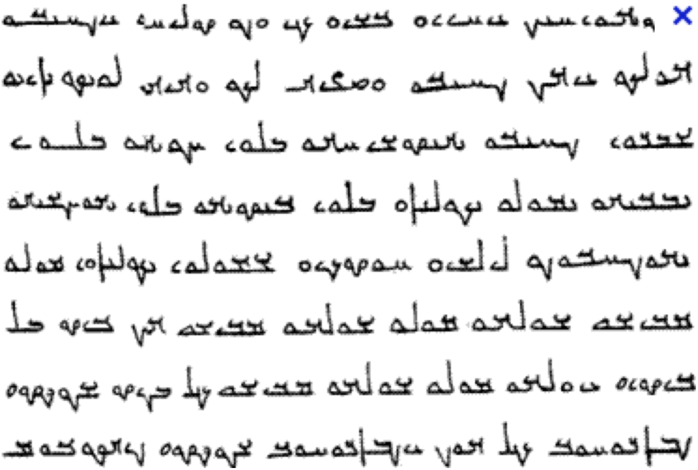
Size does not matter. Frequency does. A study of features for measuring lexical complexity

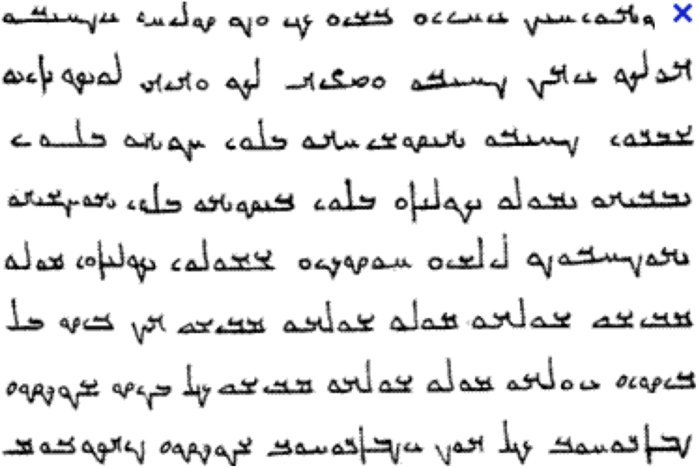
Aline Villavicencio and Marco Idiart
joint work with
Rodrigo Wilkens, Alessandro Dalla Vecchia,
Marcely Zanon Boito, Muntsa Padró

Institute of Informatics
Federal University of Rio Grande do Sul (Brazil)
avillavicencio@inf.ufrgs.br, marco.idiart@gmail.com

LIF - November, 2015

1. Introduction
2. Related Work
3. Methodology
4. Results
5. Conclusions



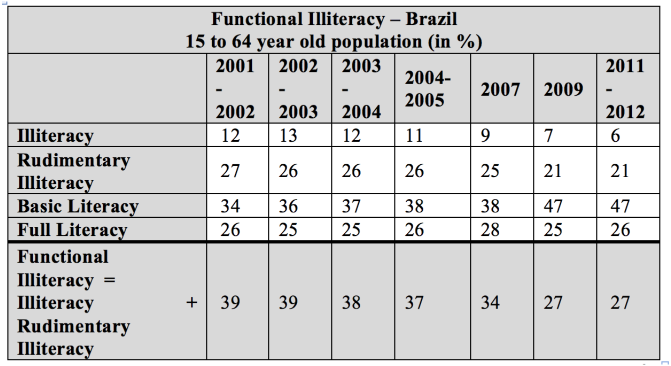


Not understanding is always a serious problem

Spelling out illiteracy in Brazil

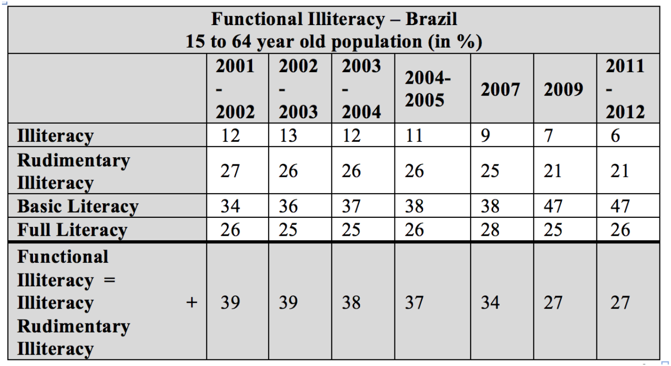
In 2012 of the total population (aged 15 or over):

- 6% is completely illiterate
- 27% is functionally illiterate



Functional Illiteracy – Brazil
15 to 64 year old population (in %)

	2001 - 2002	2002 - 2003	2003 - 2004	2004-2005	2007	2009	2011 - 2012
Illiteracy	12	13	12	11	9	7	6
Rudimentary Illiteracy	27	26	26	26	25	21	21
Basic Literacy	34	36	37	38	38	47	47
Full Literacy	26	25	25	26	28	25	26
Functional Illiteracy = Illiteracy + Rudimentary Illiteracy	39	39	38	37	34	27	27



Functional Illiteracy – Brazil
15 to 64 year old population (in %)

	2001 - 2002	2002 - 2003	2003 - 2004	2004-2005	2007	2009	2011 - 2012
Illiteracy	12	13	12	11	9	7	6
Rudimentary Illiteracy	27	26	26	26	25	21	21
Basic Literacy	34	36	37	38	38	47	47
Full Literacy	26	25	25	26	28	25	26
Functional Illiteracy = Illiteracy + Rudimentary Illiteracy	39	39	38	37	34	27	27

- Total population: 194 million
- Rudimentary to basic literacy: 145 million people
  - who may not have a full understanding of the information communicated to them.
- How can Natural Language Processing help with that?

Text Simplification

Goal:

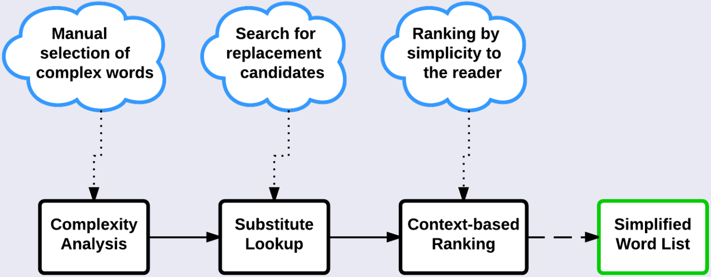
to use NLP techniques to make texts more accessible to people with comprehension limitations (e.g. language learners, clinical cases)

Text Simplification

Goal:

to use NLP techniques to make texts more accessible to people with comprehension limitations (e.g. language learners, clinical cases)

Lexical Simplification pipeline [Specia et al., 2012]


Manual selection of complex words
Search for replacement candidates
Ranking by simplicity to the reader
Complexity Analysis
Substitute Lookup
Context-based Ranking
Simplified Word List

1. Introduction
2. **Related Work**
3. Methodology
4. Results
5. Conclusions

Text Simplification

- rule-based architecture (Siddharthan, 2002)
- machine translation techniques for learning simplifications: from English Wikipedia aligned to Simple English Wikipedia (Woodsend and Lapata 2011, Biran et al. 2011)

Text Simplification

- Simplext (Saggion et al. 2011): ubiquitous text simplification for Spanish
- PorSimple (Alusio et al. 2008): text simplification for Portuguese
- FLELex (François et al. 2014): FLELex graded lexical resource for French foreign learners

Text Simplification

eXPlainText: Project funded by Samsung BR

- Lexical Simplification of Complex Expressions for Brazilian Portuguese
- Challenges
  1. to develop resources and tools for lexical simplification
  2. to investigate how to incorpora multiword expressions in the simplification pipeline (semantic vs syntactic vs distributional characteristics)

Text Simplification

- Original: The **malaria mosquito** was infected with **disease-fighting bacteria**
- Simplified: The mosquito that carries malaria was infected with a bacteria that stimulates disease-fighting
- Simplified: The mosquito that carries malaria was infected with a bacteria for fighting disease

In this work

First we want to determine if a word is complex and needs replacing.

In this work

First we want to determine if a word is complex and needs replacing.

But how do people do that?

using information about word length, frequency, polysemy,... (e.g. Chall and Dale, 1995)

In this work

First we want to determine if a word is complex and needs replacing.

But how do people do that?

using information about word length, frequency, polysemy,... (e.g. Chall and Dale, 1995)

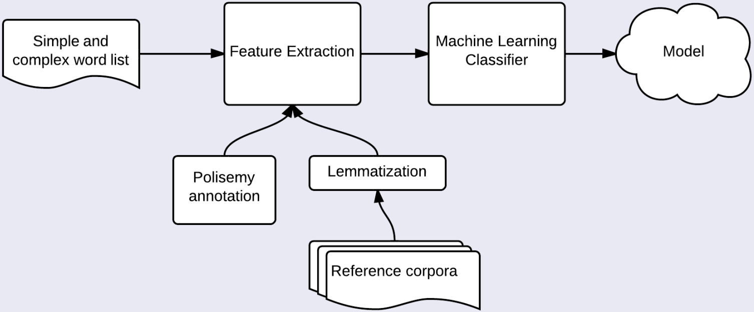
And how can we simulate that ?

1. examine if the characteristics of simple vs original texts are the same
2. use these characteristics to build classifiers for distinguishing complex from simple words and
3. determine which are the most relevant characteristics for the task.

1. Introduction
2. Related Work
3. Methodology
4. Results
5. Conclusions

Methodology

Pipeline



Original and Simple versions of classic literary books: “Coleção é só o começo”

- convert PDFs
- sentence splitting, tokenization (GATE, Cunningham et al. (2002))
- parsing (LX parser, Costa and Branco (2010))

The Corpora

book	#words (original)	#sentences	words/ sentences	#words (simple)	#sentences	words/ sentences
Alienista	16673	906	18,40	14109	1076	13,11
Cortiço	81025	5702	14,21	14958	1236	12,10
Guarani	108341	6026	17,98	19151	1571	12,19
Escrava Isaura	53503	3240	16,51	15729	1426	11,03
Policarpo Quaresma	67009	5099	13,14	19888	1560	12,75

Cross-Entropy

Cross-Entropy to cluster documents according to text simplicity

$$H(x, P, Q) = -\sum_{i} P(x_i) log_2 \frac{P(x_i)}{Q(x_i)} \tag{1}$$

Cross-Entropy

Similarity	A_o	A_s	C_o	C_s	E_o	E_s	G_o	G_s	P_o	P_s
1	G_o	E_o	P_o	P_s	P_o	P_s	C_o	E_s	C_o	E_s
2	C_o	P_o	G_o	E_s	C_o	G_s	P_o	P_s	E_o	G_s
3	P_o	C_o	E_o	A_s	G_o	C_s	E_o	C_s	G_o	C_s
4	E_o	G_o	A_o	G_s	A_o	A_s	A_o	A_s	A_o	A_s
5	A_s	C_s	A_s	E_o	A_s	E_o	A_s	E_o	A_s	E_o
6	C_s	A_o	C_s	P_o	C_s	P_o	C_s	P_o	C_s	P_o
7	P_s	P_s	P_s	C_o	P_s	C_o	P_s	C_o	P_s	C_o
8	E_s	E_s	E_s	G_o	E_s	G_o	E_s	G_o	E_s	G_o
9	G_s	G_s	G_s	A_o	G_s	A_o	G_s	A_o	G_s	A_o

Cross-Entropy

Similarity	A_o	A_s	C_o	C_s	E_o	E_s	G_o	G_s	P_o	P_s
1	G_o	E_o	P_o	P_s	P_o	P_s	C_o	E_s	C_o	E_s
2	C_o	P_o	G_o	E_s	C_o	G_s	P_o	P_s	E_o	G_s
3	P_o	C_o	E_o	A_s	G_o	C_s	E_o	C_s	G_o	C_s
4	E_o	G_o	A_o	G_s	A_o	A_s	A_o	A_s	A_o	A_s
5	A_s	C_s	A_s	E_o	A_s	E_o	A_s	E_o	A_s	E_o
6	C_s	A_o	C_s	P_o	C_s	P_o	C_s	P_o	C_s	P_o
7	P_s	P_s	P_s	C_o	P_s	C_o	P_s	C_o	P_s	C_o
8	E_s	E_s	E_s	G_o	E_s	G_o	E_s	G_o	E_s	G_o
9	G_s	G_s	G_s	A_o	G_s	A_o	G_s	A_o	G_s	A_o

Similarity matrix, where Alienista (A), Cortiço (C), Guarani (G), Escrava Isaura (E), Policarpo Quaresma (P); *s* is for simplified text; *o* is the original.

Gold Standards for English and Portuguese

Words classified as complex or simple in English and Portuguese

English

- Sentence with target word and list of synonyms with human judgments about complexity [Specia et al., 2012]
  - Remove neutral words:
    explain; tell; communicate; inform me of; inform; convey to;

Gold Standards for English and Portuguese

Words classified as complex or simple in English and Portuguese

English

- Sentence with target word and list of synonyms with human judgments about complexity [Specia et al., 2012]
  - Remove neutral words:
    explain; tell; communicate; inform me of; inform; convey to;

Portuguese

- Created from corpus “Coleção é só o começo” assuming that words that are more frequent in simplified texts are simple
  - Keyness to create the simple and complex word list

Features

Common features

W_{length} word length (number of characters of each word) [Amoia and Romanelli, 2012, Biran et al., 2011]

$Freq_{WaC}$ frequency of word in a general corpus [Devlin and Unthank, 2006].

$Freq_{Childes}$ frequency of word in corpora with children speech

$Freq_{simple}$ & $Freq_{complex}$ frequency of word in simple and complex corpora [Biran et al., 2011]

$Num_{Synsets}$ number of synsets in WordNet for polysemy

Machine Learning

Classifiers

- J48
- Naive Bayes (NB)
- Naive Bayes Network (NBN)
- Support Vector Machines (SVM)
- Ada Boost (AB)

Evaluation

10-fold cross-validation

Corpora

English corpora

General corpus ukWaC

Corpora

English corpora

General corpus	ukWaC
Complex corpus	English Wikipedia

Corpora

English corpora

General corpus	ukWaC
Complex corpus	English Wikipedia
Simple corpus	Simple English Wikipedia

Corpora

English corpora

General corpus	ukWaC
Complex corpus	English Wikipedia
Simple corpus	Simple English Wikipedia
Children corpus	English corpora in CHILDES

Corpora

Portuguese corpora

General corpus brWaC

Corpora

Portuguese corpora

General corpus brWaC

Complex corpus Folha de São Paulo, Europarl, Machado de Assis corpus and original version of Zero Hora

Corpora

Portuguese corpora

General corpus brWaC

Complex corpus Folha de São Paulo, Europarl, Machado de Assis corpus and original version of Zero Hora

Simple corpus Diário Gaúcho, simplified version of Zero Hora and books for children

Corpora

Portuguese corpora

General corpus brWaC

Complex corpus Folha de São Paulo, Europarl, Machado de Assis corpus and original version of Zero Hora

Simple corpus Diário Gaúcho, simplified version of Zero Hora and books for children

Children corpus Portuguese corpora in CHILDES

Corpora

Reference corpora

Corpus	English			Portuguese		
	Tokens	Types	TTR[a]	Tokens	Types	TTR[a]
General corpus	2,000M	3.8M	0.002	3,000M	2,7M	0.008
Complex corpus	3.0M	197K	0.065	86M	634K	0.007
Simple corpus	2.7M	173K	0.064	317K	26K	0.083
Children corpus	2.1M	35.7K	0.016	177K	5.9K	0.033

[a] Type Token Ratio

1 Introduction

2 Related Work

3 Methodology

4 Results

5 Conclusions

Results

Features	English				
	SVM	J48	NB	NBN	AB
W_{length}	*0.67*	*0.67*	0.66	*0.67*	*0.67*
$Freq_{simple}$	0.70	*0.71*	0.48	*0.71*	*0.71*
$Freq_{complex}$	0.66	0.68	0.49	0.68	*0.69*
$Freq_{simple}$ & $Freq_{complex}$	0.70	*0.73*	0.50	0.70	0.71
$Freq_{Childes}$	0.76	*0.78*	0.59	0.77	*0.78*
$Freq_{WaC}$	0.39	*0.79*	0.60	*0.79*	0.78
$Num_{Synsets}$	*0.65*	*0.65*	0.58	0.63	0.63
All features	0.42	**0.82**	0.62	0.79	0.79

Results

Features	Portuguese				
	SVM	J48	NB	NBN	AB
W_{length}	0.51	0.49	*0.53*	0.33	0.52
$Freq_{simple}$	0.61	*0.62*	0.41	*0.62*	*0.62*
$Freq_{complex}$	0.53	0.57	0.38	*0.58*	*0.58*
$Freq_{simple}$ & $Freq_{complex}$	0.53	0.62	0.40	*0.63*	0.61
$Freq_{Childes}$	0.61	*0.62*	0.41	*0.62*	*0.62*
$Freq_{WaC}$	0.49	*0.60*	0.40	*0.60*	*0.60*
$Num_{Synsets}$	*0.55*	0.54	0.50	0.53	0.54
All features	0.43	0.63	0.43	**0.64**	0.62

Features Evaluation

Feature ablation - All-1

- Test all features removing 1 each time (e.g. all-W_{length})
- Important features are $Freq_{Childes}$ and $Freq_{WaC}$

English	$Freq_{WaC} > Freq_{Childes}$
Portuguese	$Freq_{simple} \& Freq_{complex} > Freq_{Childes}, Freq_{simple}$

1. Introduction
2. Related Work
3. Methodology
4. Results
5. Conclusions

Conclusions

Goal

- investigate how to determine if a word is complex and needs replacing
- compare different characteristics as predictors of lexical complexity
- see if results are consistent for different languages

Conclusions

1. Word frequency is better predictor than length
2. Frequency in simple corpora has different predicting power for English and Portuguese
   - EN SEW can include both the original word and a paraphrase
   - PT Simple text is rewritten
3. Classifiers are better in English (82%) than in Portuguese (64%)
   - EN Gold standard manually created [Specia et al., 2012]
   - PT Gold standard automatically created

Next steps

- Refine the Portuguese gold standard
- Extend the feature set
    - Frequency in other corpora
    - Check if the word occurs in a simple list (e.g. Oxford 3000)

Size does not matter. Frequency does. A study of features for measuring lexical complexity

Aline Villavicencio and Marco Idiart
joint work with
Rodrigo Wilkens, Alessandro Dalla Vecchia,
Marcely Zanon Boito, Muntsa Padró

Institute of Informatics
Federal University of Rio Grande do Sul (Brazil)
avillavicencio@inf.ufrgs.br, marco.idiart@gmail.com

LIF - November, 2015

Acknowledgments

This research is part of the project Simplificação Textual de Expressões Complexas sponsored by Samsung Eletrônica da Amazônia Ltda., under the terms of Law number 8.248/91.

Size does not matter. Frequency does. A study of features for measuring lexical complexity

Aline Villavicencio and Marco Idiart
joint work with
Rodrigo Wilkens, Alessandro Dalla Vecchia,
Marcely Zanon Boito, Muntsa Padró

Institute of Informatics
Federal University of Rio Grande do Sul (Brazil)
avillavicencio@inf.ufrgs.br, marco.idiart@gmail.com

LIF - November, 2015