

# Simplifying Complex Textual Expressions

Rodrigo Wilkens, Marceley Boito,  
Luiza Hagemann, Alessandro Vecchia

Neurocognition and Natural Language Processing  
Research Lab

Lexical Simplification

Feature Evaluation (Wilkins et al., 2014)

Creating complex/simple list (Boito et al., 2014)

Simplification system and (S)EW

## Lexical Simplification

Feature Evaluation (Wilkins et al., 2014)

Creating complex/simple list (Boito et al., 2014)

Simplification system and (S)EW

## Project Goal

---

### Goal

Automatically simplify text documents to make them available to a large group

Example: children, second language students or clinical groups.

## Goal

---

### Warning

Lexical Simplification can't be applied indiscriminately

## Goal

---

### Warning

Lexical Simplification can't be applied indiscriminately

kicked the bucket



## System

---

### Features

- ▶ Lexical simplification
- ▶ Identification of MWEs
- ▶ Identification of compositionally and paraphrases

Lexical Simplification

Feature Evaluation (Wilkins et al., 2014)

Creating complex/simple list (Boito et al., 2014)

Simplification system and (S)EW



# What is simple?

## What is simple?

### Lack of experiments but several heuristics

- ▶ Word Length ( $W_{length}$ )
- ▶ Frequency ( $Freq_{WaC}$ )
- ▶ Frequency on Childes ( $Freq_{Childes}$ )
- ▶ Frequency on simple and Complex corpora ( $Freq_{simple}$  &  $Freq_{complex}$ )
- ▶ Number of synsets (WordNet) ( $Num_{Synsets}$ )

## What is simple?

### Lack of experiments but several heuristics

- ▶ Word Length ( $W_{length}$ )
- ▶ Frequency ( $Freq_{WaC}$ )
- ▶ Frequency on Childes ( $Freq_{Childes}$ )
- ▶ Frequency on simple and Complex corpora ( $Freq_{simple}$  &  $Freq_{complex}$ )
- ▶ Number of synsets (WordNet) ( $Num_{Synsets}$ )

So, how to measure (automatically) simple/complex words?

### ML evaluation

Creation of several ML supervised model dataset: SemEval 2012

### Algorithms

- ▶ J48
- ▶ Naive Bayes
- ▶ Naive Bayes Network
- ▶ Support Vector Machines
- ▶ Adaptive Boosting

## Results (F1)

---

Features	English	Portuguese
$W_{length}$	0.67	0.49
$Freq_{simple}$	0.71	0.62
$Freq_{complex}$	0.68	0.57
$Freq_{simple}$ & $Freq_{complex}$	0.73	0.62
$Freq_{Childes}$ (simple)	0.78	0.62
$Freq_{WaC}$	0.79	0.60
$Num_{Synsets}$	0.65	0.54
average	0.72	0.58
All features	<b>0.82</b>	<b>0.63</b>

Lexical Simplification

Feature Evaluation (Wilkins et al., 2014)

Creating complex/simple list (Boito et al., 2014)

Simplification system and (S)EW

## Goal

---

### Find out:

- ▶ the changes in the relevance of syntactic structures

### Related works

- ▶ Several works characterising different groups (Oakes, 2008, Dagan et al., 1997, Oakes and Farrow, 2007)
- ▶ Groups represented by simplified documents and their original versions
- ▶ cross-entropy

### *“Coleção é Só o Começo”*

**target** people with low literacy skills.

**simplifiers** linguists

**books** Alienista, Cortiço, Guarani, Escrava Isaura, Policarpo Quaresma



## Corpus pre-processing

Step	Example	
Input	...Havia mais de vinte anos que isso acontecia. Saindo do Arsenal de Guerra, onde era subsecretário...	
Sentence splitter	Havia mais de vinte anos que isso acontecia.	
LX parser	(ROOT (S (VP (V' (V Havia) (NP (CARD' (ADV' (ADV mais) (ADV de))) (CARD vinte)) (N anos)))) (CP (C que) (S (NP (DEM isso)) (VP (V acontecia.))))))	
Treebank splitter	ROOT ->S S ->VP VP ->V' CP V' ->V NP V ->'Havia' NP ->CARD' N CARD' ->ADV' CARD ADV' ->ADV ADV	CARD ->'vinte' N ->'anos' CP ->C S C ->'que' S ->NP VP NP ->DEM DEM ->'isso' VP ->V

## Grammatical weight

---

original	weight <sub>s</sub>	weight <sub>o</sub>	simplified	weight <sub>s</sub>	weight <sub>o</sub>
NP → N'	61098	1	ROOT → S	1	9
NP → ART N'	61097	2	NP → N	2	16
PP → P NP	61088	3	S → NP VP	3	5
VP → V NP	8	4	VP → V VP	4	31
S → NP VP	3	5	P → 'para'	5	213
N' → N A	61090	6	NP → PRS	6	99
S → VP	9	7	NP → ART N	7	18
P → 'de'	61092	8	VP → V NP	8	4
ROOT → S	1	9	S → VP	9	7
N' → N PP	61085	10	VP → V PP	10	48

Lexical Simplification

Feature Evaluation (Wilkins et al., 2014)

Creating complex/simple list (Boito et al., 2014)

Simplification system and (S)EW

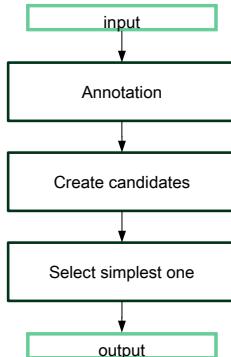
## What is simplification to non linguists?

### Simplification System

- ▶ Target: MWEs and nouns.
- ▶ Range: Wordnet (hypernym and synonym).

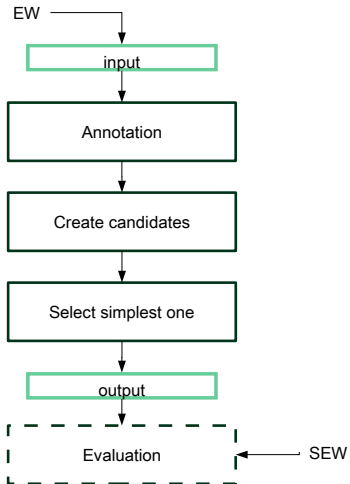
# Pipeline

---



# Pipeline

---



## Simplification “rules” from EW to SEW

---

Pairs (EW → SEW)	Freq
in → it is found in	2182
the →	833
nord-pas-de-calais region → north	608
calvados → region basse-normandie in the calvados	539
district → canton	416
also →	241
in → of kentucky in	212
former →	200
discovered → found	161
france → part of france	158

## Several maps between MWEs

---

- MWE → single word: 25,07%.  
shut down → closed
  
- single word → MWE: 16,62%.  
*machine* → *mechanical device*
  
- MWE → MWE: 22,49%.  
*lies in* → *be on*



## Future work

---

### System upgrade

- ▶ Add a WSD step;
- ▶ Improve the candidates selection step (n-gram language model)
- ▶ Analyze simplicity features to MWE

### Other steps

- ▶ Profile language groups (children, second language students, clinical groups...)
- ▶ Development of resources (Portuguese)
  - Paraphrase dictionary
  - List of simple and complex words
  - Lexical complexity scale

# Simplifying Complex Textual Expressions

Rodrigo Wilkens, Marceley Boito,  
Luiza Hagemann, Alessandro Vecchia

Neurocognition and Natural Language Processing  
Research Lab

## References:

- Boito, M. Z., Hagemann, L., Wilkens, R., and Villavicencio, A. (2014). Uma análise do perfil de entropia das estruturas sintáticas do português. *TorPorEsp*.
- Dagan, I., Lee, L., and Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.
- Oakes, M. P. (2008). Statistical measures for corpus profiling. In *Proceedings of the Open University Workshop on Corpus Profiling, London, UK (October 2008)*.
- Oakes, M. P. and Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. *Literary and linguistic computing*, 22(1):85–99.
- Wilkens, R., Vecchia, A. D., Boito, M. Z., Padró, M., and Villavicencio, A. (2014). Size does not matter. frequency does. a study of features for measuring lexical complexity. In *IBERAMIA*.