Documentation of FRITAV and RhoSquared v1.0.0

Eduard S. Lukasiewicz
(Georg-August-Universität Göttingen)

Sascha Gaglia
(Freie Universität Berlin)

Last updated: February 5, 2023

Contents

1 Introduction

The FRITAV – an acronym for *FRench and ITalian Analogies in Verbs* – database was created by Sascha Gaglia on the basis of data from a series of historical corpora of French and Italian, i.e. *Nouveau Corpus d'Amsterdam* (*NCA* in what follows; see Stein et al. 2006), *Frantext intégral* (*FI*, see "Base textuelle Frantext, 1.3.24" 2022) and *OVI-Gattoweb* (see Larson and Artale 2005). More specifically, it is a morphophonological database of certain verbs as found in Old French and Old Italian. RhoSquared is a Python script with a graphical user interface (GUI) used to communicate with the FRITAV database.

Both the FRITAV database and RhoSquared emerged in the context of the project *Temporal analysis and modelling of the paradigmatic extension of French and Italian verbal roots* funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft*, short DFG).

The aims of the project were the 1. documentation of the direction of analogies with respect to French and Italian verbal roots in the sense of paradigm cells, 2. analysis of the temporal dimension of paradigmatic analogies in terms of time spans, and 3. analysis of factors which determine the direction and the temporal dimension (i.e. accelerating or delaying) of analogies especially with respect to paradigmatic patterns such as so-called morphomes (cf. Aronoff 1994, Maiden 2018).

FRITAV was planned as a free open source to which linguists, once they have registered, may contribute with examples.

The people who directly contributed to this project are (chronologically): Manuel Möll, Noel Seger, Pascal Hornbergs, Eduard S. Lukasiewicz. The project, which was settled at the Department of Romance Philology of the Georg August University in Goettingen, ran its course from October 1st, 2017 to September 30th, 2020.

For results of the project see especially Gaglia (2020).

2 FRITAV

2.1 Theoretical Background

The history of Romance verbs is characterised by a great amount of analogical extensions of verbal root allomorphs (see the literature in Gaglia 2020, which shows that the investigation of analogies in French, especially with respect to levelling, has interested linguists and philologists since the early days of Romance philology understood as a scientific discipline). Some of the paradigms involved were levelled, e.g., Old French (henceforth OFr.) *trover* (Mod. Fr. *trouver*) 'find', *amer* (*aimer*) 'love' and *plorer* (*pleurer*) 'cry'. Other verbs like *venir* 'come' and *tenir* 'have/hold' and the corresponding Old Florentine Italian (henceforth OIt.) *venire* and *tenere* still show some significant allomorphic patterns in Modern French and Modern Standard Italian (Mod. St. It.). The paradigms mentioned here all have in common that, at some point of their development, they displayed distributional properties which can be associated with so-called morphomes: For example, OFr. *trover* followed (in terms of distribution) a morphomic L-pattern with respect to its [s]-final root *truis(s)-* in the 1SG PRES.IND and throughout the Present Subjunctive.

The investigation of analogies constitutes a longstanding tradition in inflectional morphology. Contributions that treat analogies with respect to wider areas within an inflectional paradigm (and not only as proportions between single forms) were made early (see, among others, Kurylowicz 1945-49, on levelling from Old to Middle French), but it was especially the focus on the morpho-syntactic paradigm as an object of theoretical linguistic research (see Carstairs 1987, among others) that intensified the research on different kind of patterns that serve as domains for analogical processes – see Blevins and Blevins (2009), who define the paradigm as the "central locus of analogy in grammar". Some of the major contributions that influenced, either directly or at least to a considerable extent, the field of analogical patterns are Bybee 1985, where they are treated from a frequency-based perspective, Aronoff's seminal work on so-called 'morphomes' (see Aronoff 1994; see also Maiden 2018 on morphomes in Romance languages) and Stump (2001), where paradigmatic patterns are consistently formalized.

From the literature it is well known that morphomes are prototypical domains for analogies and that they are relatively stable entities which can attract verbs which do not yet adhere to a particular distribution (Maiden 2003, 2018). Morphomes are understood as arbitrary distributions of morphological elements in a paradigm at the synchronic level, which means that their distributions can neither be explained semantically nor phonologically, although they can be originally motivated by phonology (see Aronoff 1994, who coined the term 'morphome').

As argued in Gaglia (2020), based on their stability, one expects that morphomes will contribute to analogical changes not only in the distributional-directional dimension but also in the temporal dimension of dynamics, i.e. potentially slowing down analogical levelling. Quantitative analyses of analogical extensions based on Old French corpus data are still underrepresented, which is the reason why it is still difficult to unveil the 'path' of each extension in terms of paradigmatic patterns, sub-paradigms and paradigm cells, a fact which also holds for Old Italian. One of the aims of this project was, therefore, to provide the starting point for a unified open source ?? data base, which enables linguists and philologists to contribute to this database on the grounds of different sources such as diachronic sources and to ease the problem of availability of data.

The data included in FRITAV shall provide information for each token about the following issues: a) general information on the root, b) linguistic characteristics, c) information regarding the source, and d) information about the collaborator who integrated the token into the FRITAV-database. In the following paragraphs this information shall be explained.

Warning: As of February 5, 2023, the actual names in RhoSqrd can vary slightly from those listed here. It should however be clear what corresponds to what.

a) General information

- `Language`, i.e French or Italian
- `Lemma (mod.)` (where 'mod' stands for 'modern'): gives information about the lemma in Modern French or Modern Italian, e.g. *trouver* (= Mod. French)

- `Lemma (hist.)` (where 'dia' stands for 'diachronic'): gives information about the lemma in Old French or Old Italian, e.g. *trover* (= Old French)
- `Verb Form (mod.)` : gives a verb form in Modern French or Modern Italian, e.g. *trouve* (Old French)
- `Verb Form (hist.)`: gives a verb form in Old French or Old Italian, e.g. *truis* (Old French)
- `Stem (hist.)`: gives the stem of the verb form in Old French or Old Italian, e.g. *truv-*

b) Linguistic characteristics

- `POS_morphology` (where 'pos' stands for 'part of speech' and 'm' for 'morphological'): give information about the morphological features of a verb form, e.g. pos=VER_fut_3sg (= future indicative, 3rd person singular)
- `POS_morphology (alt.)` (where 'alt' stands for 'alternative'): this information is only given if `POS_morphology` (see above) are not unambiguous
- `Orth. Context` (where 'orth.' stays for 'orthographic'): give information about the orthographic string of the whole verb form in a generalized fashion and in terms of features known from phonology, e.g. plosiv.rhotic.vowel.vowel.fricative (representing *truis*, *truef* etc.), plosiv.rhotic.vowel.vowel.fricative-vowel.fricative (representing *trueves*, *trouvez* etc.); morpheme boundaries are given as hyphens
- `Morph. Phenom.` (where 'phenom' stands for 'phenomenon'): information is given about a morpholo-phonological phenomenon displayed by the token such as 'diphthong'

c) Source

- `Date (comp.)`: gives information about the year of the composition including the given token, e.g. dateComposition="1213ca" (where 'ca' stands for 'circa')
- `Location (comp.)`: gives information about where the text including the token was composed, e.g. fior. (where 'fior.' stands for 'fiorentino', i.e. Florentine)
- `Date (manuscr.)`: gives information about the year of the manuscript including the given token, e.g. dateComposition="1213ca" (where 'ca' stands for 'circa') ['manuscript:dates' is only available for the Old French data]
- `Location (manuscr.)`: gives information about where the text including the token was composed, e.g. fior. (where 'fior.' stands for 'fiorentino', i.e. Florentine) ['manuscript:localization' is only available for Old French]
- `Reg. Codes (Dees)`: gives information about the Regional codes provided by Dees (1980) ['Regional codes_Dees' is only available for Old French]
- `Genre`: gives information about the genre of the text including a given token, e.g. lir./ lirico / lyrisch

- `Verse`: gives information about the structure of the text including a given token, i.e. if its written in verses ('yes') or in prose ('no')
- `Orig. DB Id`: gives information about the code regarding the original database of the token, e.g. ovi:Ritmo lucchese_v.43,48.7 (where 'ovi' stands for 'Opera del Vocabolario Italiano (Online)')
- `Token Sentence`: offers context in which the token is embedded, e.g. *Alli altri affagr ogn'om tenrà* (where 'tenrà' is the token)

The original version of FRITAV (published 15.09.22) includes data from the databases described in the next sections.

Nouveau Corpus d'Amsterdam

The Nouveau Corpus d'Amsterdam (NCA; Stein et al. 2006, see the chapters in Kunstmann and Stein 2007) is an annotated digital text corpus consisting of 3,2 million words covering Old French, i.e. the ancestor of Modern French spoken in the period from, roughly, 1150 to 1350. It represents an extended, modified and digital version of the *Atlas des formes linguistiques des textes litteraires de l'ancien français* provided by Dees (1987). The NCA itself does not provide morphological segmentation, i.e. verb roots are not categorized as such. Hence, we transferred data from the NCA and modified it according to our purposes. Data from the NCA given in FRITAV include the abbreviation 'nca' in the signature of each example. FRITAV includes data from all varieties represented in NCA. 2,982 tokens were extracted from NCA an included in FRITAV.

OVI-Gattoweb

The Corpus OVI dell'italiano antico (OVI; see Larson and Artale 2005) is a digital text corpus including 23,2 million words from Old Italian. The corpus is annotated but it lacks tagging. Therefore, we extracted manually the data we included in FRITAV. This extracted data concerns exclusively Old Florentine, which later on became Standard Italian. Data from OVI given in FRITAV include the abbreviation 'ovi' in the signature of each example. 3,879 tokens were extracted from OVI.

Frantext

Frantext intégral (FI; see "Base textuelle Frantext, 1.3.24" 2022) is a text corpus including more than 277 million words from different stages of French. The corpus is annotated but it lacks tagging. Data from Frantext given in FRITAV include 'Frantext' in the signature of each example. 2,172 tokens were extracted from Frantext and included in FRITAV. The time lapse is from 1100 to 1444 for `Date (comp.)` and 1225 to 1444 for `Date (manusc.)`

2.2 Implementation

The FRITAV data are stored using PostgreSQL, a popular free and open source relational database. We chose PostgreSQL based on its ease of use and availability. In addition to that,

we felt like PostgreSQL lends itself well to be used with a RESTful API (see 2.3 below).

Initially, we intended to store the data in XML format compliant with the TEI standard. However, we could not proceed due to technical limitations on the servers we would be using for hosting the data once published. The entirety of the FRITAV-RhoSqrd infrastructure was then built around PostgreSQL, the next best option. This actually facilitated the whole process thanks to PostgreSQL's server-side programming capabilities, which we used extensively while preparing the data for publication. As of now, however, the relational aspect of PostgreSQL is not fully exploited, and this remains a work-in-progress.

The data are accessed via an API based on PostgREST, a popular RESTful API framework for PostgreSQL. In the next section we present the FRITAV API and explain how one can use it to query the FRITAV database.

2.3 FRITAV API

The PostgreSQL implementation of FRITAV cannot be accessed directly over e.g. `psql` or object–relational mapping (ORM) frameworks out of safety concerns. We decided instead to use PostgREST as a RESTful API, so that data may be queried via HTTP requests. PostgREST is a Haskell framework that allow its users to quickly build RESTful APIs serving data in JSON format on top of existing PostgreSQL databases. It is simple to setup and it fits well into our infrastructure, so we did not deem it necessary to write an API from scratch.

It is possible to query FRITAV by directly sending an HTTP request. For example, if one would want to ask the API to fetch all verb forms of Old French *aimer* with stem *aing-*, it would suffice to paste `https://api.eslksw.dev/fritav/aimer?stem=fts.aing` in the browser of choice (we suggest to use the free Mozilla Firefox due its excellent JSON visualization capabilities out of the box). If on UNIX or an UNIXoid system, it is possible to use `curl` to save the result in a file or to pipe it into some JSON prettyprint script (e.g., the well-known `python3 -m json.tool`) and then save it. One could then read the JSON file into any programming language and analyze the newly fetched data.

There are however limitations to this approach. First off, it expects from the user to learn not only the database-internal terminology, but also how to send the right requests using PostgREST's language. For example, it is probably not immediately clear to the inexperienced user as to why we are sending a request using the prefix `fts` in the example above (it stands for *free text search* and PostgREST uses it to tell PostgreSQL to search for an arbitrary string in a given TEXT column). Second, visualizing the data requires a series of steps that for some may be too laborious, which in turn leads to a diminished usability of the database. These and other reasons are why we chose to write a simple front-end to FRITAV that uses the FRITAV API interactively, allowing any user to access the data in (we hope) the least complicated way possible. We called this front-end *RhoSqrd* and the few remaining sections are dedicated to its presentation.

3 RhoSqrd

RhoSqrd, which is currently available as a beta pre-release, has been developed as a front-end to the FRITAV database. It is a Python app with a Qt graphic user interface (GUI)[1]. We chose this combination for a series of reasons with varying degrees of importance, but the popularity of Python among linguists was a key point. We believe that Python has come to occupy a central position in computational linguistic research and it is, we believe, uncontroversial to state that most researchers and even advanced students have at least once come in extended contact with the programming language. This means that many people in the field could contribute to the project in order to make it both more powerful and more flexible, which could lead in the future to it having more functions and being used with other databases.

Flexibility also played a significant role in choosing the appropriate GUI framework. Inspired by other FOSS projects such as QGIS and KDE and due to its being easy to learn while retaining the power of Qt, we chose the Qt Modeling Language (short QML, not to be confused with the homonymous QGIS style language). QML is a declarative programming framework based on Qt and in essence it can be thought of as a markup language. There seems to have been a shift in programming patterns and one of them is the increasing use of declarative UI frameworks in place of the more traditional imperative ones due to the first ones being more intuitive to program (for a recent commentary on this shift see for example Peter Steinberger's article in Increment, Issue 18, August 2021). We felt that this was indeed the case when switching from Qt 5 (later Qt 6) to QML and we decided to keep the change and wrote RhoSqrd's whole GUI in QML.

Other than this, QML allows developers to embed JavaScript code into visual elements in order to implement complex interaction at the GUI level. This allows for a clean division of labor in that Python is responsible for the back-end (in this case sending requests, unpacking the JSON results and preparing them for display) and connects via Qt's signals and slots system to the GUI written in QML, whose front-end behaviour is controlled at least in part by JavaScript. This allows to first modify the GUI and then connect it to the backend or viceversa, which makes the developing process very flexible.

We do indeed hope to turn RhoSqrd into a more general information-management system (IMS) for linguistic data. This thought has driven the whole development process and led to design choices that as of now may not be completely transparent to the end user, although this should not be an issue to them.

As of now, however, RhoSqrd only works in relation to FRITAV, so in the next section we show how to use it to query data from the database.

3.1 Example of Usage

Using RhoSqrd is very simple thanks to its intuitive GUI. Upon clicking on the app's icon and thus starting the app, the user will immediately see the main dashboard, which can be

[1]The versions are 3.8.14 and 6.3, respectively. We have tested the app with Python 3.9.x and 3.10.x and found out that there compatibility issues involving the app packaging system. As soon as this is solved we will switch to the latest stable version of Python.

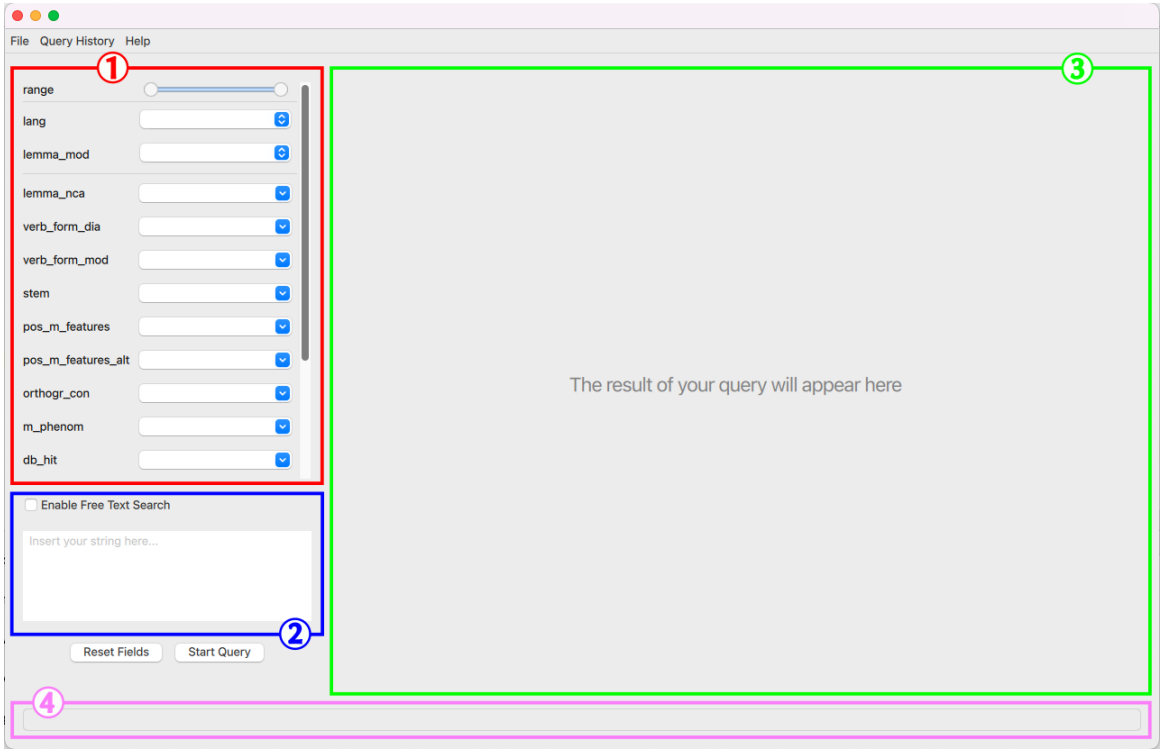

Figure A: RhoSqrd's Dashboard

subdivided into the Query Builder (1), the Free Text Search field (2), the Result Visualization (3) area and the Status Bar (4).

The user can then put together a query using the 'widgets' (technically, they are not widgets, but it is useful as a general term to refer to single graphical components as part of the GUI) found in the Query Builder. More specifically, they can set the time by using the `range` slider, decide from which language and what lemmas to query and select one or many filters to refine the results. Otherwise, if the user knows already how to use HTTP requests as briefly mentioned in §2.3, it is possible to input the GET part of the request in the Free Text Search field once it is enabled (it should be noted that enabling the Free Text Search field implies that any text that has been input in the Query Builder will be ignored when the request is sent to the FRITAV API, meaning that the two methods of querying FRITAV are mutually exclusive). For example, using the query in §2.3, the input would be `aimer?stem=fts.aing`.

Once the request is sent by clicking on Start Query and if the server is reachable, it will return a JSON object that will be then visualized in the Result Visualization area. The Status Bar tells the user how many rows of data there are in the query just made.

Bibliography

Aronoff, M. (1994). *Morphology by itself*. Cambridge, MA, MIT Press.

Base textuelle Frantext, 1.3.24. (2022). ATILF-CNRS, Université de Lorraine. Retrieved September 15, 2022, from http://www.frantext.fr

Blevins, J. P., & Blevins, J. (2009). Introduction: Analogy in grammar. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar* (pp. 3–12). Oxford, Oxford University Press.

Bybee, J. L. (1985). *Morphology. A study of the relation between meaning and form*. Amsterdam, Benjamins.

Carstairs, A. (1987). *Allomorphy in inflexion*. London, New York, Sydney, Croom Helm.

Dees, A. (1980). *Atlas des formes et des constructions des chartes françaises du 13e siècle (avec le concours de Pieter Th. Van Reenen et de Johan A. De Vries)*. Tübingen, Niemeyer.

Dees, A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français (avec le concours de Marcel Dekker, Onno Huber et Karin van Reenen-Stein)*. Tübingen, Niemeyer.

Gaglia, S. (2020). The dynamics of analogy: Old French and Old Italian verbal roots. *Lingue e Linguaggio*, *19*(1), 61–89.

Kunstmann, P., & Stein, A. (2007). *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Stuttgart, Franz Steiner.

Kurylowicz, J. (1945-49). La nature des procès dits «analogiques». *Revue internationale de linguistique structurale*, *5(1)*, 15–37.

Larson, P., & Artale, E. (2005). Corpus OVI dell'italiano antico. CNR-Istituto Opera del Vocabolario Italiano. Retrieved September 15, 2022, from http://gattoweb.ovi.cnr.it/(S(f0xfc5ybzivqwqimufjir455))/CatForm01.aspx

Maiden, M. (2003). Verb augments and meaninglessness in early Romance morphology. *Studi di Grammatica Italiana*, *22*, 1–61.

Maiden, M. (2018). *The romance verb. Morphomic structure and diachrony*. Oxford, Oxford University Press.

Stein, A. Et al. (2006). Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen. Institut für Romanistik/Linguistik. Retrieved September 15, 2022, from https://sites.google.com/site/achimstein/research/resources/nca

Stump, M. (2001). *Inflectional morphology. A theory of paradigm structure*. Cambridge, Cambridge University Press.