U.PORTO
FLUP FACULDADE DE LETRAS UNIVERSIDADE DO PORTO
Centro de Linguística da Universidade do Porto
ACADEMIA DAS CIÊNCIAS DE LISBOA
WIKIMEDIA PORTUGAL

DDHUM25

G2P-DLP: An Open-Source System for Grapheme-to-Phoneme Conversion

Leonor Martins[1,2], Ana Salgado[2,1] & Carlos Silva[2,3]
Institute of Lexicology and Lexicography of the Portuguese Language, Academy of Sciences of Lisbon (ILLLP-ACL)[1]
Centre of Linguistics of University of Porto (CLUP)[2]
Wikimedia Portugal[3]

3rd International Conference on Data & Digital Humanities
27th-28th November 2025
ELACH, University of Minho, Portugal

Agenda

DDHUM25

1. Introduction

European Portuguese currently lacks openly accessible grapheme-to-phoneme (G2P) tools, **limiting access to phonological data** and thus **hindering phonological research and education**

This work **addresses this gap** by introducing a G2P system:

Findable
Accessible
Interoperable
Reusable

2. G2P Systems

Traditionally, **G2P (grapheme-to-phoneme) systems** convert written text into phonetic representations.
But also…

- syllabic divider
- stress marker
- phonetic transcriber

Efforts to develop these systems for **European Portuguese** have produced a range of algorithms and models based on…

- **linguistic rules** (Oliveira *et al.*, 1992; Caseiro *et al.*, 2002; Teixeira, 2004; Braga *et al*, 2006)
- **statistical data** (Trancoso *et al.*, 1994; Oliveira *et al.*, 2001; Barros & Weiss, 2006)
- **hybrid combination** (Candeias & Perdigão, 2011)

2. G2P Systems

Grafone

https://portulanclarin.net/workbench/it-grafone/

absence of information about the phonetic conventions and characters (*e.g.*, representation of the central vowel [ɨ] as /ə/)

incorrections in the transcription of certain vowels (*e.g.*, bebé, */bəbˈɛ/)

Convert Text to IPA Transcription

https://european-portuguese.info/ipa#Ol%C3%A1%20mundo!

inconsistencies on the transcription of certain vowels (*e.g.*, amarelo *[ɐ.mɐ.ˈɾ**e**.lu]) and the diphthong [ɔj] (*e.g.*, comboio *[kõ.ˈboj.u])

incorrections in the transcription of certain vowels (*e.g.*, normal *[n**u**ɾ.ˈmaɬ])

They are NOT reproducible or transparent, as their underlying methodologies and datasets remain undisclosed to the public

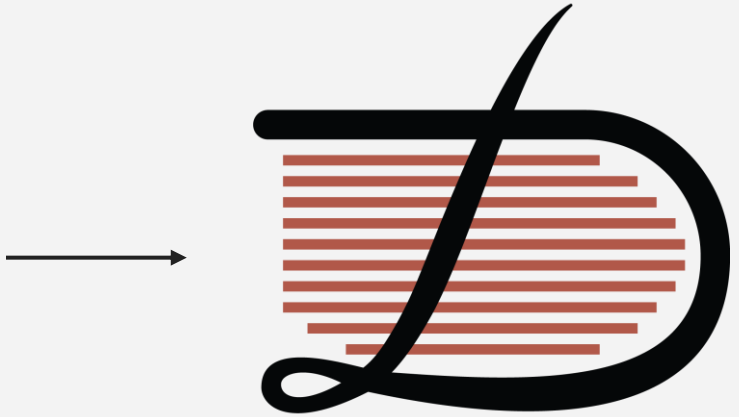
3. Methodology Overview

DDHUM25

3.1. Data

To develop and evaluate our system, we adopted a **dictionary-based approach**, which relies on a list of words (a **lexicon**) corresponding to the **lemmas** of the *Dicionário da Língua Portuguesa* of the Academy of Sciences of Lisbon (DLP-ACL)

	A	B	C	D
1	aberto			
2	alegre			
3	anterior			
4	atento			
5	barato			
6	básico			
7	brilhante			
8	central			
9	certo			
10	clássico			
11	colorido			
12	completo			

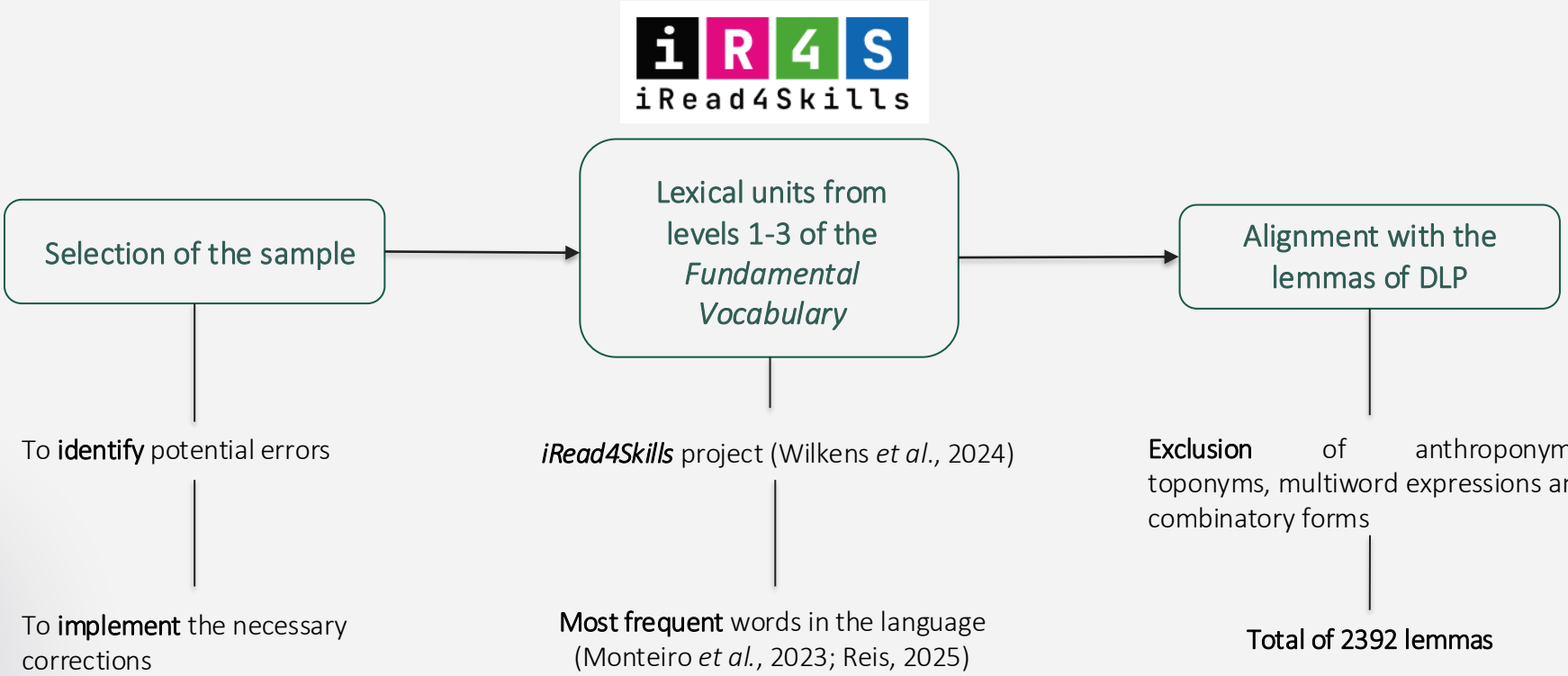


https://dicionario.acad-ciencias.pt/

3. Methodology Overview

DDHUM25

3.1. Data



3.2. Regular Expressions

DDHUM25

Regular expressions (**regex**) are versatile search patterns, effective in identifying and/or modifying data based on predefined character sequences (Goyvaerts, 2007; Goyvaerts & Levithan, 2009)

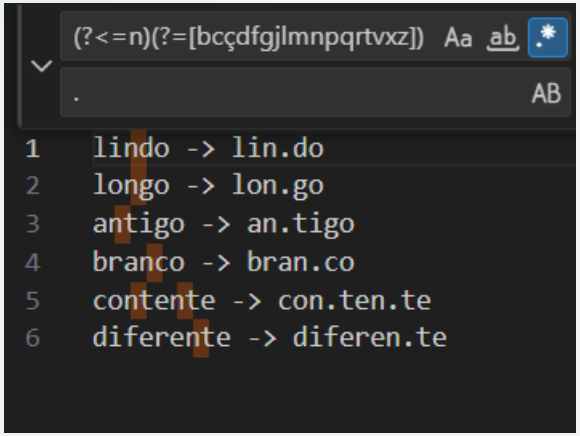
- broad applicability
- high interoperability
- great stability across software environments

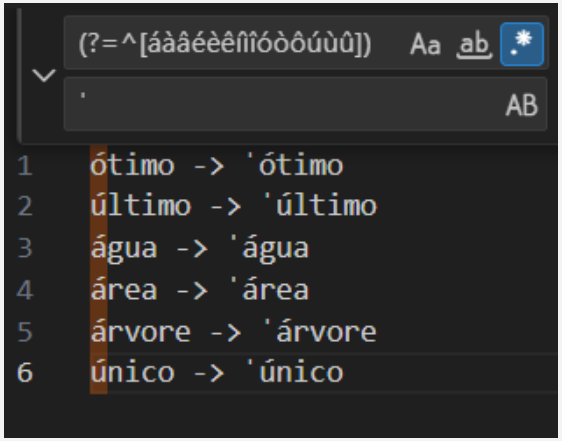
Sample of the syntax of regex characters

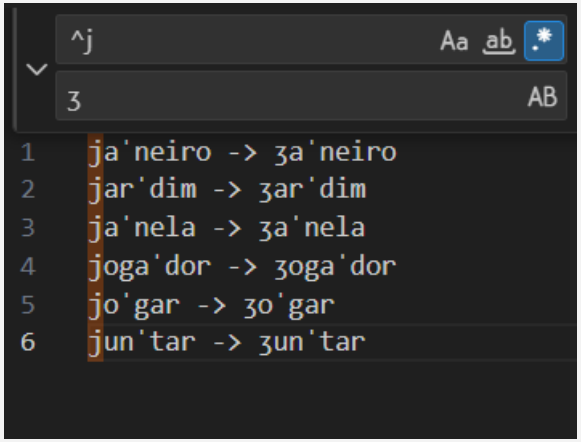
Character	Description	Example
^	Selects the initial position	^g → **g**anhar
$	Selects the final position	á$ → c**á**
(|)	Selects the character(s) on the left OR the character(s) on the right	(m|n|lh) → **m**a**lh**a; **m**a**n**a; **m**ala
(?=) (*positive lookahead*)	Selects the character when it is followed by what goes after =	p(?=ão) → p**ão**; p**ã**
(?!) (*negative lookahead*)	Selects the character when it is NOT followed by what goes after !	p(?!e) → perto; **p**reto; **p**orta
(?<=) (*positive lookbehind*)	Selects the character when it is preceded by what goes after =	(?<=o)l → v**o**ltar; possível
(?<!) (*negative lookbehind*)	Selects the character when it is NOT preceded by what goes after !	(?<!g)ui → seguir; aq**ui**lo

3.2. Pipeline architecture

DDHUM25







Syllabic Division

- selects a **boundary** between syllables
- inserts a **syllabic divider**

Stress Marking

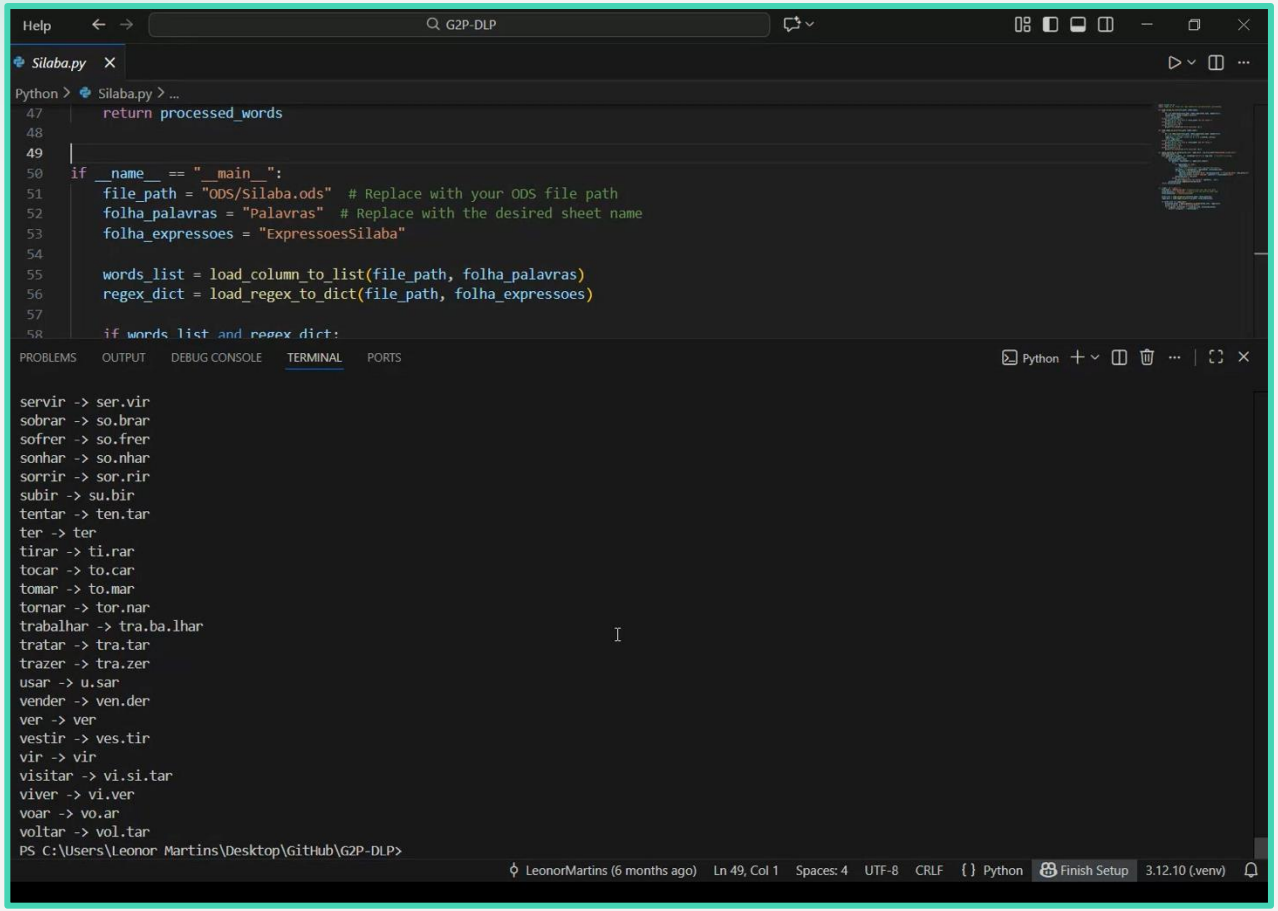
- selects a **position** before a syllable or a **boundary** between syllables
- inserts a **stress marker**

Phonetic Transcription

- **replaces** a segment with another segment

3.2. Pipeline architecture

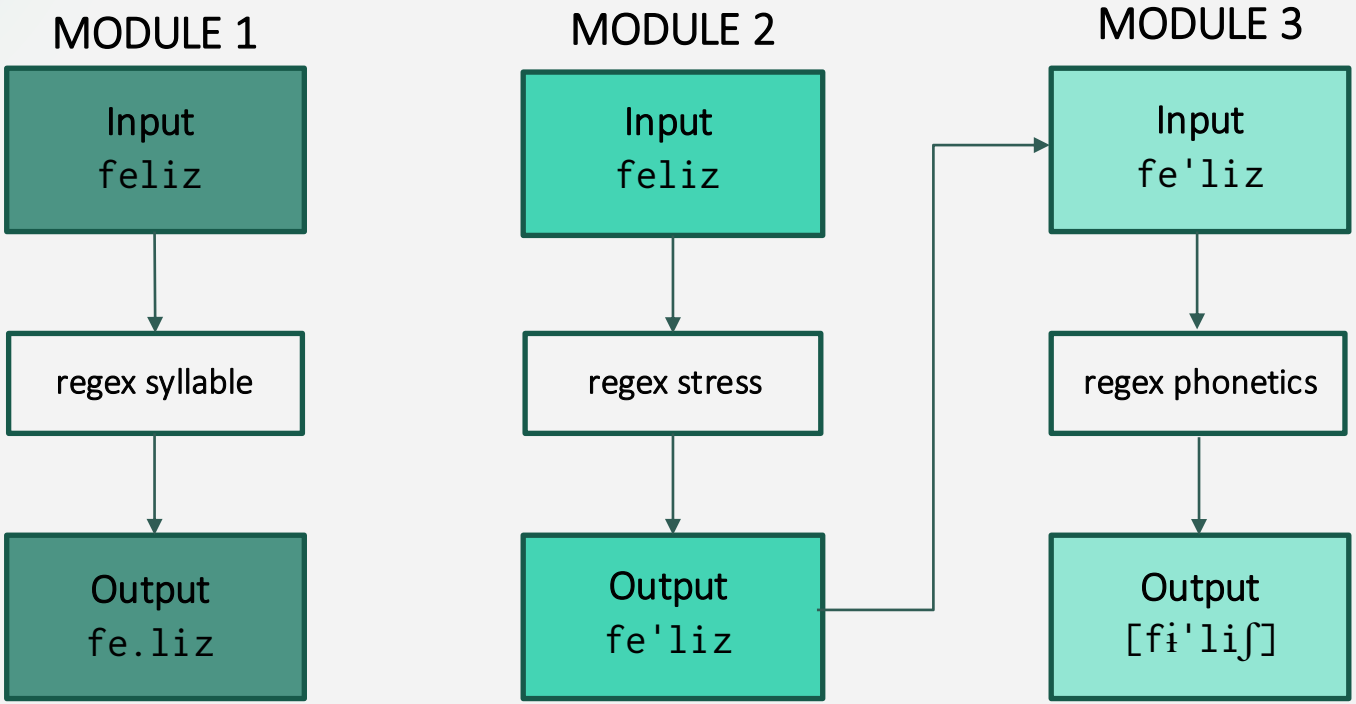
DDHUM25




```
servir -> ser.vir
sobrar -> so.brar
sofrer -> so.frer
sonhar -> so.nhar
sorrir -> sor.rir
subir -> su.bir
tentar -> ten.tar
ter -> ter
tirar -> ti.rar
tocar -> to.car
tomar -> to.mar
tornar -> tor.nar
trabalhar -> tra.ba.lhar
tratar -> tra.tar
trazer -> tra.zer
usar -> u.sar
vender -> ven.der
ver -> ver
vestir -> ves.tir
vir -> vir
visitar -> vi.si.tar
viver -> vi.ver
voar -> vo.ar
voltar -> vol.tar
```

3.2. Pipeline architecture

DDHUM25



3.4. Open-Source Implementation

DDHUM25

Advanced regular expression processing

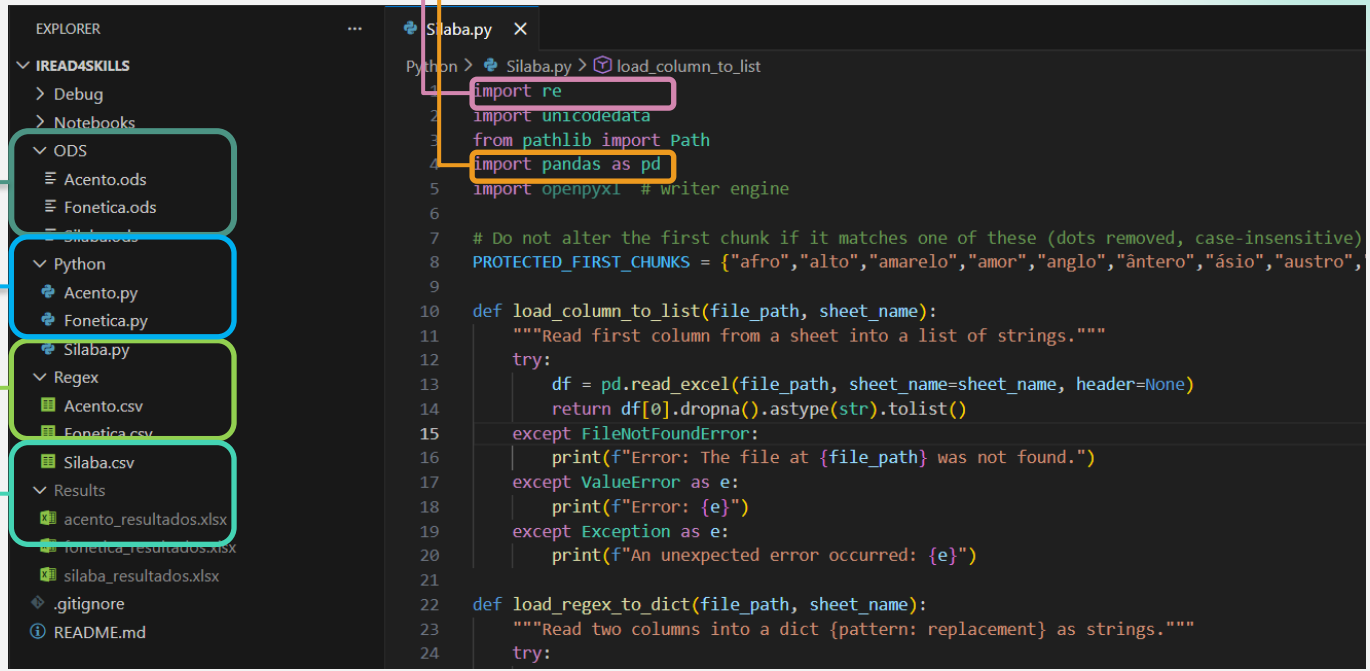
Reading and manipulating data in the ODS files

ODS
(input + regex)

Python
(script)

Regex
(ID, Examples)

Results
(ouput)



3.4. Open-Source Implementation

DDHUM25

GitHub

Repository Structure

```
├── Debug/
│   └── Log and debug files generated during the execution of the scripts and notebooks.
│
├── Notebooks/
│   └── Jupyter notebooks corresponding to the three modules (syllable, stress, phonetics).
│       They present the process of developing the Python files and the results obtained.
│
├── Python/
│   └── Standalone Python scripts for the three main modules.
│       - silaba.py
│       - acento.py
│       - fonetica.py
│
├── ODS/
│   └── ODS-format data files containing the corpora used in the modules.
│       These files specify the contexts, patterns, and substitutions applied.
│
├── Regex/
│   └── CSV tables with the regular expressions used in each module.
│       Each line includes:
│       - Regex ID
│       - Context Pattern
│       - Regex expression (Find)
│       - Applied substitution (Replace)
│       - Example of application
```

4. Results and Discussion

DDHUM25

4.1. Level 1

Syllabic Division

100%

All syllable boundaries were detected:
– C.C (partir → par.tir)
– V.C (tratar → tra.tar)
– G.C (baixo → bai.xo)
– G.V (meia → mei.a)
– V.V (voar → vo.ar)

Stress Marking

99.6%

All stress patterns were detected:
– **proparoxytones** (férias → **'fé**rias)
– **paroxytones** (festa → **'fes**ta)
– **oxytones** (flor → **'flor**)
Except *simples*, *apenas* and *menos*:
– *simˈples → ˈ simples
– *apeˈnas → a ˈ penas
– *meˈnos → ˈmenos

Phonetic Transcription

97.6%

All consonantal segments were correctly transcribed ([puˈxar] → [puˈʃar]).
The vowel segments were more problematic:
– *bebé*: *[bɨˈbɛ] → [bɛˈbɛ]
– *normal*: *[nurˈmaɬ] → [nɔrˈmaɬ]
– *ação*: *[ɐˈsɐ̃w̃] → [aˈsɐ̃w̃]
– *amanhã*: *[ɐmɐˈɲɐ̃] → [amɐˈɲɐ̃]

4.2. Levels 2 and 3: Scalability and Sustainability

DDHUM25

Syllabic Division

99.6%

A small set of **ambiguous graphemic sequences** challenged syllabification, namely **<ai>** and **<ui>**:
- rainha (r**a.i**.nha) *vs*. baixo (b**ai**.xo)
- fluidez (fl**u.i**.dez) *vs.* cuidar (c**ui**.dar)

Stress Marking

99.5%

The few deviations were limited to **predictable sources of ambiguity**:
- ambiguous graphemic sequences (*ˈjuiz → juˈiz)
- double marked words (*ˈórˈfão → ˈórfão)
- exceptions (*amˈbos → ˈambos)

Phonetic Transcription

97.7%

Errors were restricted to a small number of **well-defined contexts**:
- grapheme <x> (*exame*: *[eˈʃamɨ] → [eˈzamɨ])
- adverbs in *–mente* (*somente*: *[sᴜˈmẽtɨ] → [sɔˈmẽtɨ])
- exceptions of vowel reduction (coleção: *[kulɨˈsẽw̃] → [kulɛˈsẽw̃])

5. Future Applications and Implications

DDHUM25

Technical Framework and Sustainability

- uses **open-source**, **widely supported technologies**
- provides a **transparent**, **replicable**, and **extensible** framework
- flexibale architecture that supports **easy updates** and adaptation to **new corpora or languages**

Pedagogical Applications

- can be integrated into tools for **pronunciation training** and **language learning**
- supports **literacy development**, **first and second-language teaching** and the **creation of educational materials**

Educational Resource Development

- enables rapid production of **annotated datasets** that can be used in classroom resources
- supports teachers and learners through **standardized phonological representations**

Research and Linguistic Innovation

- facilitates **large-scale phonological analysis**
- could support **comparative studies on prosody** and serve as a preprocessing component for **speech technologies**
- contributes to **language conservation and revitalization**

6. Conclusion

DDHUM25

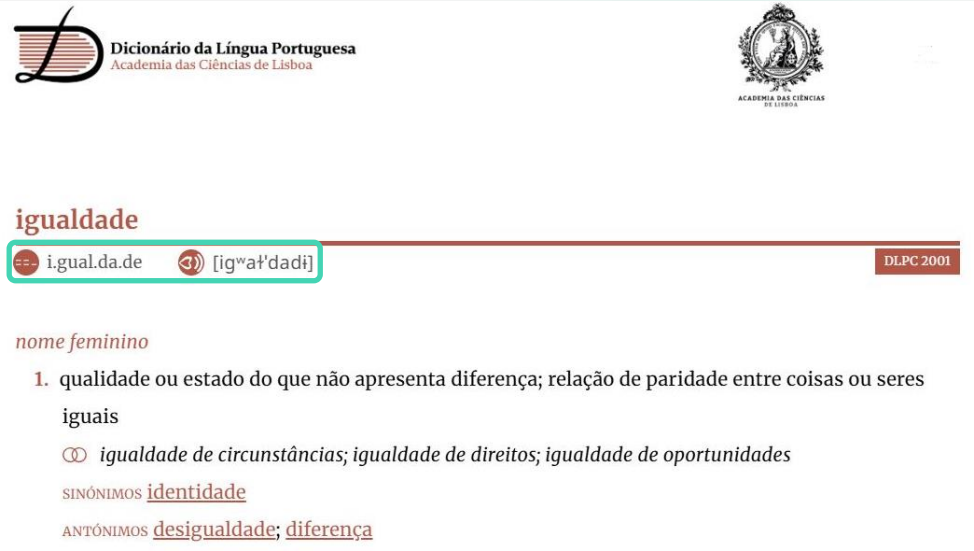
Our G2P system demonstrates **stability** and **accuracy** across various modules and dataset sizes

↓

Providing **openly available** solutions is crucial to enabling equitable and democratic participation in linguistic research and development

↓

This project offers a **FAIR resource** that illustrates the **impact of open infrastructutes**



References

DDHUM25

Aho, A. V. (1991). Algorithms for finding patterns in strings, *Handbook of theoretical computer science (vol. A): algorithms and complexity*. MIT Press, Cambridge, MA.

Braga, D., Coelho L., Resende Jr, F. G. (2006). A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese. Em *VI International Telecommunications Symposium*, 328-333.

Candeias, S. & Perdigão, F. (2011). Integração linguística em sistemas de conversão de grafema para fone[ma]. Em A. R. Luís (Ed.), *Estudos de linguística*, 1, 181-194. doi: 10.14195/978-989-26-0231-8_12

Candeias, S. & Perdigão, F. 2013. Grafone: uma ferramenta de/com recurso à informação linguística. Em *Textos Selecionados, XXVII Encontro Nacional da Associação Portuguesa de Linguística*, 189-203. doi: http://doi.org/10.1109/its.2006.4433293

Caseiro, D. A., Trancoso, I., Oliveira, L. & Viana, C. (2002). Grapheme-to-phone using finite-state transducers. Em *Proceedings of the IEEE Workshop on Speech Synthesis,* 215 218. doi: 10.1109/WSS.2002.1224412

DLP-ACL. (2025). Academia das Ciências de Lisboa. *Dicionário da Língua Portuguesa*. Salgado, A. (Coord.). Lisboa: Academia das Ciências de Lisboa. Disponível em https://dicionario.acad-ciencias.pt/

Goyvaerts, J. (2007). *Regular Expressions: The Complete Tutorial.* LuluPress, Incorporated.

Goyvaerts, J. & Levithan, S. (2009). *Regular Expressions Cookbook*. O'Reilly Media, Inc.

IPA – International Phonetic Association. (n.d.). *International Phonetic Alphabet*. Retirado de https://www.internationalphoneticassociation.org/.

Martins, L. (2025). *Desenvolvimento de um sistema de conversão grafema-fone[ma] para o vocabulário Fundamental da Academia das Ciências de Lisboa*. Dissertação de Mestrado, Faculdade de Letras da Universidade do Porto. https://repositorio-aberto.up.pt/handle/10216/168957

Martins, L. (2025). *G2P-DLP*. https://github.com/LeonorMartins/G2P-DLP

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference,* Vol. 445, pp. 51-56.

Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauchola, R., Moutinho, M., & Blanco Escoda, X. 2023. *iRead4Skills – Complexity Levels*. Zenodo. https://doi.org/10.5281/zenodo.10459090

Oliveira, L. C., Viana, M. & Trancoso, I. (1992). A rule-based text-to-speech system for portuguese. Em *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 72-76. 10.1109/ICASSP.1992.226117 Imprensa da Universidade de Coimbra. doi: 10.1109/ICASSP.1992.226117

Reis, M. L. (2025). *Vocabulário Fundamental da Academia das Ciências de Lisboa: seleção lexical, alinhamento de sentidos e codificação*. Tese de Mestrado. Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa.

Teixeira, J. (2004). *A Prosody Model to TTS Systems*. [Tese de Doutoramento, Faculdade de Engenharia da Universidade do Porto]. http://hdl.handle.net/10198/1496

Trancoso, I., Viana, M. C., Silva, F. M., Marques, G. C. & Oliveira, L. C. (1994). Rule based vs neural network-based approaches to letter-to-phone conversion for Portuguese common and proper names. Em *3rd International Conference on Spoken Language Processing* (ICSLP 1994), 1767-1770. doi: 10.21437/ICSLP.1994-197

Trigo, L. & Silva, C. (2022). Comparing lexical and usage frequencies of palatal segments in Portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference*, PROPOR 2022, Fortaleza, Brazil, March 21–23, Proceedings, pages 353–362, Berlin. Springer.

Wilkens, R., Pintard, A., François, T., Barbosa, S., Reis, M. L., Amaro, R., Ribeiro, E., Mamede, N., Baptista, J., Blanco, X., Catena, A., Gauchola, R., & Mu, K. (2024). *iRead4Skills – Basic Lexicons per Complexity Level* (v1.0) [Data set]. Zenodo. https://doi.org/10 – 81/zenodo.10889986

U.PORTO
FLUP FACULDADE DE LETRAS UNIVERSIDADE DO PORTO
clup Centro de Linguística da Universidade do Porto
ACADEMIA DAS CIÊNCIAS DE LISBOA
WIKIMEDIA PORTUGAL

DDHUM25

G2P-DLP: An Open-Source System for Grapheme-to-Phoneme Conversion

Leonor Martins[1,2], Ana Salgado[3,4] & Carlos Silva[2,5]
Institute of Lexicology and Lexicography of the Portuguese Language, Academy of Sciences of Lisbon (ILLLP-ACL)[1]
Centre of Linguistics of University of Porto (CLUP)[2]
Academy of Sciences of Lisbon (ACL)[3] & Faculty of Arts and Humanities of University of Porto (FLUP)[4]
Wikimedia Portugal[5]

3rd International Conference on Data & Digital Humanities
27-28 November 2025
ELACH, University of Minho, Portugal