NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese

Sandra M Aluisio
NILC/ICMC/USP

Leitura e(m) Interfaces: Teorias, Métodos e Aplicações

Programa de Pós-Graduação em Inglês: Estudos Linguísticos e Literários da UFSC

16/04/2021

Outline

Introducing NILC-Metrix

Motivation and Theoretical Foundations

Presentation of the Metrics

NILC-Metrix: Applications

Guidelines for using NILC-Metrix

Future Work

Introducing NILC-Metrix

What is NILC-Metrix ?

200 metrics developed over more than a decade for Brazilian Portuguese (from 2008 - 2021) at The Interinstitutional Center for Computational Linguistic (NILC) (ICMC/USP)

to extract objective information from **various linguistic levels of written and spoken language** (multilevel analysis)

Main objective: to provide proxies to assess cohesion, coherence and textual complexity,

in **descriptive analysis** and in the **creation of computational models**

Text Genre, Text Types, Authorship, Texts for Grades, Text Sources, Translated Texts ....

... Text and Sentential Complexity Predictors

Metrics may help researchers to investigate

(i) how text characteristics correlate with reading comprehension;

(ii) which are the most challenging characteristics of a given text, that is, which characteristics make a text or corpus more complex;

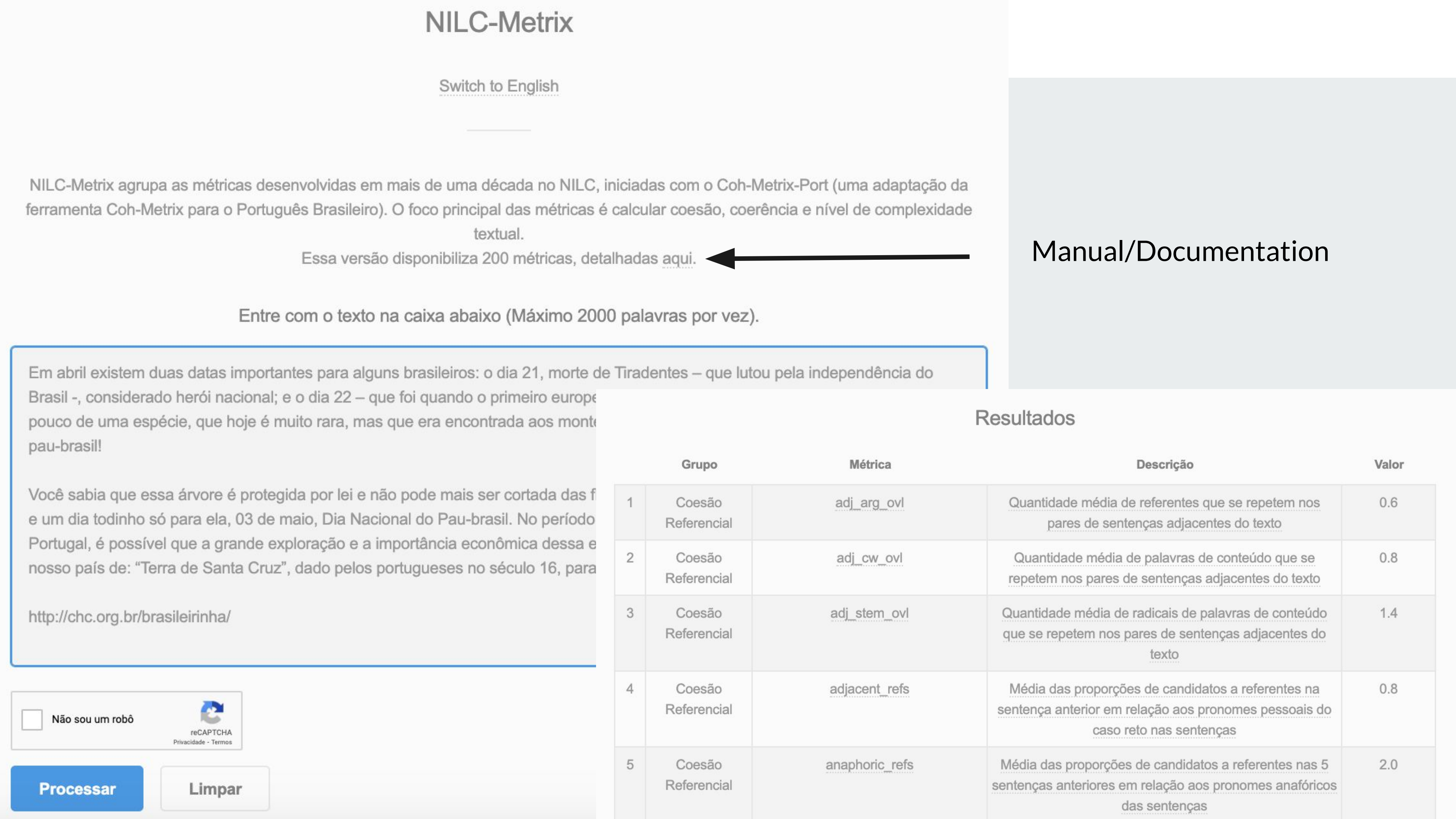
(iii) which texts have the most adequate characteristics to develop target learners' skills; and

(iv) which parts of a text are disproportionately **complex** and should be simplified to meet **a given audience**.

Where can I test NILC-METRIX ?

Web-based tool:
http://fw.nilc.icmc.usp.br:23380/nilcmetrix

Source Code of NILC-Metrix:
https://github.com/nilc-nlp/nilcmetrix (September 30th) under AGPLv3 license.



NILC-Metrix

Switch to English

NILC-Metrix agrupa as métricas desenvolvidas em mais de uma década no NILC, iniciadas com o Coh-Metrix-Port (uma adaptação da ferramenta Coh-Metrix para o Português Brasileiro). O foco principal das métricas é calcular coesão, coerência e nível de complexidade textual.
Essa versão disponibiliza 200 métricas, detalhadas aqui.

Manual/Documentation

Entre com o texto na caixa abaixo (Máximo 2000 palavras por vez).

Em abril existem duas datas importantes para alguns brasileiros: o dia 21, morte de Tiradentes – que lutou pela independência do Brasil -, considerado herói nacional; e o dia 22 – que foi quando o primeiro europe
pouco de uma espécie, que hoje é muito rara, mas que era encontrada aos monte
pau-brasil!

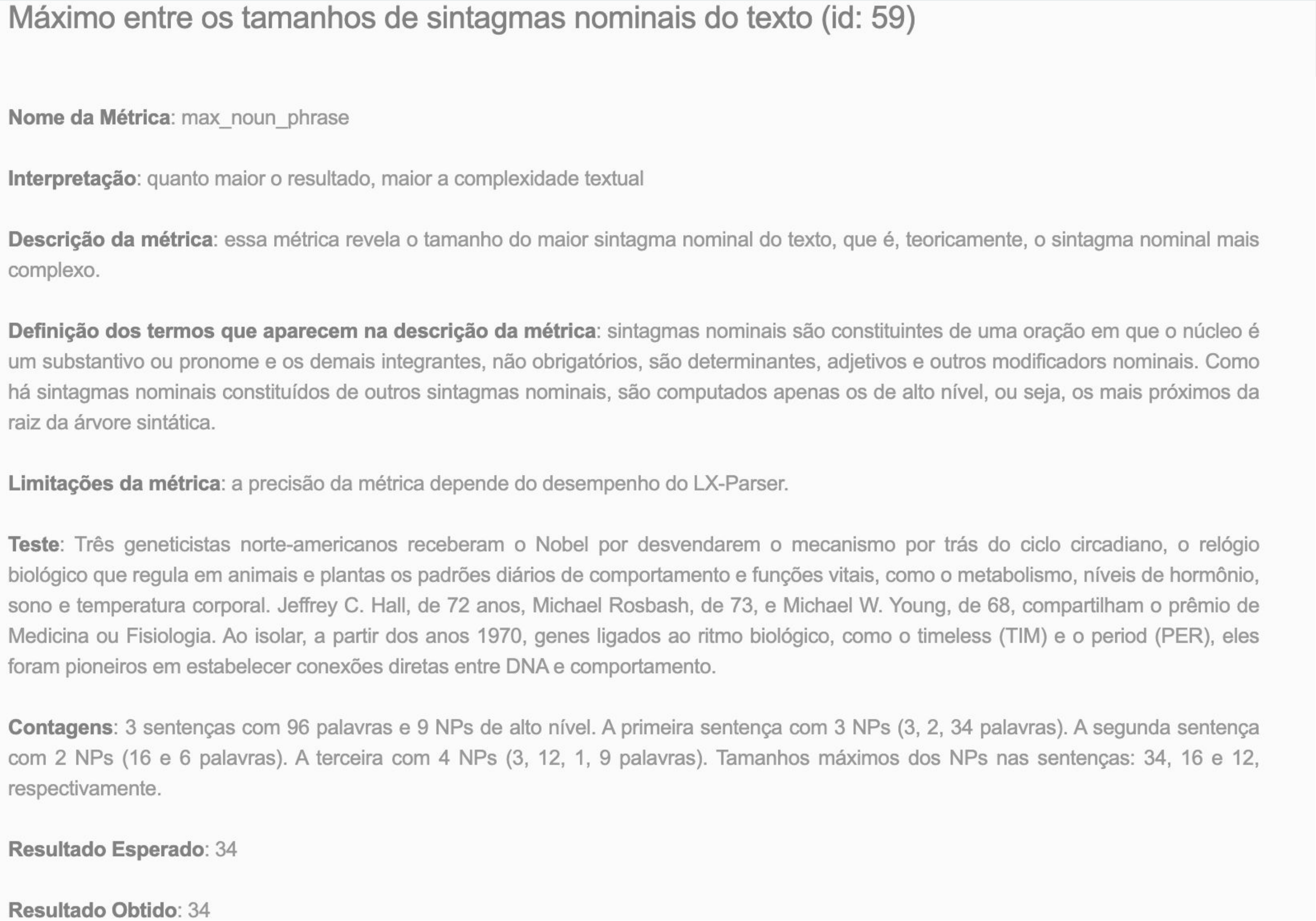
Você sabia que essa árvore é protegida por lei e não pode mais ser cortada das f
e um dia todinho só para ela, 03 de maio, Dia Nacional do Pau-brasil. No período
Portugal, é possível que a grande exploração e a importância econômica dessa e
nosso país de: "Terra de Santa Cruz", dado pelos portugueses no século 16, para

http://chc.org.br/brasileirinha/

Resultados

	Grupo	Métrica	Descrição	Valor
1	Coesão Referencial	adj_arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes do texto	0.6
2	Coesão Referencial	adj_cw_ovl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto	0.8
3	Coesão Referencial	adj_stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto	1.4
4	Coesão Referencial	adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças	0.8
5	Coesão Referencial	anaphoric_refs	Média das proporções de candidatos a referentes nas 5 sentenças anteriores em relação aos pronomes anafóricos das sentenças	2.0

Documentation



Máximo entre os tamanhos de sintagmas nominais do texto (id: 59)

Nome da Métrica: max_noun_phrase

Interpretação: quanto maior o resultado, maior a complexidade textual

Descrição da métrica: essa métrica revela o tamanho do maior sintagma nominal do texto, que é, teoricamente, o sintagma nominal mais complexo.

Definição dos termos que aparecem na descrição da métrica: sintagmas nominais são constituintes de uma oração em que o núcleo é um substantivo ou pronome e os demais integrantes, não obrigatórios, são determinantes, adjetivos e outros modificadores nominais. Como há sintagmas nominais constituídos de outros sintagmas nominais, são computados apenas os de alto nível, ou seja, os mais próximos da raiz da árvore sintática.

Limitações da métrica: a precisão da métrica depende do desempenho do LX-Parser.

Teste: Três geneticistas norte-americanos receberam o Nobel por desvendarem o mecanismo por trás do ciclo circadiano, o relógio biológico que regula em animais e plantas os padrões diários de comportamento e funções vitais, como o metabolismo, níveis de hormônio, sono e temperatura corporal. Jeffrey C. Hall, de 72 anos, Michael Rosbash, de 73, e Michael W. Young, de 68, compartilham o prêmio de Medicina ou Fisiologia. Ao isolar, a partir dos anos 1970, genes ligados ao ritmo biológico, como o timeless (TIM) e o period (PER), eles foram pioneiros em estabelecer conexões diretas entre DNA e comportamento.

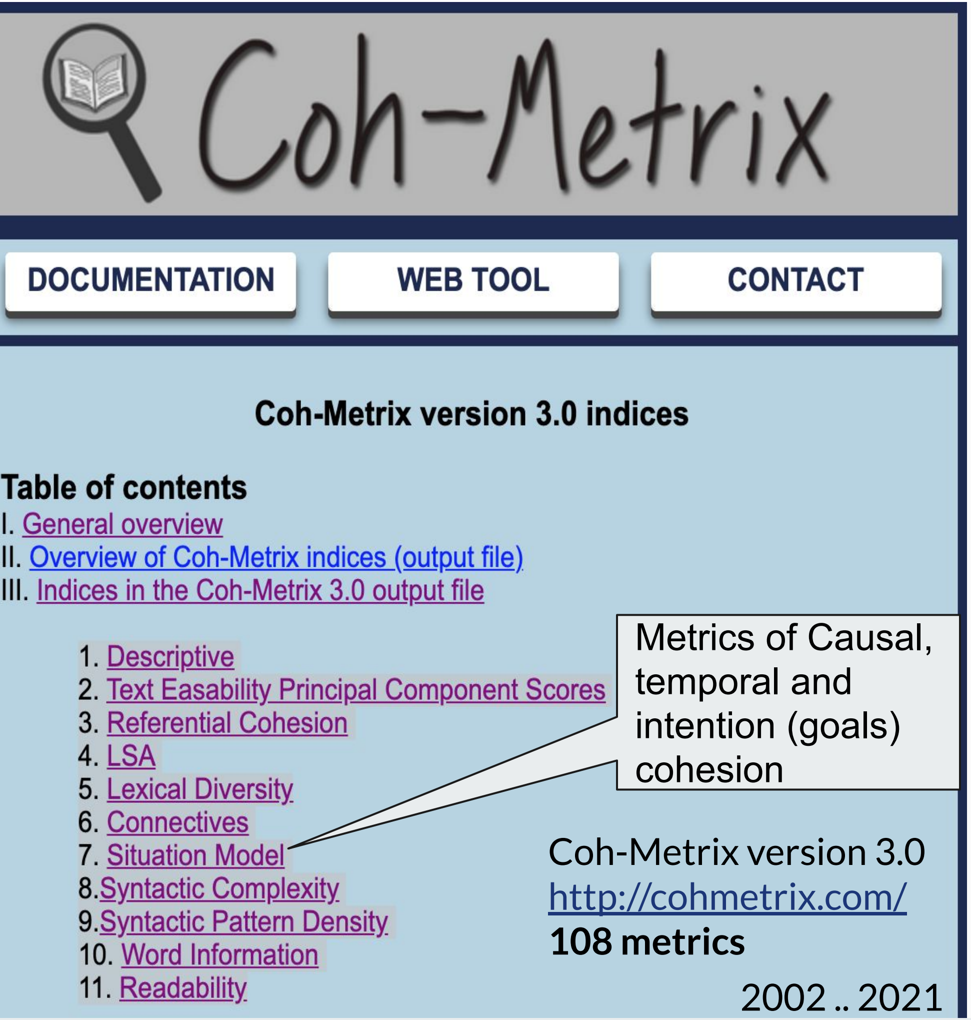
Contagens: 3 sentenças com 96 palavras e 9 NPs de alto nível. A primeira sentença com 3 NPs (3, 2, 34 palavras). A segunda sentença com 2 NPs (16 e 6 palavras). A terceira com 4 NPs (3, 12, 1, 9 palavras). Tamanhos máximos dos NPs nas sentenças: 34, 16 e 12, respectivamente.

Resultado Esperado: 34

Resultado Obtido: 34

How NILC-METRIX metrics were implemented ?

- **NLP resources**
- Lexicons
  - Positive & Negative words of LIWC Portuguese (http://143.107.183.175:21380/portlex/)
  - Discourse markers (Pardo and Nunes, 2006),
  - Repository of Psycholinguistic Properties of Brazilian Portuguese words (imageability, concreteness, subjective frequency = familiarity, AoA) (Santos et al., 2017),
  - List of simple words of the children's dictionary (Biderman, 2006)
  - Temporal lexicon (Pardo and Nunes, 2006) and (Bick, 2000)
  - Frequency Lists from 3 large BP corpora (Banco de Português, Corpus Brasileiro, BrWac)
- Thesaurus (**TeP** - Portuguese Electronic Thesaurus) and **Wordnet.Br** (Verbs)
- **NLP tools**
- Tokenizer and Tagger (nlpnet)
- 3 Parsers (constituency & dependency): Palavras, MALTParser e LXParser
- Similarity Models (LSA and Span-LSA)



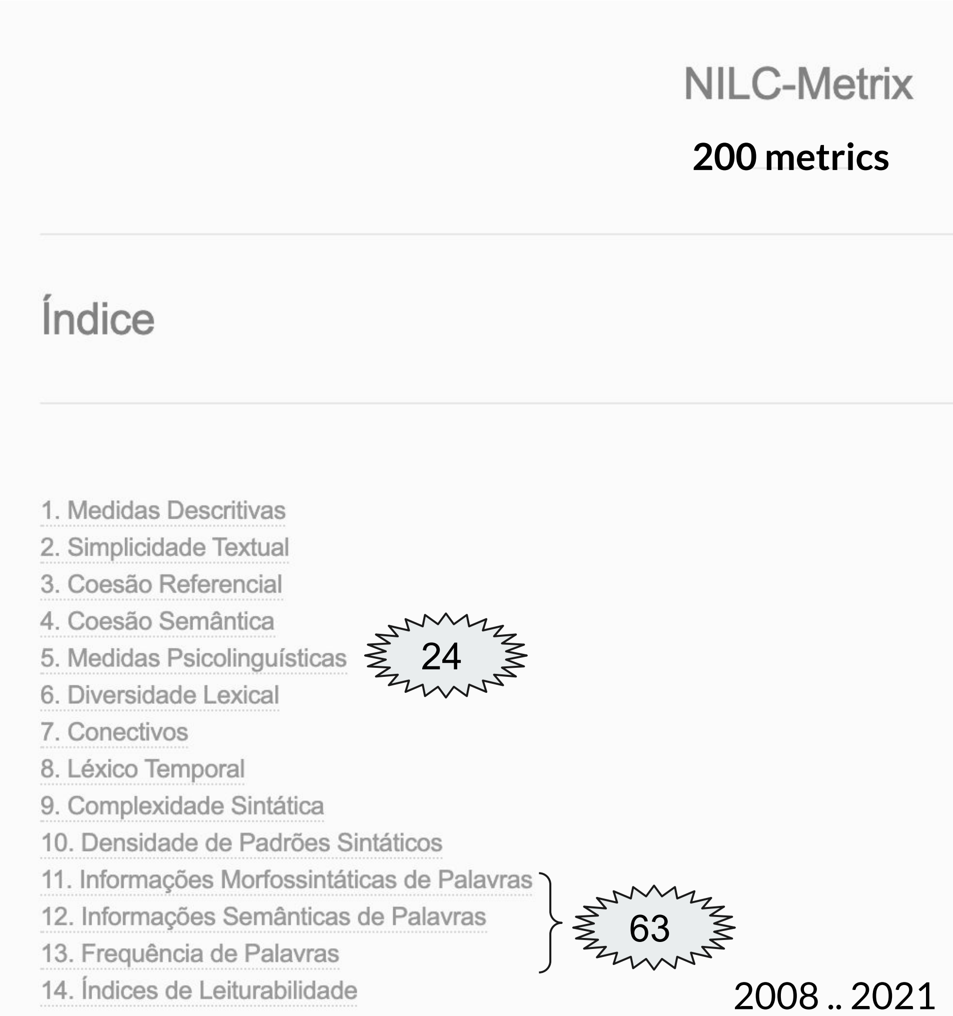
Coh-Metrix version 3.0 indices

Table of contents

Metrics of Causal, temporal and intention (goals) cohesion

Coh-Metrix version 3.0
http://cohmetrix.com/
108 metrics

2002 .. 2021



NILC-Metrix

200 metrics

Índice

24

63

2008 .. 2021

NILC-Metrix **is** not well suited to

- evaluate the basic reading components (alphabet, letter-sound correspondences, lexical decoding, morphological awareness and reading fluency --- words read per minute)
-
    - **Léxico do Português Brasileiro - LexPorBR**
      http://www.lexicodoportugues.com/
    - **Léxico do Português Brasileiro Infantil - LexPorBR-Infantil**
    - http://www.lexicodoportugues.com/infantil/

Gustavo Estivalet,
UFPB

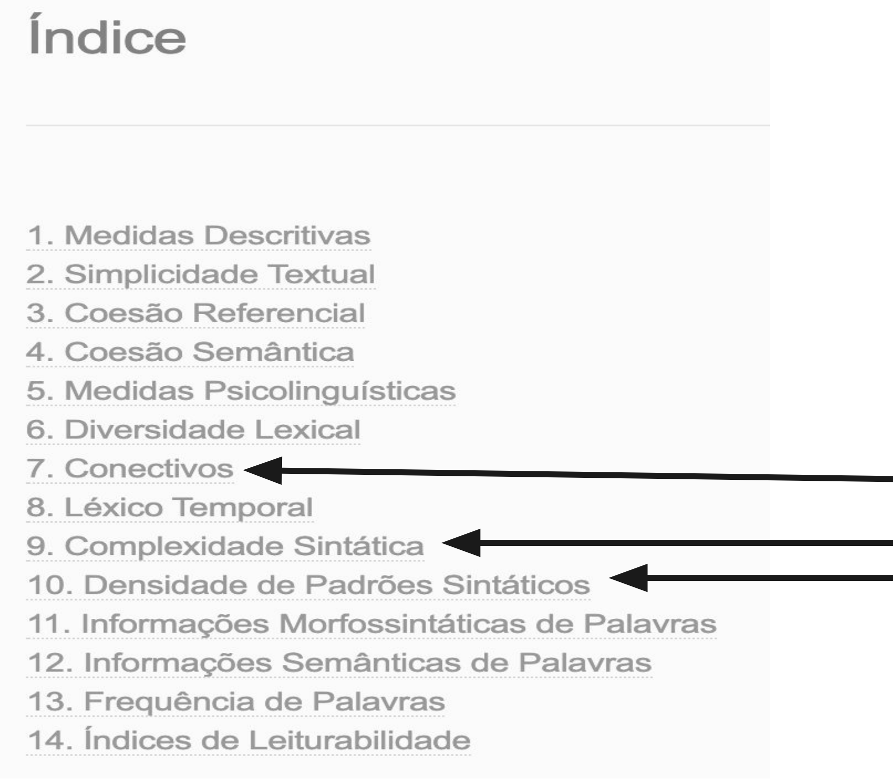
- classify words into psychological categories (LIWC - Linguistic Inquiry Word Counts, 2007 and 2015 dictionaries)
    - **Brazilian Portuguese LIWC 2007 Dictionary**
    - http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc

Oral Discourse vs. Written Texts

Spoken language has simpler syntactic structures with few embedded clauses, and active rather than passive voice

Sentences in **academic articles**, frequently have a complex, embedded syntax that creates demands on an individuals working memory

(Dowell, et al., 2016)



Índice

1. Medidas Descritivas
2. Simplicidade Textual
3. Coesão Referencial
4. Coesão Semântica
5. Medidas Psicolinguísticas
6. Diversidade Lexical
7. Conectivos
8. Léxico Temporal
9. Complexidade Sintática
10. Densidade de Padrões Sintáticos
11. Informações Morfossintáticas de Palavras
12. Informações Semânticas de Palavras
13. Frequência de Palavras
14. Índices de Leiturabilidade

Example: excerpt from an abstract of a PhD thesis

Dentre os genes altamente induzidos, destacam-se os **que codificam proteínas de choque térmico (Hsps),** que previnem a desnaturação e a formação de agregados protéicos ou degradam polipeptídeos irreversivelmente desnaturados.

A partir da determinação do início de transcrição de seis genes altamente induzidos no choque térmico, **propôs-se** um consenso para promotores dependentes do fator sigma alternativo **que controla a resposta ao choque térmico**, sigma32.

Observou-se também a indução de genes relacionados ao estresse extracitoplasmático, **que são regulados pelo fator sigma alternativo sigmaE**.

No choque osmótico e salino, os genes codificando a maioria das Hsps **foram** reprimidos na exposição prolongada a esses estresses, indicando que a resposta **não** é mediada por sigma32 ou sigmaE.

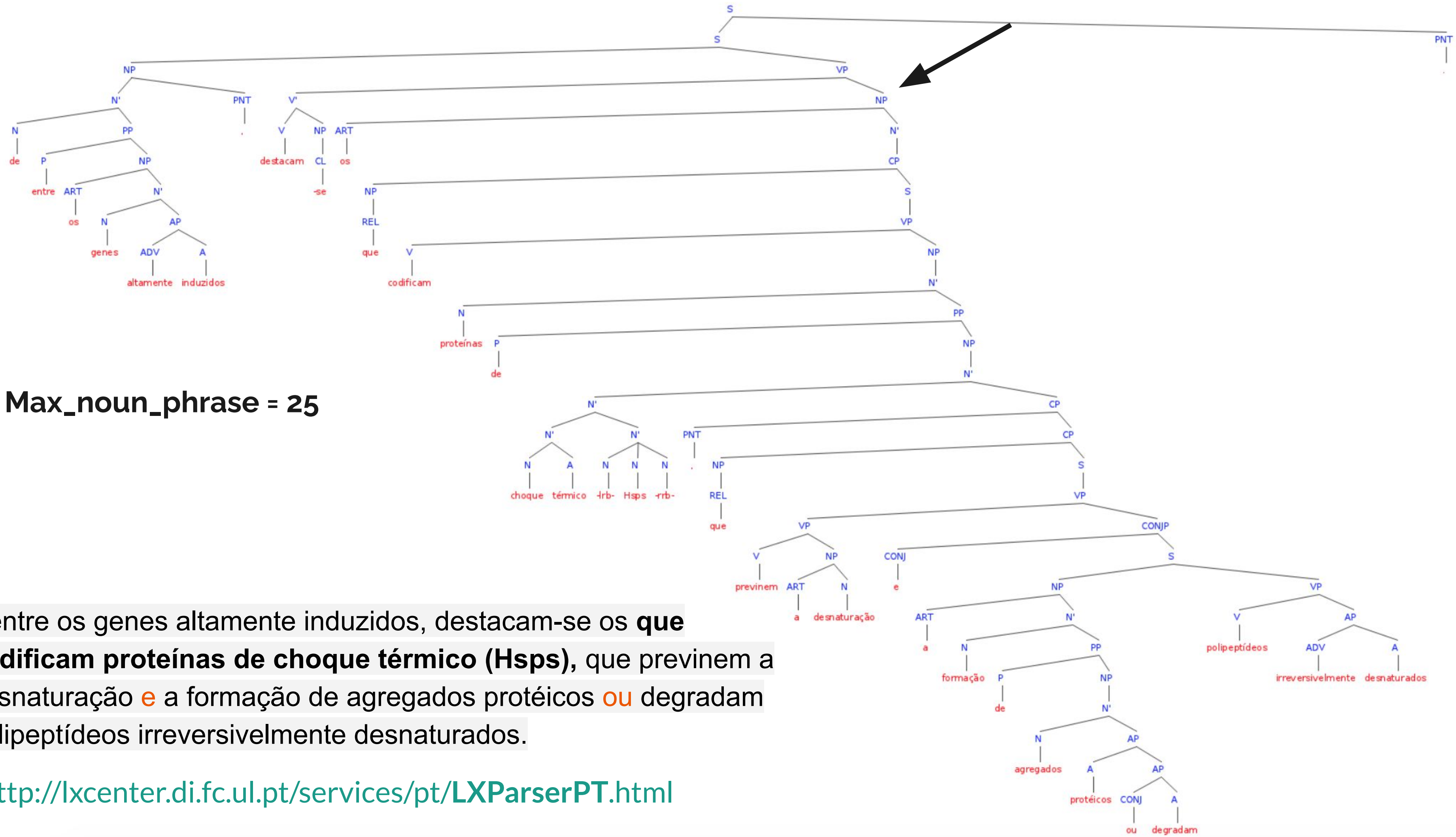
Dos 142 genes induzidos tanto no estresse salino como osmótico, 57% **codificam** proteínas hipotéticas **ou** hipotéticas conservadas, indicando uma possível função na resposta a estes estresses.

Fonte: https://teses.usp.br/teses/disponiveis/46/46131/tde-20092006-014126/pt-br.php

1. it contains **dense noun phrases** with many modifiers (Densidade de Padrões Sintáticos: max_noun_phrase, min_noun_phrase, mean_noun_phrase)
2. it places a **high number of words before the main verb** (i.e., "**propôs**") **of the main clause**, thus taxing the reader's working memory that creates demands on an individuals working memory (Complexidade Sintática: words_before_main_verb) (17)

Ex: A partir da determinação do início de transcrição de seis genes altamente induzidos no choque térmico, **propôs**-se um consenso para promotores dependentes do fator sigma alternativo que controla a resposta ao choque térmico, sigma32.

3. it requires the reader to keep track of **many combinations of meaning with logic-based words** such as "and," "or," and not (Conectivos: and_ratio, negation_ratio, or_ratio, logic_operators)
4. high frequency of **passive voice**, which is more difficult to process than active voice and **of embedded syntax** (Complexidade Sintática: passive_ratio , relative_clauses)

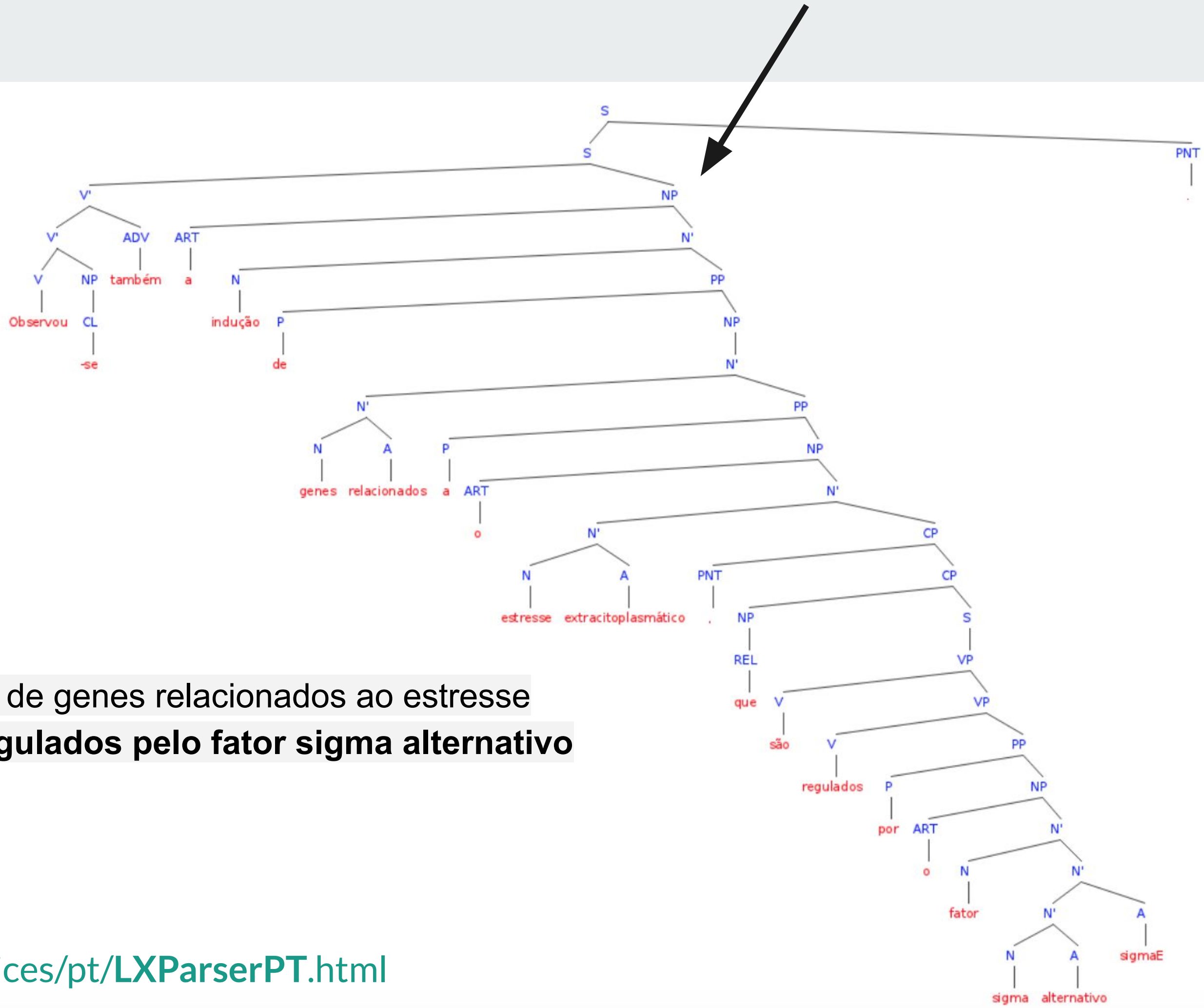


Max_noun_phrase = 25

Dentre os genes altamente induzidos, destacam-se os **que codificam proteínas de choque térmico (Hsps),** que previnem a desnaturação e a formação de agregados protéicos ou degradam polipeptídeos irreversivelmente desnaturados.

http://lxcenter.di.fc.ul.pt/services/pt/**LXParserPT**.html

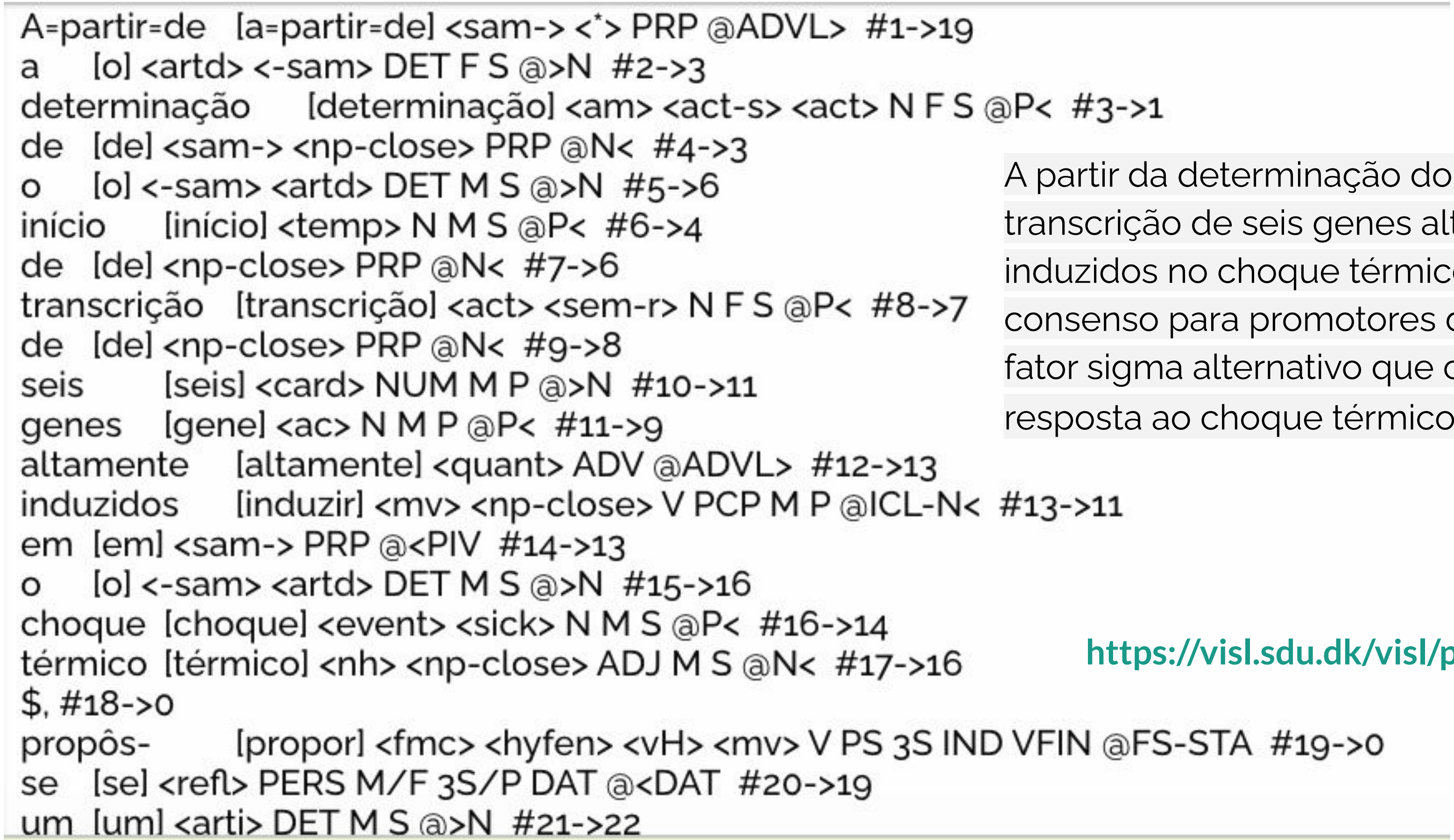
Max_noun_phrase = 16



Observou-se também a indução de genes relacionados ao estresse extracitoplasmático, **que são regulados pelo fator sigma alternativo sigmaE**.

http://lxcenter.di.fc.ul.pt/services/pt/**LXParserPT**.html

Parser Palavra - Flat output



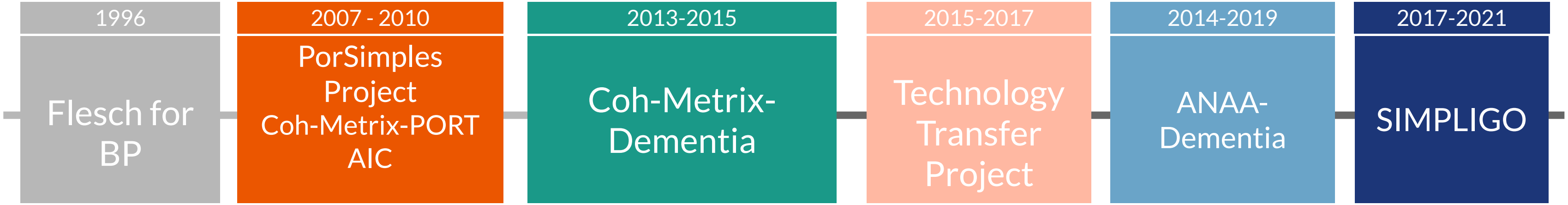
```
A=partir=de   [a=partir=de] <sam-> <*> PRP @ADVL>  #1->19
a     [o] <artd> <-sam> DET F S @>N  #2->3
determinação     [determinação] <am> <act-s> <act> N F S @P<  #3->1
de   [de] <sam-> <np-close> PRP @N<  #4->3
o     [o] <-sam> <artd> DET M S @>N  #5->6
início      [início] <temp> N M S @P<  #6->4
de   [de] <np-close> PRP @N<  #7->6
transcrição   [transcrição] <act> <sem-r> N F S @P<  #8->7
de   [de] <np-close> PRP @N<  #9->8
seis        [seis] <card> NUM M P @>N  #10->11
genes     [gene] <ac> N M P @P<  #11->9
altamente     [altamente] <quant> ADV @ADVL>  #12->13
induzidos     [induzir] <mv> <np-close> V PCP M P @ICL-N<  #13->11
em  [em] <sam-> PRP @<PIV  #14->13
o     [o] <-sam> <artd> DET M S @>N  #15->16
choque  [choque] <event> <sick> N M S @P<  #16->14
térmico  [térmico] <nh> <np-close> ADJ M S @N<  #17->16
$, #18->0
propôs-        [propor] <fmc> <hyfen> <vH> <mv> V PS 3S IND VFIN @FS-STA  #19->0
se   [se] <refl> PERS M/F 3S/P DAT @<DAT  #20->19
um  [um] <arti> DET M S @>N  #21->22
```

A partir da determinação do início de transcrição de seis genes altamente induzidos no choque térmico, **propôs-se** um consenso para promotores dependentes do fator sigma alternativo que controla a resposta ao choque térmico, sigma32.

https://visl.sdu.dk/visl/pt/parsing/automatic/

Motivation and Theoretical Foundations

Timeline



1996
Flesch for BP

2007 - 2010
PorSimples Project
Coh-Metrix-PORT
AIC

2013-2015
Coh-Metrix-Dementia

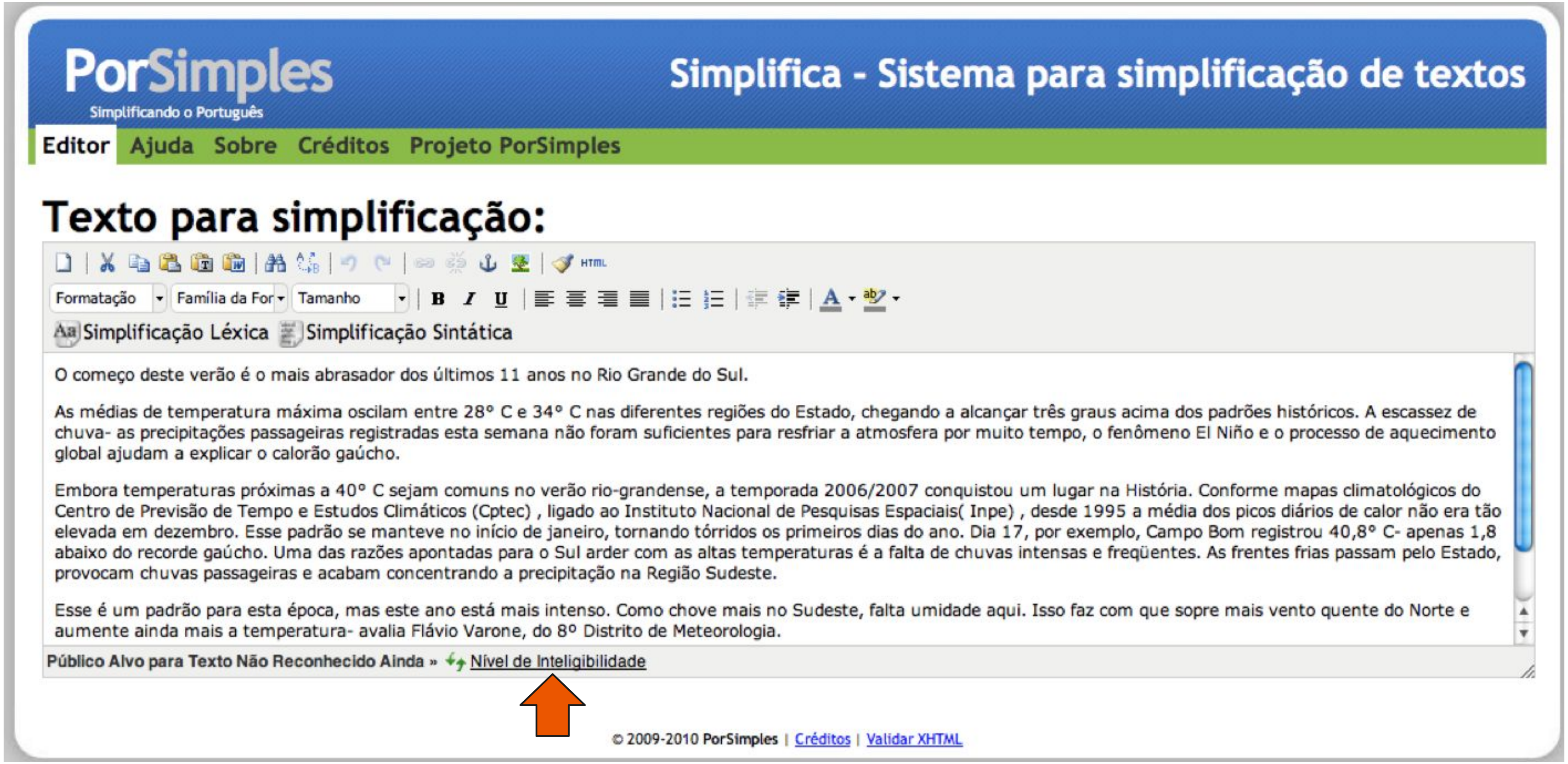
2015-2017
Technology Transfer Project

2014-2019
ANAA-Dementia

2017-2021
SIMPLIGO

Simplification of Portuguese Texts for Digital Inclusion and Accessibility (2007 - 2010)

Initial Motivation: PorSimples project (2007-2010) -- INAF (prof/básico/rudimentar)



Texto para simplificação:

O começo deste verão é o mais abrasador dos últimos 11 anos no Rio Grande do Sul.

As médias de temperatura máxima oscilam entre 28° C e 34° C nas diferentes regiões do Estado, chegando a alcançar três graus acima dos padrões históricos. A escassez de chuva- as precipitações passageiras registradas esta semana não foram suficientes para resfriar a atmosfera por muito tempo, o fenômeno El Niño e o processo de aquecimento global ajudam a explicar o calorão gaúcho.

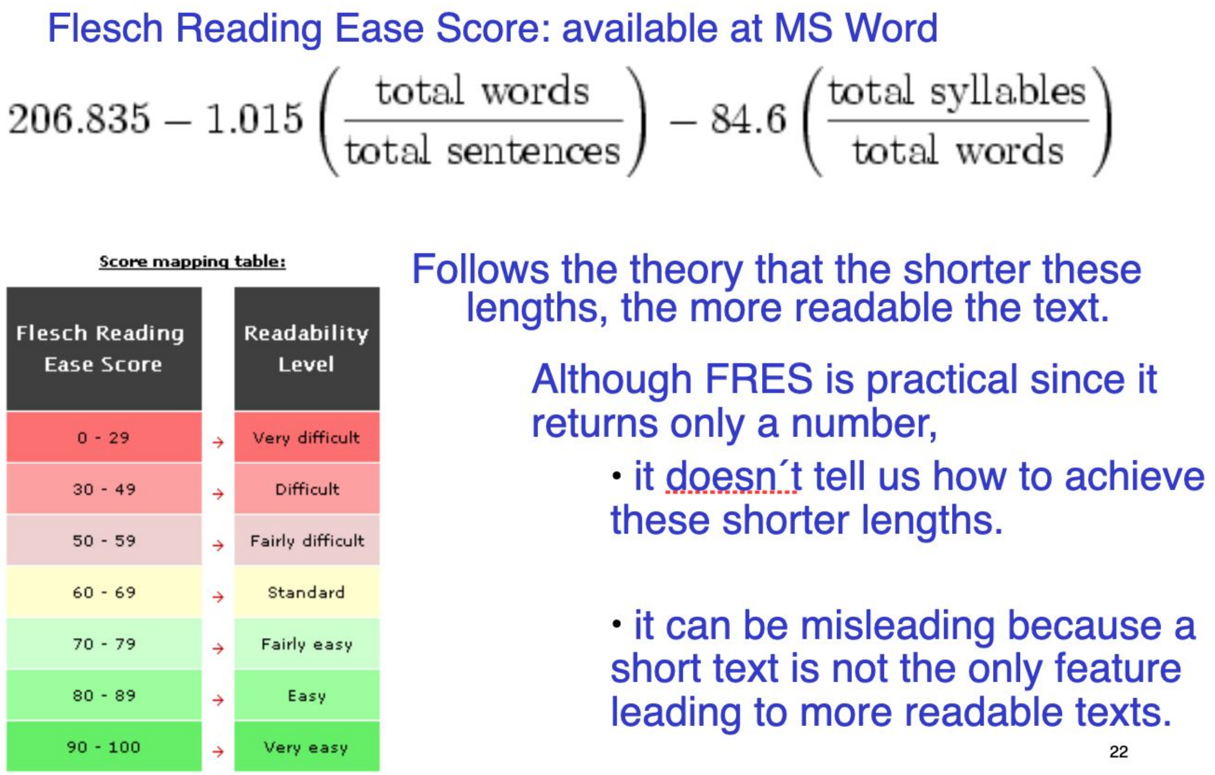
Embora temperaturas próximas a 40° C sejam comuns no verão rio-grandense, a temporada 2006/2007 conquistou um lugar na História. Conforme mapas climatológicos do Centro de Previsão de Tempo e Estudos Climáticos (Cptec) , ligado ao Instituto Nacional de Pesquisas Espaciais(Inpe) , desde 1995 a média dos picos diários de calor não era tão elevada em dezembro. Esse padrão se manteve no início de janeiro, tornando tórridos os primeiros dias do ano. Dia 17, por exemplo, Campo Bom registrou 40,8° C- apenas 1,8 abaixo do recorde gaúcho. Uma das razões apontadas para o Sul arder com as altas temperaturas é a falta de chuvas intensas e freqüentes. As frentes frias passam pelo Estado, provocam chuvas passageiras e acabam concentrando a precipitação na Região Sudeste.

Esse é um padrão para esta época, mas este ano está mais intenso. Como chove mais no Sudeste, falta umidade aqui. Isso faz com que sopre mais vento quente do Norte e aumente ainda mais a temperatura- avalia Flávio Varone, do 8º Distrito de Meteorologia.

Público Alvo para Texto Não Reconhecido Ainda » Nível de Inteligibilidade

© 2009-2010 PorSimples | Créditos | Validar XHTML

1996 - Adaptation of the Flesch Index to PB at NILC



Flesch Reading Ease Score: available at MS Word

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Score mapping table:

Flesch Reading Ease Score		Readability Level
0 - 29	→	Very difficult
30 - 49	→	Difficult
50 - 59	→	Fairly difficult
60 - 69	→	Standard
70 - 79	→	Fairly easy
80 - 89	→	Easy
90 - 100	→	Very easy

Follows the theory that the shorter these lengths, the more readable the text.

Although FRES is practical since it returns only a number,

- it doesn´t tell us how to achieve these shorter lengths.
- it can be misleading because a short text is not the only feature leading to more readable texts.

Traditional readability formulas ....

were not made to explain the reason for the difficulty of a text, as they are not based on theories of text understanding.

In PorSimples, we chose the Coh-Metrix project as a foundation for the metrics to be developed:

- Coh-Metrix-Port, 48 metrics grouped in 10 classes
- AIC tool, 39 metrics, grouped in 5 classes

Coh-Metrix

Computes metrics based on models of textual understanding and cognitive models of reading that explain:

(i) how a reader interacts with a text,

(ii) what types of memories are involved in reading, e.g., how the overload of working memory caused by using too many words before the main verb negatively influences the processing of sentences,

(iii) the role of the propositional content of the speech which means that if the coherence of a text is improved, so will its comprehension, and

(iv) how the mechanisms of cohesion, for example, **discourse markers** and **repetition of entities**, will help to create a coherent text.

In summary, Coh-Metrix tool uses a framework of multilevel analysis.

Coh-Metrix-Port: 48 metrics

1. **Basic Counts** contains 14 metrics related to basic statistics
2. **Logic operators** contains 5 metrics related to the counting of logical operators AND, OR, IF, Negation;
3. **Content word frequencies** contains 2 metrics that use the largest lexicon that existed at the beginning of PorSimples (Banco do Português)
4. **Hypernyms and Ambiguity** bring a metric that calculates the average number of hypernyms per verbs and 4 metrics that calculate the impact of the number of senses for content words
5. **Tokens** groups 3 metrics of lexical richness and level of formality
6. **Constituents** deal with 3 metrics related to the workload in working memory
7. **Connectives** brings 9 metrics related to discursive markers
8. **Coreferences and Anaphoras** bring 7 metrics that address referential cohesion

AIC Tool: 39 metrics mainly from Palavras parser

1. **Basic Counts** contains 6 metrics related to basic statistics
2. **Syntactic Information** brings 13 metrics about clause information in sentences, mainly extracted from the parser Palavras
3. **Density of Syntactic and Morphosyntactic Categories**, extracted using the parser Palavras, contains 8 metrics
4. **Personalisation** contains 10 metrics related to the number of personal and possessive pronouns and their division by person and number
5. **Discourse Markers** contains two metrics related to discursive markers: number of discursive markers and number of ambiguous discursive markers in the text.

Coh-Metrix-Dementia (73 metrics) (2013-2015)

During the implementation of Coh-Metrix-Dementia, the first re-implementation of Coh-Metrix-Port was done to standardise interfaces and the use of NLP tools.

25 new metrics, for measuring:

- **lexical diversity** (Brunet and Honoré indexes)
- **syntactic complexity** (Yngve and Frasier, Mean Clauses per Utterance)
- **idea density** (number of propositions of a text, divided by its number of words)**, cross entropy and**
- **text cohesion through latent semantics analysis** (LSA)**.**

Psycholinguistic Metrics

In 2017, during a NILC student's PhD:

a large lexical base with 26,874 words in BP was automatically annotated with concreteness, age of acquisition, imageability and subjective frequency (similar to familiarity), enabling the implementation of 24 psycholinguistic metrics.

- **word imageability** is the ease and speed with which a word evokes a mental image;
- **concreteness** is the degree to which words refer to objects, people, places, or things that can be experienced by the senses;
- **subjective frequency (familiarity)** is the estimation of the number of times a word is encountered by individuals in its written or spoken form, and
- **AoA** is the estimation of the age at which a word was learned

Technology transfer project (2015-2017)

72 new metrics, many of them related to **lexical and syntactic simplicity**, were added to the already extensive set of metrics built by NILC.

2016 - 2017: Linguistic Revision & Documentation

Some metrics were rewritten, others discarded, several others had their NLP resources updated and documented.

This documentation is available on the project's website.

Simpligo project (2021)

- 10 metrics based on semantic cohesion, via Latent Semantic Analysis (LSA) trained on the large corpus BrWac with 2.68 billion words,
- 8 lexical frequency metrics from large corpora (BrWac and Corpus Brasileiro, a billion word corpus), now normalised.

Code will be publicly available for download --- https://github.com/nilc-nlp/nilcmetrix --- with an AGPLv3 license.

Presentation of the Metrics

1. Descriptive Index (10 metrics)

- Metrics that describe basic text statistics: use only a tokeniser and sentence segmentation
  - length of words, sentences and paragraphs **correlates with** the effort required to read a text
  - standard deviation of words per sentence and maximum and minimum number of words per sentence, indicate **how homogeneous a text is** under this parameter
  - large standard deviation is suggestive of large variations in terms of the number of words per sentence

number of words in the text	mean number of words per sentence
number of paragraphs in the text	maximum number of words per sentence
number of sentences in the text	minimum number of words per sentence
mean number of sentences per paragraph	standard deviation of number of words per sentence
mean number of syllables per content word	proportion of subtitles in relation to the number of sentences in the text.

2. Text Easability Metrics (8 metrics)

- Proportion of short, medium, long and very long **sentences** in relation to all sentences in the text.
- Proportion of easy and difficult **conjunctions** to total words.
- Proportion of **first-person personal** pronouns in relation to all personal pronouns in the texts. **First-person personal pronouns indicate proximity to the reader.**
- Proportion of **simple content words** to all content words in the text. Content words (nouns, verbs, adjectives and adverbs) constitute the variable vocabulary a reader has to know to understand the text. The greater the proportion, the simpler the text.

3. Referential Cohesion (9 metrics)

Capture the presence of elements necessary to construct coreference chains. Coreference occurs when a noun, pronoun, or NP refers to another constituent in the text.

These metrics calculate the overlap of content words in adjacent sentences (**local cohesion**) and among all sentences of the text (**global cohesion**).

> The **longer the text**, the **greater the need of coreference chains** to help the reader to make connections between parts of the text, rendering the text easier to understand

4. LSA-Semantic Cohesion (11 metrics)

The metrics that calculate semantic cohesion are grounded in Latent Semantic Analysis (LSA) --- http://lsa.colorado.edu/ --- which considers the overlap of semantically related words.

Six of them calculate the mean and the standard deviation of **semantic overlap between**:

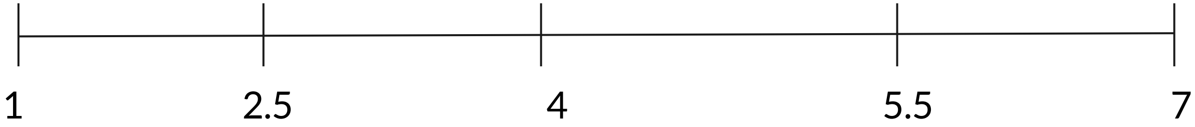
adjacent sentences, adjacent paragraphs and all sentence pairs in the text.

Sentence to paragraph: this measures how similar each sentence is to its paragraph.
Paragraph to paragraph: this measures how similar a paragraph is to the other paragraphs in the text
Sentence to text: this measures how similar a sentence is to the text.

Cross entropy measures the "surprise" level of the language model for the sentence. Higher values of cross entropy mean that a sentence has unusual word combinations: an indication of complexity.

5. Psycholinguistic Measures (24 metrics)

NILC-Metrix brings six indices for each of the following psycholinguistic measures: age of acquisition, concreteness, familiarity and imageability, totalling 24 metrics.



These measures are related to text easability:

the **lower** the words' age of acquisition, the **easier** the text, and the **higher** the words' concreteness, familiarity and imageability, the **easier the text**.

6. Lexical Diversity (15 metrics)

Lexical diversity is a measure obtained through the **type-token ratio (TTR)**, that is, the number of types (all words, disregarding repetitions) divided by the number of tokens (all words, considering repetitions).

Lexical diversity is inversely proportional to cohesion: the lower the lexical diversity, the higher the cohesion.

NILC-Metrix includes TTR for: all words, content words, function words, nouns, verbs, adjectives, pronouns, indefinite pronouns, relative pronouns, prepositions and punctuation.

The detailed metrics are intended to investigate where the difficulty of the text lies.

7. Connectives (12 metrics)

Connectives are words that help the reader to establish cohesive links between parts of the text.

Proportion of four different types of connectives: **additive, causal, logical and temporal.**

Temporal connectives, however, are within the temporal lexicon category.

For each type, there is a distinct metric specifying the positive and negative ones.

positive connectives (also, moreover) & negative connectives (however, but)

The most frequent connectives, “e” (and), “ou” (or) and “se” (if) are focused on specific metrics.

8. Temporal Lexicon (12 metrics)

The indices detail the relative occurrences of each **verb tense and mood** in relation to the total verb tenses and moods in the text.

Temporal connectives, positives and negatives, are also included in this category.

The temporal lexicon is the first step towards enabling the construction of **temporal cohesion metrics.**

9. Syntactic Complexity (27 metrics)

Metrics use both dependency and constituency parsers.

Measures the demand on working memory: **the number of words before the main verb.**

Using data from **the constituency parser LX-parser**: two syntactic complexity indexes: Yngve and Frazier.
Using data from the the dependency tree of **Maltparser**: distance in the dependence tree.

Various proportion measures involving clauses, enabling an in-depth investigation on **where the complexity of a text lies**:

clauses with postponed subject	relative clauses
clauses in non-canonical order (SVO)	adverbial clauses
clauses in passive voice	infinite verb clauses
subordinate clause	sentences with n clauses (n = 0 ..7+)

10. Syntactic Pattern Density (4 metrics)

In this category, there are four metrics correlated with text processing difficulty:

gerund clauses,

mean number of words per noun phrase,

maximum and

minimum number of words per noun phrase.

11. Morphosyntactic Word Information (42 metrics)

Measures of content and functional word densities, in the text and per sentence, as well as a series of breakdowns of these densities: **useful to investigate in detail where the difficulty of a text lies:**

adjectives,

adverbs,

verbs (inflected and non-inflected),

nouns,

prepositions,

pronouns (detailed by type and inflection).

12. Semantic Word Information (11 metrics)

Two of them use Brazilian Portuguese LIWC 2007 Dictionary to calculate the proportion of words with negative/positive polarity in relation to all words in the text.

Five measures of ambiguity (of content words, and in detail by nouns, adjectives, verbs and adverbs) are calculated according to their respective number of senses in TeP (Portuguese Electronic Thesaurus).

The average amount of hypernyms per verb in sentences uses information extracted from Wordnet.Br.

Three metrics relating to the proportion of abstract nouns and proper nouns in sentences and in the text (named_entity_ratio_sentence, named_entity_ratio_text, abstract_nouns_ratio)

The greater the number of named entities, the greater the memory load required and, therefore, the greater the textual complexity.

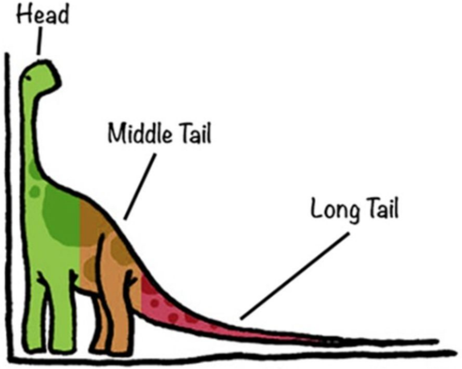
13. Word Frequency (10 metrics)

Average frequency of content words and rare content words: not normalised, fpm and logarithmic scale (Zipf scale).

Corpus do Português, which was the largest corpus at that time, with **700 thousand words**: **two** oldest present frequencies (**not normalised**) of **all content words** and of the **rarer words** in the text.

Four frequency measures were extracted from **Corpus Brasileiro (1 billion words)** , which has around one billion tokens

Four from **BrWaC**, which has around **2.68 billion tokens**



https://i.workana.com/

14. Readability Formulas (5 metrics)

The Brunet readability index a kind of type/token ratio that is less sensitive to the text length.

The **Dale Chall adapted formula** --- combines the percentage of unfamiliar words with the average number of words per sentence. Unfamiliar words are those not included in the **Dictionary of Simple Words** (Biderman, 2006).

The **Flesch readability index adapted for BP**: 248.835 - [1.015 x (average words per sentence)] - [84.6 x (average syllables per word)].

Honore's Statistics, a type/token ratio that takes into account, besides the number of types and tokens, the number of hapax legomena.

Gunning's Fog index adds the average sentence length to the percentage of difficult words (2 ou + syllables) and multiplies this by 0.4.

NILC-Metrix: Applications

Natural Language Processing (NLP)
Neuropsychological Language Tests
Education
Language Studies
Translation Studies

NLP: Detection of Fake News (LREC 2020)

Measuring the Impact of Readability Features in Fake News Detection

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, Carolina Scarton

- 17 metrics from 4 categories (Readability Formulas, Referential Cohesion, Text Easability Metrics and Psycholinguistics), named as **readability features** by the authors.

- Corpus: open access and balanced corpus called Fake.Br corpus, with aligned texts totalling 3,600 false and 3,600 true news.

- Classifier model: SVM, with the standard parameters of Scikit-learn

The results of their study showed that readability features were relevant for detecting fake news in BP, achieving, alone, up to 92% classification accuracy

Tests: Dementia Diagnosis (Propor 2016)

Evaluating Progression of Alzheimer's Disease by Regression and Classification Methods in a Narrative Language Test in Portuguese

Sandra Aluísio, Andre Cunha, Carolina Scarton

- Evaluated methods to **identify linguistic features for dementia diagnosis**, focusing on Alzheimer Disease (AD) and Mild Cognitive Impairment (MCI), to distinguish them from Control Patients (CT).
- Narrative language test was used based on sequenced pictures (Cinderella story --- 20 AD, 20 MCI and 20 CT) corpus
- 73 features were extracted from the resulting **transcriptions**, using the Coh-Metrix-Dementia tool (part of its metrics is included in NILC-METRIX)
- **Results:** **0.82 F1-score** in the experiment with three classes (AD, MCI and CT), and **0.90 for two classes CT versus (MCI+AD)**, both using the CFS-selected features method.

Education: Text Complexity in Open Educational Resources (STIL 2019)

Predição da complexidade textual de recursos educacionais abertos em português

Murilo Gazzola, Sidney Leal, Sandra Aluísio

- Investigated the impact of textual genre in assessing text complexity in BP educational resources (**MEC-RED - https://plataformaintegrada.mec.gov.br/**)
- A corpus with **2076** extracts from textbooks for Elementary School I, Elementary School II, Secondary School and Higher Education (**https://github.com/gazzola/corpus_readability_nlp_portuguese**)
- was compiled.
- A set of 79 metrics from NILC-Metrix was selected, based on the study by Grasser et al., (2011)
- 5 Machine Learning methods were tested: SVM, MLP, Logistic Regression and Random Forest from scikit learn
- SVM performed better with **0.804 F-Measure**; therefore it was used in an **extrinsic evaluation** with two sets of OER, reaching **0.518 F-Measure in the set with text genres similar** from the training set (textbook corpus) and **0.389 F-Measure for the animation/simulation and practical experiment resources**, which are very common in the MEC-RED platform.

Language Studies: Characterization of Popular News (STIL 2011)

Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero

Maria José Finatto, Carolina Scarton, Amanda Rocha, Sandra Aluísio

- Evaluated (contrastive analysis) the differences in text complexity of popular Brazilian newspapers (public with a lower education) with traditional ones (more educated readers), using cohesion, syntax and vocabulary metrics, including ellipsis.
- 48 metrics from Coh-Metrix-Port and included 5 new ones related to the co-reference of ellipses, based on a corpus annotation.
- ellipses of three types: nominal, verbal and sentential
- balanced corpus of texts --- 80 texts from the traditional Zero Hora newspaper from 2006 and 2007 and 80 texts from the popular Diário Gaucho from 2008
- The most discriminative features: a set of 14 features grouped into 5 classes: Referential Cohesion, Word Frequency, Syntactic Complexity, Descriptive Index, Morphosyntactic Word Information, extracted using Coh-Metrix-Port, but ellipsis did not have a distinctive role.

Translation studies: TI evaluation of sources and translations using readability tools (STIL 2011)

Comparando Avaliações de Inteligibilidade Textual entre Originais e Traduções de Textos Literários

Bianca Franco Pasqualini , Carolina Evaristo Scarton , Maria José B. Finatto

- Textual intelligibility evaluation of a set of selected source texts and translations in Portuguese (PT) and English (EN) using the Coh-Metrix (60) and Coh-Metrix-Port (48) tools.
- Objective: contrast the textual intelligibility of short stories in English language and in Brazilian Portuguese produced between **1830 and 1940** and its translations, in the **English-Portuguese and Portuguese-English direction.**
- Balanced Corpus: 14 short stories in English and 14 in BP
- 31 metrics were directly compared from these classes: Basic Counts, Flesch Index, Constituents, Connectives, Logic operators, Coreferences, Pronouns and TTR, Anaphoras.

Context: recent initiatives of the Ministry of Education of Brazil (MEC) aim to popularize access to classics of national and international literature for new readers

Results:

Lexical metrics: In this regard, when translated into Portuguese, we have texts that require more effort to understand.

Syntactic Metrics: i) Connectives: the results point to a higher cohesion index in PT and, therefore, greater readability; ii) NP, NP modifiers, pronouns by NP and personal pronouns: number of modifiers per NP is inferior in PT in both directions, which indicates greater readability in this item in PT

Semantic Metrics: Overlapping content words (adjacent), word stem overlap and word stem overlap (adjacent): these metrics concern the flow of the text and the maintenance of topics. All texts in EN, both translated and original, have lower indices in these three metrics, which indicates less repetition of words and stem and a possible greater difficulty in reading.

Guidelines for using NILC-Metrix

Corpora studies and NILC-Metrix

Representative and balanced corpus for a research question

- Pre-processing: "garbage in, garbage out"
  - Transcripts: segmentation
  - Movie subtitles
  - Figure/table captions
  - Previous annotations
  - Mathematical symbols
- Use metrics that describe basic text statistics to describe the chosen corpus
- Select a group of Nilc-Metrix, based on literature review to answer the research question

Future Work

What next?

- **Future works depend on the union between areas**

- **NLP Tools:**

  Instead of using **three parsers** (LX-Parser, Malt and Palavras) when implementing syntactic metrics, in the near future we will be able to use robust parsing models for Portuguese, available in the **POeTiSA project**.

- **New metrics:**

- Idea Density is a metric that computes the number of propositions of a text, divided by its number of words; it was implemented in Coh-Metrix-Dementia using a set of rules over dependency parsing by Cunha et al. (2015). Once a robust parsing model is made available, this metric can be implemented in the NILC-Metrix.

- Temporal cohesion (Duran et al, 2007), Causal and Intentional cohesion, responsible for capture the **situational model** (comprehender's mental representation) when a given context is activated, are available in Coh-Metrix and **should be studied and evaluated for BP**.

Thanks !

To all the members of the PorSimples project who provided the basis for building Coh-Metrix-Port and AIC metrics. Also, to all the students who contributed (after PorSimples finished) to enlarging the set of metrics, revising it, applying it in various NLP tasks and, finally to making NILC-Metrix publicly available.

Special thanks to Sidney Leal, Magali Duran, Carol Scarton and Nathan Hartmann

Questions?

References

Duran Nicholas D., McCarthy Philip M, Graesser Art C, McNamara Danielle S (2007) Using temporal cohesion to predict temporal coherence in narrative and expository texts. Behavior Research Methods, Instruments, & Computers 39:212—-223, DOI 10.3758/BF03193150

CUNHA, André ; Sousa, Lucilene. B ; MANSUR, Leticia ; ALUÍSIO, Sandra Maria. Automatic Proposition Extraction from Dependency Trees: Helping Early Prediction of Alzheimer's Disease from Narratives. In: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, 2015, São Carlos. Proceedings of the 2015 IEEE 28th International Symposium on Computer-Based Medical Systems. Washington DC: IEEE Computer Society, 2015. v. 1. p. 127-130.

Pardo TAS, Nunes MGV (2006) Review and evaluation of dizer - an automatic discourse analyzer for Brazilian Portuguese. In: Vieira R, Quaresma P, das Graças Volpe Nunes M, Mamede NJ, Oliveira C, Dias MC (eds) Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings, Springer, Lecture Notes in Computer Science, vol 3960, pp 180–189, DOI 10.1007/11751984\19, URL https://doi.org/10.1007/11751984_19

Bick E (2000) The Parsing System “Palavras”. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Arhus.

Santos LB dos, Duran MS, Hartmann NS, Candido Junior A, Paetzold GH, Aluísio SM (2017) A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. International Conference on Text, Speech, and Dialogue - TSD 2017, Proceedings, Springer, Lecture Notes in Artificial Intelligence, vol 10415, pp 281–28, DOI 10.1007/978-3-319-64206-232

Sardinha APB (2004) Corpus brasileiro. URL http://corpusbrasileiro.pucsp.br/cb/Inicial.html, [Online; accessed 2021.03.21]

Wagner Filho JA, Wilkens R, Idiart M, Villavicencio A (2018) The brWaC corpus: A new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, URL https://www.aclweb.org/anthology/L18-1686

Biderman MTC (2006) Dicionário Ilustrado de Português. Editora Ática, São Paulo

McNamara DS, Graesser AC, McCarthy PM, Cai Z (2014) Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, DOI 10.1017/CBO9780511894664

Landauer TK, Laham D, Rehder B, Schreiner ME (1997) How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In: Shafto MG, Langley P (eds) Proceedings of the 19th annual meeting of the Cognitive Science Society, pp 412–417

Hu X, Cai Z, Louwerse M, Olney A, Penumatsa P, Graesser A (2003) A revised algorithm for latent semantic analysis, Morgan Kaufman Publishers, pp 1489–1491. 18th International Joint Conference of Artificial Intelligence, IJCAI'03 ; Conference date: 09-08-2003 Through 15-08-2003

Dowell, N.M., Graesser, A. C. , Cai , Z. (2016). Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale. In: Journal of Learning Analytics, 3 (3), 72–95. http://dx.doi.org/10.18608/jla.2016.33.5