Qumín : from a bundle of scripts to a scientific toolkit

Sacha Beniamine[1] Jules Bouton[2]
(1) Surrey Morphology Group, School of Literature and Languages, University of Surrey, Guildford, GU2 7XH, United Kingdom
(2) Université Paris-Cité, Laboratoire de Linguistique Formelle, CNRS, 8 Place Paul Ricœur, 75013 Paris, France
s.beniamine@surrey.ac.uk, jules.bouton@u-paris.fr

MOTS-CLÉS : morphologie quantitative, bonnes pratiques, ingéniérie logicielle, python.

KEYWORDS: quantitative morphology, good practices, software development, python.

Scientific software development involves finding a balance between the demands of fundamental research and code maintenance. This contribution focuses on the example of *Qumín*, a python package for studying inflectional paradigms. *Qumín* was initially released in 2018 as a collection of scripts supporting a computational linguistics PhD thesis. While this release was only meant as a guarantee for replicability, the toolbox encountered an unexpected popularity among morphologists. As new features started to accumulate, it became necessary to reorganize the code base and to create a cleaner environment for contributors. The history of *Qumín*'s development illustrates the challenges that researchers may face while sharing their code : making the code easily reusable by others, finding a balance between new features and long-term stability, ensuring long term maintenance beyond initial funding. We discuss two facets of *Qumín*'s recent history : first, the scientific benefits of open-source software (1) ; second, the important amount of engineering required to achieve this (2). We conclude with some perspectives on *Qumín*'s future (3).

1 *Qumín* for linguists

Qumín (**Qu**antitative **M**odelling of **In**flection) is a software dedicated to studying inflectional morphology, and more specifically paradigm structure. The core of *Qumín* is an algorithm for inferring morphologically meaningful analogical patterns from pairs of forms. Further analytical steps take analogical patterns as an input to provide various insights : measures of morphological complexity using Shannon's entropy (Ackerman *et al.*, 2009; Bonami & Beniamine, 2016), clustering of lexemes in hierarchical inflection classes (Beniamine, 2021), etc. The initial success of *Qumín* was due to the ease with which it allowed for quantitative measurements of inflectional systems, requiring only unsegmented paradigms and a description of the phonemes. Its ability to model similarly very different morphological systems (whether concatenative or not, affixal or suprasegmental) made it extremely

useful for morphological typology. All of its outputs (analogical patterns, measurements, inflectional classifications) are easily interpretable in linguistic terms. This allows linguists to pin-point the exact sources and mechanisms leading to structural properties. After being used to compare Russian, French, English, Chatino, Arabic, Navajo and Portuguese in Beniamine's PhD (Beniamine, 2018), *Qumín* was employed for monolingual descriptions of the inflectional systems of Latin (Pellegrini, 2023), Portuguese (Beniamine *et al.*, 2021), Catalan (Herce & Pricop, 2024a), Romanian (Herce & Pricop, 2024b), Pame (Herce, 2024). Studies beyond Romance are on-going. In other words, *Qumín* allowed researchers less familiar with programming to use quantitative methods to investigate morphology.

Since its first release, *Qumín* had to be updated to support this successful uptake by improving its FAIR compliance (Wilkinson M. D. et al, 2016). While its first version relied on ad-hoc file structures for paradigms and sound descriptions, version 2.0 dropped support for this format in favour of the Paralex standard (Beniamine *et al.*, 2023)[1]. Paralex standardizes machine-readable, morphologically rich inflected lexicons. This enabled a straightforward pipeline from dataset creation to its analysis, bringing significant improvements in research worflow. It is not a surprise for instance that novel Paralex datasets were released simultaneously with an analysis based on *Qumín* (Portuguese, Catalan, Romanian, Pame). Second, version 2.0 ships computation results as a frictionless DataPackage, which includes an extensive metadata description of the inputs and outputs (Fowler *et al.*, 2017). This design choice makes downstream consumption of computation outputs easier. Furthermore, the DataPackage includes all the configuration settings and command line arguments that were passed to *Qumín*, thus enabling strong reproducibility.

Finally, the increasing usage of *Qumín* led to numerous new features, which illustrate the virtuous circle of back-and-forth interaction between research and software development. A notable addition enabled fixed predictors in complexity measures, an addition which supported the work of a PhD dissertation (Pellegrini, 2023).

2 *Qumín* for developers

While the scientific success of *Qumín* made it a promising tool, its backbone, in 2018, was not appropriate for long-term stability. It was made available as a collection of scripts more than as a consistent package, which made further evolutions less obvious. During the subsequent years, the code structure was slowly improved. Starting from 2023, development became more active, with the removal of exploratory features that became unnecessary and made maintenance more difficult. At the same time, several modules were rewritten in an object-oriented approach to increase the code modularity and to increase the ability to handle non-canonical morphology, in particular defectiveness and overabundance. Finally, as the amount of available morphological datasets started to grow (21 lexicons to date, covering 8

1. https://paralex-standard.org/

language families), it became necessary to introduce multiprocessing. This resulted in an enormous reduction of computation times, opening the path to large scale comparisons.

While these improvements were welcome, massive changes are also likely to introduce new bugs and issues. The second step in securing the code-base involved the development environment. After the project had been moved to Gitlab, the continuous development pipeline was reinforced with a set of integration tests to ensure stability of the main scripts. Unfortunately, to date, unit tests only cover a small part of the code : future plans have been laid to extend them systematically.

Although all these changes were necessary and beneficial scientifically, they were also time-consuming. Unfortunately, and this type of work tends to be valued less than production of new scientific outputs. Beyond the initial PhD project (ending in 2018), no specific funding has targeted *Qumín*'s extension and maintenance. Yet, continued engineering support is essential for the long-term success of linguistic software. To mitigate this, *Qumín* was recently archived on Zenodo[2] and on Software Heritage, making it more easily citable.

3 The future of *Qumín*

Although *Qumín* has now become a mature research software, stable and accessible to users not familiar with programming, it is not clear how it will evolve in the future. There are at least two potential scientific perspectives. First, the fast growth of the Paralex ecosystem makes it increasingly feasible to employ *Qumín* to conduct typological studies with a high coverage. Second, the massive changes in the code structure made *Qumín* ready to work with non-canonical morphological phenomena in a more meaningful way.

Simultaneously, there are some clear scientific limits to how *Qumín* works : the patterns module, which inherits a reimplementation of minimal generalization (Albright & Hayes, 2003), sometimes over- or under-generalizes ; the complexity measures, being holistic by design, fail to capture orthogonal complexity layers (eg. cross-cutting affixal and supra-segmental morphology). Detailed plans have been laid out to improve the situation by implementing (a) novel analogical patterns enabling better treatement of tiered phenomena while maintaining interpretability ; (b) a radical review of the pattern selection algorithm, favoring standard machine-learning classifiers ; (c) further extensions of the complexity measures, beyond conditional entropy, to integrate analyses of non-canonical phenomena and facilitate human interpretation.

To summarize, on-going development efforts on *Qumín*'s code and data environment had clear scientific benefits, making it a popular tool for quantitative morphology. The next challenge for *Qumín*'s development will be to ensure increased stability while rewriting some of the central algorithms to enabling novel analyses.

2. https://doi.org/10.5281/zenodo.15008373

Références

ACKERMAN F., BLEVINS J. P. & MALOUF R. (2009). Parts and wholes : Implicative patterns in inflectional paradigms. In J. P. BLEVINS & J. BLEVINS, Éds., *Analogy in grammar : form and acquisition*, Oxford linguistics, p. 54–82. Oxford : OUP.

ALBRIGHT A. C. & HAYES B. P. (2003). Rules vs. analogy in english past tenses : A computational/experimental study. *Cognition*, **90**, 119–161.

BENIAMINE S. (2018). *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. Thèse de doctorat, Université Sorbonne Paris Cité.

BENIAMINE S. (2021). One lexeme, many classes : Inflection class systems as lattices. In B. CRYSMANN & M. SAILER, Éds., *One-to-many-relations in morphology, syntax, and semantics*, volume 7 de Empirically Oriented Theoretical Morphology and Syntax, p. 23–51. Language Science Press. DOI : 10.5281/ZENODO.4729789.

BENIAMINE S., ANDERSON C., CARROLL M., GUZMÁN NARANJO M., HERCE B., PELLEGRINI M., ROUND E., SIMS-WILLIAMS H. & TRESOLDI T. (2023). Paralex : a DeAR standard for rich lexicons of inflected forms. In *International Symposium of Morphology (ISMo 2023)*, Nancy, France.

BENIAMINE S., BONAMI O. & LUÍS A. R. (2021). The fine implicative structure of European Portuguese conjugation. *Isogloss. Open Journal of Romance Linguistics*, **7**, 1–35. DOI : 10.5565/rev/isogloss.109.

BONAMI O. & BENIAMINE S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure*, **9**(2), 156–182. DOI : 10.3366/word.2016.0092.

FOWLER D., BARRATT J. & WALSH P. (2017). Frictionless Data : Making Research Data Quality Visible. *International Journal of Digital Curation*, **12**(2), 274–285. DOI : 10.2218/ijdc.v12i2.577.

HERCE B. (2024). VeLePa : Central Pame verbal inflection in a quantitative perspective. *Morphology*, **34**(3), 281–319. DOI : 10.1007/s11525-024-09426-x.

HERCE B. & PRICOP B. (2024a). VeLeCa : A verbal lexicon of Catalan with PCFP analysis. *Isogloss. Open Journal of Romance Linguistics*, **10**(1), 1–17. DOI : 10.5565/rev/isogloss.457.

HERCE B. & PRICOP B. (2024b). VeLeRo : an inflected verbal lexicon of standard Romanian and a quantitative analysis of morphological predictability. *Language Resources and Evaluation*. DOI : 10.1007/s10579-024-09721-3.

PELLEGRINI M. (2023). *Paradigm Structure and Predictability in Latin Inflection*. Volume 6 de Studies in Morphology. Cham : Springer International Publishing. DOI : 10.1007/978-3-031-24844-3.

WILKINSON M. D. ET AL (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**(1), 160018. DOI : 10.1038/sdata.2016.18.