

## 1. Introduction

**Text readability assessment** measures how easy it is for a reader to understand a text, e.g. for student reading material selection (Petersen and Ostendorf, 2009) and text simplification (Aluisio et al., 2010).

- ▷ Viewed as a **text classification task** (Petersen and Ostendorf (2009), Vajjala and Meurers (2014)).
- ▷ **However**: lack of availability of reliably annotated data, and possibly subjective task.

This work presents a **framework for the automatic construction** of large Web corpora classified by readability level, comparing the performance of different classifiers for the task of readability assessment.

- ▷ Hypothesis (**H1**): the use of deeper (syntactic) attributes contributes to a better classification than just shallow attributes and
- ▷ Hypothesis (**H2**): a classifier trained on a reference annotated corpus is able to capture significant linguistic differences among classes.

Evaluation focuses on **Portuguese and English** corpora.

## 2. Materials and Methods

**Corpora:**

| Language | Corpus     | Classes | Documents | Sentences |
|----------|------------|---------|-----------|-----------|
| PT       | Wikilivros | 3       | 78        | 38,865    |
|          | ESOC       | 2       | 130       | 21,667    |
|          | PSFL       | 2       | 259       | 3,075     |
|          | ZH         | 3       | 279       | 7,127     |
|          | BrEscola   | 2       | 9,083     | 200,132   |
| EN       | Wikibooks  | 4       | 35        | 65,704    |
|          | SW         | 2       | 4,480     | 515,230   |
|          | BB         | 3       | 2,385     | 101,149   |

Wide range of **classifiers** to evaluate any possible algorithm bias in the task:

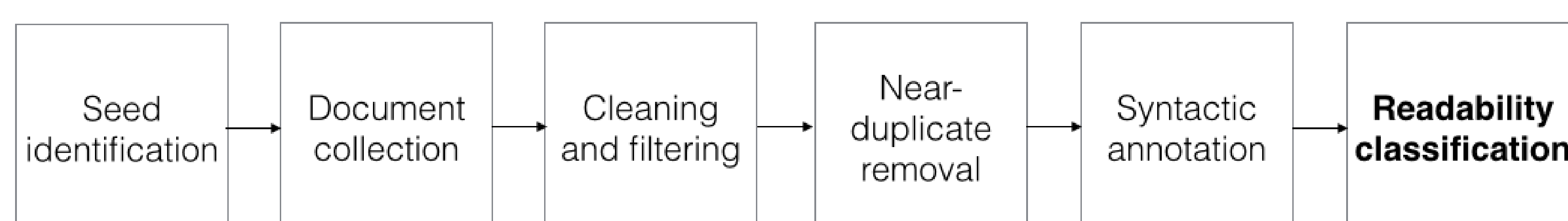
- ▷ **SMO**, **SimpleLogistic**, **DecisionStump** and **RandomForest** from Weka, using 10 fold cross-validation.

**Features:**

- ▷ Varying types of information: **sub-categorization**, **readability formulas**, **text descriptors** and **corpora-based**.
- ▷ Varying depth of processing: **shallow** (counts and lists), **medium** (POS tags) and **deep** (pareses and WordNet information).
- ▷ Total of **134** for Portuguese and **89** for English.

## 3. Web corpus collection framework

- ▷ Web as a Corpus crawling extends Bernardini et al. (2006) with a readability classifier.



- ▷ 6k random pairs of average frequency words used as input to a search engine API → 60k URLs expanded by BFS in two levels → **24 million seeds for Brazilian Portuguese**.
- ▷ Final corpus with **1.56 billion tokens** and **4.15 million types (TTR 0.0026)**.
- ▷ All documents annotated with Palavras parser.

## 4. H1 Feature analysis

- ▷ Evaluation based in the **average features rank** with **information gain**.
  - **Classical formulas** relevant for English but not for Portuguese;
  - **Textual descriptors** informative for both languages;
  - For both languages **shallow features** outperform medium and deep features.

## 5. H2 Model performance analysis

- ▷ The linear **logistic regression** algorithm got best results for both languages.
- ▷ Intermediary classes more challenging.
- ▷ Shallow features informative for classification with low computational cost.
- ▷ In over 62% of the cases combination of shallow features with deep features brought increase in performance.
- ▷ As classifiers have complementary performance, use agreement among them for more generalisable classification.

## 6. Web corpus classification

- ▷ Using SimpleLogistic classifier, trained in two class scenarios with all features.
- ▷ All classifiers resulted in **significant differences** between simple and complex documents for all sets of features of varying depth.
- ▷ The **agreement** among 3 classifiers (12.5% of the corpus as simple and 8.8% as difficult) brought increase in performance.

| Category    | PSFL train | PSFL | ZH   | Wikil. | 3-model agreement |
|-------------|------------|------|------|--------|-------------------|
| Shallow     | 0.15       | 0.15 | 0.08 | 0.11   | 0.27              |
| Medium      | 0.30       | 0.06 | 0.09 | 0.09   | 0.17              |
| Deep        | 0.29       | 0.10 | 0.10 | 0.16   | 0.23              |
| Subcat      | 0.07       | 0.03 | 0.06 | 0.05   | 0.10              |
| Formulas    | 0.19       | 0.06 | 0.09 | 0.08   | 0.20              |
| Descriptors | 0.16       | 0.62 | 0.27 | 0.11   | 0.82              |
| C-based     | 0.09       | 0.04 | 0.01 | 0.12   | 0.09              |

## 7. Conclusions and Future work

- ▷ Hypotheses **H1** and **H2** were validated. **H1**: using deeper (syntactic) features improved performance in most cases. **H2**: significant differences found between classes in automatically readability assessed Web corpus.
- ▷ Using classifier agreement resulted in stricter classification, less prone to over-fitting.
- ▷ Collected **Web corpus available** at <http://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC>
- ▷ Future Work: new analysis, including a manual sample assessment by linguists; approach can be straightforwardly expanded for other languages.