Assessing the Efficiency of Suffix Stripping Approaches for Portuguese Stemming

Wadson G. Ferreira Willian A. Santos Breno M. P. Souza
Tiago M. M. Zaidan Wladmir C. Brandão

Department of Computer Science (DCC)
Pontifical Catholic University of Minas Gerais (PUC Minas)
Belo Horizonte, Brazil

22nd International Symposium on String Processing and Information Retrieval
September 1-4, 2015
London, UK

Problem

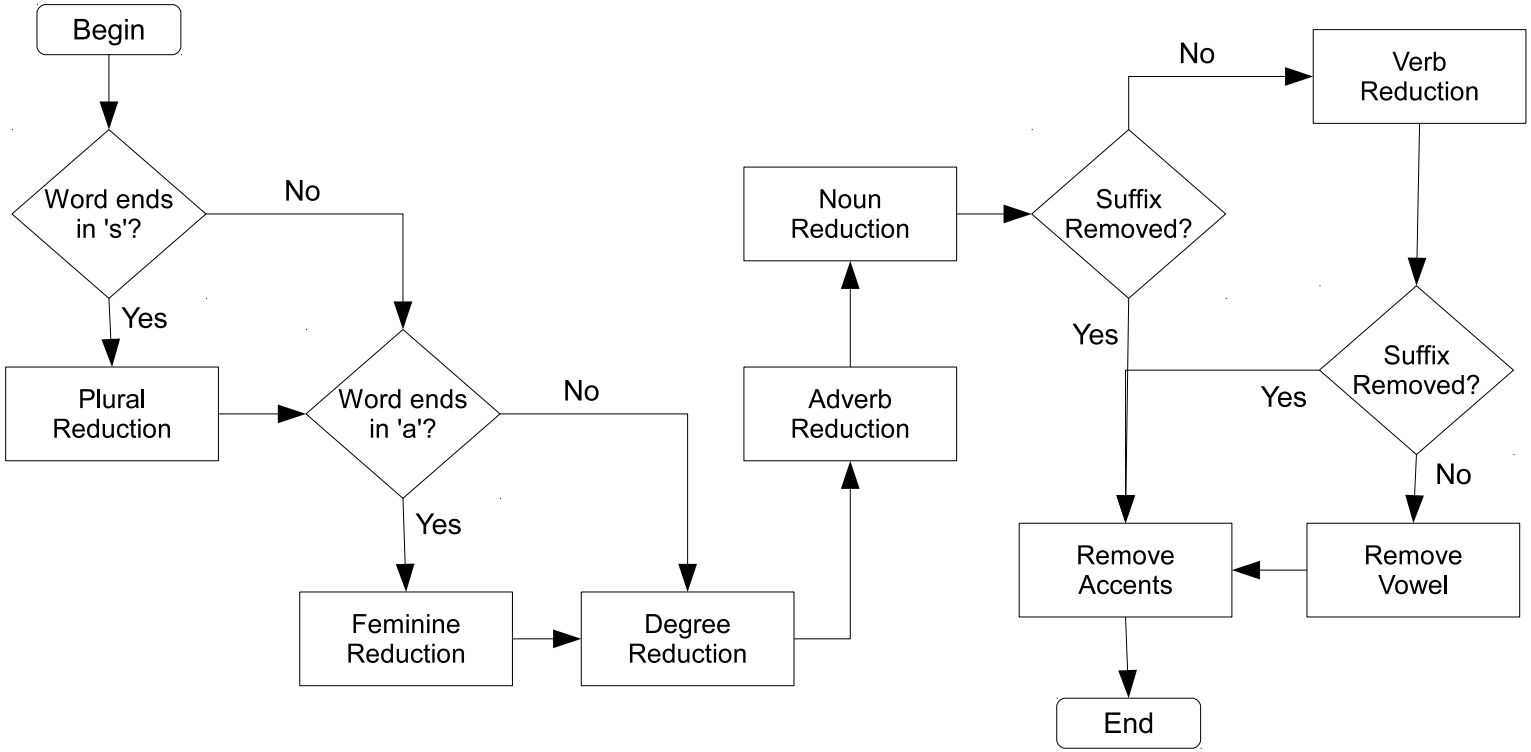
Perform Portuguese Stemming Efficiently

- State-of-the-Art Algorithm $\rightarrow$ RSLP
- RSLP $\rightarrow$ Suffix stripping by processing a list of rules
- Time complexity $\rightarrow O(WR)$
  - $W \rightarrow$ Number of characters to compare (all characters of the word in the worst case)
  - $R \rightarrow$ Number of suffix rules to process
- Objective $\rightarrow$ Improve processing time by reducing R

Stemming Strategies

1. Table Lookup → Lookup of a word in a table to retrieve the stem. Simple, but strongly dependent on the language vocabulary
2. N-grams → Word clustering procedure to identify bigrams and trigrams in text
3. Successor Variaty → Morpheme boundaries recognition within the language organization to produce the stem
4. Affix Removal → Word affixes removal or replacement, following well defined affix rules for a language

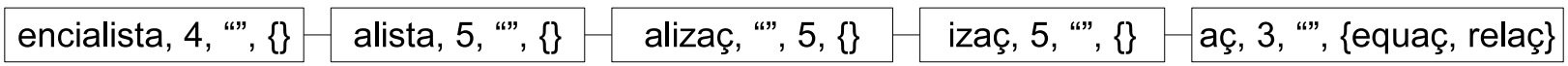
Suffix Stripping Strategy



The List-Based Approach (LBA)

For each step (type of reduction), process a single list of suffix rules to iteratively reduce the word or not

$$Rule(suffix, remainingsize, replacement, exceptions)$$

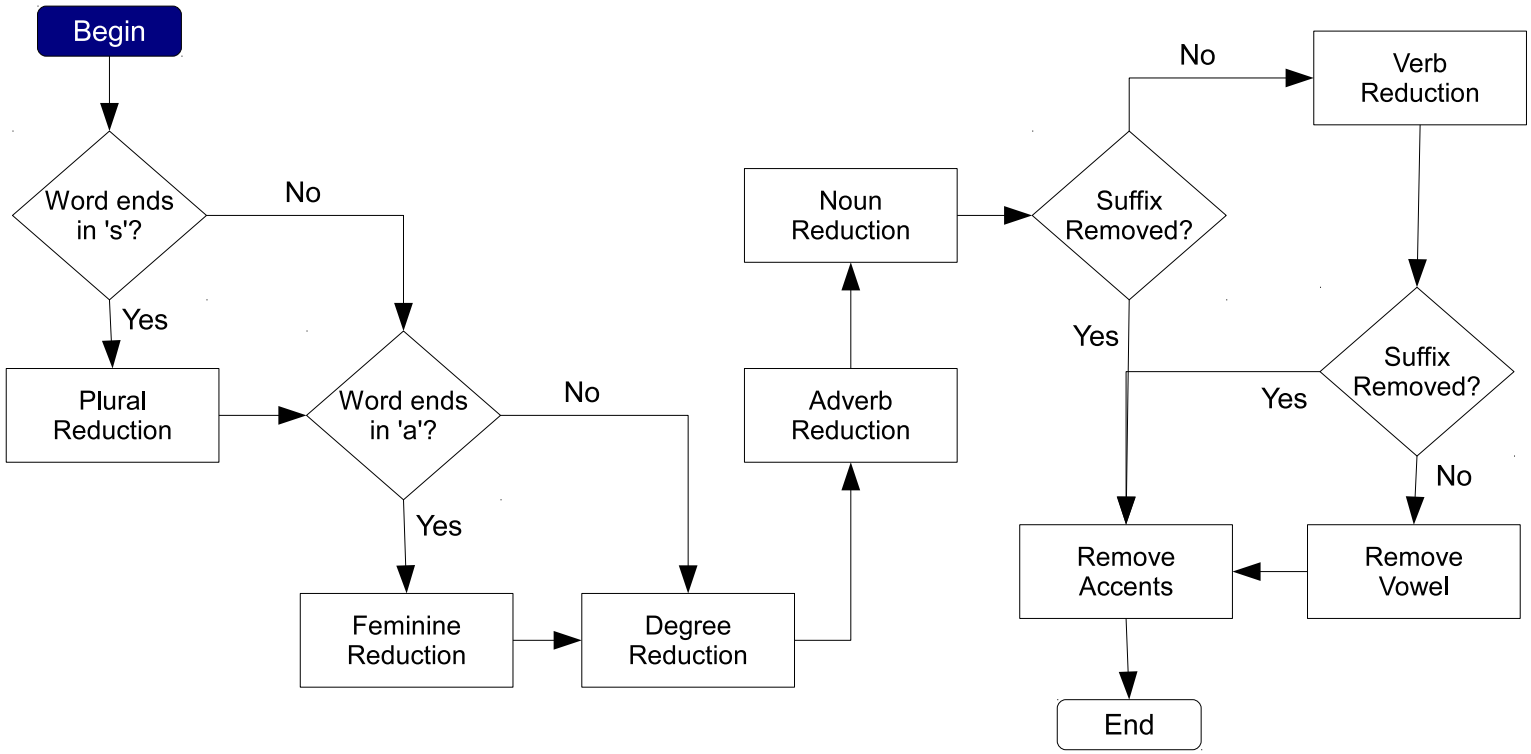


- Worst case → All suffix rules are processed in each step

Plural	Feminine	Degree	Adverb	Noun	Verb
11	15	18	1	61	89

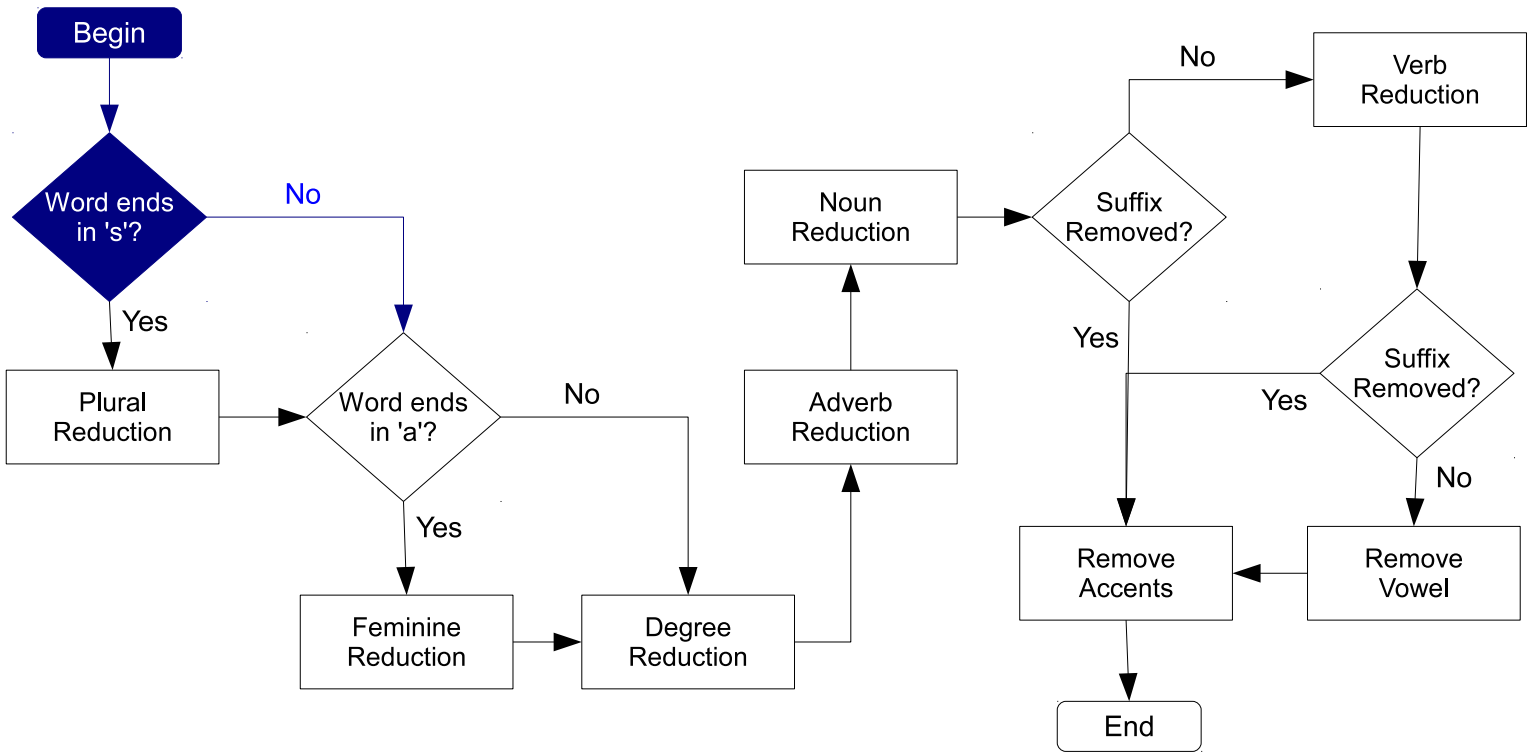
LBA Example

- subutilização (underutilization)



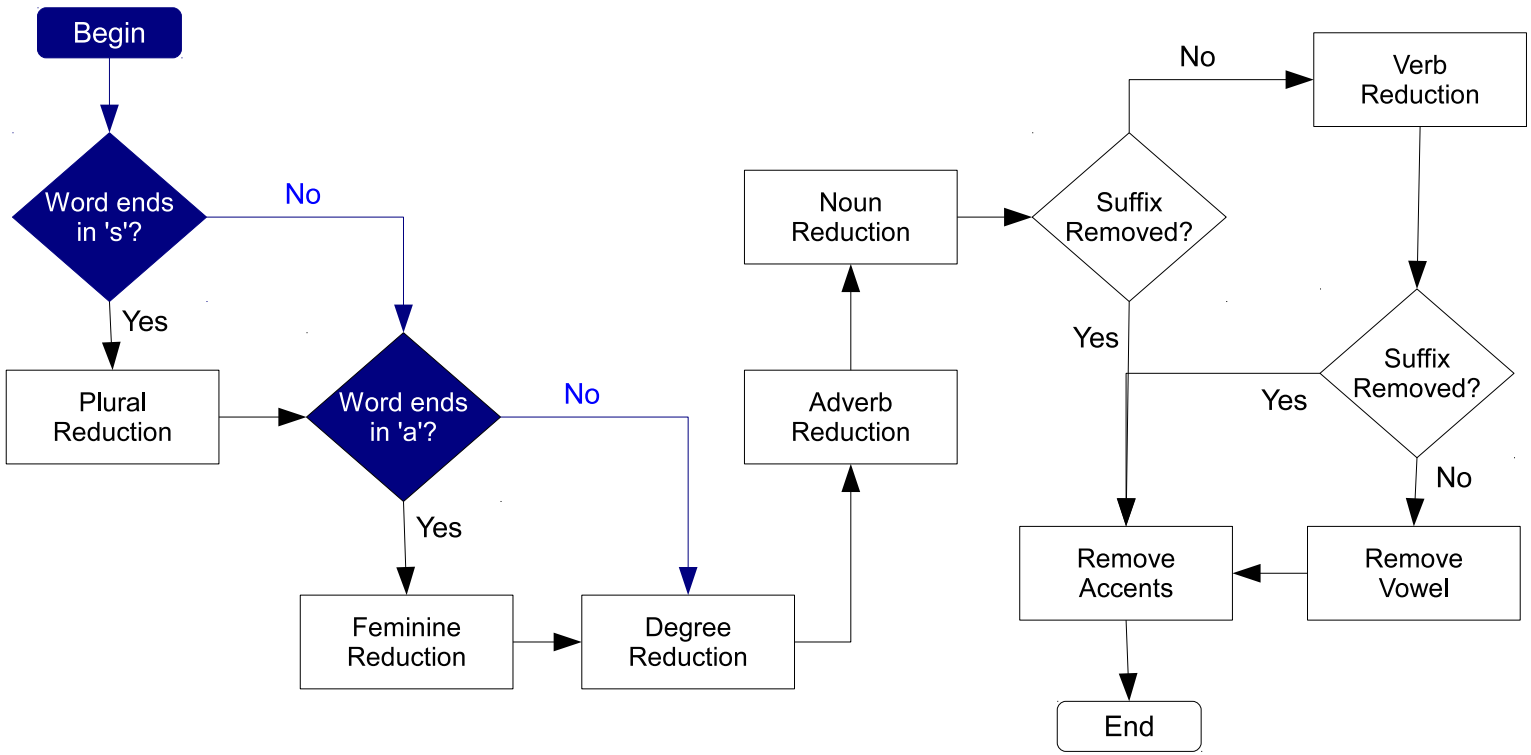
LBA Example

- subutilizaçã**o**



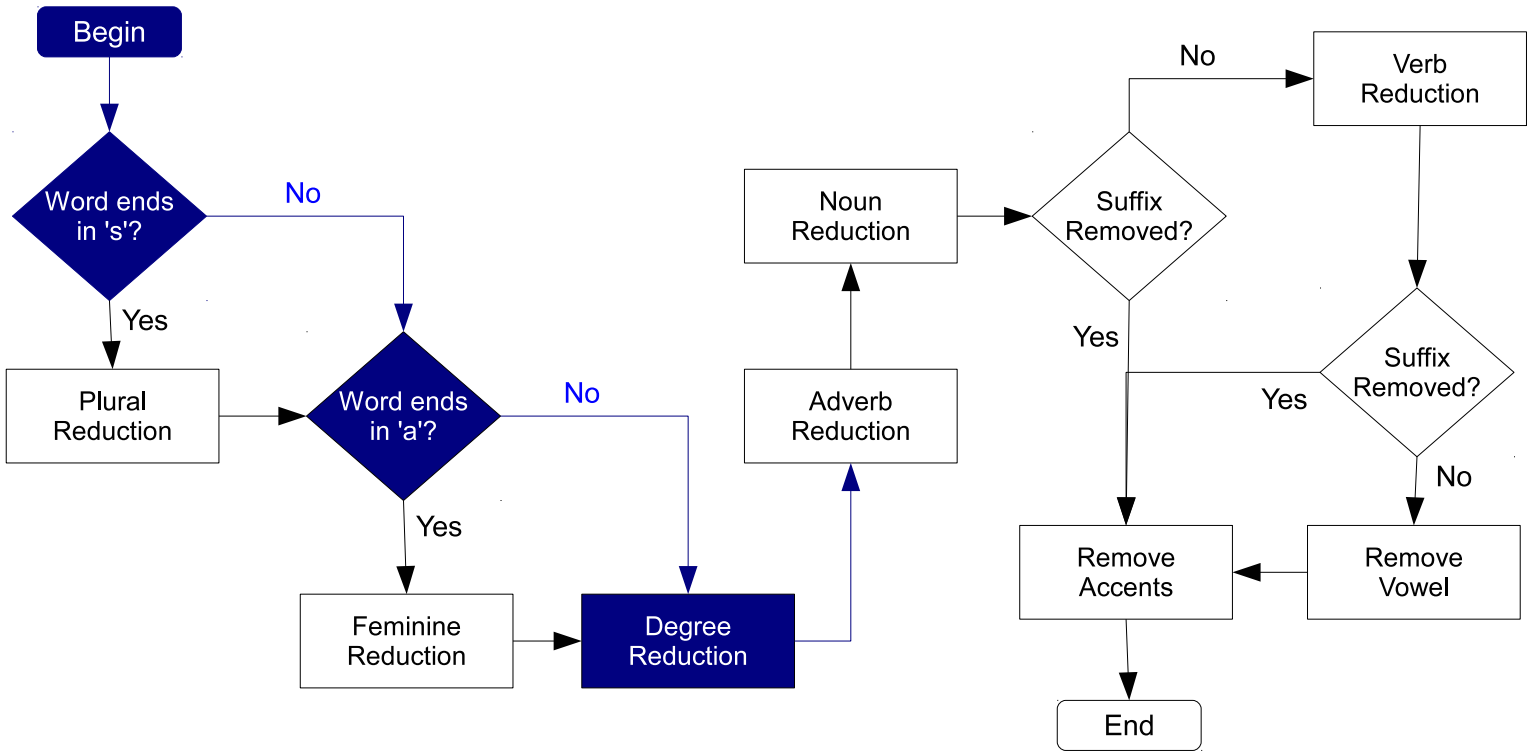
LBA Example

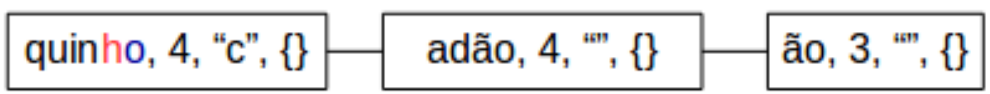
- subutilizaçã**o**



LBA Example

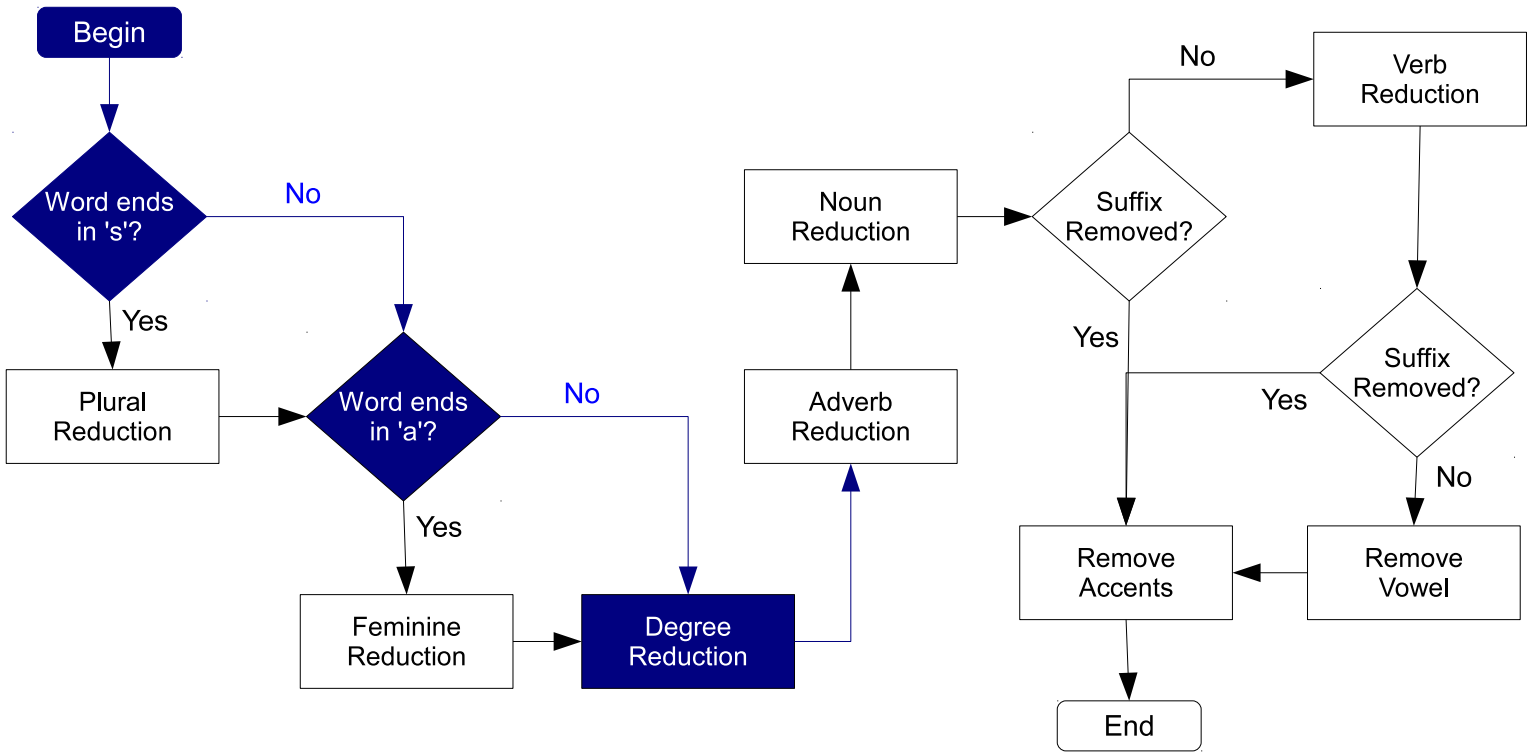
- subutilização

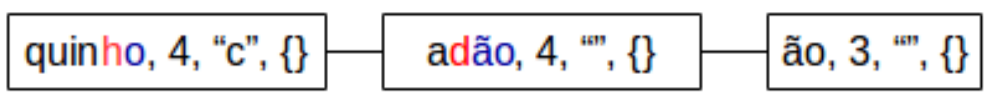




LBA Example

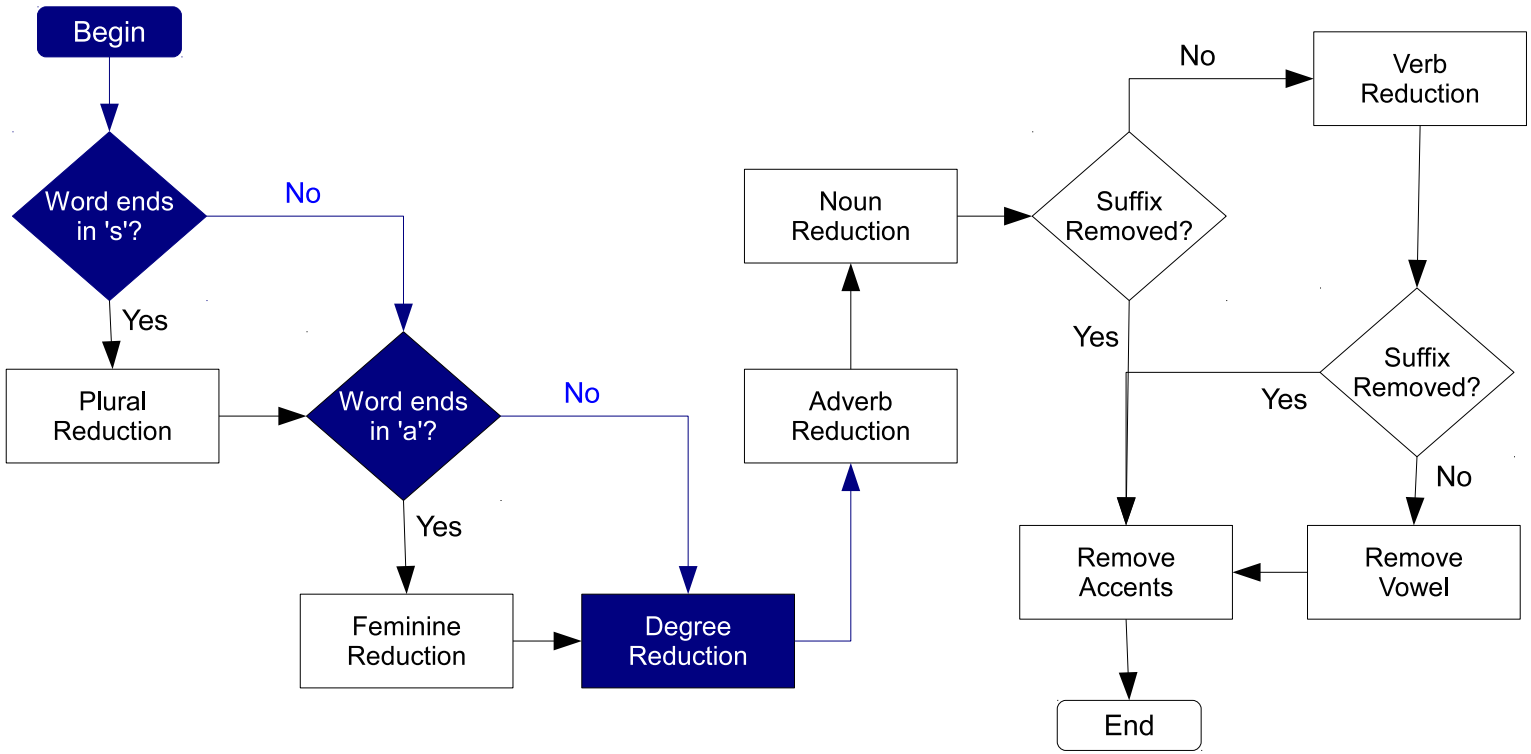
- subutilização

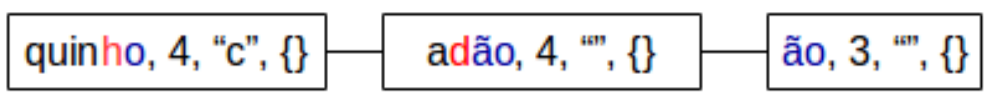




LBA Example

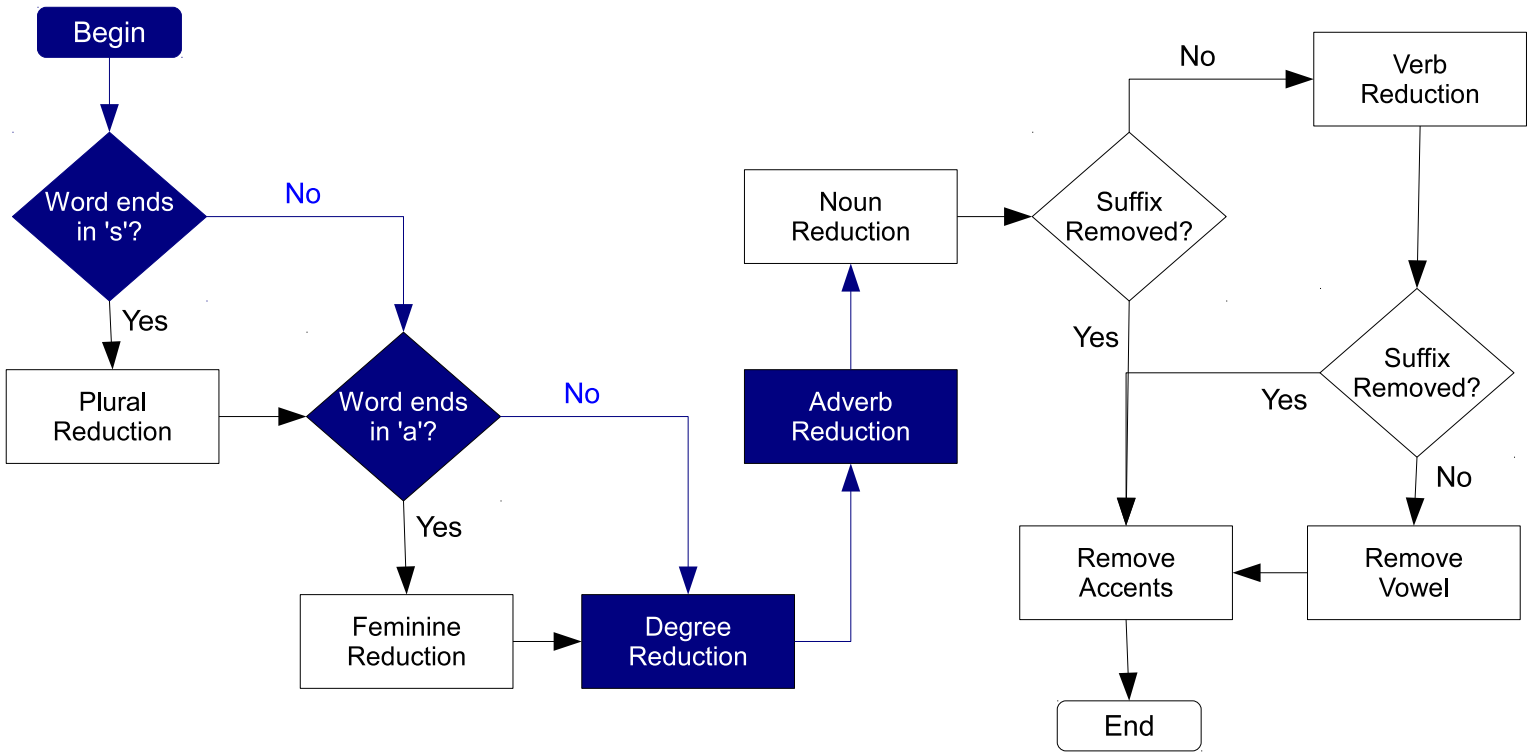
- subutilização → subutilizaç





LBA Example

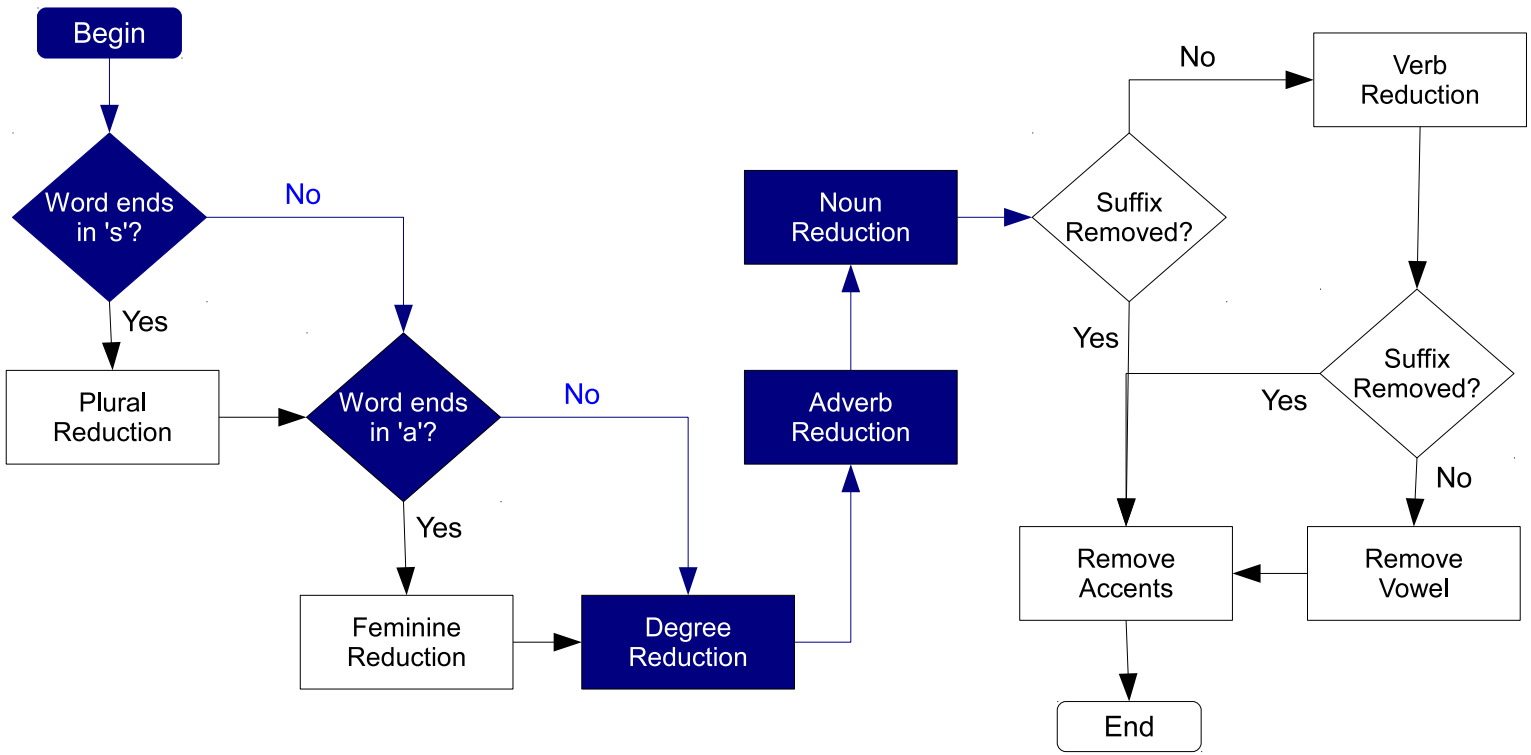
- subutilizaç



mente, 4, "", {}

LBA Example

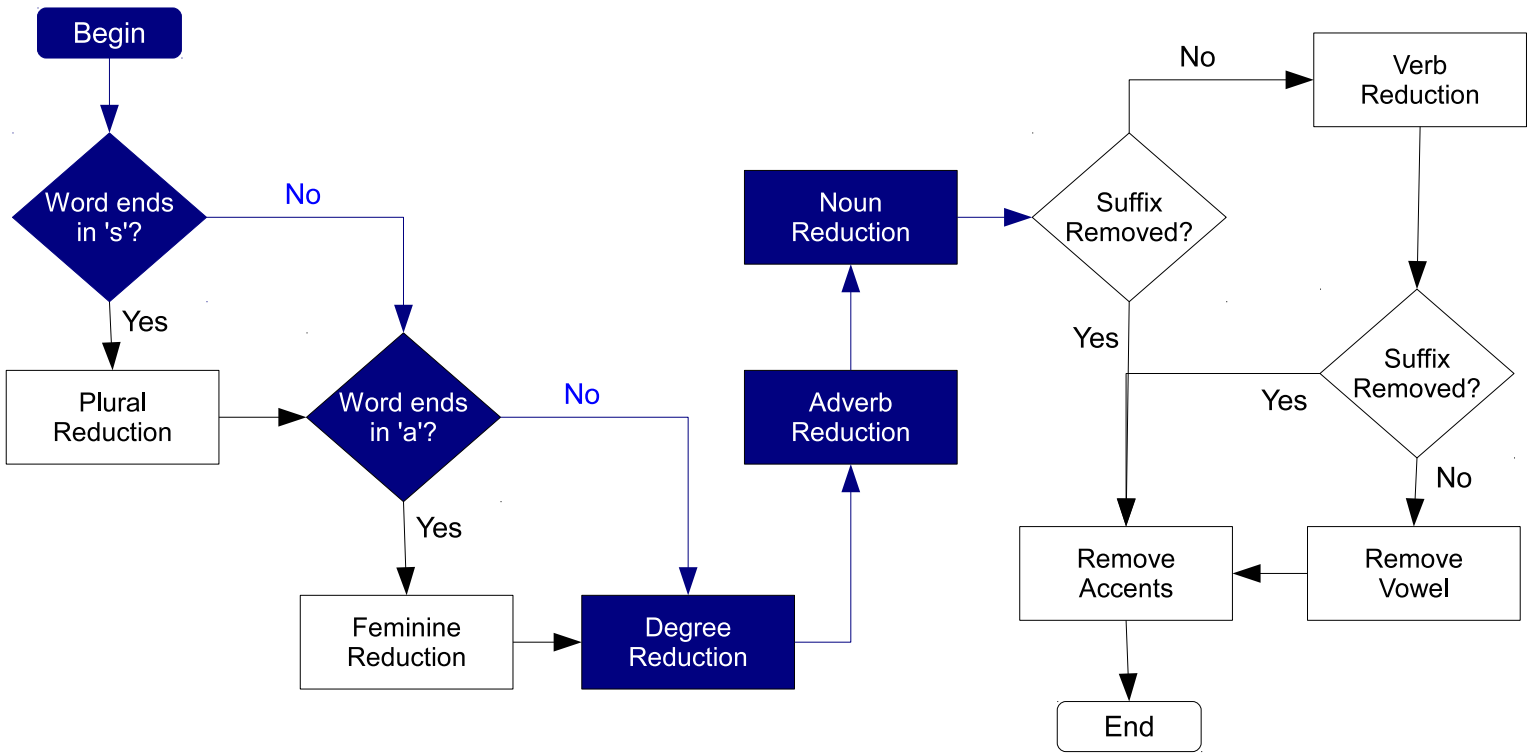
- subutilizaç



encialista, 4, “”, {} — alista, 5, “”, {} — alizaç, “”, 5, {} — izaç, 5, “”, {} — aç, 3, “”, {equaç, relaç}

LBA Example

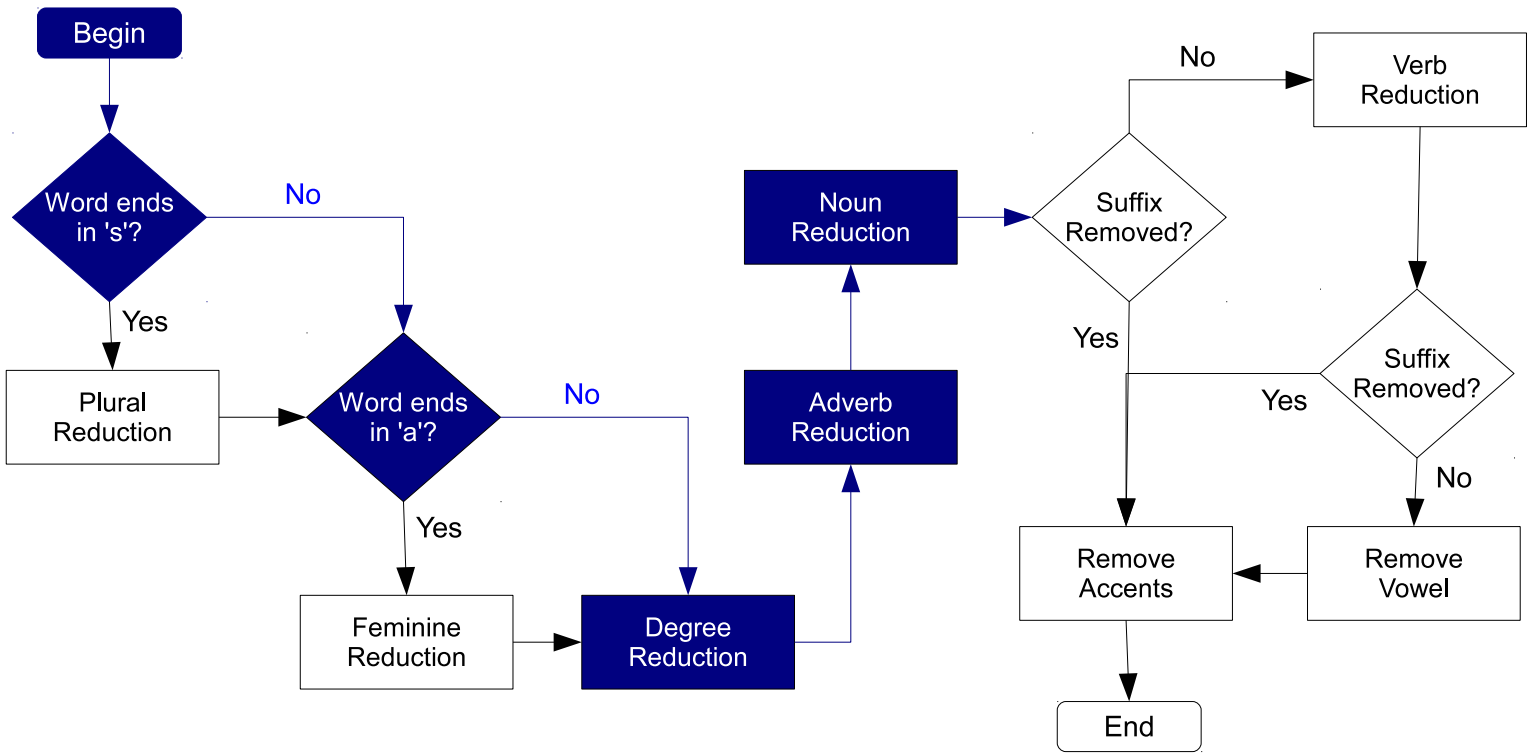
- subutilizaç



encialista, 4, “”, {} — alista, 5, “”, {} — alizaç, “”, 5, {} — izaç, 5, “”, {} — aç, 3, “”, {equaç, relaç}

LBA Example

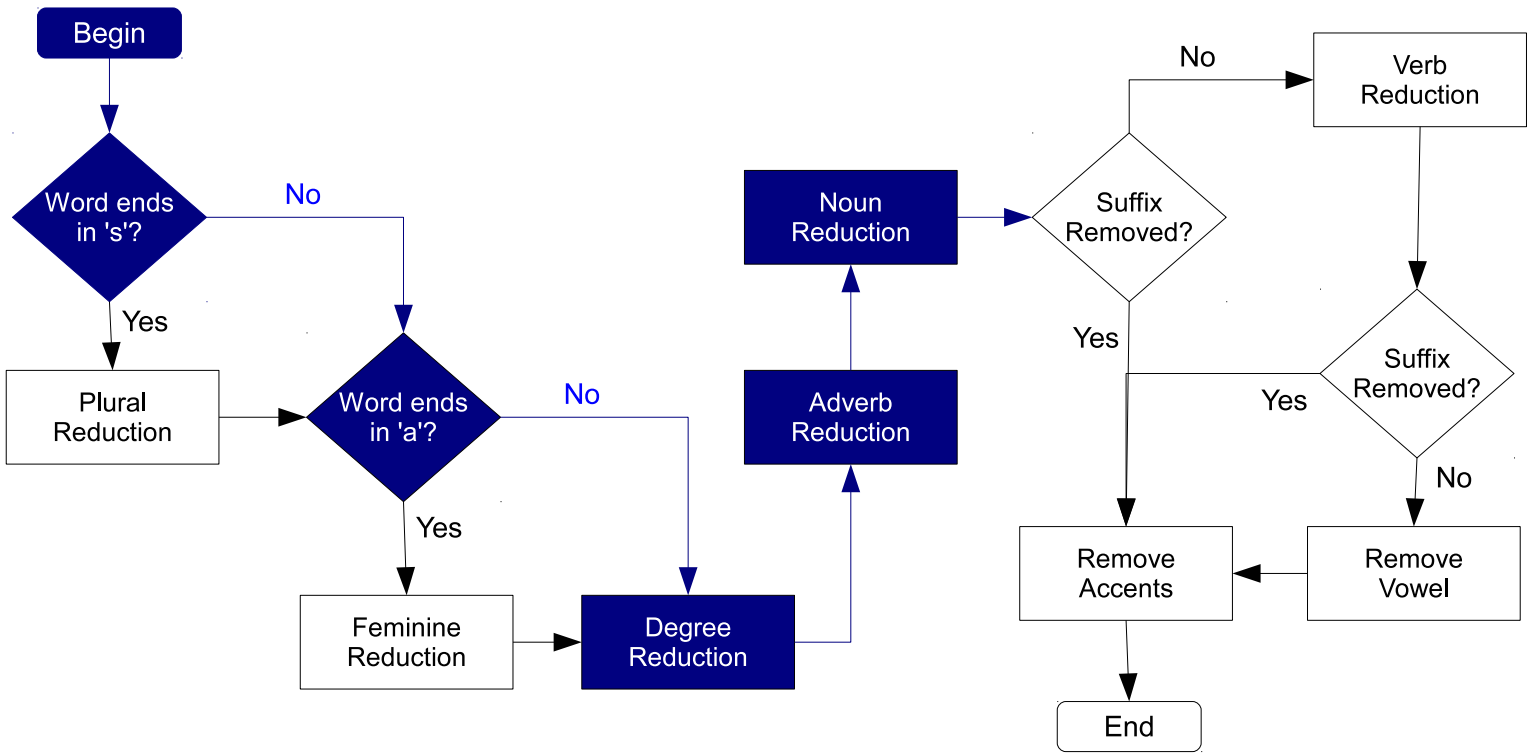
- subutilizaç



encialista, 4, “”, {} — alista, 5, “”, {} — alizaç, “”, 5, {} — izaç, 5, “”, {} — aç, 3, “”, {equaç, relaç}

LBA Example

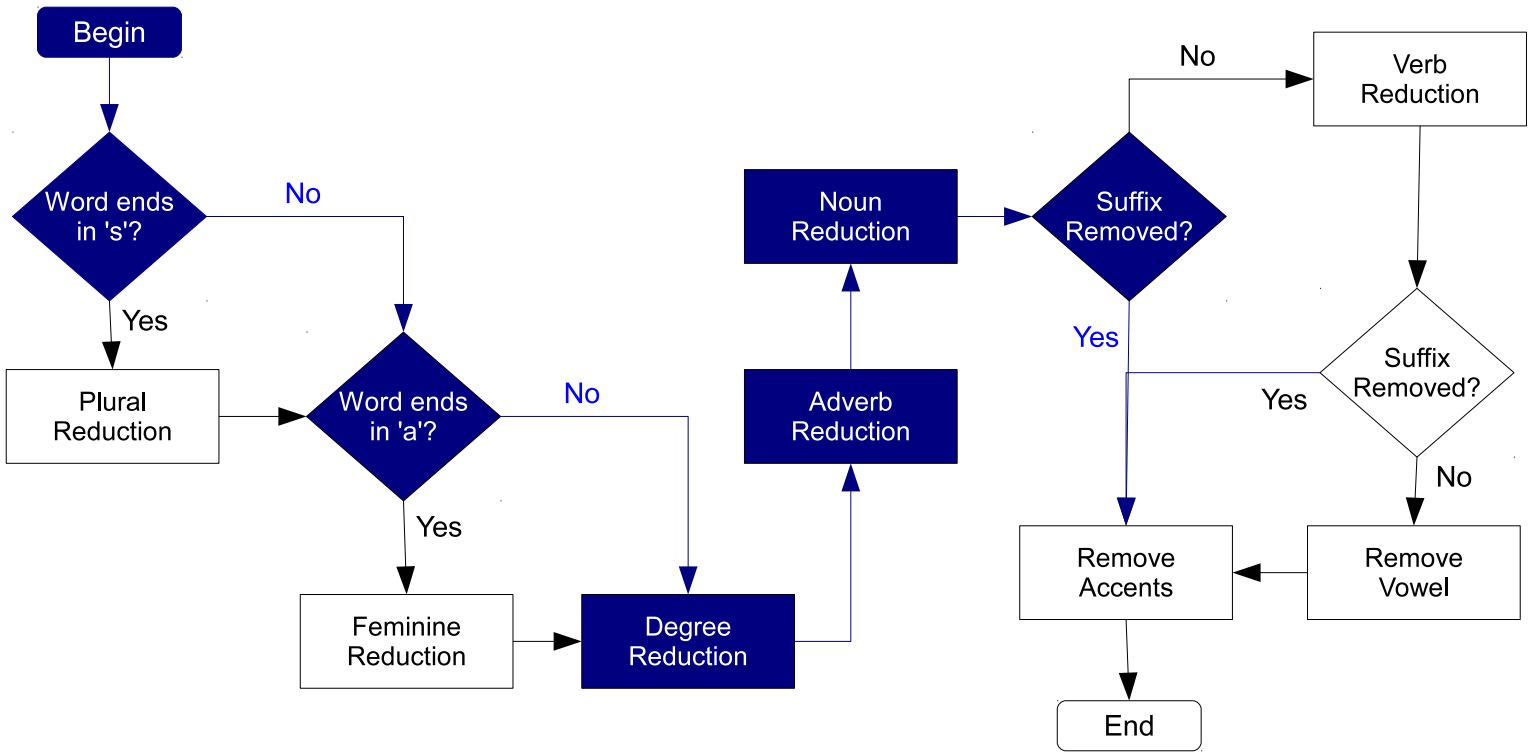
- subutil**izaç** → subutil



encialista, 4, “”, {} — alista, 5, “”, {} — alizaç, “”, 5, {} — izaç, 5, “”, {} — aç, 3, “”, {equaç, relaç}

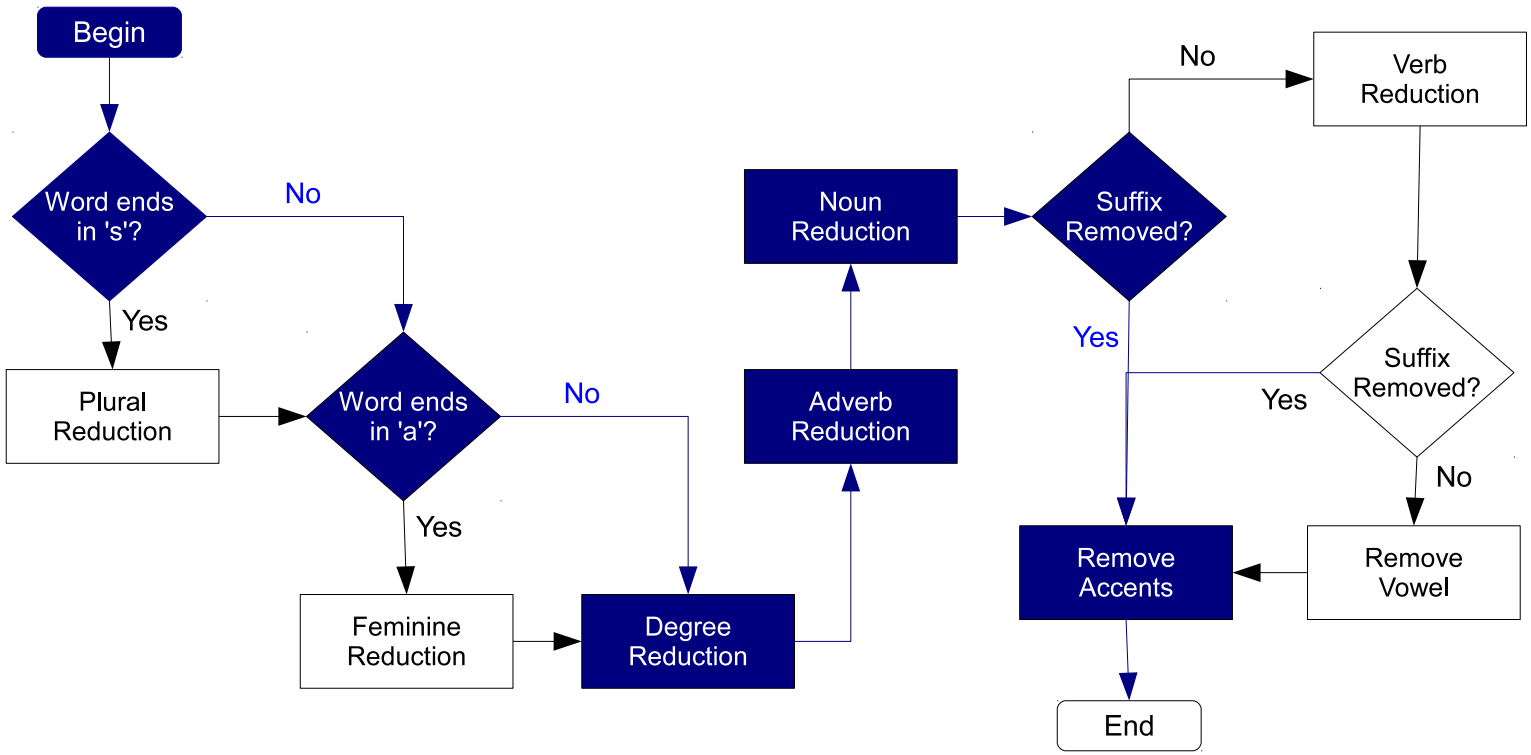
LBA Example

- subutil



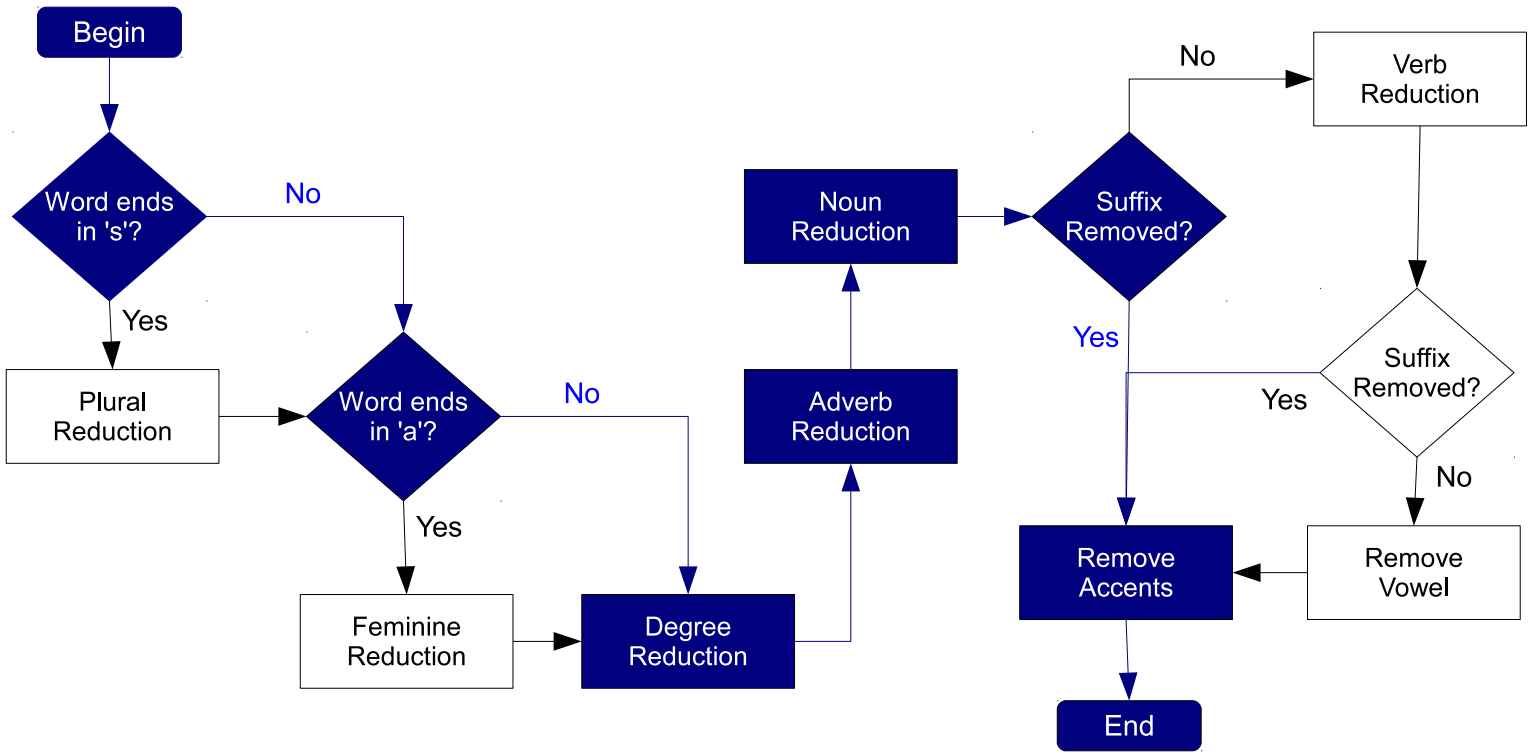
LBA Example

- subutil



LBA Example

- subutil



Many "inappropriated" suffix rules must be checked

The Hash-Based Approach (HBA)

For each step (type of reduction), a hash table breaks the single list of suffix rules into smaller lists

- Worst case → Hash entry pointing to the longest list of suffix rules

Plural	Feminine	Degree	Adverb	Noun	Verb
11	15	18	1	24	25

HBA Example

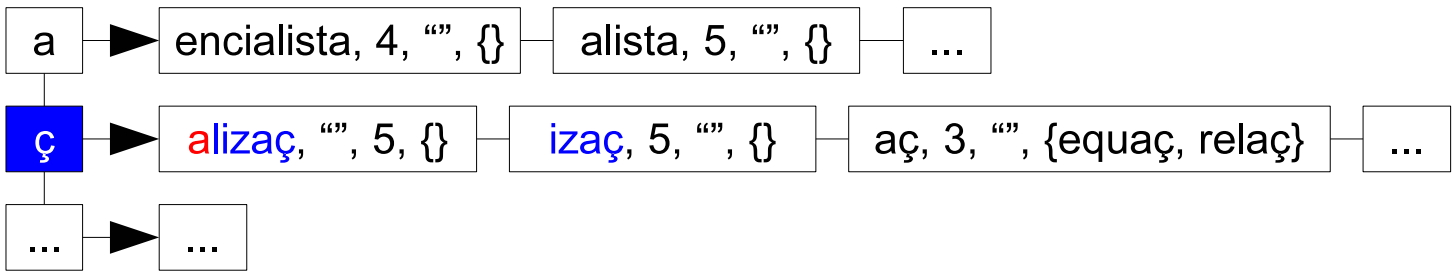
- subutilizaç

HBA Example

- subutilizaç

HBA Example

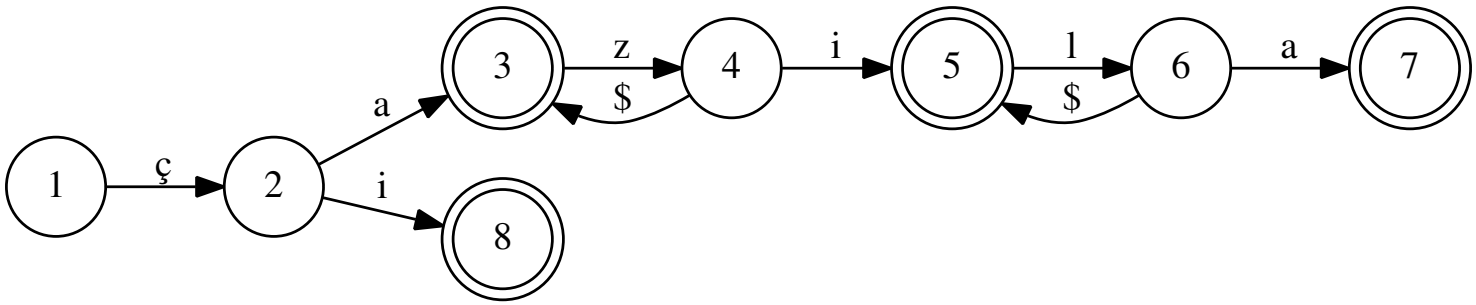
- subutil**izaç**



Less than LBA, but still many “inappropriated” suffix rules must be checked

The Automata-Based Approach (ABA)

For each step (type of reduction), a deterministic finite automata (DFA) reduces the number of character comparisons to the minimum, processing only the appropriate suffix rules

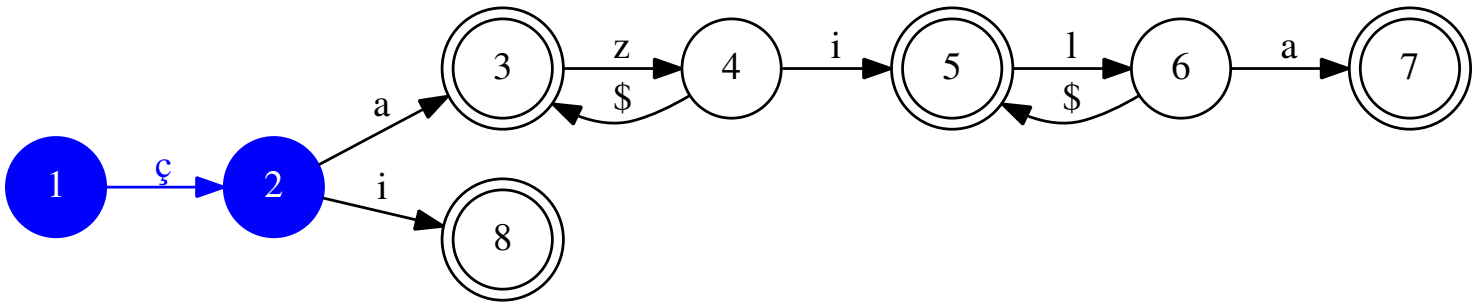


- Worst case → Checking the biggest path of the DFA

Plural	Feminine	Degree	Adverb	Noun	Verb
3	2	3	1	3	4

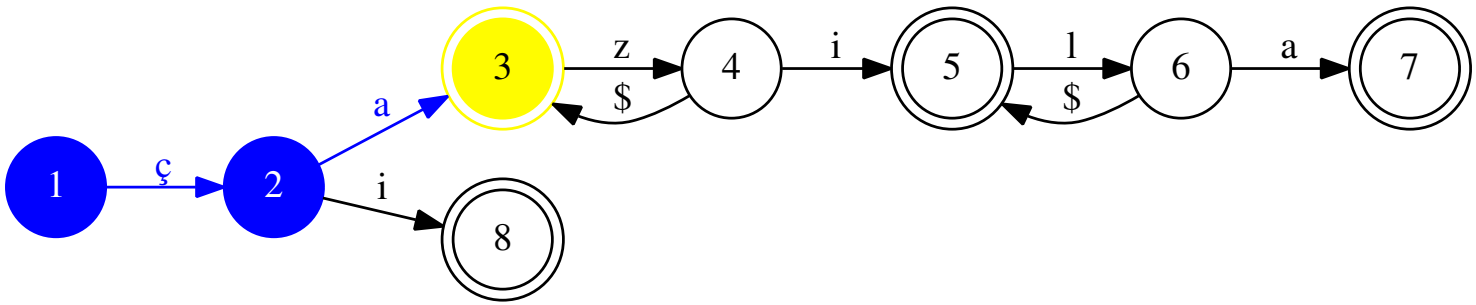
ABA Example

- subutilizaç



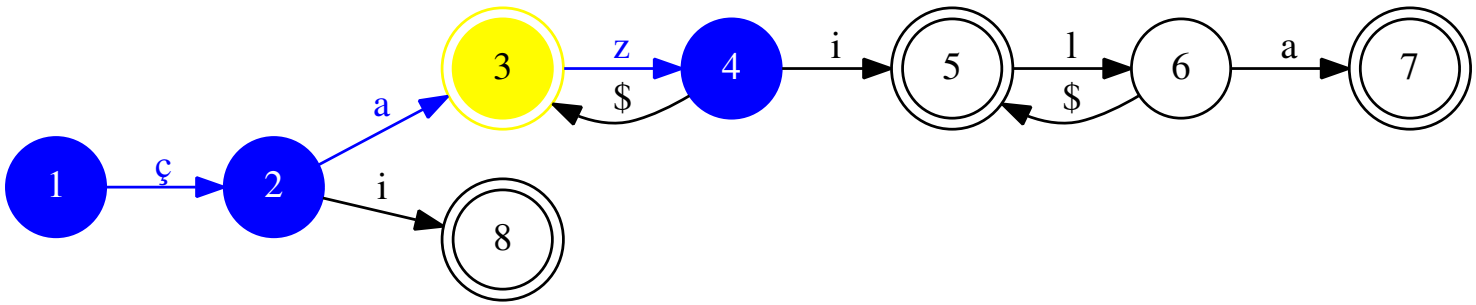
ABA Example

- subutiliz**aç**



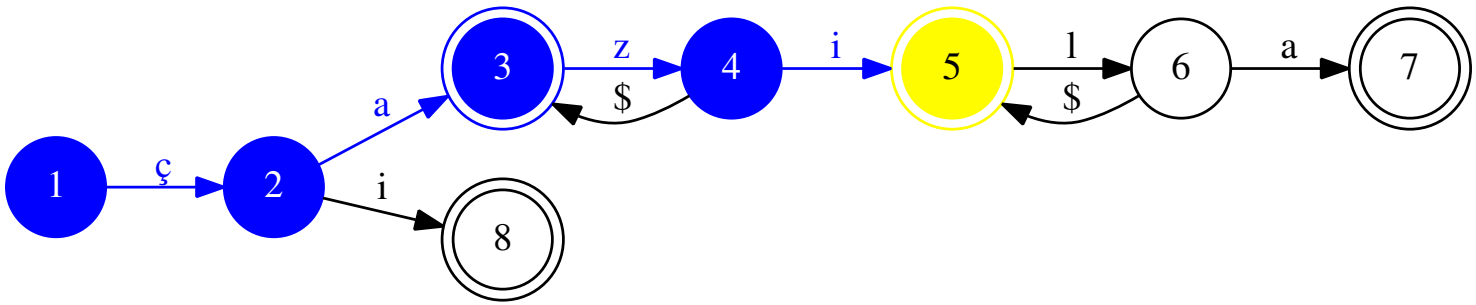
ABA Example

- subutili**zaç**



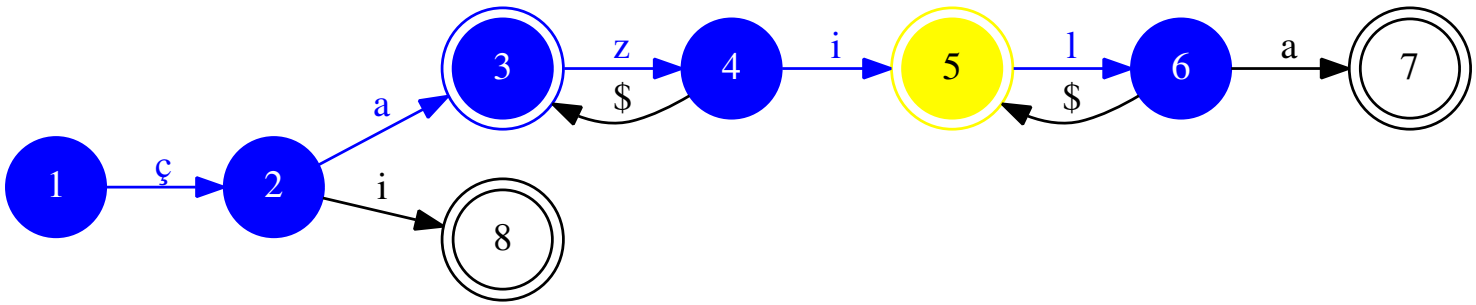
ABA Example

- subutil**izaç**



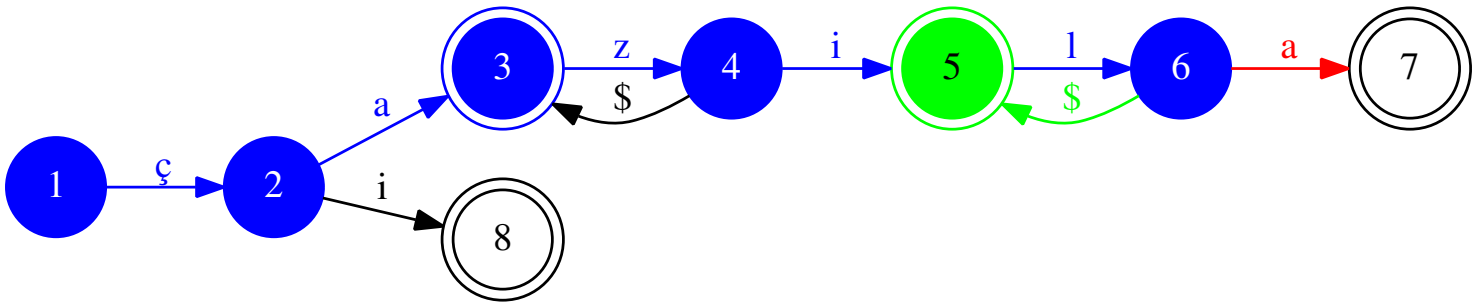
ABA Example

- subutil**izaç**



ABA Example

- subutilizaç



A small number of character comparisons and suffix rules must be performed and checked

Experiments

- Research Questions:
  - Performance by Step → How the LBA, HBA and ABA perform in each step of suffix stripping?
  - Overall Performance → How efficient (in time) are they?
- Datasets:
  - RSLP → 198 Portuguese words (one for each suffix rule in each step)
  - WBR99 → 206 millions Portuguese words extracted from approximately 6 millions Web pages crawled from the Brazilian Web (.br domain), and distributed into 6 collections.

Experiments: Setup

- Evaluation metric → Stemming time, i.e., the average elapsed time in microseconds (μs) for stemming a word, ignoring the time to load the suffix rules and hash entries in memory
- Environment → Single computer running Linux kernel version 3.16, 64-bit Intel Core i3 2.13 GHz processor, 3 GB of main memory, and 1 SATA II disk of 320 GB
- Implementation → C++
- Results → Two-tailed paired t-test at $p < 0.01$ level

Experiments: Performance by Step

Stemming time (μs) by reduction step on the RSLP dataset

Reduction Step	LBA	HBA	ABA
Noun	2.20733	1.07529 (51.28%) ▼	**0.46008 (79.15%) ▼▼**
Verb	3.63152	1.54178 (57.54%) ▼	**0.68953 (81.01%) ▼▼**
Plural	**0.16546**	**0.16546 (00.00%) •**	0.21505 (-29.97%) ▲▲
Feminine	0.28770	0.28770 (00.00%) •	**0.20967 (27.12%) ▼▼**
Degree	0.22786	0.22786 (00.00%) •	**0.18010 (19.64%) ▼▼**

Experiments: Overall Performance

Stemming time (μs) on WBR99 collections

Collection	LBA	HBA	ABA
AmostRA-NILC	1.37346	0.57674 (58.00%) ▼	**0.20779 (84.87%) ▼▼**
CETEMPúblico	0.03657	0.01338 (63.41%) ▼	**0.00504 (86.21%) ▼▼**
Museu da Pessoa	0.23419	0.09632 (58.87%) ▼	**0.03729 (84.07%) ▼▼**
ReLi	0.75644	0.31680 (58.11%) ▼	**0.12272 (83.77%) ▼▼**
Tycho Brahe	0.53799	0.19651 (63.47%) ▼	**0.07759 (85.57%) ▼▼**
Vercial	0.18901	0.06561 (65.28%) ▼	**0.02555 (86.48%) ▼▼**
All	0.52128	0.21090 (59.54%) ▼	**0.07933 (84.78%) ▼▼**

Contributions

- ABA → Simple and effective approach for Portuguese stemming
- Adaptable to work with different languages, such as English and Spanish
- Outperforms the baseline with gains of up to 86.48% in stemming time

Thank You

Questions?

Wladmir Cardoso Brandão
www.wladmirbrandao.com
wladmir@pucminas.br

"Science is a way of thinking...
much more than it is a body of knowledge."
Carl Sagan