

LEARNING PHONOLOGY WITH DATA IN THE CLASSROOM:  
ENGAGING STUDENTS IN THE CREOLISTIC  
RESEARCH PROCESS

---

LUÍS TRIGO , CARLOS SILVA  and  
VERA MOITINHO DE ALMEIDA 

**Abstract** *Phonology is a linguistic discipline that is naturally computational. However, as many researchers are not familiar with the use of digital methods, most of the computation required is still performed by humans. This article presents a training experiment of master's students of the phonology seminar at the University of Porto, bringing the research process directly to the classroom. The experiment was designed to raise students' awareness of the potentialities of combining human and machine computation in phonology. The Centre for Digital Culture and Innovation (CODA) readily embraced this project to showcase the application of digital humanities as humanities in both research and training activities. During this experiment, students were trained to collect and process phonological data using various open-source and free web-based resources. By combining a strict protocol with some individual research freedom, the students were able to make valuable contributions towards Creolistic Studies, while enriching their individual skills. Finally, the interdisciplinary nature of the approach has demonstrated its potential within and beyond the humanities and social sciences fields (e.g., linguistics, archaeology, history, geography, ethnology, sociology, and genetics), by also introducing the students to basic concepts and practices of Open Science and FAIR principles, including Linked Open Data.*

**Keywords:** project-based learning, humanities research, phonology teaching, Open Science, Linked Open Data

---

*International Journal of Humanities and Arts Computing* 18.1 (2024): 40–57

DOI: 10.3366/ijhac.2024.0320

© Edinburgh University Press 2024

[www.eupublishing.com/ijhac](http://www.eupublishing.com/ijhac)

## I. INTRODUCTION

The phonology of Portuguese-based creoles is a field of linguistics that is still little explored. On the one hand, the linguists who collect data in the field are not, for the most part, phonologists or phoneticians. Consequently, the descriptions of Portuguese creoles report little more than the segmental inventory of these languages in terms of their sound systems. On the other hand, theoretical phonology-feature systems, suprasegmental structures, and other phonological conjectures are based on the study of European languages, which leads phonologists to also focus on analysing creole languages in order to corroborate or rebut previous proposals. Moreover, in historical phonology studies, contact languages such as creoles and pidgins are often avoided given the lack of genetic classification<sup>1</sup> and the high number of loanwords they incorporate, which often display irregularities.<sup>2</sup>

These reasons, as well as the fact that many Portuguese-based creoles are endangered languages, make it urgent to systematically collect sound phonological data and compose databases. In addition, there is a critical need to resume these data in phonological studies that describe and give visibility to the so-called ‘Phonology of Creoles’.

The CreoPhonPt project was initiated at the Max Planck Institute within Silva’s ongoing PhD research.<sup>3</sup> The project continued growing with contributions from further research centres and universities, while serving as a launchpad for other creole-related projects at the Centre of Linguistics of the University of Porto (CLUP).<sup>4</sup> The CreoPhonPt database incorporates the FAIR (findable, accessible, interoperable, reusable) principles,<sup>5</sup> whenever possible. The ultimate goal of this database is to help document and preserve Portuguese-based creole languages and enable further studies related to ‘creole phonology’, as well as to disseminate it among the scientific community and the general public.

The Centre for Digital Culture and Innovation (CODA)–based at the Faculty of Arts and Humanities of the University of Porto (FLUP)–identified CreoPhonPt as an opportunity to promote a demonstration project for its goals:

- To foster interdisciplinary studies and collaboration between the FLUP’s R&D units;
- To showcase the benefits of applying digital methods in research and in the classroom;
- To engage with Citizen Science initiatives, such as the Wikimedia Foundation, through its Wikidata project;<sup>6</sup>
- To normalize computational methods as another available method in humanities – digital humanities as humanities.

In view of these, the CODA developed a collaboration with CreoPhonPt's leading researcher and lecturer of the Phonology I seminar from the master's course in linguistics, at FLUP, to train the students in the use of digital research methods and techniques, while contributing to the database. During the lectures, the students became actively engaged in tasks like data collection and processing, as they were guided by the lecturer and CODA's members through each step of the data collection and processing stages.

The use of computational methods in phonology comprises theoretical and technological knowledge, whose primary function is to support the phonological description of field data.<sup>7</sup> The combination of theory with hands-on analysis makes it a pedagogical opportunity, although the advantages go beyond the classroom. The focus on data accessibility, accountability, and stability of empirical research turns the use of digital humanities in the classroom into a good introduction to research practices in general. These methods are particularly important for the documentation and preservation of endangered languages, such as Portuguese-based creoles.<sup>8</sup>

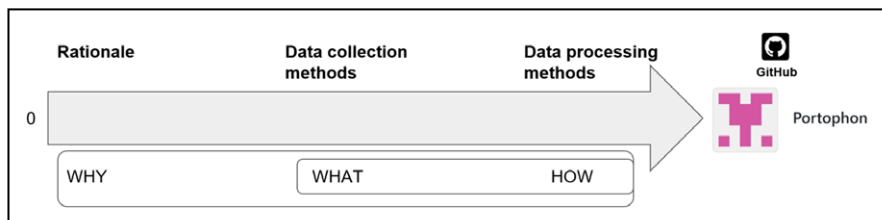
This article describes the pilot project in a chronological way. Sections 2, 3, and 4 report the data collection and processing methods and provide a detailed view of the research protocol. Section 5 describes the experiments to connect the data to other freeware and open-access online databases. In section 6, we present some of the project's outcomes. Finally, in section 7, the main achievements (and limitations) of the pilot project are presented, followed by prospects for future works.

## 2. PARTICIPANTS AND CONTEXT

The introduction to computational methods was conducted within the Phonology I seminar of the master's course in linguistics, at FLUP. The seminar students shared two main characteristics: (1) low, to any, computational knowledge; and (2) a diverse bachelor background, mostly from other disciplines within the humanities (e.g. law, information sciences, communication). Moreover, approximately 40 per cent of them were international students.

With these in mind, we adopted the Project-Based learning principles<sup>9</sup> and posed the following driving question to attract the students' interest: 'How is the phonology of Portuguese creoles formed?' This question motivated the students to acquire knowledge in the two key subjects of the seminar (i.e., phonological theory and computational methods) and apply it to a specific object (i.e., a creole language). To increase their interest, students were given the freedom to select the creole language with which they could best identify. For instance, a Chinese student selected Macao Creole and Brazilian students preferred African creoles.

Another objective of this approach was to explore and demonstrate automation issues following an Augmented Intelligence perspective<sup>10</sup> that can support



**Figure 1.** Protocol structure.

students in their future research and professional careers. For instance, a traditional analysis of a language explores the phonological properties of small sets of words, usually targeting approximately 20–30 words with particular features. However, computational methods may be used to enlarge the word pool sample by enabling the automation of several analyses and speeding up the verification of linguistic/phonological hypotheses.

### 3. DATA COLLECTION

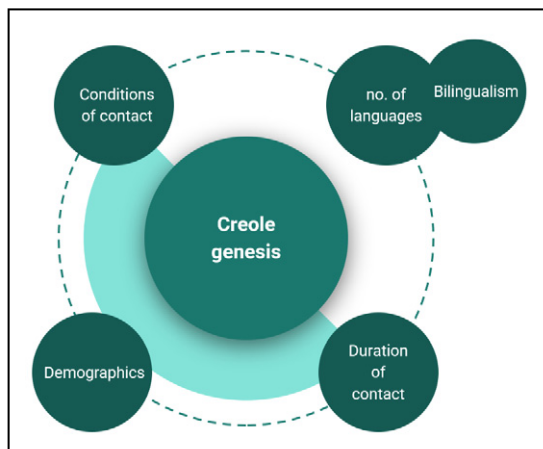
Regarding the research methodology, several steps were taken. Both CLUP and CODA leading researchers developed a protocol describing the procedures for data collection and manual transformation (Figure 1), which would also be the basis for the computational processing in the tutorial classes.

This section illustrates the rules for collecting and preparing lexical data from creole languages for a medium-scale comparative phonological analysis, according to the CreoPhonPT prototype.<sup>11</sup> The data to be collected are of two types: (1) phonological words, which should allow an analysis at the level of phoneme selection, distinctive features and syllable patterns;<sup>12</sup> and (2) socio-historical characteristics of each creole, which should precede the study of any contact language.<sup>13</sup>

#### 3.1 Socio-historical Data

The formation of creole languages involves the contact of at least two communities that come together to form a new language.<sup>14</sup> However, there are several factors that can affect this phenomenon and determine the degree of proximity or distance of the languages in contact with creole.

Figure 2 illustrates some of the variables that are used to fill in the metadata fields for each creole. The students had to fill in a metadata table concerning the creole that they chose, as exemplified in Table 1. Beyond the example table, they were provided with a detailed description for each variable, as follows:



**Figure 2.** Socio-historical features with impact on language creation.

**Table 1.** Example of table for the language metadata, as provided to the students.

<b>Language</b>	Indo-Portuguese Diu	Cardoso2013
<b>Area</b>	Northern India	CardosoHagemeijerAlexander2015
<b>Lexifier</b>	Portuguese	Cardoso2013
<b>FirstMajorSettlement</b>	1535	Cardoso2013
<b>EndOfInfluence</b>	1740	Cardoso2009
<b>ContactConditions</b>	Fort Creole	Cardoso2009
<b>LanguageContact</b>	Gujarati, English, Hindi, Konkani	Cardoso2013
<b>NumberOfLanguages</b>	5	Cardoso2013
<b>Bilingualism</b>	Yes	Cardoso2009
<b>Notes</b>	Gujarati is dominant	Cardoso2013

*Notes:*

\* H. Cardoso, ‘Diu Indo-Portuguese’, in S. M. Michaelis, P. Maurer, M. Haspelmath and M. Huber, eds, *The survey of pidgin and creole languages. Volume 2: Portuguese-based, Spanish-based, and French-based languages* (Oxford, 2013), 90–101.

\*\* H. C. Cardoso, T. Hagemeijer and N. Alexandre, ‘Crioulos de base lexical Portuguesa’, in *Manuel des anthologies, corpus et textes romans* (Berlin and Boston, MA, 2015), 670–92.

\*\*\* H. C. Cardoso, *The Indo-Portuguese language of Diu* (Utrecht, 2009).

1. **Language:** name of the language as it appears in APiCS – Atlas of Pidgin and Creole Language Structures.<sup>15</sup> When this is not possible, the name of the language is used as it appears in the bibliographic source to be indicated.
2. **Area:** Creole linguistic area.<sup>16</sup> Upper Guinea, Gulf of Guinea, Northern India, Southern India, South-East Asia, East Asia.

3. **Lexifier:** corresponds to the variable ‘Major Lexifier’ as defined in APiCS.
4. **First Major Settlement:** year of the first significant settlement, that is, the demographic event that brought together enough people to form a linguistic community (may not be the year of arrival).
5. **End Of Influence:** year of independence or year in which another colonial power took possession of the territory.
6. **Contact Conditions:** definition of contact conditions in creole formation:<sup>17</sup> ‘plantation creole’, ‘maroon creole’, or ‘fort creole’.
7. **Language Contact:** corresponds to the variable ‘Other contributing languages’ as defined in APiCS. This includes substrates and adstrates.
8. **Number of Languages:** corresponds to the numeric value of the variable ‘Language Contact’ added to the ‘Lexifier’.
9. **Bilingualism:** defines the existence of situations of bilingualism vs monolingualism in creole communities, especially in relation to substrate languages.
10. **Notes:** Additional socio-historical data on creole.

### *3.2 Phonological Data*

The phonemic inventory corresponds to the set of speech sounds that are distinctive in a language. Since the creoles under study import most of the Portuguese lexicon, it would be expected that they would import the same phonemic inventory. However, as contact languages, creoles include in their lexicon a large number of borrowings from both substrate and adstrate languages, which leads to a phonological realignment of the recipient language.<sup>18</sup>

Phonological inventories are, however, highly susceptible to discussion, since the time in which they are collected and the theoretical orientation of those who collect them can lead to divergent phoneme casts.<sup>19</sup>

Instead, we use Swadesh lists as a basis for analysis.<sup>20</sup> The Swadesh list is the most common starting point for lexical-statistical analysis and the chronological study of relationships between languages, as it compiles terms that tend to be universal and resistant to substitution by borrowing.<sup>21</sup>

Swadesh lists collected for analysis in the classroom must comply with two fundamental principles: (i) provenance from credible and citable sources and (ii) phonetic transcription from original sources in the International Phonetic Alphabet (IPA) or reference to instructions that make possible transcription to IPA. To this end, a new table (Table 2) was provided to the students, along with a detailed description of each variable:

1. **Language:** name of the language as it appears in APiCS. When this is not possible, the name of the language is used as it appears in the bibliographic source to be indicated.

Table 2. Example of table for the phonological data, as provided to the students.

Language	SourceWord	TargetWord	TargetOrthography	Translation	Reference	Notes
Timor Pidgin	kor'saŋ	kurɛ sɛw̃	coração	heart	Baxter1990*	
Timor Pidgin	'ɬua	'juvɛ	chuva	rain	Baxter1990	
Timor Pidgin	ka'tforu	ke'foru	cachorro	dog	Baxter1990	
Timor Pidgin	'ɬeru	'ɬɛru	cheiro	smell	Baxter1990	

Note:

\* A. N. Baxter, 'Notes on the creole Portuguese of Bidau, East Timor', *Journal of Pidgin and Creole Languages*, 5, no. 1 (1990), 1–38, <https://doi.org/10.1075/jpcl.5.1.02bax>.

2. **SourceWord:** IPA transcription of the word in creole, as in the source or converted.
3. **TargetWord:** IPA transcription of the Portuguese word.
4. **TargetOrthography:** orthographic transcription of the Portuguese word.
5. **Translation:** Translation of the word into English.
6. **Reference:** BibTex key of the work from which the word included in the list was taken.
7. **Notes:** Additional notes on the word.

Data insertion in MS Excel spreadsheets was not recommended in the classroom because this software does not read IPA symbols well. Alternatively, students were instructed to use freeware software, such as GoogleSheets or LibreOffice, for creating, editing, and exporting tables in a CSV (comma-separated values) file format. Contrarily to MS Excel, this software uses UTF-8 as default character encoding. This way, the students would not take any risks of having their IPA symbols deleted or mis-formatted.

#### 4. DATA PROCESSING

After the data extraction from the CreoPhonPt project, we presented basic concepts of human and machine computation to students.<sup>22</sup> They learned that these concepts are not mutually exclusive, but complementary. Next, the phonological data were divided into two levels of analysis: words and syllables. At each level, human annotation was performed prior to automatic processing. The latter was also presented as a tool to audit and detect potential manual annotation errors, while visual inspection of the variables that were automatically generated was important to detect errors in this kind of processing.

##### *4.1 Spreadsheet Operations*

We also presented some basic data concepts and processing techniques, and related them with the linguistic and phonological knowledge that was being taught. Regarding the tabular organization of data, students were introduced to short and long table formats. For organizing metadata, a short table format is more suitable (e.g. Table 1). For listing words and their features or matching items and concepts, a long table format should be preferred (e.g. Table 2).

Considering that both human and automatic data collection and processing are prone to error, students were taught to clean the collected data (e.g. removing white spaces and other spurious characters such as punctuation), normalize it (e.g. using regular expressions to standardize phonetic symbols), and perform auditing (e.g. sorting and filtering values to check non-alphabetic or unexpected characters; checking frequencies to identify outliers).



Students were introduced to spreadsheet formulas to automatically transform strings and calculate string length. After distributing functions through multiple cells, for a better understanding of a sequential algorithm for string processing, we replaced this operation with formulas containing nested functions. As a step forward to demonstrate algorithmic thinking, we introduced decision structures (e.g. ‘if’ statements functions) for converting numeric positions into initial/medial/final positions (not so trivial for words with fewer than three syllables) and loop structures (e.g. ‘from i to j’) for row iterations within the spreadsheet. Thus, we integrated spreadsheet concepts with programming concepts as a way to give a more understandable visual learning.

#### 4.1.1 Word processing

Phonetic transcriptions under the columns *SourceWord* and *TargetWord* were manually divided into syllables under two new columns – *SourceWordSyl* and *TargetWordSyl*. From the former pair of columns, we demonstrated the automatic generation of two columns, namely *LengthSourceWord* and *LengthTargetWord*. These variables are an output of data processing and allow learners to see an immediate by-product of their work. The observations correspond to the number of phonemes of the Portuguese words and their cognates in creoles. From the latter pair, instead of generating observations regarding the number of phonemes, we manually extracted the number of syllables in the words of both languages (Table 3).

These numeric variables show the most identical and different words in the relationship between Portuguese and each creole (*WordDiff* and *SylDiff*) and, more broadly, the distance between Portuguese and the creole selected by each student. They serve as a basis for correlation with socio-historical data and, more importantly, to answer a central question of the creole debate:<sup>23</sup> ‘Are creoles more complex/simple than their lexifiers and other natural languages?’

#### 4.1.2 Syllable processing

Some syllabic positions are more prone to change than others. For example, the initial position is more likely to maintain consonants or undergo fortition processes. The medial and final positions commonly undergo lenition (weakening) or even erasure processes.<sup>24</sup> The syllable treatment prepares the data for the positional analysis of phonemes (initial, medial, or final). Thus, the rows of the table must be manually unfolded according to the number of syllables of each word in order to create and fill in the following fields:

1. **SourceSyllable:** IPA transcription of each syllable of the word in creole.
2. **TargetSyllable:** IPA transcription of each syllable of the word in Portuguese.

3. **PositionSource:** numerical value corresponding to the position that the syllable represented in the Syllable for the creole.
4. **PositionTarget:** numerical value corresponding to the position that the syllable represented in the Syllable for Portuguese.
5. **Process:** when relevant, indicates the transformation process that occurs between Target (the Portuguese) and Source (the creole). Each cell can receive none or one of these three labels:
  - a. **assimilation** (alteration of a phoneme in order to approach phonetically to that which is adjacent to it);
  - b. **dissimilation** (alteration of a phoneme in order to distance itself phonetically from that which is adjacent to it);
  - c. **metathesis** (transposition of a phoneme from one syllable to another).

The individualization of syllables in different cells allows us to apply a phonetic simplification algorithm<sup>25</sup> that will automatically provide two new columns (*TemplateSource* and *TemplateTarget*) with the syllable template in Portuguese and creole (e.g. CV vs CVC). This, in turn, will enable us to infer if the transfer from Portuguese to creoles reduces (or increases) the complexity in the constitution of syllables. The manual processing of the data in the spreadsheet had the goal of aligning syllables. It consisted of multiplying the rows associated with each word for the number of its syllables; filling in the aligned syllable strings in the proper fields; and assigning the absolute position of the syllable inside the word (Table 4).

Assigning a numerical value to the syllabic positions enables us to apply a simple algorithm that automatically calculates the syllable position (initial, medial, or final), either in creole (*SimplifiedPositionSource*) or in Portuguese (*SimplifiedPositionTarget*). Again, these variables are an output of data processing and allow learners to see an immediate by-product of their work.

It is also possible to add other processes to the *Process* column that can be detected without the need for manual annotation—for example, deletion (aphaeresis, syncope, apocope) and insertion (prothesis, epenthesis, proparalepsis).

#### *4.2 Enhancing Script Processing with Python and RegEx*

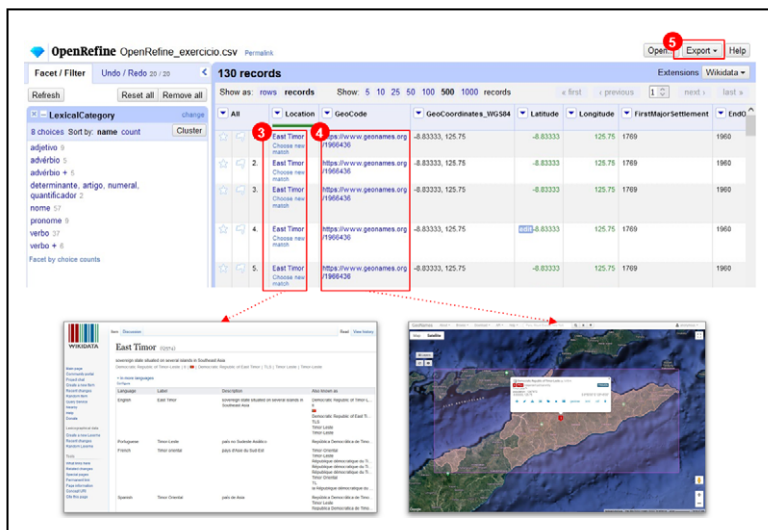
As an example of spreadsheet limitations, we observed that character replacement could be unfeasible with nesting formulas with larger numbers of characters to be replaced. A gradual step for more complex text treatment was to explore the use of RegEx in other creole-based projects,<sup>26</sup> as well as to provide and explore with students basic Python scripts for data transformation, such as using mapping tables for multiple substitutions. For this task, we used

**Table 3.** Example of table after manual and automatic processing for words.

Source- WordSyl	Target WordSyl	Source- Word Length	Target- Word Length	Source- WordSyl Length	Target- WordSyl Length	Word difference	Word bigger	Syl difference	Syl bigger
ka.tfo.ru	ke.'fo.ru	6	3	3	0	Equal	Equal	0	Equal
dʒa	ʒa	2	1	1	0	Equal	Equal	0	Equal
'tju.a	'ju.ve	3	4	2	-1	Target	Target	0	Equal
'tje.ru	'fej.cu	4	5	2	-1	Target	Target	0	Equal
n.tji.du	ẽ.'ji.du	5	6	3	-1	Target	Target	0	Equal

**Table 4.** Example of table after manual and automatic processing of syllables.

Source WordSyl	Target WordSyl	Source- Syllable	Target Syllable	Position Source	Position- Target	Position- Source Simple	Position- Target Simple	Previous SourceSyl- lable	Next SourceSyl- lable	Previous TargetSyl- lable	Next Tar- getSyl- lable	Process
ma.ɔ̃ʒi.s̃t̃	ma.t̃di.s̃w̃	's̃w̃	's̃t̃	3	3	Final	Final	ɔ̃ʒi	3	di	3	
go.vo.'j̃ẽ	vir.'go.je	vir	go	1	1	Initial	Initial	0	vo	0	'go	metathesis
go.vo.'j̃ẽ	vir.'go.je	'go	vo	2	2	Medial	Medial	go	'j̃ẽ	vir	je	
go.vo.'j̃ẽ	vir.'go.je	je	'j̃ẽ	3	3	Final	Final	vo	3	'go	3	
'b̃ẽ.ŋ.ku	'br̃ẽ.ku	'br̃ẽ	'b̃ẽ:ŋ	1	1	Initial	Initial	0	ku	0	ku	
'b̃ẽ.ŋ.ku	'br̃ẽ.ku	ku	ku	2	2	Final	Final	'b̃ẽ:ŋ	2	'br̃ẽ	2	



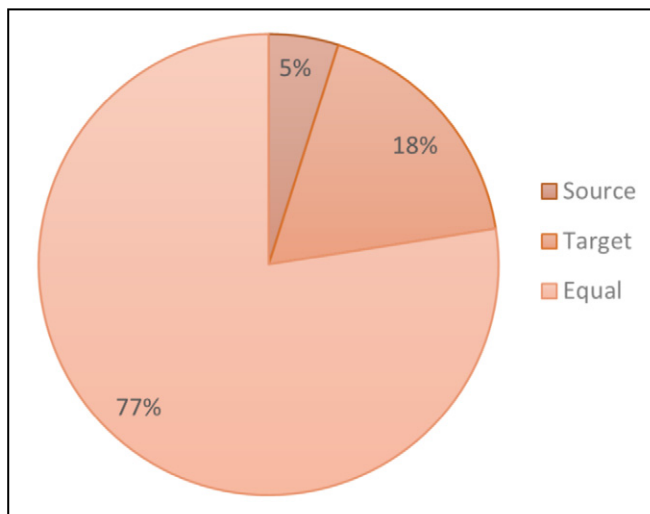
**Figure 3.** Using OpeRefine to reconcile the data with Wikidata and GeoNames, and extract further data.

Python notebooks that enable visual and modular organization, which are easier to explain. For implementing the notebooks, we used the online Google Colaboratory (a hosted Jupyter Notebook service) to avoid the hassle of installing Python and other required libraries on local computers. The cases that required string replacement were the IPA character normalization to APiCS standard and the format conversion from IPA to CV (Consonants and Vowels).

### 5. MAKING DATA AVAILABLE AND CONNECTED

In the spirit of Open Science and Citizen Science,<sup>27</sup> the students had a hands-on session with OpenRefine.<sup>28</sup> This free and open-source software was mainly used for: (1) data wrangling, specifically, data cleaning, transforming, and filtering; and (2) data reconciliation and matching.

Reconciling data also enabled to introduce the students to Linked Open Data (LOD) and the FAIR principles (findable, accessible, interoperable, and reusable)—in particular, the I2 ‘(meta)data use vocabularies that follow the FAIR principles’, the I3 ‘(meta)data include qualified references to other (meta)data’ and the R1.3 ‘(meta)data meet domain-relevant community standards’.<sup>29</sup> Hence, their data became more consistent, namely, by extending it to other publicly available web services, external data, knowledge bases and persistent identifiers (PIDs)—for example APiCS, Wikidata for lexemes, and GeoNames for geographical places and coordinates (Figure 3).<sup>30</sup> During this entire process, every student integrated a GitHub team and project,



**Figure 4.** Word-size differences (number of syllables) between creole (Source) and Portuguese (Target) – generated by a student.<sup>31</sup>

where they all published their data and continuously kept updating it ([https://github.com/Portophon/CreoPhonPt\\_classes](https://github.com/Portophon/CreoPhonPt_classes)).

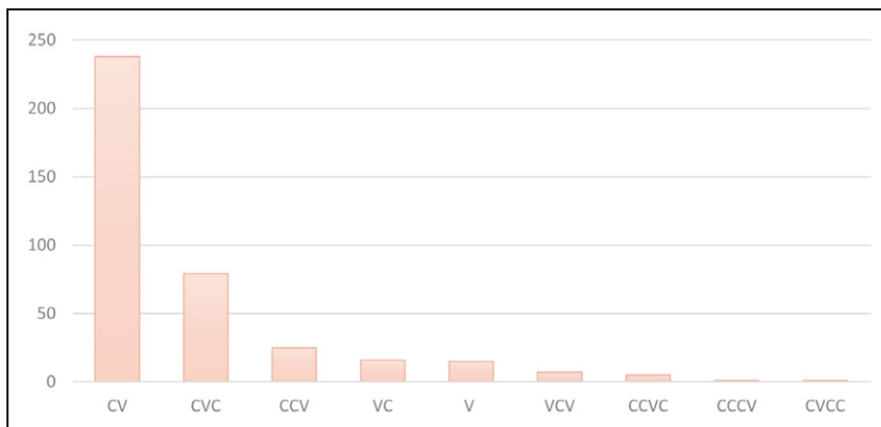
## 6. RESEARCH OUTCOMES

For presenting Open Science concepts, students were also introduced to GitHub as a public repository where they can collaboratively publish their data and access the scripts that were prepared for them. To this end, each student had to deliver four tables:

1. *creole\_metadata.csv*: raw language metadata.
2. *creole\_data.csv*: raw phonological data after collection.
3. *creole\_words.csv*: phonological data with word manual and automatic processing.
4. *creole\_syyls.csv*: phonological data with syllable manual and automatic processing.

Moreover, students received an automatic record of their contributions and could easily reference their work (and versioning) for a scientific impacting project.

The students were also provided with some knowledge about descriptive statistics and visualisation techniques that could be applied in the spreadsheets. To complete the evaluation process, the students presented and published their own work based on a bibliographic collection regarding the findings from their data analysis. Figures 4 and 5 show some results presented by students in



**Figure 5.** Syllable patterns from a creole language visualization – generated by a student.<sup>32</sup>

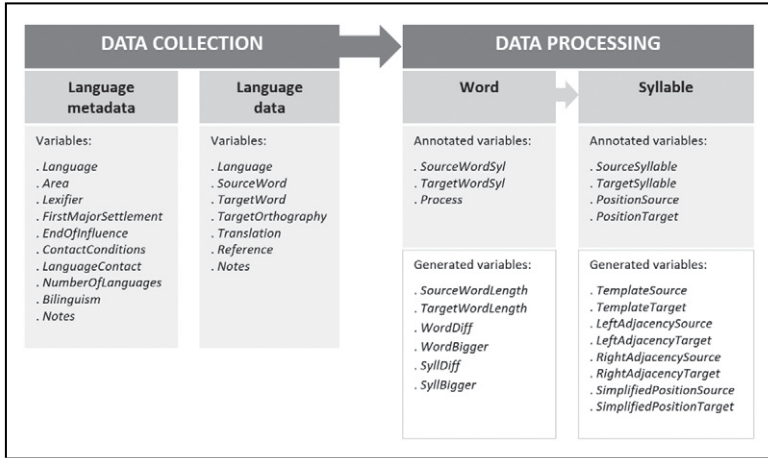
their final essays.<sup>33</sup> It is worth mentioning that patterns with very low frequency may indicate that there could have been some mistake in string processing – hence students checked the words with the outlier pattern and confirmed that they were well processed.

## 7. CONCLUSION

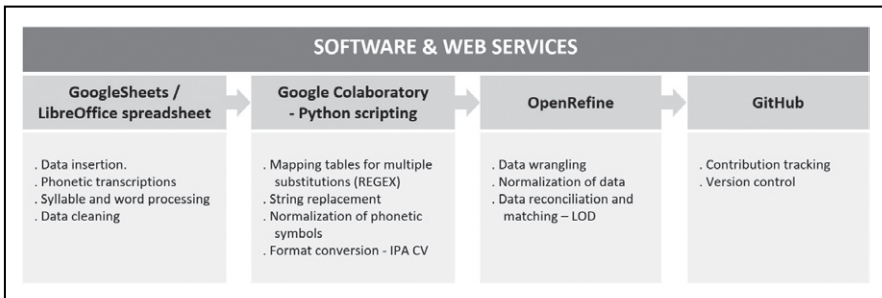
This project reached several results that were very encouraging. The phonology class became more, dynamic and students reported a very positive evaluation of the method. The data output has also a great value for scientific research in an underexplored area – that is, Creolistic Studies. Finally, we believe this may be very useful for the CreoPhonPt leading researchers who also acquired very useful insights regarding the workflow (Figure 6), as well as the methods and tools (Figure 7) for future research and teaching integration in linguistics and the digital humanities.

We have already started measuring distances between creole languages and the lexifier. Distance, clustering and connected visualization techniques were already used in the Romance Linguistics class from the Language Science bachelor degree at FLUP, to demonstrate its potential use for improving comparative studies between languages.

As a study of phonetic change, Creolistic Studies may also contribute to the fields of language acquisition and learning. We are also exploring the potential use of this database beyond academia to develop creole orthographic norms, which are a crucial tool for language revitalization and preservation.<sup>34</sup>



**Figure 6.** Workflow for collecting data and preparing lexical data of creole languages for a medium-scale comparative phonological analysis, according to the CreoPhonPT prototype.



**Figure 7.** Freeware software and web services used during the demonstration project.

As creoles are born from unique situations of contact, their birth and development can never be fully understood without historical, geographical, ethnological and genetic data.<sup>35</sup> These data should be included in CreoPhonPt in the near future. As such, conversations with sociology researchers to collect phonological and linguistic data within African communities in Porto, Portugal, have just started taking place.

#### ACKNOWLEDGEMENTS


The CreoPhonPt repository has been created within Carlos Silva's PhD project 'Consonant stability of Portuguese-based creoles', which was funded by the Portuguese Foundation for Science and Technology (FCT) (SFRH/BD/2020.07466.BD) and supported by the Center of Linguistics of the University of Porto (FCT-UIDB/00022/2020). The Centre for Digital Culture and Innovation (CODA) is also funded by the FCT, under the CEECINST/00050/2021 contract programme. Special thanks are extended to Steven Moran for his help in defining the data structure of CreoPhonPt at its early stages. We thank the anonymous reviewers for their valuable comments and suggestions, which have helped improve the quality of this article.

#### SUPPLEMENTARY DATA

The data collection to feed into the CreoPhonPt repository and data analysis training (including scripts and conversion tables) can be found online at [https://github.com/Portophon/CreoPhonPt\\_classes](https://github.com/Portophon/CreoPhonPt_classes) under a GNU General Public License Version 3, and cited as supplementary material to this article.

#### ORCID

Luís Trigo  <https://orcid.org/0000-0002-3772-7081>

Carlos Silva  <https://orcid.org/0000-0002-8052-4271>

Vera Moitinho de Almeida  <https://orcid.org/0000-0003-4979-8247>

#### END NOTES

- <sup>1</sup> S. Wichmann and E. W. Holman, *Temporal stability of linguistic typological features* (Munich, 2009).
- <sup>2</sup> S. J. Greenhill, C. H. Wu, X. Hua, M. Dunn, S. C. Levinson and R. D. Gray, 'Evolutionary dynamics of language systems', *Proceedings of the National Academy of Sciences (PNAS)*, 114, no. 42 (2017), E8822–E8829, <https://doi.org/10.1073/pnas.1700388114>.
- <sup>3</sup> C. Silva and S. Moran, 'Stability drivers in the emergence of Portuguese-based creoles', *The Fifth Edinburgh Symposium on Historical Phonology. Abstracts booklet* (Edinburgh, 2021), <http://www.lel.ed.ac.uk/symposium-on-historical-phonology/pdf/eshp5-abbk.pdf>, last accessed 13 November 2023; C. Silva and S. Moran, CreoPhonPt 1.0 [Dataset] (2022); C. Silva, *Consonant stability of Portuguese-based creoles* (Porto, 2023). <https://doi.org/10.5281/zenodo.7575862>, last accessed 13 November 2023.
- <sup>4</sup> L. Trigo and C. Silva, 'PtIanka: an online corpus of Sri Lanka Portuguese lexicon and phonology', unpublished paper presented at *OpenCor*, 3 December 2021.
- <sup>5</sup> M. D. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, A. Gonzalez-Beltran, A. Gray, P. Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hoof, T. Kuhn, R. Kok, J. Kok, S.



- Lusher, M. Martone, A. Mons, A. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3 (2016), 160018, <https://doi.org/10.1038/sdata.2016.18>.
- <sup>6</sup> D. Vrandečić and M. Krötzsch, 'Wikidata: a free collaborative knowledgebase', *Communications of the ACM*, 57, no. 10 (2014), 78–85.
- <sup>7</sup> S. Bird, 'Computational phonology', *arXiv* (2002), 1–4, <https://doi.org/10.48550/arXiv.cs/0204023>.
- <sup>8</sup> Trigo and Silva, 'Ptlanka: an online corpus of Sri Lanka Portuguese lexicon and phonology'; C. Silva and L. Trigo, 'Exploring consonant frequency in Sri Lanka Portuguese', *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022)* (Fortaleza, 2022), 1–6, <https://ceur-ws.org/Vol-3128/paper12.pdf>, last accessed 13 November 2023.
- <sup>9</sup> D. Li, Z. Chunling and H. Yuanjian, 'Project-based learning in teaching translation: students' perceptions', *The Interpreter and Translator Trainer*, 9, no. 1 (2015), 1–19.
- <sup>10</sup> K.-L. A. Yau, H. J. Lee, Y.-W. Chong, M. H. Ling, A. R. Syed, C. Wu and H. G. Goh, 'Augmented Intelligence: surveys of literature and expert opinion to understand relations between human intelligence and artificial intelligence', *IEEE Access*, 9 (2021), 136744–61.
- <sup>11</sup> C. Silva and S. Moran, *CreoPhonPt 1.0*; C. Silva, *Consonant stability of Portuguese-based creoles* (Porto, 2023). <https://doi.org/10.5281/zenodo.7575862>, last accessed 13 November 2023.
- <sup>12</sup> P. de Lacy, *The Cambridge handbook of phonology* (Cambridge, 2007), <https://doi.org/10.1017/CBO9780511486371>; D. Odden, *Introducing phonology*, 2nd ed. (Cambridge, 2013), <https://doi.org/10.1017/CBO9781139381727>.
- <sup>13</sup> S. G. Thomason and T. Kaufman, *Language contact, creolization, and genetic linguistics* (Berkeley, CA, 1988), <https://doi.org/10.1525/9780520912793>.
- <sup>14</sup> S. Romaine, *Pidgin and creole languages* (London, 1988), <https://doi.org/10.4324/9781315504971>.
- <sup>15</sup> S. M. Michaelis, P. Maurer, M. Haspelmath and M. Huber, *Apics Online* (Leipzig, 2013), <https://apics-online.info/>, last accessed 13 November 2023.
- <sup>16</sup> H. C. Cardoso, T. Hagemeyer and N. Alexandre, 'Crioulos de base lexical Portuguesa', in *Manuel des anthologies, corpus et textes romans* (Berlin and Boston, MA, 2015), 670–92, <https://doi.org/10.1515/9783110333138-043>.
- <sup>17</sup> P. Bakker, A. Daval-Markussen, M. Parkvall and I. Plag, 'Creoles are typologically distinct from non-creoles', *Journal of Pidgin and Creole Languages*, 26 (2011), 5–42, <https://doi.org/10.1075/jpcl.26.1.02bak>.
- <sup>18</sup> I. Maddieson, 'Borrowed sounds', in J. A. Fishman, ed., *The Fergusonian impact* (Berlin and Boston, MA, 1986), 1–16, <https://doi.org/10.1515/9783110873641-002>.
- <sup>19</sup> S. Moran, *Phonetics information base and lexicon* (Washington, DC, 2012), <https://digital.lib.washington.edu/researchworks/handle/1773/22452>.
- <sup>20</sup> M. Swadesh, 'Towards greater accuracy in lexicostatistic dating', *International Journal of American Linguistics*, 21, no. 2 (1955), 121–37.
- <sup>21</sup> M. Dunn, 'Language phylogenies', in C. Bowern and B. Evans, eds, *The Routledge handbook of historical linguistics* (London, 2014), <https://doi.org/10.4324/9781315794013.ch7>.
- <sup>22</sup> D. A. Grier, *When computers were human* (Princeton, NJ, 2013); P. Naur, *Computing: a human activity* (New York, 1992).
- <sup>23</sup> J. H. McWhorter, 'The world's simplest grammars are creole grammars', *Linguistic Typology*, 5, no. 2–3 (2001), <https://doi.org/10.1515/lity.2001.001>.

- <sup>24</sup> N. Gurevich, *Lenition and contrast: the functional consequences of certain phonetically conditioned sound changes* (Boca Raton, FL, 2004).
- <sup>25</sup> C. Silva and L. Trigo, 'Phonetic simplification for automatic syllable division in Sri Lanka Portuguese', *Demonstration Session at the 15th Edition of the International Conference on Computational Processing of Portuguese (PROPOR)*, Fortaleza, Brazil, 21–23 March 2022.
- <sup>26</sup> Silva, 'Phonetic simplification for automatic syllable division in Sri Lanka Portuguese'.
- <sup>27</sup> Foster, *Fostering the practical implementation of Open Science in Horizon 2020 and beyond* (2017), <https://www.fosteropenscience.eu/>, last accessed 13 November 2023; European Commission, *Open Science* (2019-), [https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science\\_en](https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en), last accessed 13 November 2023; R. Kerson, 'Lab for the environment', *MIT Technological Review*, 92, no. 1 (1989), 11–12; Societize and European Commission, *Green paper on Citizen Science. Citizen Science for Europe: towards a society of empowered citizens and enhanced research* (2014), [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=4122](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=4122), last accessed 13 November 2023.
- <sup>28</sup> OpenRefine, *OpenRefine. Code for science and society* (2012–), <https://openrefine.org/>, last accessed 13 November 2023.
- <sup>29</sup> T. Berners-Lee, *Linking Open Data* (2008), <https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>, last accessed 13 November 2023; Bird, 'Computational phonology'.
- <sup>30</sup> Michaelis, <https://apics-online.info/>; Wikimedia Foundation, *Wikidata* (2012–), <https://www.wikidata.org/>, last accessed 13 November 2023; F. Nielsen, 'Lexemes in Wikidata: 2020 Status', *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (online, 2020), 82–6; GeoNames, *GeoNames* (2005), <http://www.geonames.org/>, last accessed 13 November 2023.
- <sup>31</sup> Pinto, *O crioulo de Sri Lanka e as suas propriedades fonológicas*.
- <sup>32</sup> Pinto, *O crioulo de Sri Lanka e as suas propriedades fonológicas*.
- <sup>33</sup> M. Pinto, *O crioulo de Sri Lanka e as suas propriedades fonológicas (Phonology I–MSc. Linguistics)* (Porto, 2023).
- <sup>34</sup> H. Rutkowska, 'Orthography', in L. J. Brinton and A. Bergs, eds, *The history of English: historical outlines from sound to text* (Berlin and Boston, MA, 2017), 200–17, <https://doi.org/10.1515/9783110525281-011>.
- <sup>35</sup> T. Hagemeyer and J. Rocha, 'Creole languages and genes: the case of São Tomé and Príncipe', *Faits de Langues*, 49, no. 1 (2019), 167–82.