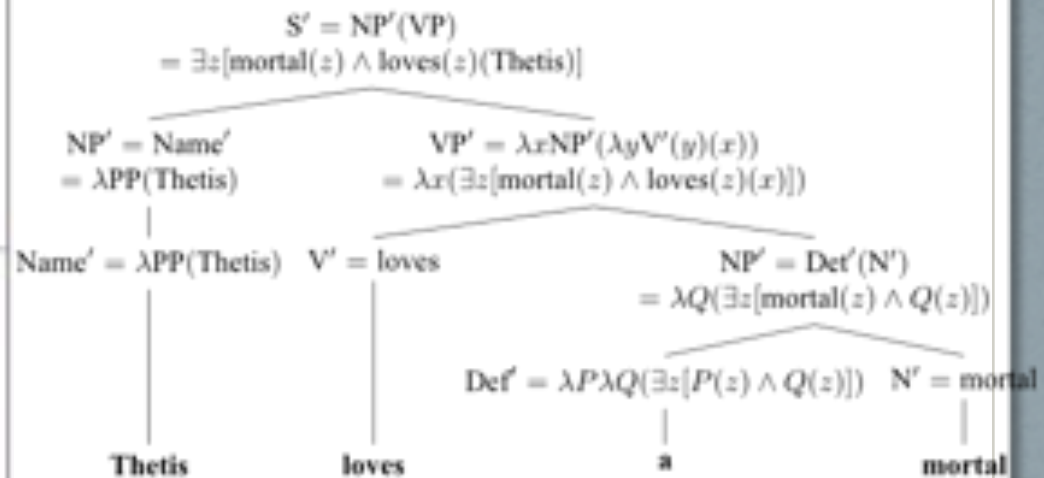
# NLP x Ontologies



© 2004 - 2006 whutbird.com



© 2004 - 2006 whutbird.com





# NLP x Ontologies

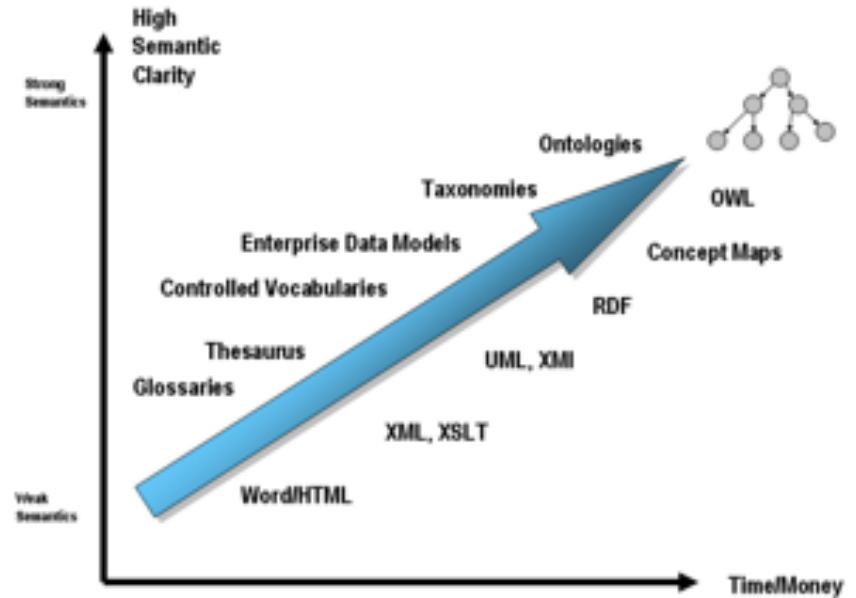
- How do they converge, need/influence each other?
  - NLP for building ontologies from textual knowledge
  - Ontologies to make more semantically oriented NLP

# NLP for Ontologies

Ontology extraction/learning from texts

# Ontologies

- Spectrum
  - Terms
  - Glossary
  - Thesaurus (narrower term)
  - Is-a hierarchies
  - Properties
  - Instances
  - Logical constraints
  - Axioms



[wiki.opensemanticframework.org](http://wiki.opensemanticframework.org)

**SEMANTICS**

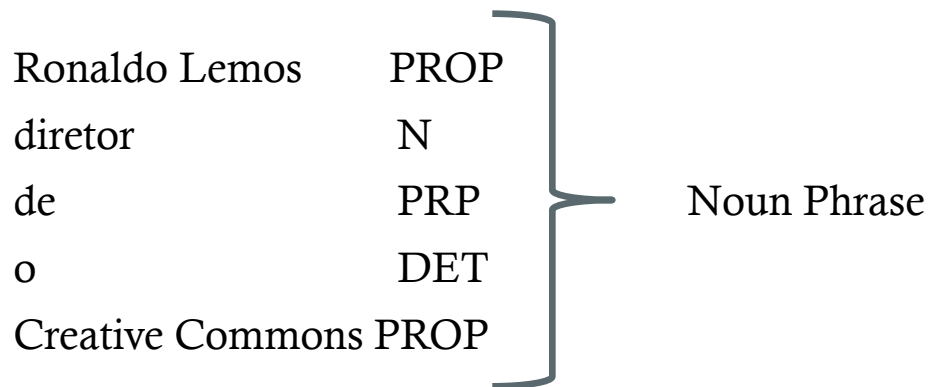
# Ontology learning from text

- Ontology components - NLP
  - Concepts – term extraction
  - Hierarchy – is-a relation
  - Properties – other relations
  - Instances – named entities
- Basic NLP needed for ontology learning
  - POS tagging (word classes: verbs, nouns, adjectives, etc.)
  - Parsing (word groups: noun phrases, verb phrases, etc.)
  - PLUS – statistical processing and machine learning



# Basic NLP: POS and Parsing

Ronaldo Lemos, diretor do Creative Commons aprovou ontem ....



# NLP for Ontologies

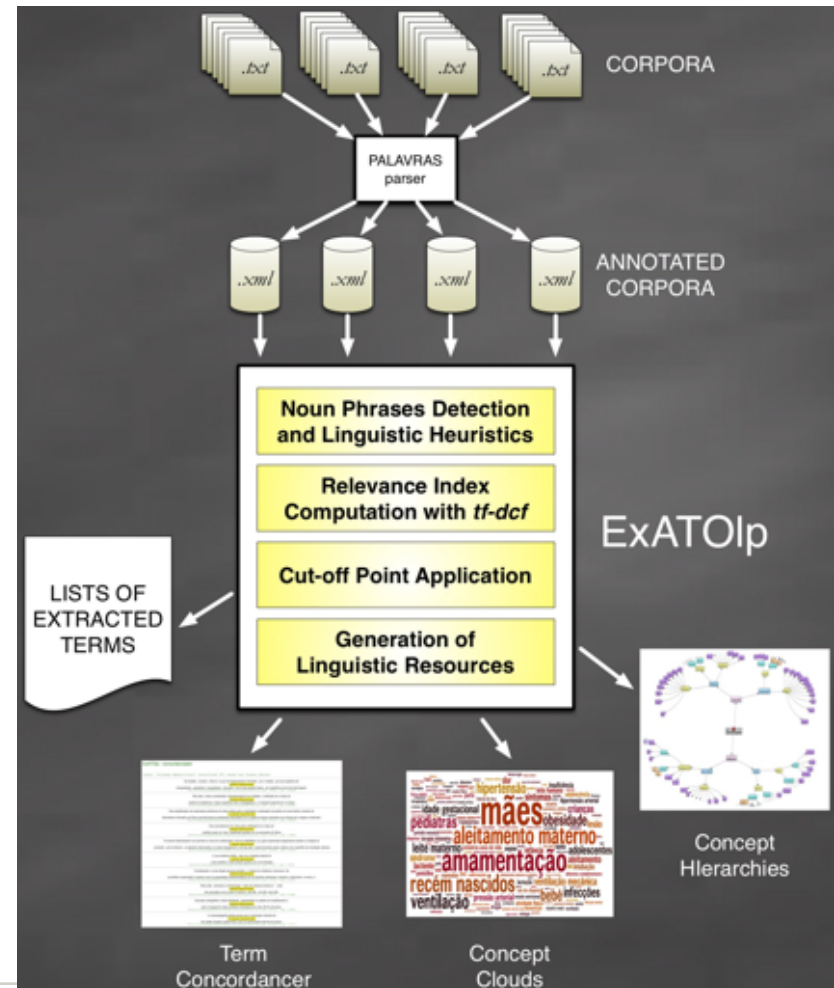
Related research at PUCRS

# NLP for Ontologies

- Ontology learning layer by layer
  - **Concepts (Lucelene Lopes)**
  - Hierarchy
  - Properties
  - Instances

# Concept Extraction

- Input: Parsed Corpora
- Term Extraction
- Relevance Computation
- Concept Identification
- Resources Generation
  - Lists
  - Concordancer
  - Clouds
  - Hierarchies





# ExATO/p – Portuguese term extraction

	A	B	C	D	E	F
	leia	leitura	leitor	leitura	le	leitor
1	leia de refrigeração	leia de refrigeração	leia	leitor	4	4,0
2	leia elétrica	leia elétrica	leia	leitor	3	3,0
3	leia antiga	leia antiga	leia	leitor	2	2,0
4	leia australiana	leia australiana	leia	leitor	2	2,0
5	leia crescente	leia crescente	leia	leitor	2	2,0
6	leia europeia	leia europeia	leia	leitor	2	2,0
7	leia hinduista	leia hinduista	leia	leitor	2	2,0
8	leia interior	leia interior	leia	leitor	2	2,0
9	leia meridional	leia meridional	leia	leitor	2	2,0
10	leia profunda	leia profunda	leia	leitor	2	2,0
11	leia técnica	leia técnica	leia	leitor	2	2,0
12	leia transgressiva	leia transgressiva	leia	leitor	2	2,0

(28)

ExWJOg 1, 2, 4 - concordado

Labels: **parent** is constructed if  $\text{parent} \neq \text{null}$

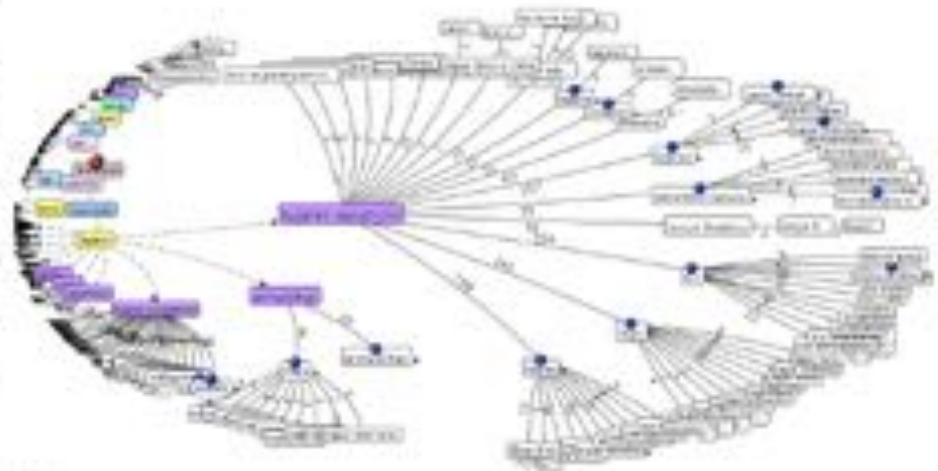
©2004 by the author. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or by any information storage or retrieval system, without permission in writing from the author.

[illegible]

(b)



(c)



(d)

# Extraction Heuristics

Nosso petróleo é uma riqueza mineral e abundante, considerando depósitos marinhos.

SN tagged by PALAVRAS	Heuristic	SN extracted by ExATOlp
nosso petróleo	pronoun removal	petróleo
uma riqueza mineral	article removal adjective removal adjective conjunction	riqueza mineral riqueza riqueza abundante
depósitos marinhos	lemma adjective removal	depósito marinho depósito

# Extracted Term Lists

Statistically chosen relevant terms according to *tf-dcf* index  
(using contrastive corpora)

term	lemma	head	sem	tf	tfdcf
arenitos	arenito	arenito	mat	516	516.00
granito	granito	granito	anbo mat cc drink	374	374.00
fácies	fácies	fácies	percep-f	817	272.33
Grupo Itararé	grupo itararé	Grupo Itararé	inst	236	236.00
crosta	crosta	crosta	lcover	212	212.00
sistemas deposicionais	sistema deposicional	sistema	ac	202	202.00
feições	feição	feição	f-an percep-f	377	188.50
lagoa	lagoa	lagoa	lwater	188	188.00
biotita	biotita	biotita	hifam	182	182.00
planícies	planície	planície	ltop	174	174.00
estratificação cruzada	estratificação cruzado	estratificação	event	151	151.00
folhelhos	folhelho	folhelho	anbo con	150	150.00
Bacia de Campos	bacia de campos	Bacia de Campos?		150	150.00
feldspato	feldspato	feldspato	cm-chem	142	142.00
margem continental	margem continental	margem	am ltop Labs cc-r	141	141.00
litologias	litologia	litologia	?	361	139.65

# Evaluation of the new proposed relevance index

Pediatric corpus and reference lists - 15% of the extracted terms

	P	R	F	P	R	F
PALAVRAS SN Detection	12%	13%	13%	13%	7%	9%
ExATOlp Linguistic Heuristics	60%	68%	64%	68%	40%	50%
ExATOlp Relevance Index - <i>tf-dcf</i>	72%	93%	81%	77%	93%	84%
	bigrams			trigrams		



# Proposed Index – *tf-dcf*

Top ranked bigrams for Pediatrics corpus

rank	according to <i>tf</i>	according to <i>tf-dcf</i>
1	aleitamento materno	aleitamento materno
2	recém nascido	recém nascido
3	faixa etária	leite materno
4	presente estudo	idade gestacional
5	leite materno	ventilação mecânica
6	idade gestacional	via aérea
7	ventilação mecânica	pressão arterial
8	via aérea	leite humano
9	pressão arterial	hipertensão arterial
10	sexo masculino	terapia intensiva

# Concordancer

## Terms occurrences with context information

ExATOL v. 2.0 - concordanciador

Termo **arenito** encontrado 516 vezes nas frases abaixo

Clique em um termo para ver detalhes da sua ocorrência na frase

---

Em contraste com as rochas sedimentares de a Formação Águas Claras , constituídas predominantemente por rochas textural e composicionalmente maduras , formadas por quartzo arenitos ,

**arenitos**

sublíticos e argilitos depositados em ambiente de plataforma marinha a litorânea e fluvial entrelaçado , ocorrem arenitos com baixa maturidade composicional e textural .

função gramatical: CJT - núcleo: **arenito** - etiqueta sintática: n - etiqueta semântica: mat

Ocorrência 3 (rank: 1) retirada do arquivo

/Users/lucelene/Documents/POB\_DOC/corpora/geo/Txts71.txt.xml (frase: 135)

---

Coefficientes de a função discriminante utilizado para a identificação de

**os arenitos**

e carbonatos de o Campo\_de\_Namorado , Bacia\_de\_Campos , Brasil .

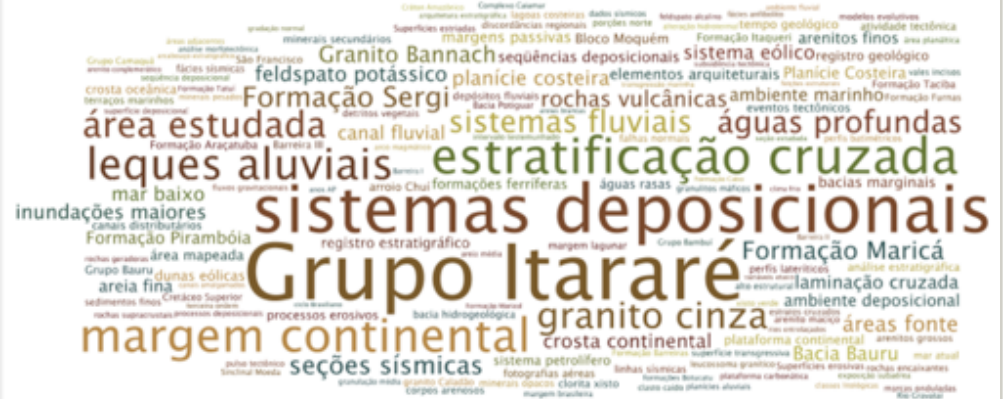
função gramatical: CJT - núcleo: **arenito** - etiqueta sintática: n - etiqueta semântica: mat

Ocorrência 4 (rank: 1) retirada do arquivo

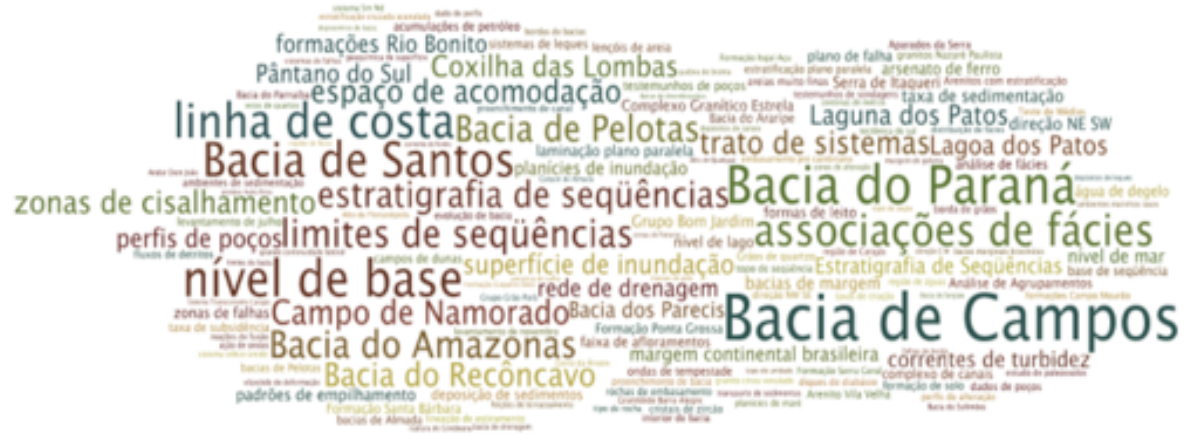
/Users/lucelene/Documents/POB\_DOC/corpora/geo/Txts85.txt.xml (frase: 75)

---

# Concept Clouds



# Representation according to relevance



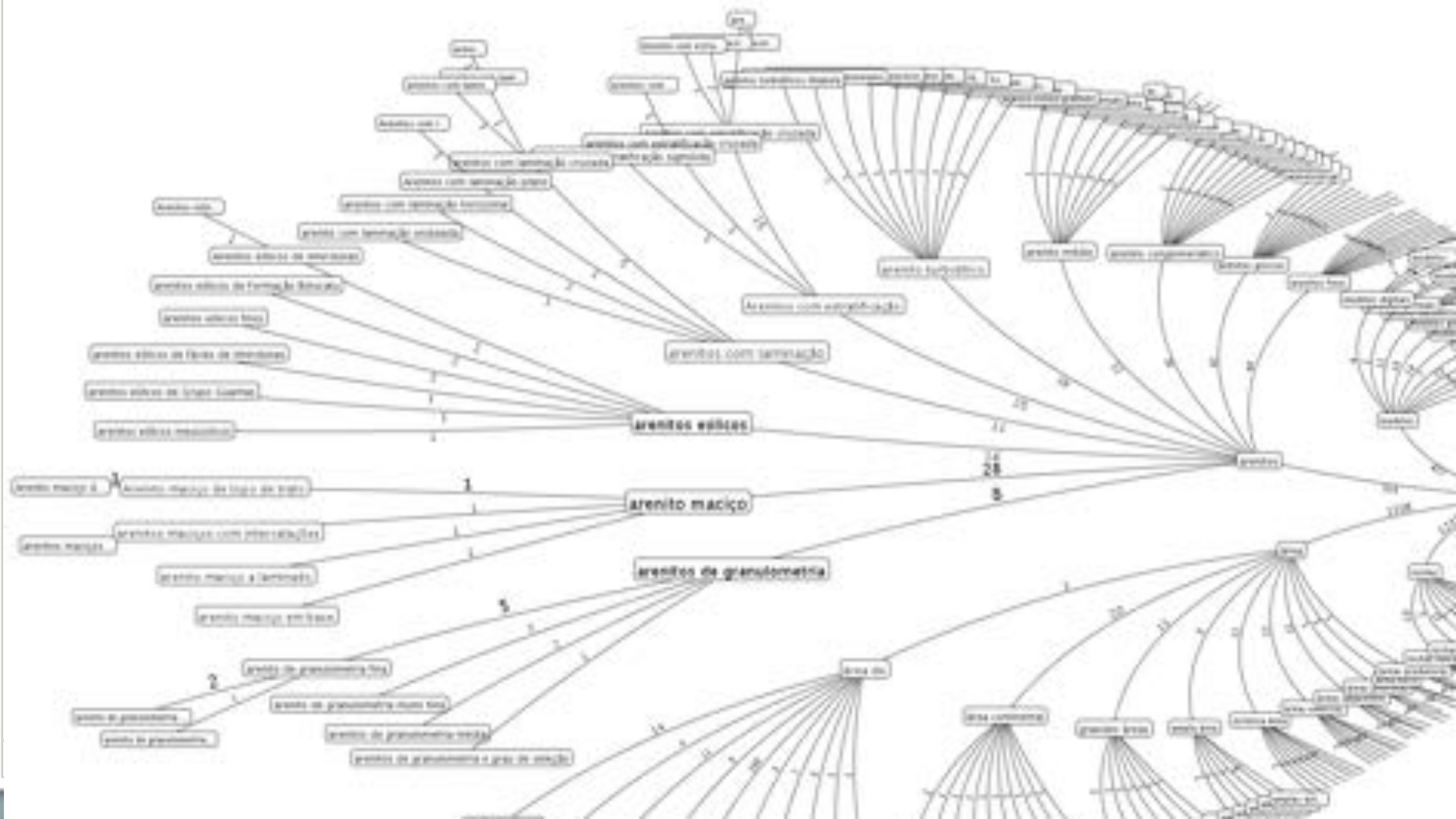
# Hierarchies

- Some hierarchical relations are also given by the tool
  - Semantic classes (parser)
  - Noun phrase structure
    - Arenito
      - Arenito maciço



# Concept Hierarchies

Based on NP structure: head modifiers



## Based on semantic categories

# References

**Lopes, L.** Extração Automática de Conceitos a partir de Textos em Língua Portuguesa - Tese de Doutorado. Porto Alegre: PUCRS, 2012.

**Lopes, L. ; Fernandes, P. H. L. ; Vieira, R .** Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. Knowledge-Based Systems, 2016.

# NLP for Ontologies

- Ontology learning
  - Concepts
  - **Hierarchy (Roger Granada)**
  - Properties
  - Instances

# Hierarchy

PhD Student Roger Granada

- Comparison of several methods of hierarchy extraction from texts
  - 2 Rule-based methods
  - 2 Statistical-based methods



# Hierarchy extraction

## Lexico-syntactic patterns

“...os vários ambientes que compõem os rios, tais como planícies de inundação, canais, macroformas e depósitos de transbordamento.”

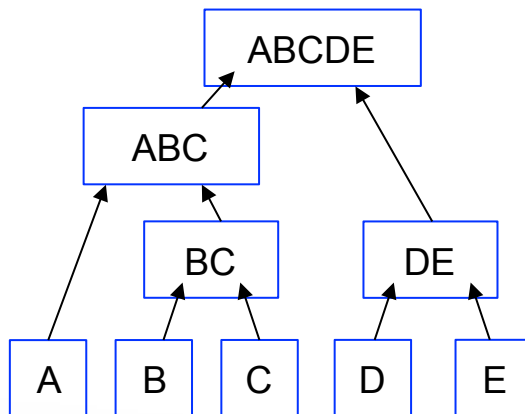
## Head modifier

Arenito

arenito aeolico

arenito macico

## Hierarchical clustering



Clusters are generated based on the contexts of each word

## Co-occurrence analysis

A term  $x$  subsumes  $y$  if the documents in which  $y$  occurs are a subset of the documents in which  $x$  occurs.

$$P(x|y) > P(y|x) \text{ and } P(x|y) > \text{threshold}$$

# Hierarchy extraction

## Lexico-syntactic patterns

Only extracts relations inside the same phrase.

High precision, low recall

## Head modifier

Only extracts relations inside a noun phrase.

High precision, low recall

## Hierarchical clustering

Uses contexts to extract relations.  
May generate other semantic relations, like synonymy, meronymy, etc.

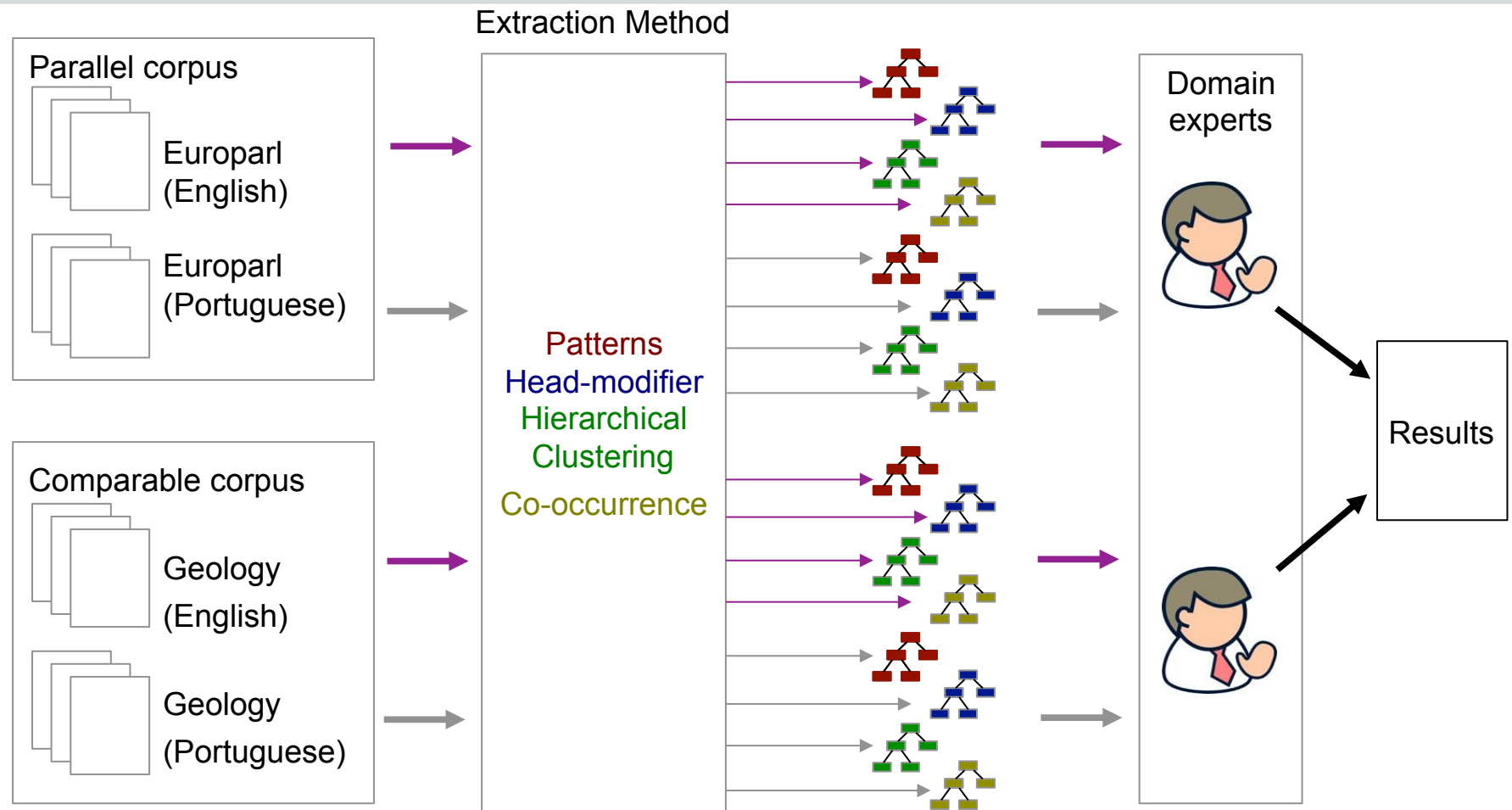
Low precision, high recall

## Co-occurrence analysis

Uses the co-occurrence of terms in documents, generates relations even if the terms are not semantic related.

Low precision, high recall

# Hierarchy extraction



# References

---

**Granada, R.** Evaluation of Methods for Taxonomic Relation Extraction from Text. PhD Thesis, Porto Alegre:PUCRS, 2015.

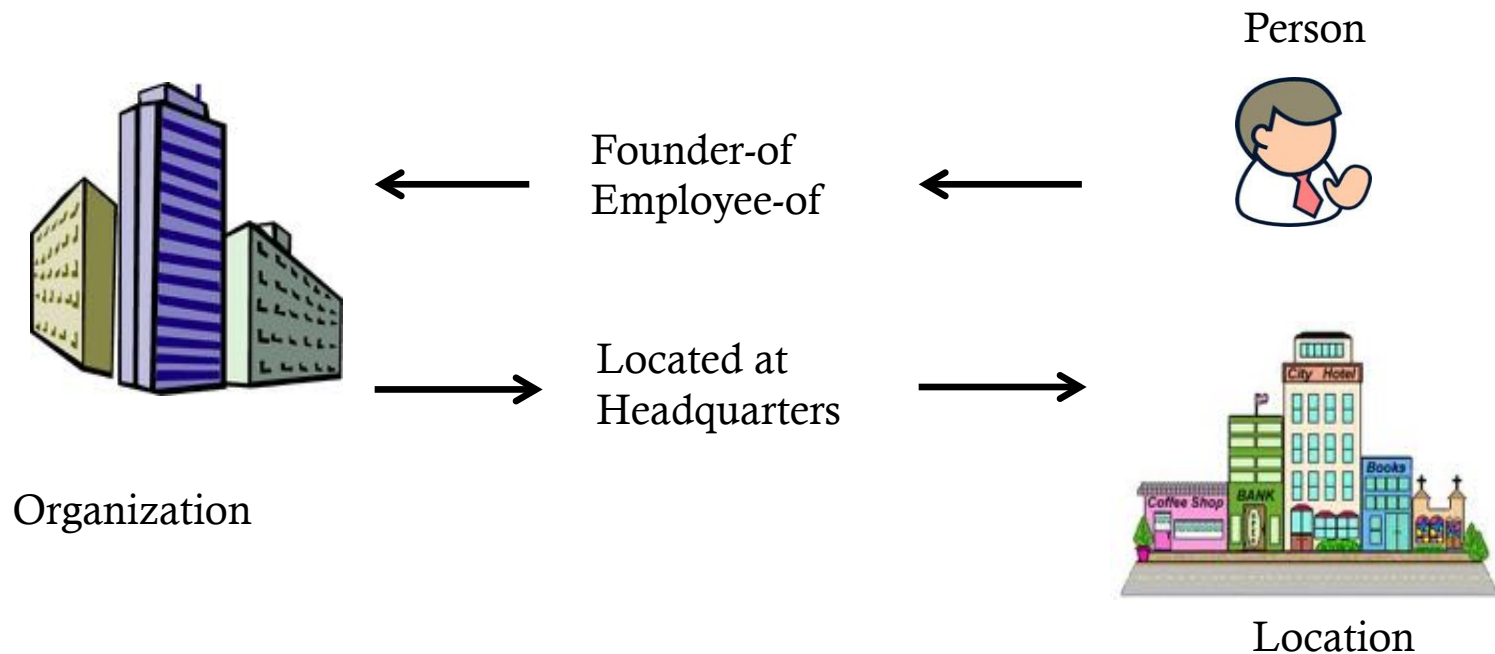
# NLP for Ontologies

- Ontology learning
  - Concepts
  - Hierarchy
  - **Properties/Relations (Sandra Collovini)**
  - Instances



# Relation Extraction

Explicit relations between entities:  
restricted by relation type; **by entity type**; open



# Relation Extraction

## ORG-PES

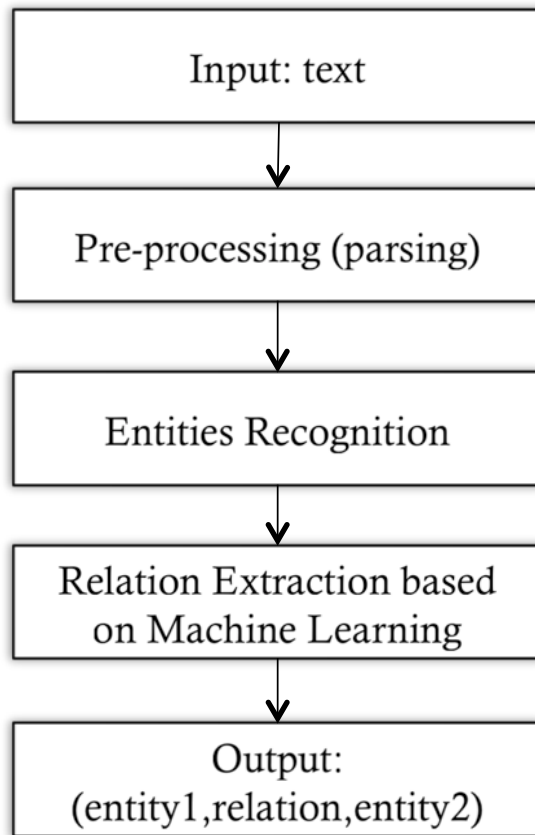
Relation Instances	Relation Descriptor
<p>Fernando Gomes, <b>presidente da</b> Câmara Municipal do Porto</p> <p><i>Fernando Gomes, <b>president of the</b> Câmara Municipal do Porto</i></p>	<p><b>presidente da</b></p> <p><b>(president of the)</b></p>
<p>A Legião da Boa Vontade, instituição educacional, cultural e beneficente, <b>foi fundada pelo</b> jornalista Alziro Zarur</p> <p><i>Legião da Boa Vontade, an educational, cultural and beneficent institution, <b>was founded by</b> journalist Alziro Zarur</i></p>	<p><b>foi fundada pelo</b></p> <p><b>(was founded by)</b></p>

# Relation Extraction

## ORG-LOCAL

Relation Instances	Relation Descriptor
Hospital de São João, <b>no</b> Porto	<b>no</b>
<i>Hospital de São João, <b>at</b> Porto</i>	<b>(at)</b>
Departamento Municipal de Limpeza Urbana <b>de</b> Porto Alegre	<b>de</b>
<i>Departamento Municipal de Limpeza Urbana <b>of</b> Porto Alegre</i>	<b>(of)</b>

# Relation Extraction



Ronaldo Lemos, diretor do Creative Commons

Ronaldo Lemos <hum> PROP @SUBJ>  
diretor <Hprof> N @N<PRED  
de PRP @N<  
o ART @>N  
Creative Commons <org> PROP @P<

Ronaldo\_Lemos <PROP, PER>  
Creative\_Commons<PROP, ORG>

Annotated corpus with Features

(Ronaldo\_Lemos, diretor-de, Creative\_Common)

# References

**Abreu, S. C.** Extração de relações do domínio de organizações para o português. Tese de doutorado, Porto Alegre: PUCRS, 2014.

**Abreu, S. C. ; Vieira, R .** RelP: Portuguese Open Relation Extraction. Knowledge Organization, v. 44, p. 163-177, 2017.



# NLP for Ontologies

- Ontology learning
  - Concepts
  - Hierarchy
  - Properties
  - **Instances (Named entities/Daniela Amaral, Evandro Fonseca)**

# Named Entity Recognition

A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina). Guerra participou do debate "Biotecnologia para uma Agricultura Sustentável", realizado ontem durante a 52ª Reunião Anual da SBPC (Sociedade Brasileira para o Progresso da Ciência), sobre as biotecnologias apropriadas ao desenvolvimento do país. Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare. Para o agrônomo...

The input/output vector

A opinião é do agrônomo Miguel Guerra da UFSC..."

O, O, O, O, O, PESS PESS, O, LOCAL...

# Named Entity Recognition

---

## Features

the word itself

POS tag

the word begins contains lowercase or uppercase

the previous/next word contains lowercase or uppercase

# Geology domain

Este trabalho utiliza o arcabouço bioestratigráfico anteriormente estabelecido para o **Quaternário** da margem continental do Sudeste do Brasil, para mostrar como as relações quantitativas entre as associações de **Foraminíferos planctônicos**, que caracterizam os intervalos zonais e subzonais dos últimos 1,8 milhão de anos, podem auxiliar no posicionamento cronoestratigráfico de amostras de testemunhos e pistão oriundos das **bacias de Santos, Campos, Espírito Santo e Iguatemi**.

É o resultado de centenas de cálculos sobre as variações percentuais do grupo *Globorotalia senardi*, dos gêneros *Pulleniatina* e *Orbulina* e das espécies *Globorotalia inflata* e *Globorotalia truncatulinoides*.

Relacionaram-se os percentuais médios de cada grupo, gênero ou espécie em cada intervalo zonal e subzonal e seus marcos locais e globais diante de uma escala de tempo, de acordo com o zoneamento definido inicialmente para o **Quaternário Superior da Bacia de Campos**.

O presente trabalho, porém, abrange todo o **Quaternário** e é válido para as bacias marginais do Sudeste brasileiro.

**Palavras-chave:** **Quaternário** | bioestratigrafia | **Foraminíferos planctônicos**. Inspirado em Ericson e Hollin, Vitalvi elaborou um arcabouço bioestratigráfico para o **Quaternário Superior da Bacia de Campos**, com base na sucessão vertical de associações de **Foraminíferos planctônicos**.

As associações que permitiram reconhecer cada um dos intervalos zonais e subzonais foram descritas, mas com a advertência de que estes intervalos definidos não são unidades bioestratigráficas stricto sensu, pois se baseiam em eventos climáticos recorrentes.

Por serem recorrentes, as espécies que compõem a maioria das associações que caracterizam estes intervalos são, de modo geral, sempre as mesmas.

Este trabalho procura mostrar que, embora esta afirmativa seja verdadeira, a proporção entre as espécies é diferente para cada intervalo.

Quando se trabalha com eventos climáticos de natureza recorrente, o ideal seria analisar seções completas através de uma sequência de amostras.

Como nem sempre isto é possível, obrigando em algumas ocasiões a utilização de amostras isoladas, a tarefa de identificação dos intervalos bioestratigráficos propostos torna-se muito difícil.

Para facilitar o posicionamento cronoestratigráfico de uma determinada amostra isolada procurou-se mostrar as características faunais de cada associação para cada intervalo zonal, utilizando-se de média da frequência das espécies para cada um deles.

Para uma melhor compreensão e visualização da distribuição dessas associações, utilizou-se a figura 1, onde as espécies características dos intervalos zonais e subzonais do **Quaternário Superior** eram apresentadas com suas variações quantitativas estimadas e marcos locais diante de uma escala de tempo correspondente aos últimos 188 mil anos.

Esta figura 1, que teve a sua eficiência comprovada durante vários anos de aplicação

**Filtro:**

- ☒ EON
- ☒ ERA
- ☒ PERÍODO
- ☒ ÉPOCA
- ☒ IDADE
- ☒ ROCHA SEDIMENTAR SILICILÁSTICA
- ☒ ROCHA SEDIMENTAR CARBONÁTICA
- ☒ ROCHA SEDIMENTAR QUÍMICA
- ☒ ROCHA SEDIMENTAR ORGÂNICA
- ☒ BACIA SEDIMENTAR
- ☒ CONTEXTO GEOLÓGICO DE BACIA
- ☒ UNIDADE LITOE STRATIGRÁFICA
- ☒ OUTRO

Tempo Geológico

- Eon
- Era
- Período
- Época
- Idade

Rochas Sedimentares

- Rocha Silicilástica
- Rocha Carbonática
- Rocha Química
- Rocha Orgânica

Bacia Sedimentar

- Contexto Geológico de Bacia
- Unidade Litoestratigráfica

# Portuguese geo entities corpus

Classes	Entidades Geológicas
Eon	288
Era	326
Período	637
Época	650
Idade	796
Rocha Sedimentar Siliciclástica	743
Rocha Sedimentar Carbonática	240
Rocha Sedimentar Química	5
Rocha Sedimentar Orgânica	22
Bacia Sedimentar	243
Contexto Geológico de Bacia	262
Unidade Litoestratigráfica	581
Outro	739



# References

**Amaral, D. O. F.** Reconhecimento de entidades nomeadas na área de geologia: bacias sedimentares brasileiras. Tese de Doutorado, Porto Alegre: PUCRS, 2017.

**Amaral, D. O. F.; Fonseca, E. B.; Lopes, L.; Vieira, R.,** Comparative Analysis of Portuguese Named Entities Recognition Tools. In: Proceedings of IX International Conference on Language Resources and Evaluation - LREC, 1: 2554-2558, Iceland, 2014.

# Co-reference resolution

A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina). Guerra participou do debate "Biotecnologia para uma Agricultura Sustentável", realizado ontem durante a 52ª Reunião Anual da SBPC (Sociedade Brasileira para o Progresso da Ciência), sobre as biotecnologias apropriadas ao desenvolvimento do país. Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare. Para o agrônomo...

# Correferece

- Corpus annotation
  - Linguistic information
  - Summ-it
    - Pos, syntax, coreference chains, text summaries

# Co-reference resolution

Same entity:

[NP: Guerra ]

[NP: o agrônomo ]

[NP: Miguel\_Guerra ]

[NP: o agrônomo ]

# Correferencia and Semantics

- Semantic bases for matching entities
  - o dinossauro
  - o animal
- pessoas plugadas
- vizinhos conectados



# Correferencia tool

Bem comum de a hu... A ministra de a J... a sequência de um...  
**genes** a União Européia o genoma o sequenciamento a  
**França** patenteamento de ... diretiva favorável a  
 determinação eu... o CCNE Todas as Cadeias

nício de o sequenciamento de o genoma , em a semana passada , [a França [5]] resiste como [único país de  
 não\_Européia a não permitir patenteamento de genes . A UE adota , desde junho de 1998 , diretiva favorável a o  
 nento de genes . O texto , redigido por o Parlamento\_Europeu , Comissão\_Européia\_e\_Conselho\_de\_Ministros ,  
 princípio de que o genoma não é patenteável , mas a sequência de um gene pode ser . em o entanto , há restrições .  
 amento só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de o gene é  
 ) . [A França [5]] é [o único país [5]] que se recusa a aceitar a determinação européia . A ministra de a  
 e [o país [5]] , Elisabeth\_Guigou , disse que a norma é incompatível com as leis francesas de bioética . em o  
 o mês , o CCNE ( Comitê\_Consultivo\_Nacional\_de\_Ética ) , órgão que orienta o governo francês sobre aspectos  
 a biotecnologia , reforçou a posição de a ministra , alegando que o conhecimento de a sequência de um gene não  
 assimilado como produto patenteado e , portanto , não é patenteável . Bem comum de a humanidade , ( o  
 umento de genes ) não pode ser limitado por patentes que pretendem , em nome de o direito de propriedade  
 | , proteger a exclusividade de esse conhecimento , diz parecer de o CCNE . O assunto deve ser debatido durante a  
 ia francesa de a UE , em o segundo semestre .

CADEIA : [5]
a França
único país de a União Européia
A França
o único país
o país
CADEIA : [6]
a União Européia
A UE
a UE
CADEIA : [7]
patenteamento de genes
o patenteamento
O patenteamento
CADEIA : [22]
A ministra de a Justiça de o país
Elisabeth Guigou
a ministra
CADEIA : [28]
o CCNE
Comitê Consultivo Nacional de l
a CCNE

# References

**Fonseca E. B.**, Resolução de correferência nominal usando semântica em Língua Portuguesa. Tese de Doutorado. Porto Alegre: PUCRS, 2018.

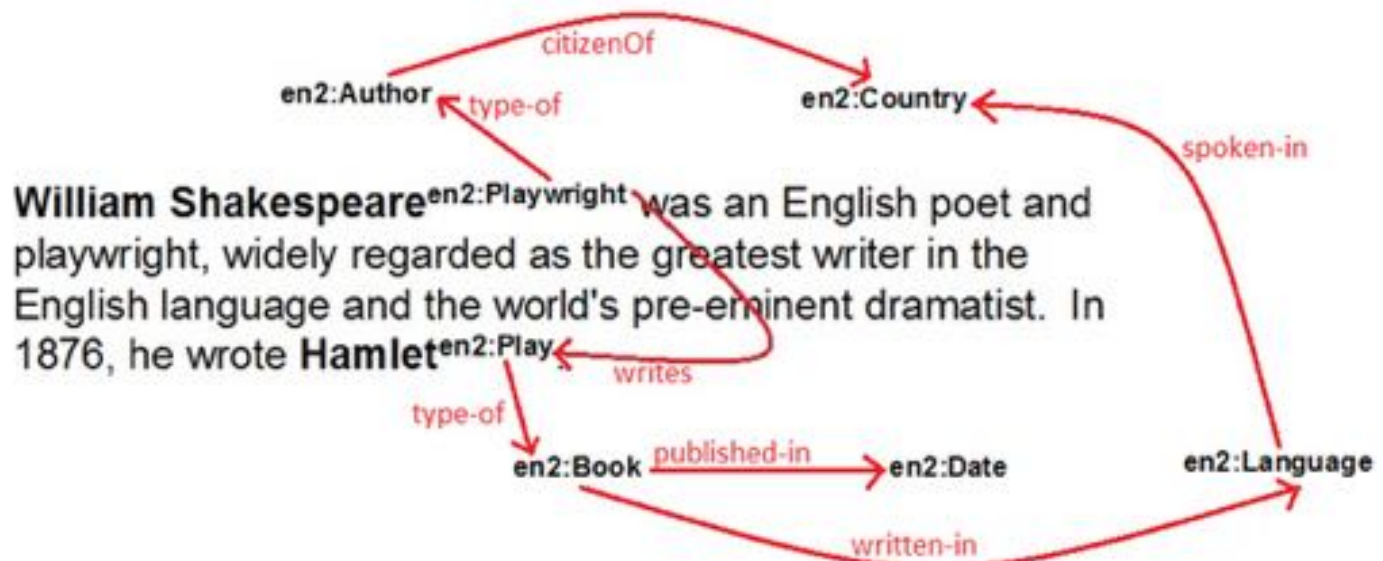
**Fonseca, E. B ; Sesti, V. ; Antonitisch, A. ; Vanim, A. ; Vieira, R.** CORP: Uma Abordagem Baseada em Regras e Conhecimento Semântico para a Resolução de Correferências. Linguamática, v. 9, p. 3-18, 2017.

# Ontologies for NLP

Improving NLP with richer semantics

# Ontologies for NLP

- Semantics
- A play is a type of book, has an author, has a language



# Lexicon x Ontologies

- NLP: is based on lexicons
- Lexicon: conventional inventory of words
- Ontology formalizes concepts and their logical relations
- Computational linguistics used to accurately map the relations between words and the concepts that they can be linked to
- Integration between lexical and semantic resources

# Lexicon x Ontologies

- WordNet
  - Semantic lexical database widely used in NLP
- Projects for linking upper level ontologies and WordNet
  - SUMO
  - DOLCE
- Current project in linking domain ontologies to top ontologies via WordNet



# Ontologies for NLP

Related research

# Ontology based Sentiment Analysis in Aspect Level (Larissa Freitas)

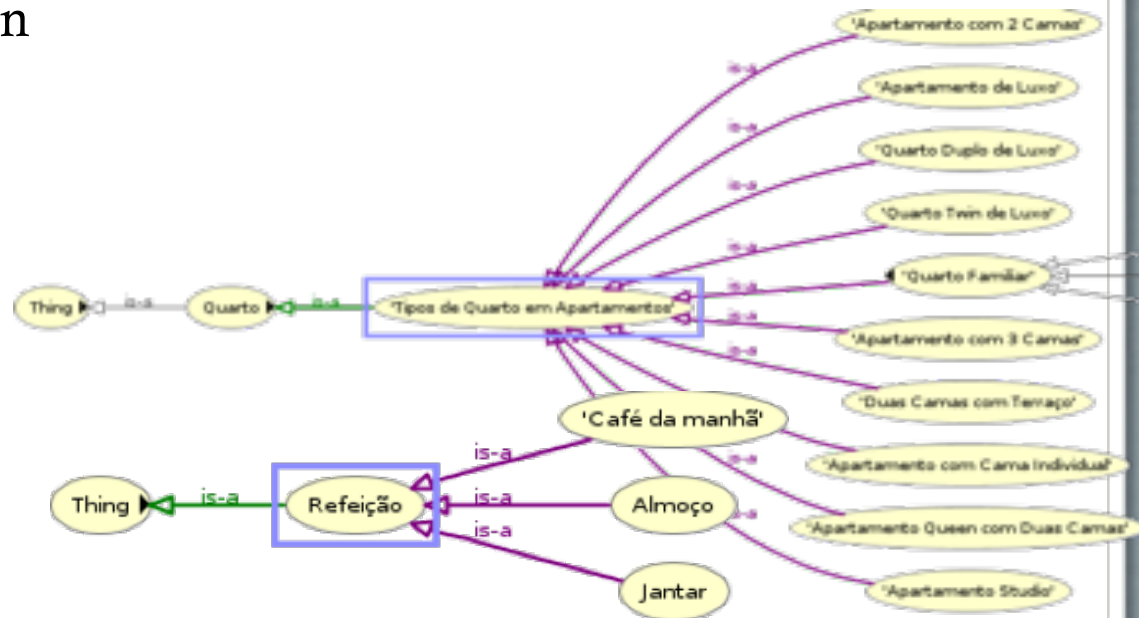
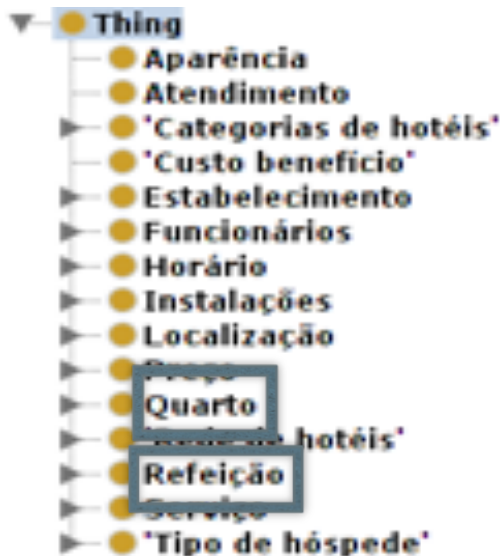
- Sentiment Analysis to infer people's

- opinions,
- sentiments
- evaluations
- emotions

towards entities or their aspects (parts and attributes)

# HOntology

- HOntology<sup>1</sup> is a **multilingual** (English, **Portuguese**, Spanish and French) ontology for the hotel domain



<sup>1</sup><http://ontolp.inf.pucrs.br/Recursos/downloads-Hontology.php>

# Ontology based Sentiment Analysis in Aspect Level

- Explicit and Implicit Aspect

## Explicit - Rooms

“Os **quartos** e banheiros são bons”

## Implicit - Value

“Apesar da **taxa de estacionamento** ser salgada”



# References

---

**Freitas, L. A. ; Vieira, R.** Ontology based Feature-Level Sentiment Analysis in Portuguese Reviews. International Journal of Business Information Systems, 2019.

**Freitas, L. A.** Feature-level sentiment analysis applied to brazilian portuguese reviews. PhD Thesis, Porto Alegre: PUCRS, 2015.

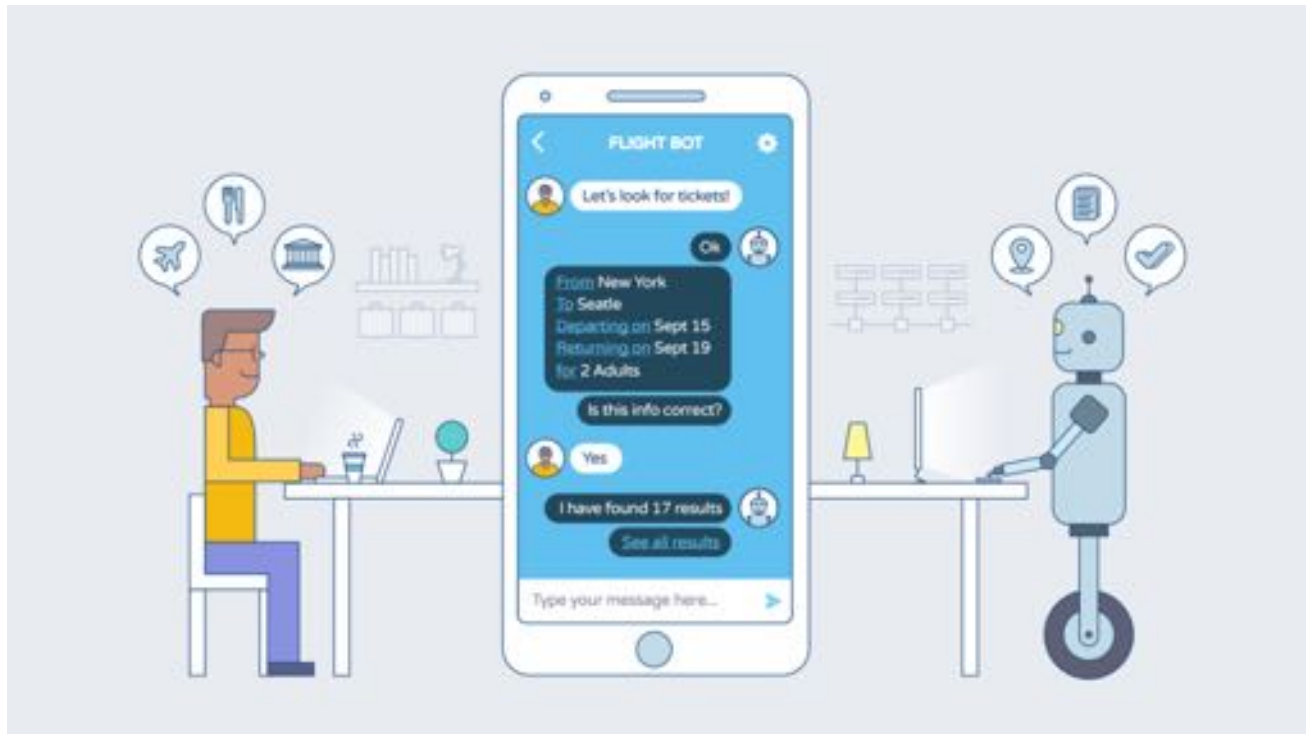
# Conclusion





# Language industry

- Chat bots



# Dandelion API

## Semantic Text Analytics

- The entire analytical process takes just a few minutes and:
  - It identifies mentions of banks in official company documents;
  - Effectively disambiguates these mentions to link them to the correct bank (banks may have similar-sounding names and the same bank may have more than one name);

# Dandelion API

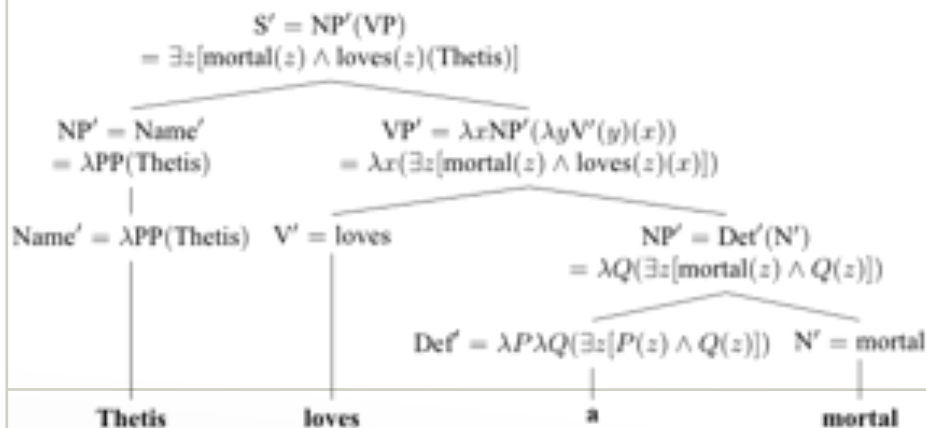
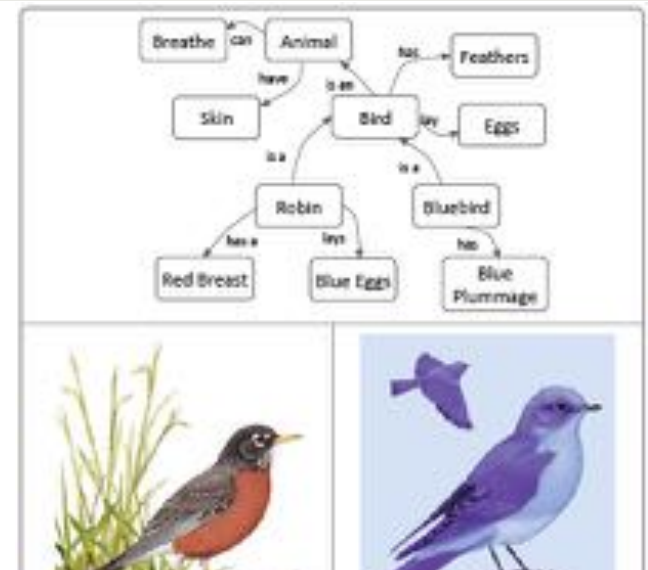
## Semantic Text Analytics

- The entire analytical process takes just a few minutes and:
  - Identifies which kind of relationship a company has with a bank (is it some kind of shareholding? A current account? A loan? Which of the many different kind of loans?);

# Machine communication



# NLP and Ontologies are the basis



# Technological evolution is also human evolution





# Challenges

---

- Considering Portuguese is relevant/strategic
- Multidisciplinary approach is required
  - Linguistic knowledge
  - Mathematical models
  - Humanities

# Such is the complexity of human communication

metaphor

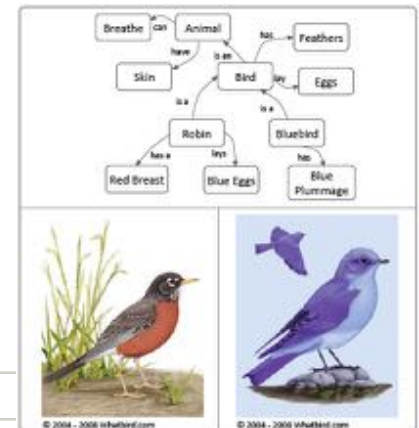
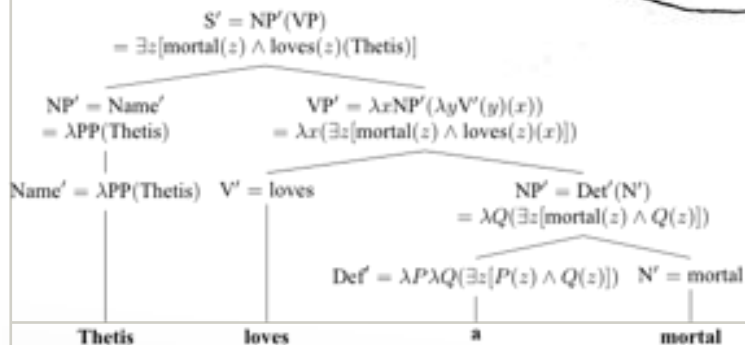
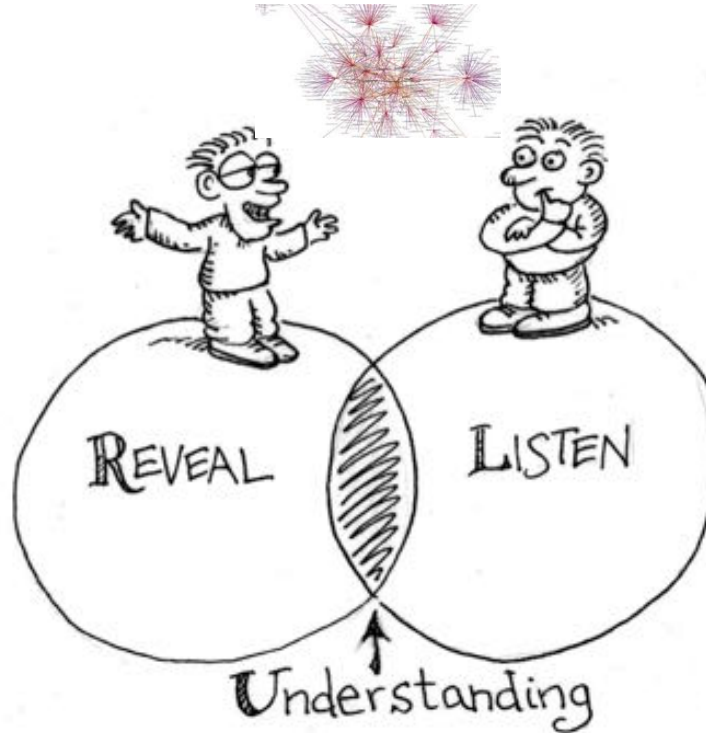
irony

literacy

cultural  
background

intonation

previous  
knowledge





Pensamos e falamos

Colocamos os pensamentos no mundo

Escrevemos, armazenamos e compartilhamos

Criamos mais sobre o que pensar

E há muito para ler

Pensamos sobre a maneira como pensamos e falamos

E assim construimos maquinas para nos ajudar a comunicar

E mudamos nossa forma de pensar, de falar e de comunicar