# Reshaping the Haphazard Folksonomy of the Semantic Domains of the French *Wiktionary*

**Noé Gasparini[1], Cédric Tarbouriech[2], Sébastien Gathier[3], Antoine Bouchez[4]**

[1,3,4] Institut international pour la Francophonie - Université Jean Moulin Lyon 3, 1C avenue des Frères Lumière CS 78242 - 69372 Lyon Cedex 08 France

[2] Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse & CNRS, France

E-mail: noe.gasparini@univ-lyon3.fr, cedric.tarbouriech@irit.fr, sebastien.gathier@univ-lyon3.fr, antoine.bouchez19@gmail.com

**Abstract**

Semantic domains are a source of headaches in dictionary projects, and one was built haphazardly in the French edition of the collaborative online project *Wiktionary* called *Wiktionnaire*. *Wiktionnaire* is a lexicographical project that started 17 years ago. It is hosted by the Wikimedia Foundation and edited by a community of volunteers that made it a mature project, but with lacunas, with semantic domains being one of these. Between January 2019 and December 2020, this nomenclature of semantic domains was transformed by a small team with complementary expertise and skills. The team consisted of four people with academic knowledge in linguistics, lexicography and information science, as well as technical skills for coding, proofreading and community management. The strategy was the following: mapping the existing terminology, comparison and extension of the list, documentation, structuring, discussions with the community, deployment, cleaning of remaining irregularities, and monitoring the changes after this process. The result of this two-year operation is a complete reshaping of a messy folksonomy into an innovative lattice nomenclature fully integrated into the *Wiktionnaire* and adopted by the community, but also used in an RDF-based dictionary reusing that data, the *Dictionnaire des francophones*. This paper outlines the context of this work on continually changing content and presents the strategy used by the team, including the major issues and choices encountered during the process.

**Keywords:** semantic domains; *Wiktionnaire*; *Wiktionary*; folksonomy; collaborative lexicography.

## 1. Background

*Wiktionary* is a collaborative multilingual open online collection of lexicographical information (Murano, 2014). The edition in French, *Wiktionnaire*, commenced in March 2004 and has since then seen a constant growth in content and quality (Sajous *et al.* 2014). The population of regular contributors is about 100 people, each contributing on average more than 100 edits monthly. Between 1,500 and 2,000 volunteers edit at least once a month. Most of the regular contributors acquired lexicographic skills as they contributed, without any academic background (Meyer &

Gurevych, 2012a), developing a community-based practice. For an overview of studies about *Wiktionnaire*, see Sajous *et al.* (2020).

One aspect of crafting definitions is to indicate semantic domains for technical terms. In *Wiktionnaire*, such domains are presented at the beginning of definitions, between parentheses. When an editor adds a definition for a lemma, a dedicated code indicating its semantic domain is also added, written between curly brackets. This is a way to transclude a subpage named a *Modèle* (*Template* in English). These templates serve to display a text and categorise pages in *Categories*. For example, the code {{anatomie|fr}} was used to insert the content of the template Modèle:anatomie, resulting in the text "anatomie" being displayed and a link to the page with the lemma being included in Category:Anatomie, forming the French lexicon of anatomical terms.

Before 2020, indicators of semantic domains at the beginning of definitions in *Wiktionary* projects were irregular, and the granularity of subdomains showed a large heterogeneity (Meyer & Gurevych, 2012b). This aspect of *Wiktionnaire* or *Wiktionaries* is rarely studied, and most research on folksonomy focuses on *Wikipedia* and tries to construct an external and independent ontology (Macías-Galindo, 2011).

The evolution, maintenance and reuse of these domain indicators were made complex by the existence of over 400 templates, sometimes with aliases. Most of these templates were poorly documented. There were also more than 10,000 entries with a plaintext indicator rather than a dedicated template. Some pages were added manually to categories, instead of by using a template. The category pages displaying the lists of terms associated with a domain were poorly documented. The structure was not standardised between languages described in *Wiktionnaire*.

*Wiktionary* was previously used as a corpus to create external tools like the XML-encoded machine-readable version GLAWI (Sajous & Hathout, 2015), or the comparison of new words in dictionaries listed in the DiCo corpus (Martinez, 2013). The results of such studies are rarely shared with contributors and rarely injected back into the *Wiktionnaire* (or *Wiktionary* in another language). This project on semantic domains not only produced an independent taxonomy, but improved the existing structure of *Wiktionnaire* itself. It was crowdsourced applied lexicography.

## 2. Motivations

In March 2018, the French president Emmanuel Macron presented a plan for promoting the French language and multilingualism. This included a project funded by the Ministère de la Culture [Ministry of Culture] and managed by the Institut International pour la Francophonie at the Université Jean Moulin Lyon 3. The ambition was to create a new dictionary for varieties of French, the *Dictionnaire des francophones* (DDF). It was to be structured as an RDF-based lexicographical database furnished by existing lexicographical resources, including the French entries of *Wiktionnaire*

(Dolar et al., 2020; Steffens et al., 2020). Linked data for lexicography opened a new field, and this project was going to be a first for the French language. A short explanation of this way of organising data is presented by Klimek and Brümmer (2015).

In 2020, a 'Wiktionarian in residence' was recruited to clean up *Wiktionnaire's* content to help the integration of this resource. The purpose was to undertake corrections of general issues but also to clean information structures, including the semantic domains. A dedicated task force emerged to fulfil this mission.

Sébastien Gathier, the resident, is a senior wiki proofreader, mostly for French Wikipedia, Wikidata and Open Food Facts. Noé Gasparini, the DDF project manager, has training in linguistics and language documentation. Antoine Bouchez, an intern for four months, was a lexicographer by training. A skilled Wiktionary contributor, Cédric Tarbouriech, was invited. He is a contributor trained in coding and ontology modelling. The group set regular meetings to work together remotely.

## 3. Strategy

The strategy developed by the team had several steps: map existing terminology (labels and their aliases); compare and augment this list with domains used in other referential works; build a structure; define each label with short glosses to document and disambiguate domains; discuss with the community to obtain consent to implement this solution; implement the list in the Lua language; prepare scripts to deploy this new code in more than 20,000 pages; correct uncountable irregularities that may remain after deployment; build a 'lexicovigilance' similar to a pharmacovigilance to monitor any adverse effects subsequent to this large transformation of *Wiktionnaire*.

### 3.1 Mapping existing terminology in *Wiktionnaire*

The *Wiktionnaire* uses templates to transclude content into other pages. These templates may include parameters, such as the language to use to categorise the content. Before 2020, when a contributor wanted to add a new domain for one language they had to create a new template, which was not an easy task. This new template should be documented but in practice they rarely were. Most of these creations were made by a couple of experienced users. Additional isolated templates were created to cater for very specific needs in some languages.

The list of templates used for semantic domain indicators was augmented by some aliases, i.e. shorter names based on traditional abbreviations from printed dictionaries such as 'hist' for 'history'. Some of them were opaque and could be misinterpreted and wrongly used, e.g., 'litt' could be read as indicating the domain of literature or of literary language; 'comm' could be read as the vocabulary of commerce or of communication.

Glossaries are lists of pages gathered around a common hypernym, such as lists of rivers, birds, languages, etc. There are more than 2,000 glossaries in *Wiktionnaire.* Most of them are included under a semantic domain, such as ornithology for the glossaries of bird names. The distinction between glossaries and domains is a grey area. This may lead to confusion when defining new templates. For example, 'graph theory' was seen as a glossary but is in fact a lexicon, and vice versa for 'feelings'.

### 3.2 Comparison and extension of the list

The original list of domains contained about 400 items. Some very specific subdomains were covered, influenced by contributors' interests. Other domains stayed barely explored due to the lack of interested contributors. Subdomains of computer science were well described but some sports or scientific domains were missing. Some new domains were added thanks to a comparison with other sources in French, such as the *Larousse illustré* (2014), the *Dictionnaire universel* (2008), and the *Dewey Decimal Classification.* Some of these new domains are banking, bryology, clothing, electromagnetism, geomorphology, leather crafting, immunology, petrology, puppetry, and speleology.

A second phase was initiated with the alignment of items from *Le Grand Dictionnaire Terminologique.* More than 4,000 lexical entries from this resource were given to the *Dictionnaire des francophones* by the *Office québécois à la langue française* [Quebec Board of the French Language], and they were willing to share their own terminology with the team. The *Commission d'enrichissement de la langue française* [Commission to enrich the French language], responsible for the content of the website *FranceTerme*, also shared their classifications to prepare for the integration of this resource in the *Dictionnaire des francophones.* With both of these works more domains were added, such as advertising, archery, brewery, flour production, materials science and engineering, spatial planning, woodworking, and waste management.

New domains were also added by exploring definitions with domain indicators that had no dedicated template in *Wiktionnaire.* The initial list contained 390 domains (April 2019) and the final one contains 615 domains (April 2021). Of these, 578 domains are for the French language and others. Some domains are used for only one or a couple of specific languages, such as 'cleaning' (only used for German), or 'Estonian mythology' (only for Estonian). Some new domains were suggested but in this new taxonomy are not in use for any languages yet.

### 3.3 Documenting the lattice structure

Most of the domain templates in *Wiktionnaire* had a short documentation explaining the technical use of the template but not the definition of the concept itself. In order to clarify the terms, short glosses in French were written for each domain.

The initial list was structured with a ramification of categories, in a repeated process of grouping categories together in supercategories or splitting categories into subcategories, with some branches being diversely connected depending on the language described. This structure was regularised and developed as a lattice structure – or, more precisely, a directed acyclic graph – rather than a tree structure, as some domains could have more than one domain above them.

This structure was planned to be explored in DDF through a contributive interface, from the top to the bottom. There are seven top-level domains: technology, arts, alimentation, sports, politics, science and society. They are not supposed to be used as such, but serve as a coarse division of domains to assist in navigating the lattice in the *Dictionnaire des francophones*. In *Wiktionnaire*, this structure has been fully implemented, but is not directly visible to readers.

Direct subdomains include industrial activities, types of arts, types of sports, and academic domains. Those high-level domains are considered as perennial and less inclined to change in the future in comparison with more nested domains. There are five to 26 subdomains directly under the top-level domains. More specific subdomains were not expanded in detail, considering that editors will add new domains when they want to gather the related vocabulary to build them.

### 3.4 Discussions with the *Wiktionnaire* community

The first step to engage the discussion was to publish a page in *Wiktionnaire* titled Projet:Informations lexicographiques[1] [Project:Lexicographical information] to describe the existing structure. This led to some general observations and offered a way to include more contributors for future discussions. Some parts of the process were presented to the community, mostly when it seemed better to split existing domains to follow the tendency observed in other sources. For example, we suggested a division between psychology and psychoanalysis.

A long-term discussion concerned the lexicons of French law and French history, as we felt they should be separated with a combination of domain indicators and geographical indicators (Law+France and History+France) instead. The community rejected the proposal and both lexicons remained.

Another issue was about the vocabulary of the European Union, considered as a technolect or jargon, depending on the analysis (Gardner, 2016). One of the *Wiktionnaire* contributors was a professional translator, so he had enough knowledge of this vocabulary to reorganise the entries and solve this issue. A dozen contributors

[1] https://fr.wiktionary.org/wiki/Projet:Informations_lexicographiques

shared insights on specific issues such as heraldic subdivision, organ building or social justice. Some comments were about the definition while others were about the structure.

Between January 2020 and May 2021, 18 contributors made at least one modification to the list of domains, in order to correct sentences, add new domains or slightly modify the structure.

### 3.5 Lua implementation

Lua is a lightweight high-level programming language. It is one of the few programming languages available in a MediaWiki environment, the technical software used by *Wiktionnaire*. It was needed to program advanced behaviours.

The list of domains is written as a Lua table, a structured map linking semantic domains to their information. Each item has a name, a description, an indication of the phrasing to write to make sentences readable by humans, and supercategories in which the domain has to be included. This page is called Module:Lexique/data[2] and it is the unique list of domains for *Wiktionnaire*.

Definition of the domain 'boulangerie' (bakery).

```
['boulangerie'] = {
    ['description'] = 'La boulangerie désigne la fabrication et la vente de pain et de viennoiseries.',
    ['determiner'] = 'de la ',
    ['super_categories'] = { 'cuisine' }
  },
```

### 3.6 Deployment with a dedicated script

A Python script was used to deploy the new system in both articles and categories. Specific issues were documented in maintenance categories. A couple of new domains had been created in the meantime and were added. At the end of the deployment, all old templates and their aliases were deleted to avoid further use, which would result in a confusing coexistence of two incompatible systems. Only the new template *lexique* is now used, and included in 34,000 pages. 96,860 domain tags were added to over 87,692 French definitions.

[2] https://fr.wiktionary.org/wiki/Module:lexique/data

### 3.7 Dealing with irregularities

After the deployment, a large number of definitions still displayed a free-text domain indicator rather than the new template for the domain. This meant that they were not included in the lexicons. More than 5,000 of those were corrected manually in 2020; this task is still ongoing.

Some cases needed a special investigation. For example, some templates had a parameter that indicated a subdomain, i.e. 'Canadian football' was using the same template as 'Football' with an additional parameter written as "spéc=canadien". It was changed to be two separate parameters in the new template. The same situation occurred with religions, mythologies, and subdomains of law and sports.

Another task was to ferret out missed domains. Some had been added by other contributors during our work and others were used in only a set of languages having very few words. They had to be included or recategorised. An example is the vocabulary of Palaic mythology for a couple of words.

Some *Category* pages had an introduction in plain text that conflicted with the deployment of *Modèle:catégorisation lexique* but included useful information, thus requiring careful revision.

### 3.8 Monitoring and accompanying the community to change its habits

To avoid a negative response from contributors, careful vigilance was maintained during the six months that followed the deployment to correct any mistakes and explain the changes when necessary.

In addition, a new function was developed to suggest semantic domains with an autocompletion while contributing. The function was developed by a contributor named Darmo117 after the idea was suggested by the community.

After only a few weeks, this lexicovigilance became less necessary as new semantic domains were created and added correctly by contributors outside the team. The new documentation led to the involvement of new contributors into the definition of lexicons.

## 4. Discussion

This whole process was a success, and the new taxonomy was adopted by both *Wiktionnaire* and *Dictionnaire des francophones*. The lattice allows the inclusion of new branches easily, by any contributor of *Wiktionnaire*. It is dense but still shows some irregularities due to its origins and choices made by the contributors during the process and after. The plan was to structure the semantic domains, considering it might

help the readers explore the content and the contributors add new domains. If readers' experience was not monitored, this new organisation seems to have an impact on the addition of new domains during the following months.

This strategy did not come out of the blue, it was based on previous changes of policies encouraged in *Wiktionnaire* such as, for instance, the description of protolanguages or how to describe prototypical pronunciations. The wiki workflow creates a vivid space to discuss the editorial choices made in the past and suggest transformations for any aspect of the project. The refinement of every semantic domain was nonetheless a large-scale change, more ambitious than any previous initiatives, and it had an impact on almost every contributor's habits. As such, it was a first for *Wiktionary* as it was for lexicography.

In 3.6 some metrics were given. Those were not accessible before the inclusion of *Wiktionnaire* in the *DDF* database as the wiki structure is not that easy to query. The adaptation of *Wiktionary* content into an RDF database made possible new exploration of language-centric metrics, impossible with the original multilingual pages.

This new taxonomy is mainly based on the existing one and had to remain close TO IT as the *Dictionnaire des francophones* will import updates of *Wiktionnaire* in the future. Despite a large comparison with existing dictionaries for the French language, our taxonomy remains to be compared and aligned with taxonomy in other languages, such as *SIL Semantic Domains* and *WordNet Domains*. A terminological comparison with more resources could be a way to improve the structure and relations between domains and glossary. It may also help to find any blank spots and suggest new domains to cover.

The structure of the controlled vocabulary in *Wiktionnaire* does not readily allow alignment with other taxonomies. There is no unique identifier for each domain. Nonetheless, each page of category for vocabulary is connected with an entity in the *Wikidata* database, and these could be linked to the related concepts described by the semantic domains. The concepts could then become connected with several ontologies and databases. However, as of April 2021, less than 20% of semantic domains are fully connected with the related concepts. This possibility is still on the roadmap for the team to enhance the semantic domains of *Wiktionnaire*.

This lexicographic process is focused on knowledge engineering and information science, but it aims at producing a semantic structure that is easy to explore rather than a complex graph of relations among domains with a semantic elevation to offer ways to explore the data. This controlled vocabulary of domains is one facet of definition and the semantic structure of the entries. A possible future step may include qualifiers for the relations among the domains, but also with glossaries and significant entries.

## 5. Conclusion

This experience was successful and could be reproduced on other collaborative crowdsourced projects, such as the *Wiktionaries* written in other languages. In *Wiktionnaire*, a similar process of cleaning, documenting and structuring geographical information started in 2021.

As a conclusion, we want to point out that this strategy would not have been possible without the dedication of a multidisciplinary and multicompetent taskforce during an extensive period of time. This two-year undertaking allowed the cleaning of most of the data, the identification of areas of improvement by comparison with other resources, and the involvement of the *Wiktionnaire* community in the course of the project.

## 6. Acknowledgements

The authors thank the contributors of *Wiktionnaire,* the French Ministry of Culture for the initiative and funding of *Dictionnaire des francophones* project, Wikimédia France for their funding and the Institut international pour la Francophonie for managing the project and hosting a large part of the team. The authors also want to thank their anonymous reviewers and Robert Scott, Nadia Sefiane, Marya Lambia and Damien Vergnet for their useful comments on the draft.

## 7. References

Achard-Bayle, G. & Paveau, M.-A. (2008). La linguistique 'hors du temple'. *Pratiques*, 139/140, pp. 3-16. Available at: https://hal.archives-ouvertes.fr/hal-00516249/document

Dolar, K., Steffens, M. & Gasparini, N. (2020). Dictionnaire des Francophones: A New Paradigm in Francophone Lexicography. In: *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Thrace: Democritus University of Thrace, pp. 23-30. Available at: https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p023-030.pdf

Gardner, J. (2016). *Misused English words and expressions in EU publications.* Available at: https://www.eca.europa.eu/Other%20publications/EN_TERMINOLOGY_PUBLICATION/EN_TERMINOLOGY_PUBLICATION.pdf

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. Kernerman Dictionary News. Available at: https://www.kdictionaries.com/kdn/kdn23_2015.pdf#page=5

Macías-Galindo D., Wong W., Cavedon L., Thangarajah J. (2011). Using a Lexical Dictionary and a Folksonomy to Automatically Construct Domain Ontologies. In: Wang D., Reynolds M. (eds) AI 2011: *Advances in Artificial Intelligence.* AI 2011. Lecture Notes in Computer Science, vol 7106. Springer, Berlin, Heidelberg.

Martinez, C. (2013). La comparaison de dictionnaires comme méthode d'investigation

lexicographique. N. Gasiglia (ed.). *Lexique*. 21. Villeneuve-d'Ascq, Presses universitaires du Septentrion, pp. 193-220.

Meyer, C.M. & Gurevych, I. (2012a). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger & Magali Paquot (eds), *Electronic Lexicography*, Oxford University Press, pp. 259-292.

Meyer, C.M. & Gurevych, I. (2012b). OntoWiktionary: Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In: *Semi-Automatic Ontology Development: Processes and Resources*. Maria Teresa Pazienza & Armando Stellato (eds). Information Science Reference, pp. 131-161.

Murano, M. (2014). La lexicographie 2.0 : nous sommes tous lexicographes ?. *Cahiers de recherche de l'École doctorale en linguistique française*, 8, pp. 147-162 Available at: https://www.openstarts.units.it/bitstream/10077/10767/1/9Murano.pdf

Sajous, F., Hathout, N. & Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. In: *4e Congrès Mondial de Linguistique Française*. Les Ulis: EDP Sciences, pp.663-680. Available at: https://halshs.archives-ouvertes.fr/halshs-00969260/document

Sajous F., Navarro E., Gaume B., Prévot L. & Chudy Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In: H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (eds). *Advances in Natural Language Processing*, vol. 6233 of Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 332-344. Available at: https://hal.archives-ouvertes.fr/hal-00625326

Sajous F. & Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. *Proceedings of the eLex 2015 conference*, Herstmonceux, England, pp. 405-426.

Sajous F., Calderone, B. & Hathout, N. (2020). Extraire et encoder l'information lexicale de Wiktionary : quel boulot pour étrangler le goulot ! Lexique 27, pp. 121-144. Available at: http://fsajous.free.fr/papers/Lexique27_2020/SajousEtAl2020_Lexique27_ExtraireInformationLexicaleWiktionary.pdf

Steffens, M., Dolar, K., & Gasparini, N. (2020). Structuration de données pour un dictionnaire collaboratif hybride. In: *Terminologie & Ontologie: Théories et Applications. Actes de la conférence TOTh 2019*. Chambéry: Presses Universitaires Savoie Mont Blanc, pp. 413-426.