Semantic Role Labeling and Lexical Simplification: two samples of NLP applications

Leonardo Zilio

CENTAL - UCL

Semantic Role Labeling for Portuguese

Leonardo Zilio (Instituto de Letras – UFRGS)

Maria José Bocorny Finatto (Instituto de Letras – UFRGS)

Aline Villavicencio (Instituto de Informática – UFRGS)

Objectives

- To understand how the semantic structure of Portuguese works in specialized and non-specialized contexts
- To further describe the Portuguese language in terms of generic and descriptive semantic roles
  - In 2011, there was only one project on semantic role labeling (FrameNet Brasil)

Semantic Roles

Mary broke the vase.

AGENT + PATIENT

The vase broke.

PATIENT

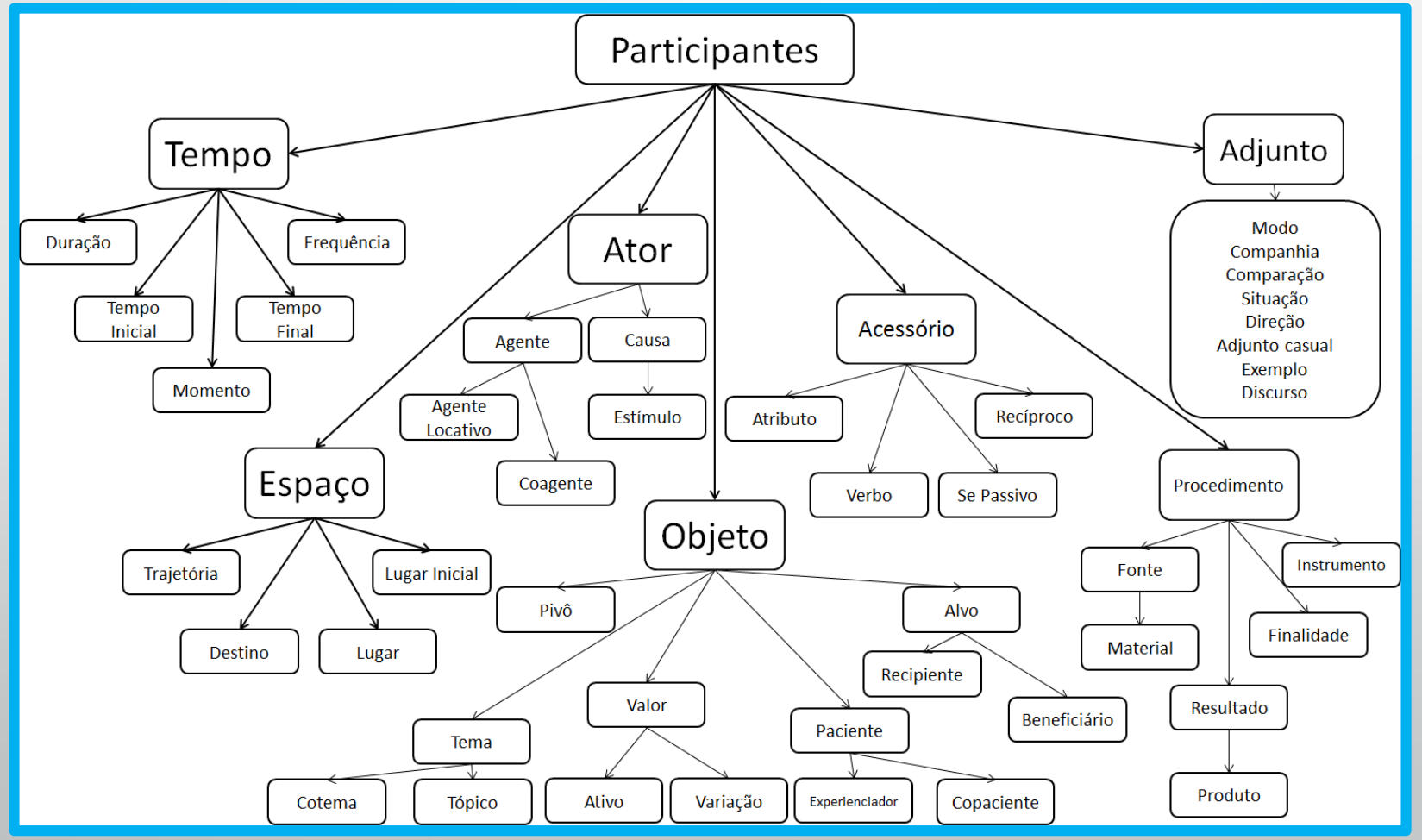
Related Work

- FrameNet
  - Descriptive semantic roles associated to a specific communicative scenario (e.g., PLAYER, REFEREE [in a soccer match context])
- VerbNet
  - Generic descriptive semantic roles (e.g. AGENTE, THEME, PATIENT)
- PropBank
  - Numbered semantic roles (e.g., A1, A2) + roles for adjuncts (e.g., TIME, PLACE)

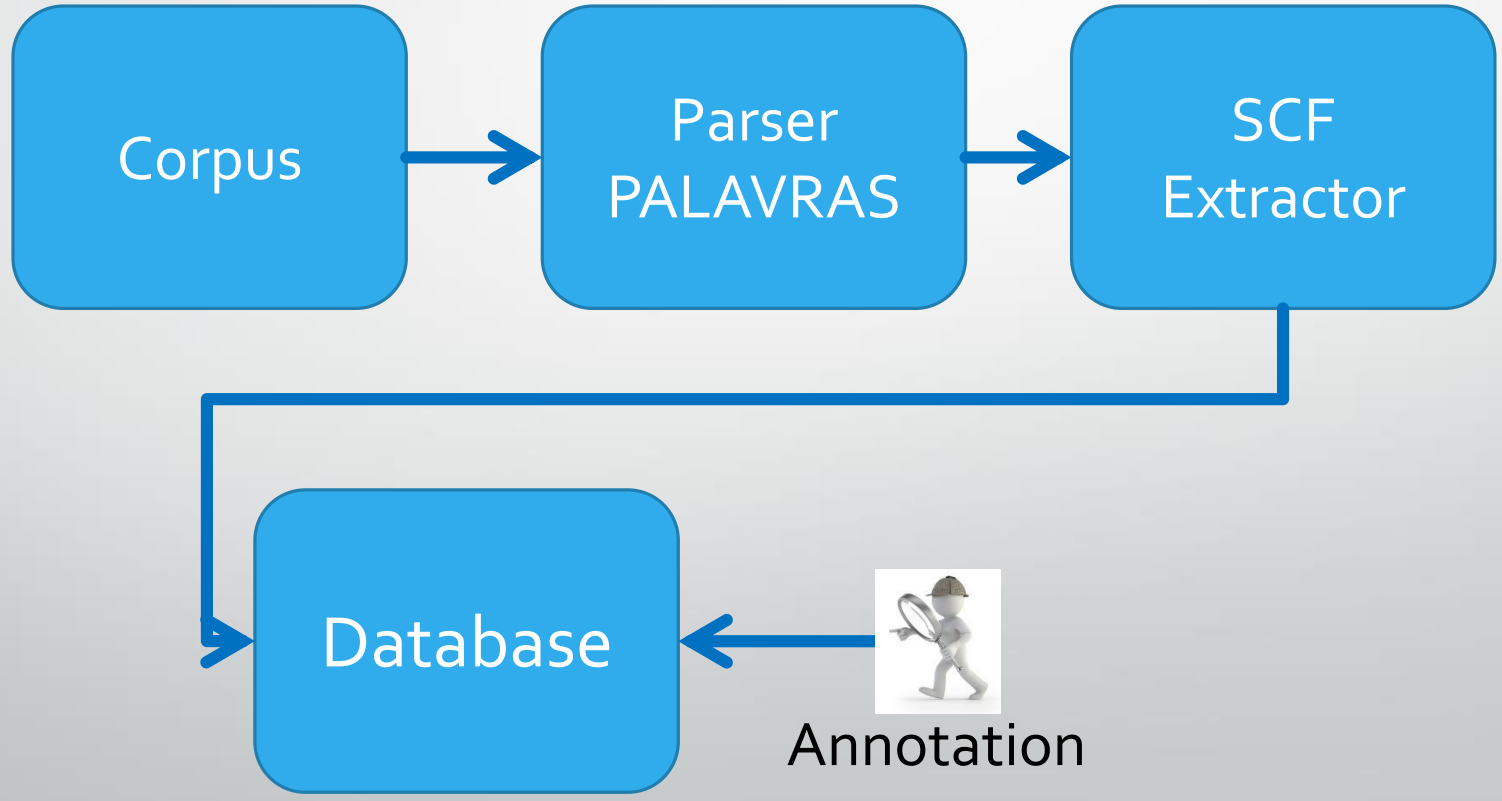
Our Choice

- VerbNet Semantic Roles + PropBank Adjunct Roles

46 Semantic Roles



Metodology


Corpus
Parser
PALAVRAS
SCF
Extractor
Database
Annotation

Corpora

	Diário Gaúcho Newspaper	Cardiology Papers
Date	2008	2005-2007
# of Tokens	1M	1,4M

Parsing

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2

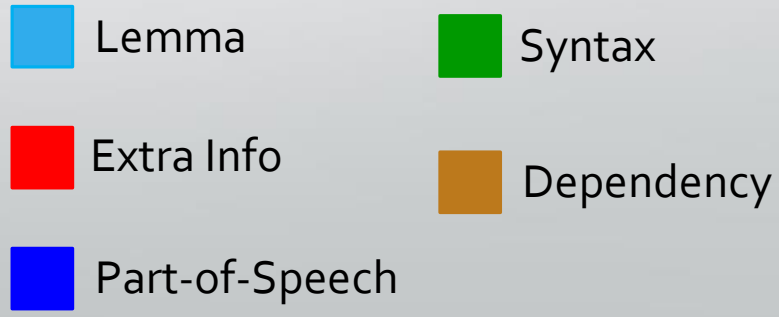
viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0

o [o] <artd> DET M S @>N #3->4

cachorro [cachorro] <Azo> N M S @<ACC #4->2

$. #5->0

</s>



Dependency Tree

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2

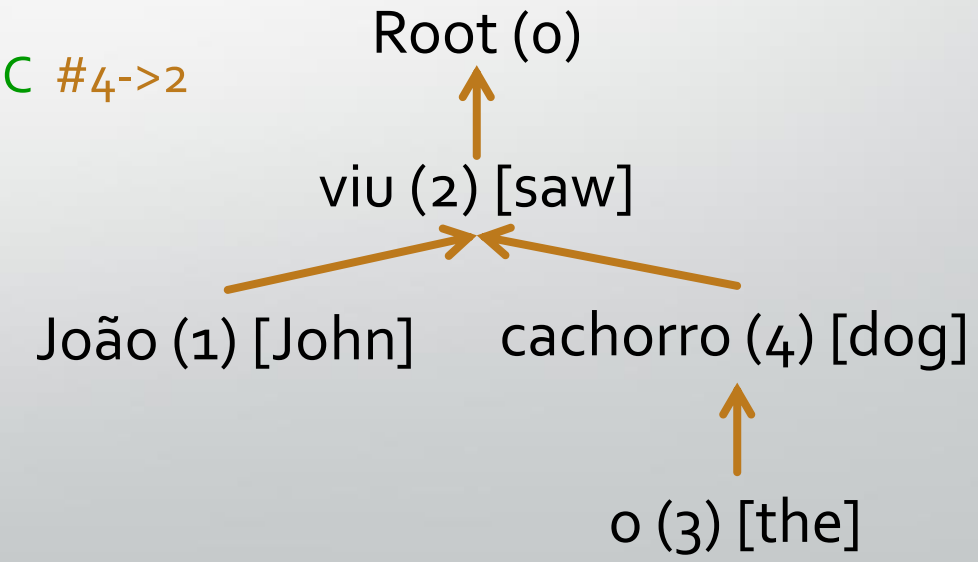
viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0

o [o] <artd> DET M S @>N #3->4

cachorro [cachorro] <Azo> N M S @<ACC #4->2

$. #5->0

</s>

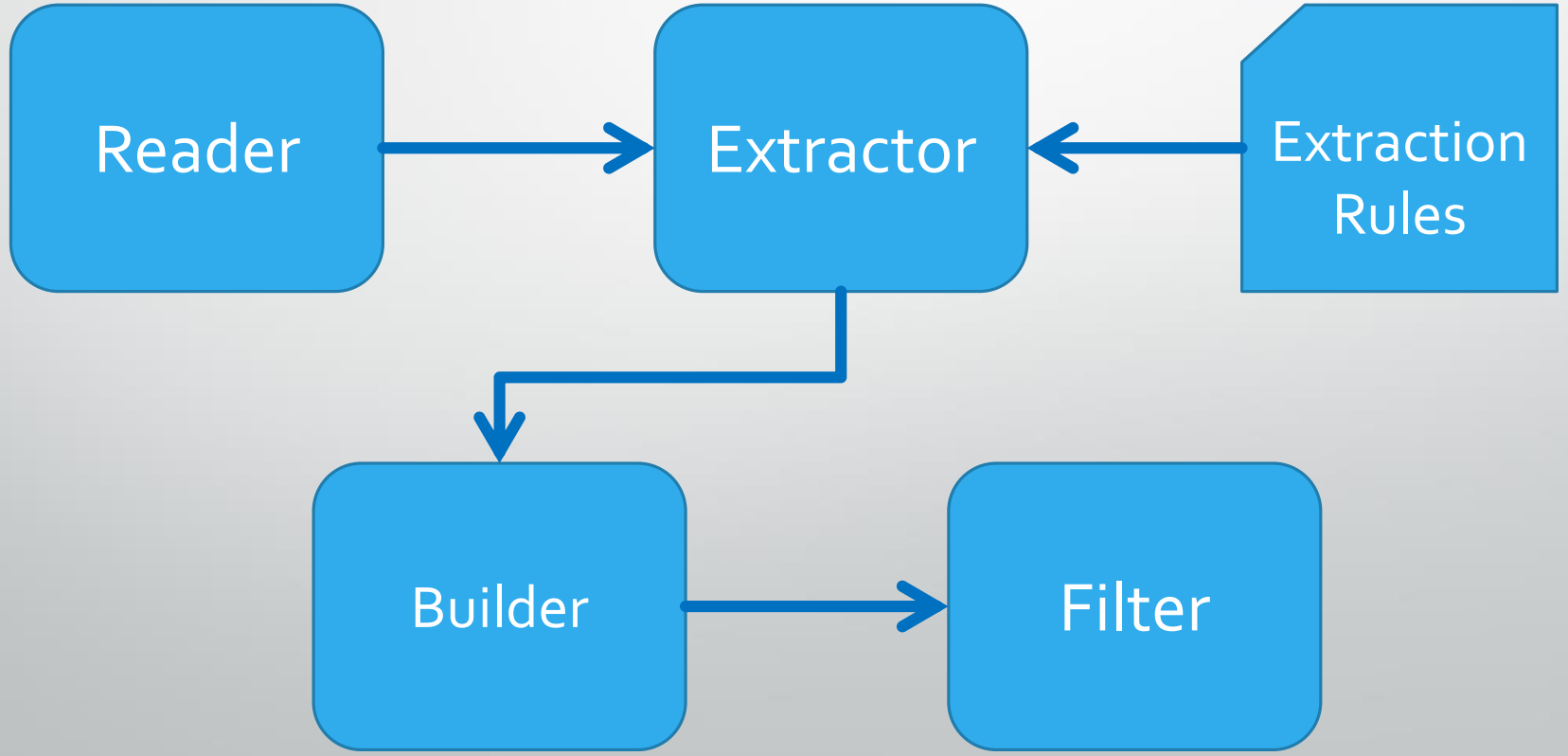


Extraction of SCFs

- Subcategorization Frames are simpler syntactic representations of sentences
- SCFs represent sentences in terms of their phrases:
  - NP_V_NP = Martin has a car.
  - NP_V_NP_PP = Martin bought a car from Paul.
  - NP_V_PP = Martin goes to the library.
- For us, SCFs help organizing sentences in the database

Subcategorization Frames (SCF) Extractor

Adriano Zanette (Instituto de Informática – UFRGS)
Leonardo Zilio (Instituto de Letras – UFRGS)



Reader Module

- Receives the parsed text
- Separates every sentence

Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

What is an argument?

- Complicated question
- For the purposes of the subcategorization frames extractor, there is no distinction between argument and adjunct
- It extracts phrases that are directly dependents of the verb, according to a set of rules

Builder Module

- Puts everything together (according to the relevance index)
- Builds the subcategorization frame for each verb and sentence
- Stores information on the database

Filter

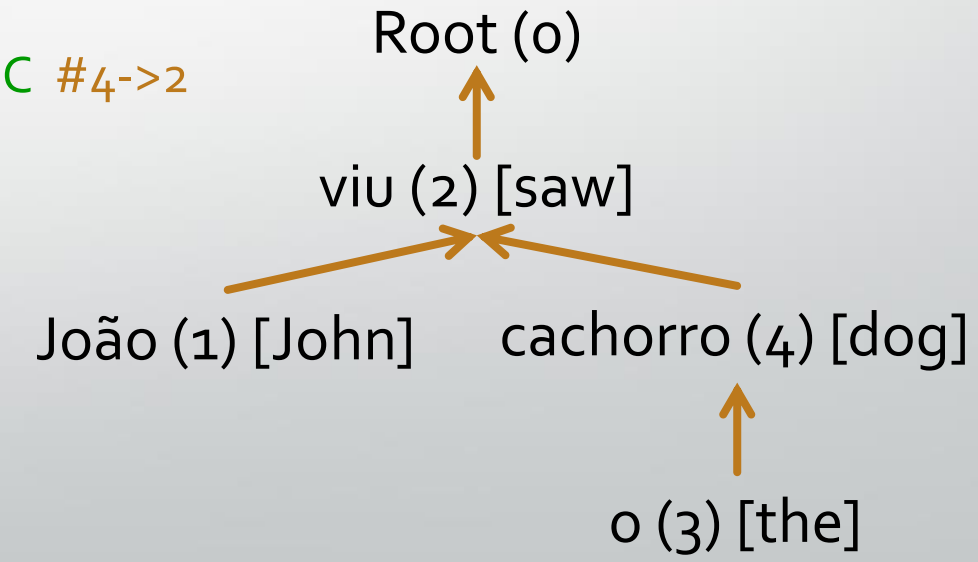
- Not mandatory
- It can filter subcategorization frames based on frequency (or frequency-like parameters)

Let’s see how it works

Simple Example (Again)

João viu o cachorro. (John saw the dog.)

```
João        [João] <hum> PROP M S @SUBJ>  #1->2
viu         [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA  #2->0
o           [o] <artd> DET M S @>N  #3->4
cachorro    [cachorro] <Azo> N M S @<ACC  #4->2
$.          #5->0
</s>
```



Reader Module

- Simply recognizes that there is a full sentence; and
- Passes it to the Extractor Module

Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist ←
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

Extractor Module

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2

viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0

o [o] <artd> DET M S @>N #3->4

cachorro [cachorro] <Azo> N M S @<ACC #4->2

$. #5->0

</s>

Extractor Module

João viu o cachorro. (John saw the dog.)

Tag for Conjugated Verbs

João [João] <hum> PROP M S @SUBJ> #1->2

viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0

o [o] <artd> DET M S @>N #3->4

cachorro [cachorro] <Azo> N M S @<ACC #4->2

$. #5->0

</s>

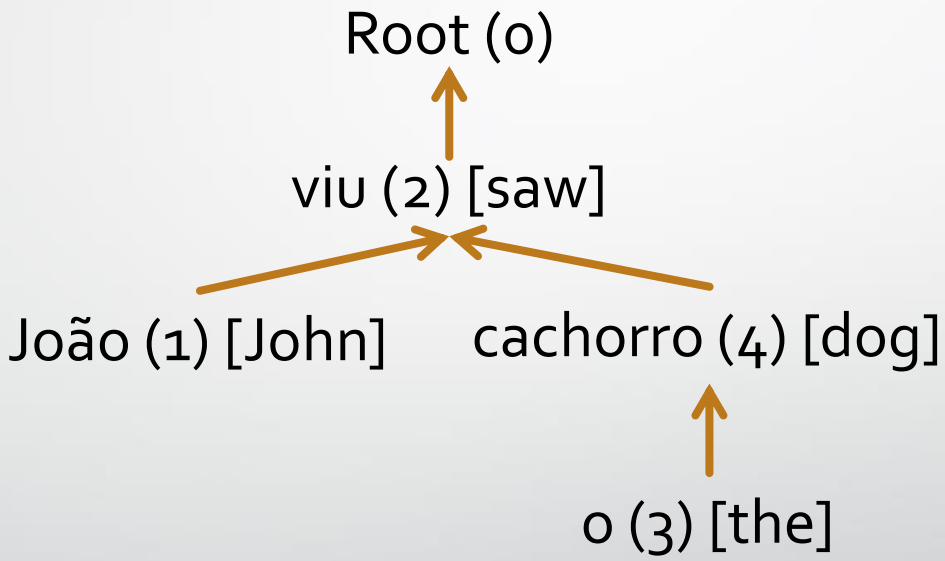
Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb ←
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

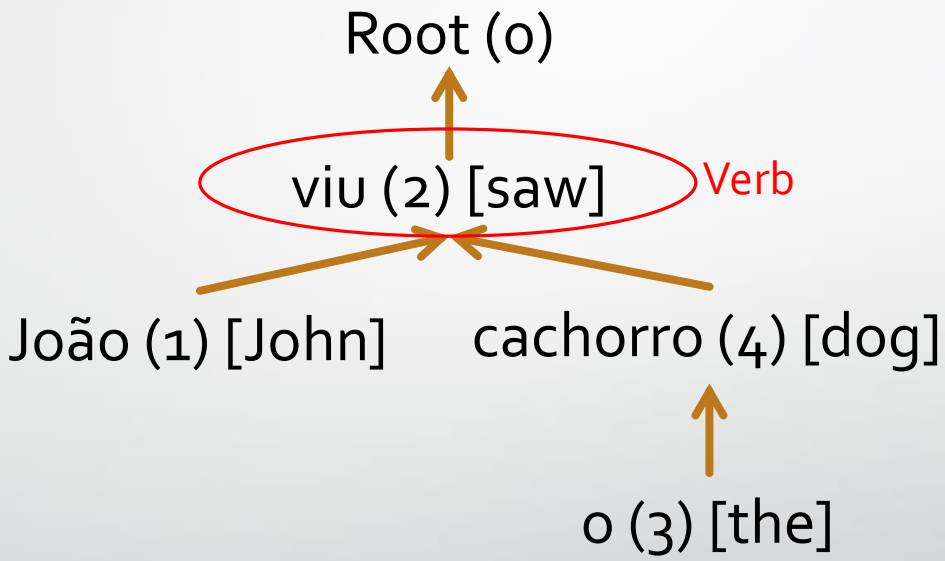
Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb ←
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

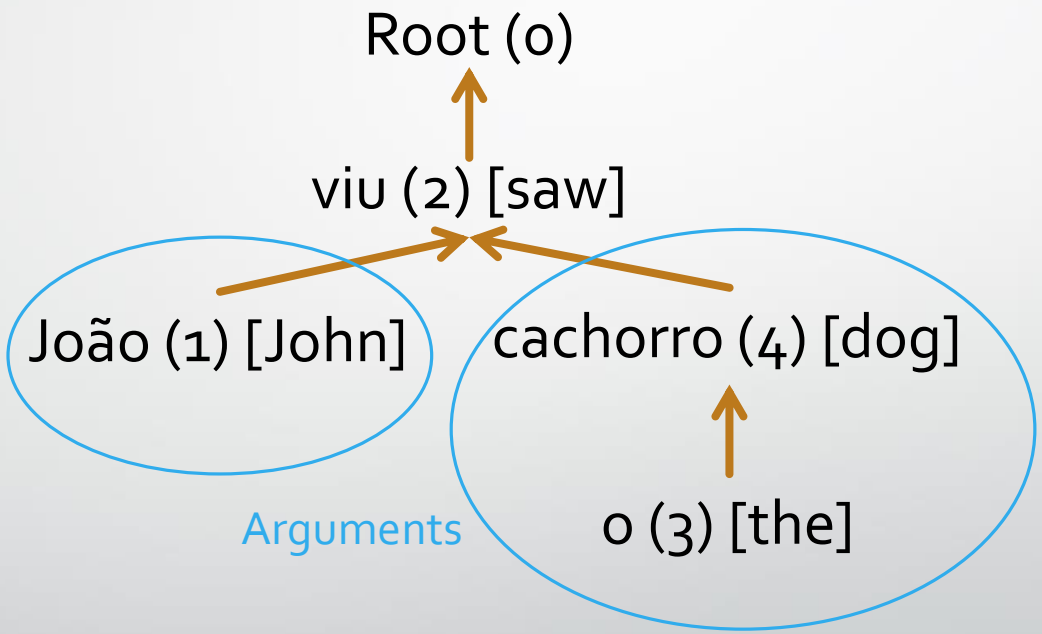
Simple Example (Again)



Simple Example (Again)



Simple Example (Again)



Extractor Module

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic cathegory of each argument and attributes a relevance index (for organization purposes)

Rules

- Format:
  - If [tag], then [argument_type]

- If SUBJ, then Subject (Relevance Index: 1)
- If ACC, then Direct Object (Relevance Index: 3)
- If ACC-PASS, then Reflexive Object (Relevance Index: 3)
- Etc.

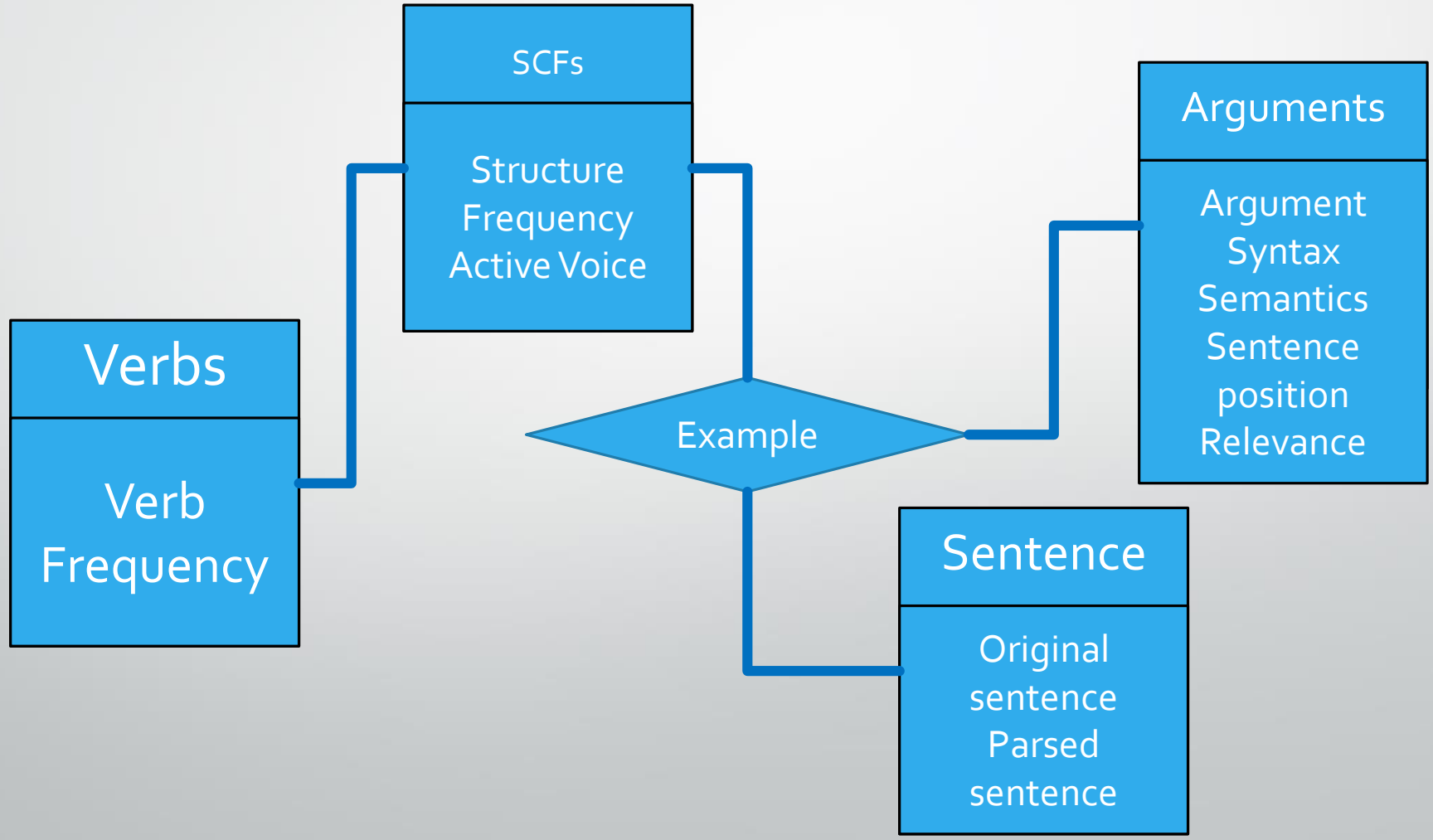
Builder Module

- Verb: ver
- Sentence: João viu o cachorro.
- SCF: SUBJ_V_NP
- SUBJ: João
- NP: o cachorro

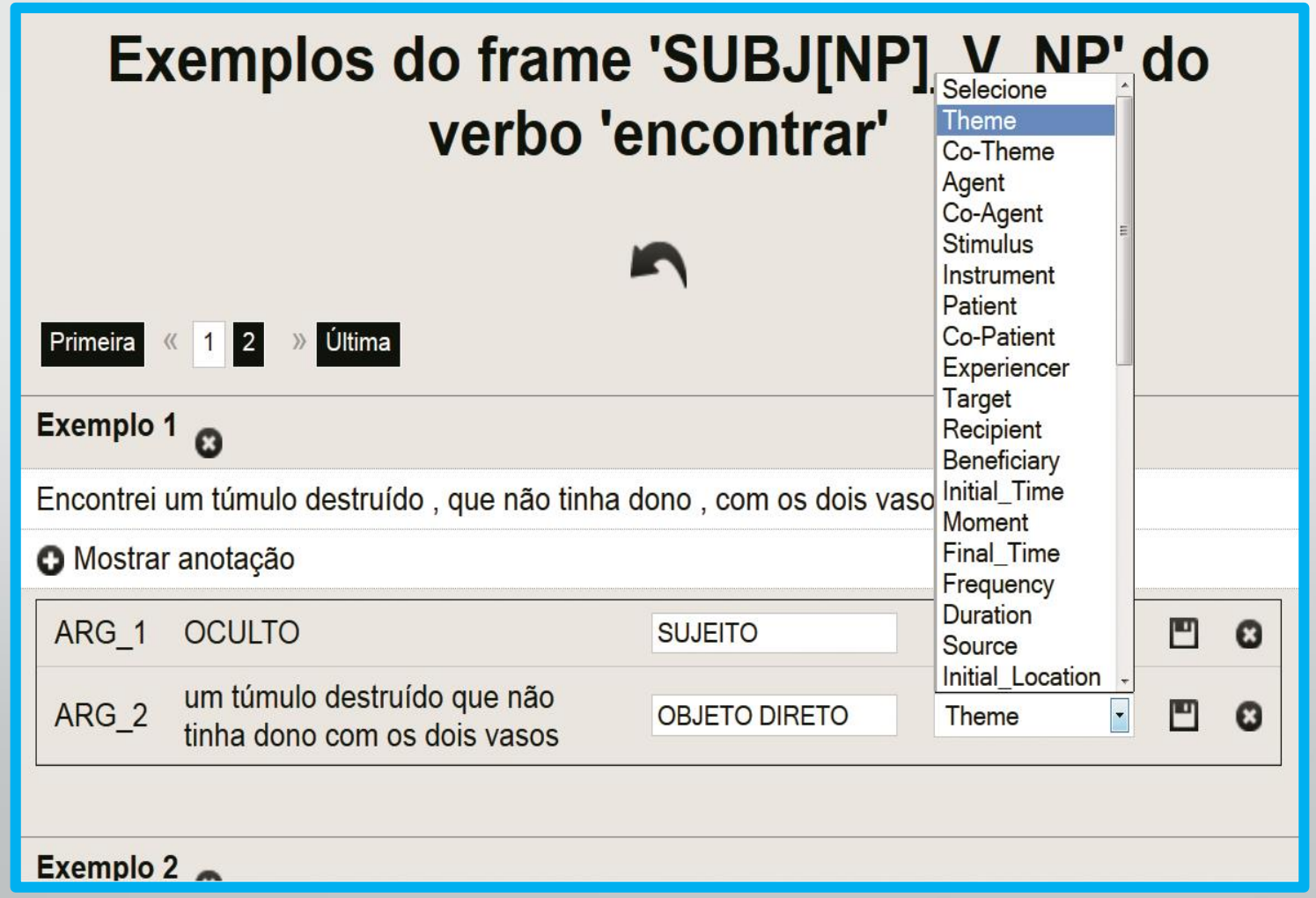
Filter

- After processing all data, if the number of similar SCFs do not amount to a certain threshold, the SCF is excluded from the database

Database (SQL)



Annotation Interface



VerbLexPor

Diário Gaúcho	Cardiologia
191 verbs	77 verbs
5.301 instances	**1.931 instances**
11.089 arguments	4.192 arguments

Availability

- XML and SQL
- Website Project CAMELEON
  - http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic%20role%20labels%20corpus%20-%20Brazilian%20Portuguese

Text Simplification

Semantic Relations

Rodrigo Wilkens

Leonardo Zilio

Eduardo Ferreira

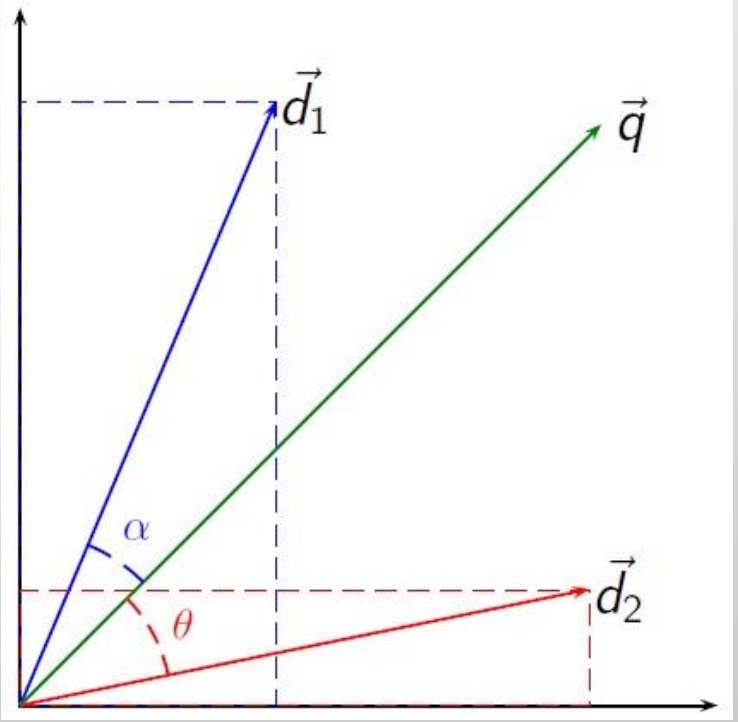
Aline Villavicencio

Objective

- To build a lexical resource with synonyms, antonyms and hypernyms
  - Distributional thesaurus + BabelNet
- Evaluate the resource against a gold standard

Distributional Thesaurus

- Distributional hypothesis:
  - You can know a word by the company it keeps
- Words can be represented as vectors in a multidimensional space

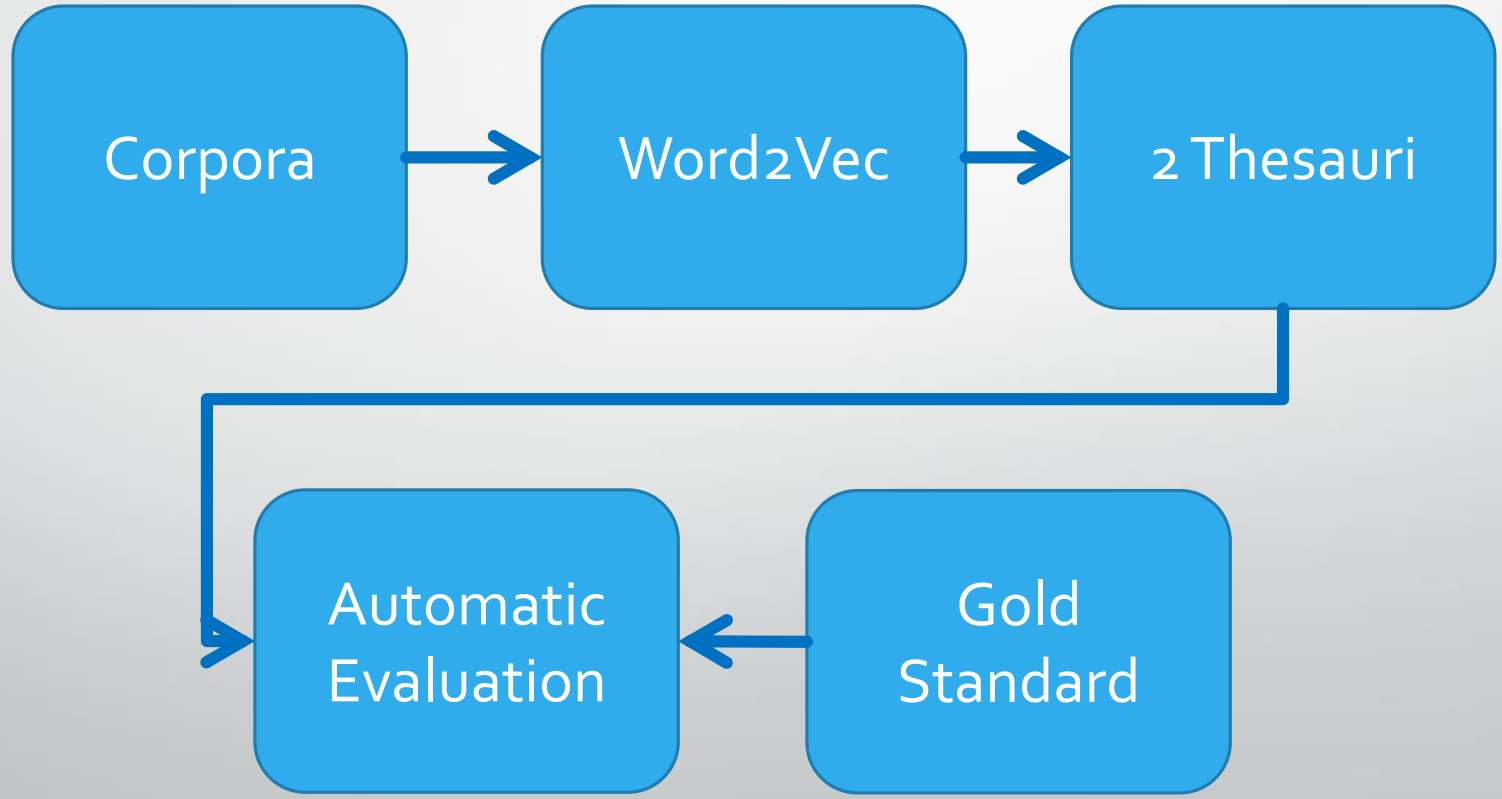


Distributional Thesaurus

- Presents pairs of words, indicating how related they are to each other

Word 1	Word 2	Relatedness
Joy	Happiness	48,5%
Joy	Smile	32,8%
Joy	Scream	15,0%
Joy	Brick	3,9%

Metodology


Corpora
Word2Vec
2 Thesauri
Automatic Evaluation
Gold Standard

Gold Standard

- AC/DC corpus = Word frequency list
- BabelNet = Resource similar to WordNet
  - Word polysemy
  - Semantic relations

Evaluation

- Which of these words is most related to "author"?
  - Poet
  - Parts
  - Patron
  - Board

Methodology

- Groups of words containing: 1 target word, 1 related word, and 3 non-related words

- TOEFL e WordNet-Based Synonymy Test (WBST)

Target Word	Alternative 1: Related Word	Alternative 2	Alternative 3	Alternative 4

Methodology

- Each word from AC/DC corpus was annotated:
  - with frequency (from AC/DC); and
  - with polysemy (from BabelNet)
- Words that were not in BabelNet were excluded

Methodology

- Target word: medium frequency in AC/DC
- Related word: closest to the target word in terms of frequency and polysemy
- Non-related words: farthest average distance from the target word

Initial Resource

	Synonym	Antonym	Hipernym	Total
Verbs	500	200	500	**1200**
Nouns	1667	200	1667	**3534**
Total	**2167**	**400**	**2167**	**4734**

Automatic Validation

- All instances of target and related word were validated agains **Onto.PT**

Manual Validation

- All instances that were **NOT** automatically validated were manually verified

BabelNet-Based Semantic Gold standard (B²SG)

	Antonym		Synonym		Hypernym		
	N	V	N	V	N	V	Total
Initial	200	200	1667	500	1667	500	**4734**
Onto.PT	40	51	676	244	191	0	**1202**
Human Judges	105	116	495	191	568	198	**1673**
Total Validated	145	167	1171	435	759	198	**2875**
% Correct	*72.5*	*83.5*	*70.2*	*87.0*	*45.5*	*39.6*	***60.7***

BabelNet-Based Semantic Gold standard (B²SG)

	Antonym		Synonym		Hypernym		
	N	V	N	V	N	V	Total
Initial	200	200	1667	500	1667	500	**4734**
Onto.PT	40	51	676	244	191	0	**1202**
Human Judges	105	116	495	191	568	198	**1673**
Total Validated	145	167	1171	435	759	198	**2875**
% Correct	*72.5*	*83.5*	*70.2*	*87.0*	*45.5*	*39.6*	***60.7***

Corpora

	TOKENS	TYPES
Surface	1.5G	3.7M
Lemma	409M	1.5M

Corpus Brasileiro was not used in the lemmatized corpus, because it is not annotated with lemmata

Distributional Thesauri

- Word2Vec
- Strict Evaluation (target word and all alternatives must be in the corpus)

Evaluation

- Which of these words is most related to "author"?
  - Poet
  - Parts
  - Patron
  - Board

Target	Alternative	Relatedness
Author	**Poet**	**24,8%**
Author	Parts	0,3%
Author	Patron	0,6%
Author	Board	0,2%

Strict Evaluation

		Surface			Lemma		
		Instances	Correct	% Correct	Instances	Correct	% Correct
Antonym	N	105	90	**85.7**	98	82	83.7
	V	143	100	69.9	141	110	**78.0**
Hipernym	N	545	432	79.3	525	425	**81.0**
	V	167	115	68.9	166	118	**71.1**
Synonym	N	861	726	84.3	832	721	**86.7**
	V	366	275	**75.1**	366	267	73.9

Next Step

- Mixing Distributional Thesaurus with BabelNet for creating a larger dictionary of synonyms

Dictionary of Complex Words

Leonardo Zilio

Susana Bautista

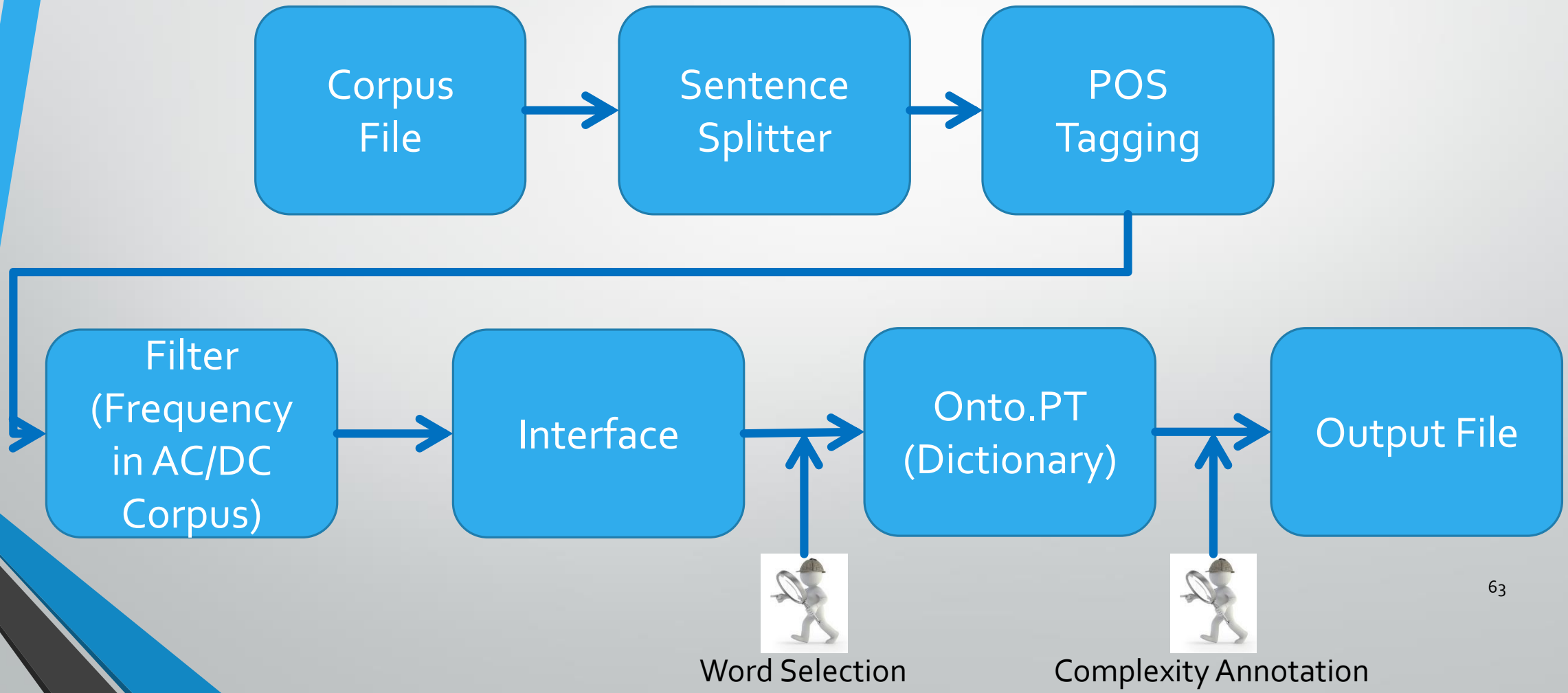
Objective

- Dictionary of complex words from Classic Literature
  - Simpler and complexer alternatives

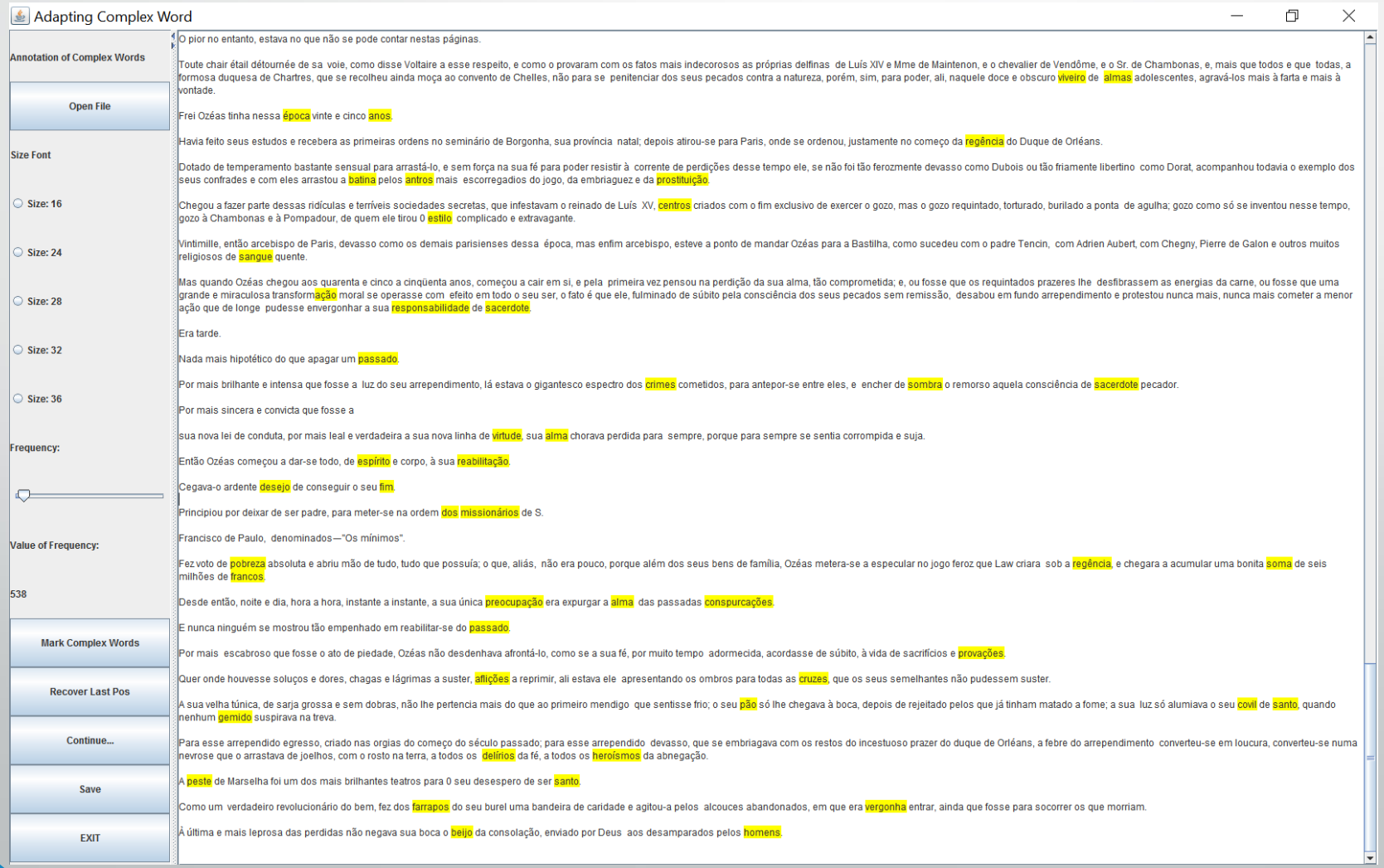
Corpus

Author – Book	Tokens	Types	Type/Token Ratio
Aluísio Azevedo – O Cortiço	81.8K	11.2K	0.137
Joaquim Manuel de Macedo – A Moreninha	47.2K	6.9K	0.147
José de Alencar – Lucíola	46K	7.8K	0.169
Machado de Assis – Memorial de Aires (*Ce que les Hommes Appellent Amour*)	51.1K	6.3K	0.123

Methodology



Interface


Adapting Complex Word
Annotation of Complex Words
Open File
Size Font
Size: 16
Size: 24
Size: 28
Size: 32
Size: 36
Frequency:
Value of Frequency:
538
Mark Complex Words
Recover Last Pos
Continue...
Save
EXIT
O pior no entanto, estava no que não se pode contar nestas páginas.
Toute chair étail détournée de sa voie, como disse Voltaire a esse respeito, e como o provaram com os fatos mais indecorosos as próprias delfinas de Luís XIV e Mme de Maintenon, e o chevalier de Vendôme, e o Sr. de Chambonas, e, mais que todos e que todas, a formosa duquesa de Chartres, que se recolheu ainda moça ao convento de Chelles, não para se penitenciar dos seus pecados contra a natureza, porém, sim, para poder, ali, naquele doce e obscuro viveiro de almas adolescentes, agravá-los mais à farta e mais à vontade.
Frei Ozéas tinha nessa época vinte e cinco anos.
Havia feito seus estudos e recebera as primeiras ordens no seminário de Borgonha, sua província natal; depois atirou-se para Paris, onde se ordenou, justamente no começo da regência do Duque de Orléans.
Dotado de temperamento bastante sensual para arrastá-lo, e sem força na sua fé para poder resistir à corrente de perdições desse tempo ele, se não foi tão ferozmente devasso como Dubois ou tão friamente libertino como Dorat, acompanhou todavia o exemplo dos seus confrades e com eles arrastou a batina pelos antros mais escorregadios do jogo, da embriaguez e da prostituição.
Chegou a fazer parte dessas ridículas e terríveis sociedades secretas, que infestavam o reinado de Luís XV, centros criados com o fim exclusivo de exercer o gozo, mas o gozo requintado, torturado, burilado a ponta de agulha; gozo como só se inventou nesse tempo, gozo à Chambonas e à Pompadour, de quem ele tirou 0 estilo complicado e extravagante.
Vintimille, então arcebispo de Paris, devasso como os demais parisienses dessa época, mas enfim arcebispo, esteve a ponto de mandar Ozéas para a Bastilha, como sucedeu com o padre Tencin, com Adrien Aubert, com Chegny, Pierre de Galon e outros muitos religiosos de sangue quente.
Mas quando Ozéas chegou aos quarenta e cinco a cinqüenta anos, começou a cair em si, e pela primeira vez pensou na perdição da sua alma, tão comprometida; e, ou fosse que os requintados prazeres lhe desfibrassem as energias da carne, ou fosse que uma grande e miraculosa transformação moral se operasse com efeito em todo o seu ser, o fato é que ele, fulminado de súbito pela consciência dos seus pecados sem remissão, desabou em fundo arrependimento e protestou nunca mais, nunca mais cometer a menor ação que de longe pudesse envergonhar a sua responsabilidade de sacerdote.
Era tarde.
Nada mais hipotético do que apagar um passado.
Por mais brilhante e intensa que fosse a luz do seu arrependimento, lá estava o gigantesco espectro dos crimes cometidos, para antepor-se entre eles, e encher de sombra o remorso aquela consciência de sacerdote pecador.
Por mais sincera e convicta que fosse a
sua nova lei de conduta, por mais leal e verdadeira a sua nova linha de virtude, sua alma chorava perdida para sempre, porque para sempre se sentia corrompida e suja.
Então Ozéas começou a dar-se todo, de espírito e corpo, à sua reabilitação.
Cegava-o ardente desejo de conseguir o seu fim.
Principiou por deixar de ser padre, para meter-se na ordem dos missionários de S.
Francisco de Paulo, denominados—"Os mínimos".
Fez voto de pobreza absoluta e abriu mão de tudo, tudo que possuía; o que, aliás, não era pouco, porque além dos seus bens de família, Ozéas metera-se a especular no jogo feroz que Law criara sob a regência, e chegara a acumular uma bonita soma de seis milhões de francos.
Desde então, noite e dia, hora a hora, instante a instante, a sua única preocupação era expurgar a alma das passadas conspurcações.
E nunca ninguém se mostrou tão empenhado em reabilitar-se do passado.
Por mais escabroso que fosse o ato de piedade, Ozéas não desdenhava afrontá-lo, como se a sua fé, por muito tempo adormecida, acordasse de súbito, à vida de sacrifícios e provações.
Quer onde houvesse soluços e dores, chagas e lágrimas a suster, aflições a reprimir, ali estava ele apresentando os ombros para todas as cruzes, que os seus semelhantes não pudessem suster.
A sua velha túnica, de sarja grossa e sem dobras, não lhe pertencia mais do que ao primeiro mendigo que sentisse frio; o seu pão só lhe chegava à boca, depois de rejeitado pelos que já tinham matado a fome; a sua luz só alumiava o seu covil de santo, quando nenhum gemido suspirava na treva.
Para esse arrependido egresso, criado nas orgias do começo do século passado; para esse arrependido devasso, que se embriagava com os restos do incestuoso prazer do duque de Orléans, a febre do arrependimento converteu-se em loucura, converteu-se numa nevrose que o arrastava de joelhos, com o rosto na terra, a todos os delírios da fé, a todos os heroísmos da abnegação.
A peste de Marselha foi um dos mais brilhantes teatros para 0 seu desespero de ser santo.
Como um verdadeiro revolucionário do bem, fez dos farrapos do seu burel uma bandeira de caridade e agitou-a pelos alcouces abandonados, em que era vergonha entrar, ainda que fosse para socorrer os que morriam.
À última e mais leprosa das perdidas não negava sua boca o beijo da consolação, enviado por Deus aos desamparados pelos homens.

Annotation

-Como quiserem, continuou Filipe, pondo-se em hábitos menores; mas, por minha vida, que a carraspana de hoje ainda me concede apreciar devidamente aqui o meu amigo Fabrício, que talvez acaba de chegar de alguma visita diplomática, vestido com esmero e alinho, porém, tendo a cabeça encapuzada com a vermelha e velha carapuça do Leopoldo; este, ali escondido dentro do seu robe-de-chambre cor de burro quando foge, e sentado em uma cadeira tão desconjuntada que, para não cair com ela, põe em ação todas as leis de equilíbrio, que estudou em Pouillet; acolá, enfim, o meu romântico Augusto, em ceroulas, com as fraldas à mostra, estirado em um canapé em tão bom uso, que ainda agora mesmo fez com que Leopoldo se lembrasse de Bocage.

Annotation

-Como quiserem, continuou Filipe, pondo-se em hábitos menores; mas, por minha vida, que a carraspana de hoje ainda me concede apreciar devidamente aqui o meu amigo Fabrício, que talvez acaba de chegar de alguma visita diplomática, vestido com esmero e alinho, porém, tendo a cabeça encapuzada com a vermelha e velha carapuça do Leopoldo; este, ali escondido dentro do seu robe-de-chambre cor de burro quando foge, e sentado em uma cadeira tão desconjuntada que, para não cair com ela, põe em ação todas as leis de equilíbrio, que estudou em Pouillet; acolá, enfim, o meu romântico Augusto, em ceroulas, com as fraldas à mostra, estirado em um canapé em tão bom uso, que ainda agora mesmo fez com que Leopoldo se lembrasse de Bocage.

Annotation

- **carraspana:** 2 - carraspana, 1 - bebedeira, 1 - porre
- **esmero:** 3 - aprumo, 3 - asseio, 2 - alinho, 2 - esmero, 1 - elegância, 1 - perfeição, 1 - primor
- **alinho:** 3 - apuro, 3 - asseio, 2 - alinho, 2 - esmero , 1 - decência, 1 - dignidade
- **fraldas:** 2 - fralda, 1 - aba
- **canapé:** 2 - canapé, 1 - sofá

Instance of Annotation Output

- 16,{Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, 0, arregaço, 2, carraspana, 0, carão, 0, chegadela, 0, esbregue, 0, esfrega, 0, pito, 0, ralhação, 0, ralho, 0, repreensão, 0, reprimenda, 0, tosa, 0, tunda, , carraspana, 1, bebedeira, 1, porre, 0, , 0,]}

Instance of Annotation Output

Sentence number in the corpus file

- 16,{Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, 0, arregaço, 2, carraspana, 0, carão, 0, chegadela, 0, esbregue, 0, esfrega, 0, pito, 0, ralhação, 0, ralho, 0, repreensão, 0, reprimenda, 0, tosa, 0, tunda, , carraspana, 1, bebedeira, 1, porre, 0, , 0,]}

Instance of Annotation Output

Original sentence

- 16,{Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, 0, arregaço, 2, carraspana, 0, carão, 0, chegadela, 0, esbregue, 0, esfrega, 0, pito, 0, ralhação, 0, ralho, 0, repreensão, 0, reprimenda, 0, tosa, 0, tunda, , carraspana, 1, bebedeira, 1, porre, 0, , 0,]}

Instance of Annotation Output

Selected word and complexity annotation

- 16,{Bocage, quando tomava carraspana, descompunha os médicos.=[**carraspana**, 779, 789, 0, arregaço, 2, carraspana, 0, carão, 0, chegadela, 0, esbregue, 0, esfrega, 0, pito, 0, ralhação, 0, ralho, 0, repreensão, 0, reprimenda, 0, tosa, 0, tunda, , carraspana, 1, bebedeira, 1, porre, 0, , 0,]}

Instance of Annotation Output

Word position in the corpus file

- 16,{Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, 0, arregaço, 2, carraspana, 0, carão, 0, chegadela, 0, esbregue, 0, esfrega, 0, pito, 0, ralhação, 0, ralho, 0, repreensão, 0, reprimenda, 0, tosa, 0, tunda, , carraspana, 1, bebedeira, 1, porre, 0, , 0,]}

Results

- Dictionary of Complex Words:
  - 3720 annotations: 790 different word senses
- Simplification gold standard for Literary Texts

Merci!

Leonardo Zilio

leonardo.zilio@uclouvain.be