Adding Glawinette into Démonette: pratical consequences and theoretical questions

Nabil Hathout[a] and **Fiammetta Namer**[b]

[a]CLLE, CNRS & Université Toulouse Jean Jaurès
[b]Université de Lorraine & ATILF, CNRS

DeriMo 2021. 9-10 September 2021

Outline

Démonette. Table of lexemes

Démonette (Hathout & Namer, 2014; Namer et al., 2019) is a large French derivational database. It is currently made up of a **table lexemes** and a **table of relations**.

The **table of lexemes** contains information on

- the category of the lexemes (N, V, Adj, Adv),
- their gender if they are nouns (Nf, Nm),
- their morphosyntactic features and the phonological representation of their inflected forms,
- their ontological category (person, artefact, process, etc.)

Démonette. Table of relations

The other information is recorded in the **table of relation**. The entries are pairs of lexemes. The information on an entry is divided into 6 subsections:

1. identification of the pair of lexemes
   indexes in the table of lexemes)
2. copy of the information on the **category** and the **written forms** of the lexemes
3. morphological description (**type** and **exponents**)
4. **complexity** and **orientation**
5. copy of the ontological information of the lexemes complemented by **fine-grained semantic categories** (eg. *fruit* for *cerise* 'cherry')
6. description of the semantic relation (**gloss**)
   *quelque chose est **lavable** si on peut la **laver***
   'something is washable if we can wash it'.

Démonette. Included datasets

Démonette 2.0 contains **63641 entries** (pairs of lexemes)
48843 individual lexemes

Démonette is a cumulative repository of morphological datasets created as part of PhDs and academic studies. Its feeding is mainly "depth first".

Démonette 2.0 has been fed by 5 datasets:

- **Démonette 1.2**: **DériF** (Namer, 2009) + **Morphonette** (Hathout, 2011) + **Lexeur** (Wauquier et al., 2020)
- **Convers** (Tribout, 2010)
- **DeNom** (Strnadová, 2014)
- **DériF** (*anti-*, *dé-*, *en-*, *-able*, *-iser*)
- **DiMoC** (Roché, 2004, 2006, 2008a,b; Roché et al., 2011; Lignon & Roché, 2011; Roché & Plénat, 2014)

Démonette. Content

derivation	lex	rel	derivation	lex	rel
convertion$_{N/V}$	3032	3243	Xable$_{A}$	1336	1432
Xeur$_{N}$	2747	8157	Xif$_{A}$	555	1495
Xeuse$_{N}$	1157	3781	Xal$_{A}$	651	897
Xrice$_{N}$	1220	359	Xique$_{A}$	3123	4433
Xage$_{N}$	2958	7874	Xaire$_{A}$	977	1474
Xment$_{N}$	2492	7078	Xier$_{A}$	898	1008
Xion$_{N}$	3003	8290	Xet$_{A}$	104	106
Xette$_{N,Npr}$	2723	2898	antiX$_{A}$	357	786
Xet$_{N,Npr}$	572	590			
Xaie$_{N}$	359	360			
Xier$_{N}$	1632	1901	déX$_{V}$	907	1725
Xat$_{N}$	1314	1545	enX$_{V}$	47	50
Xaire$_{N}$	550	588	Xiser$_{V}$	1071	5567

Démonette. How to meet a wider set of needs?

Démonette mainly contains **direct relations** (base→derivative pairs).

Many affixations are still **missing** (*re-*, *-isme*, *-iste*, *-eux*, *-u*, etc.).

French **derivational families** are not properly described in Démonette.

There is a need for a more varied content, with a larger number of affixations, a better representation of frequent French derived words.

We need to complement Démonette in order to make it more similar to the standard morphological databases like CELEX (Baayen et al., 1995) in order to meet the needs of

- statistical studies; experimental linguistics;
- speech therapy (eg. creation of remediation exercises);
- psycholinguistics (eg. mental lexicon).

Outline

Glawinette. The needs and objectives

We need a database with a wider coverage, a data base more representative of French derivational morphology.

The main objective of Glawinette (Hathout et al., 2020):

1. collecting a **large set** of pairs of lexemes
2. connected by a **large variety** of derivational relations (no restriction on the derivational processes),
3. that could be **easily** injected into Démonette (reliable + intuitive linguistic characterization).

This is a joint work done by **Basilio Calderone**, **Franck Sajous**, **Fiammetta Namer** and **Nabil Hathout**.

Glawinette. The needs and objectives

The constraint of size requires the acquisition method to be **automatic**

We used **GLAWI** (Sajous & Hathout, 2015; Hathout & Sajous, 2016), the largest available French machine readable dictionary

We designed a method capable of identifying with a **high precision** the pairs related by valid derivational relations.

We designed a method that assigns to each pair of lexemes their most **"natural" exponents**.

Glawinette. Content

97293	lexemes
47717	pairs of lexemes
15904	derivational families
5400	derivational series

Derivational family of *serrer* 'to tie'

desserrage=N:desserrer=V **desserre=N:desserrer=V**
desserrement=N:desserrer=V desserrer=V:desserrage=N
desserrer=V:desserre=N desserrer=V:desserrement=N
desserrer=V:desserroir=N **desserrer=V:indesserrable=A**
desserrer=V:redesserrer=V **desserrer=V:serrer=V** desserroir=N:desserrer=V
enserrement=N:enserrer=V enserrer=V:enserrement=N
enserrer=V:renserrer=V enserrer=V:réenserrer=V enserrer=V:serre=N
indesserrable=A:desserrer=V redesserrer=V:desserrer=V
renserrer=V:enserrer=V reresserrer=V:resserrer=V resserrage=N:resserrer=V
resserrement=N:resserrer=V resserrer=V:reresserrer=V
resserrer=V:resserrage=N resserrer=V:resserrement=N
resserrer=V:resserreur=N resserrer=V:serrer=V resserreur=N:resserrer=V
réenserrer=V:enserrer=V serrage=N:serrer=V serre=N:enserrer=V
serre=N:serrer=V serre=N:serriste=N serrement=N:serrer=V
serrer=V:desserrer=V serrer=V:resserrer=V serrer=V:serrage=N
serrer=V:serre=N serrer=V:serrement=N serrer=V:serrure=N
serrer=V:serré=A serriste=N:serre=N serrure=N:serrer=V
serrure=N:serrurerie=N serrure=N:serrurier=N serrurerie=N:serrure=N ...

Characterization of the pairs. BAP and FAP

The pairs are characterized by a broad alternation pattern (**BAP**), the most general pattern that relates the 2 lexemes.

Each pattern match one form (lemma) in the pair.

enserrer=V	enserrement=N	^(.+)r$	V	^(.+)ment$	N
serrer=V	serrure=N	^(.+)e(.+)$	V	^(.+)u(.+)e$	N

The (.+) sequences in the 2 patterns represent the same strings (the stem).

The pairs are also characterized by a fine-grained alternation pattern (**FAP**).

enserrer=V	enserrement=N	^(.+)er$	V	^(.+)ement$	N
serrer=V	serrure=N	^(.+)er$	V	^(.+)ure$	N

They describe the alternation with more "natural" exponents, eg. *-er* is the exponent of infinitive for the French verbs of the first conjugation.

The stem is one contiguous string.

Derivational series. FAPs

enserrer=V	enserrement=N	^(.+)er$	V	^(.+)ement$	N
desserrer=V	desserrage=N	^(.+)er$	V	^(.+)age$	N
resserrement=N	resserrer=V	^(.+)ement$	N	^(.+)er$	V
serrure=N	serrurier=N	^(.+)e$	N	^(.+)ier$	N
desserrage=N	desserrer=V	^(.+)age$	N	^(.+)er$	V
serrurerie=N	serrurier=N	^(.+)erie$	N	^(.+)ier$	N
serrer=V	serrure=N	^(.+)er$	V	^(.+)ure$	N
desserre=N	desserrer=V	^(.+)e$	N	^(.+)er$	V
resserrer=V	resserrement=N	^(.+)er$	V	^(.+)ement$	N
serrure=N	serrurerie=N	^(.+)e$	N	^(.+)erie$	N
réenserrer=V	enserrer=V	^ré(.+)er$	V	^(.+)er$	V
serrer=V	serrement=N	^(.+)er$	V	^(.+)ement$	N
serrurier=N	serrure=N	^(.+)ier$	N	^(.+)e$	N
serrurerie=N	serrure=N	^(.+)erie$	N	^(.+)e$	N
enserrer=V	renserrer=V	^(.+)er$	V	^r(.+)er$	V
enserrer=V	serre=N	^en(.+)er$	V	^(.+)e$	N
indesserrable=A	desserrer=V	^in(.+)able$	A	^(.+)er$	V
desserrer=V	desserroir=N	^(.+)er$	V	^(.+)oir$	N

Glawinette. Content

The **5400 series** actually include

- written form **variants**
  teck / *tek* ‘teak’
  kleptomanie / *cleptomanie* ‘cleptomanie’
  complétement / *complètement* ‘completely’
- **compounds**
  therapeute ‘therapist’ / *psychotherapeute* ‘psychotherapist’
  zone / *eurozone*

Glawinette. Content

- **derivations**
  - conversion
    plume ‘feather’ / *plumer* ‘to pluck the feathers off’
  - prefixation
    graisse ‘fat’ / *engraisser* ‘to fatten’
  - suffixation
    flemme ‘laziness’ / *flemmard* ‘lazy’
  - multiple suffixation
    stable ‘stable’ / *stabilisation* ‘stabilization’
  - prefixation + suffixation
    minéral ‘mineral’/ *déminéralisation* ‘demineralization’

Glawinette. Reliability

200 pairs randomly extracted from Glawinette have been checked manually

100% of the pairs contain morphologically related pairs

84% of the pairs have valid FAPs

31 erroneous FAPs correspond mainly to overspecified patterns and to pairs of lexemes that contain rare allomorphies

réenserrer=V	enserrer=V	^ré(.+)er$	V	^(.+)er$	V
sarkosyste=N	sarkosysme=N	^(.+)ste$	N	^(.+)sme$	N
réenserrer=V	enserrer=V	^ré(.+)$	V	^(.+)$	V
sarkosyste=N	sarkosysme=N	^(.+)yste$	N	^(.+)ysme$	N

Very few pairs are erroneous.

FAPs must be checked. The revision can be done in batches (pairs characterized by the same FAP are processed in the same batch).

Outline

Basic Principles

Glawinette is created from

- the morphological sections of GLAWI

  parachutage: *parachute*, *parachuter*, *parachutisme*, *parachutiste*

- the morphological definitions = definitions that relate the headword to a member of its derivational family (Martin, 1992).

  clocheton: *petit bâtiment en forme de* ***clocher***, *de tourelle, dont on orne les angles ou le sommet d'une construction* 'small building in the shape of a bell tower, that decorate buildings corners or tops'

  glaçon: *morceau de* ***glace*** 'piece of ice'

Problem: we cannot automatically identify these definitions.

Brute force method

1. create pairs made up of the headword and each word of the definition

   glaçon morceau
   glaçon glace

2. remove the ones that does not enter into an analogical series of at least 5 pairs. (Lepage, 1998, 2004; Stroppa & Yvon, 2005; Hathout, 2008; Langlais & Yvon, 2008; Fam & Lepage, 2021)

   No other pair that can form an analogy with *glaçon* / *morceau*

   glaçon / *glace* forms proportional analogies with:

   garçon garce
   pinçon pince
   façon face
   tierçon tierce

Brute force method

Analogies can be efficiently computed. We assign a signature to each pair of lexemes; pairs with identical signature form analogies.

$$\sigma(A, B) = (d(A, B), |A|_{a_1} - |B|_{a_1}, \ldots, |A|_{a_n} - |B|_{a_n})$$

where $d(A, B)$ is the Levenshtein distance between A and B and $\{a_1, \ldots, a_n\}$ is the alphabet of the language.

3. identify the patterns that characterize the series of words that make up the series of pairs and match these patterns.

^(.+)eur$		^(.+)age$	
allumeur	'igniter'	allumage	'ignition'
atterrisseur	'lander'	atterrissage	'landing'
balayeur	'sweeper'	balayage	'sweeping'
carreleur	'tiler'	carrelage	'tiling'
épandeur	'spreader'	épandage	'spreading'

Brute force method

4. remove the pairs described by alternation patterns that have a low coverage (less than 10% of the initial series of pairs)
5. select the most "connecting" alternation patterns of each pair of lexemes. This alternation patterns is made up of the most "natural" exponents of these lexemes.

		APs	FAP
verbaliser=V	verbalisation=N	`^(.+)er$:^(.+)ation$`	←
		`^(.+)iser$:^(.+)isation$`	
proverbial=A	proverbialement=R'	`^(.+)ial$:^(.+)ialement$`	
		`^(.+)al$:^(.+)alement$`	
		`^(.+)l$:^(.+)lement$`	
		`^(.+)$:^(.+)ement$`	←
féministe=A	féminisme=N	`^(.+)niste$:^(.+)nisme$`	
		`^(.+)iste$:^(.+)isme$`	←
		`^(.+)ste$:^(.+)sme$`	
sarkozysme=N	sarkozyste=N	`^(.+)ste$:^(.+)sme$`	←

Outline

Injection of Glawinette into Démonette

1. The **largest series** are the most reliable (strong regularity) and the most profitable ones (injection of a large number of entries at once).

^re(.+)$	^(.+)$	3624	refinancer	financer
^(.+)r$	^(.+)$	1999	amortir	amorti
^(.+)er$	^(.+)e$	1958	peigner	peigne

Long exponents are usually erroneous.

^cocycl(.+)ique$	^cycl(.+)ique$	1	cocyclomatique	cyclomatique
^(.+)omatique$	^(.+)omate$	2	diplomatique	diplomate

2. Ignore the series that describe compounding and formal variation

3. Screen the pairs in order to identify possible erroneous ones

Injection of Glawinette into Démonette

4. Add the gender to the description of the nouns

5. Correct the exponents of the series globally

^re(.+)er$	^(.+)er$	→	^re(.+)$	^(.+)$
^(.+)sme$	^(.+)ste$	→	^(.+)ïsme$	^(.+)ïste$

6. Add the other information:

type1	type2	complexity	orient
pre	NA	simple	des2as
suf	suf	simple	indirect

Outline

Evolution of Démonette

The families of Glawinette will feed a table of families in Démonette.
The families will be described as a set of pairs of lexemes.
The families will only contain pairs that have been added to the table of relations.

The compounds made up of two components will be included in a separate table.
The table of compounds will be similar to the table of relations with one difference. The entries are triple

(compound, component-lexeme1, component-lexeme2)

Evolution of Démonette

Many valid pairs of lexemes have been discarded during the creation of Glawinette.
We will complement Démonette by means of the FAPs.
The FAPs will help us recover some of the discarded pairs, and especially relevant indirect relations.

The indirect relations originating from definitions are semantically motivated.
Some indirect relations imported from other resources are too complex.
Glawinette could provide an estimate of the level of motivation of the indirect relations.

Références

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

Fam, Rashel & Yves Lepage. 2021. A study of analogical density in various corpora at various granularity. *Information* 12(8).

Hathout, Nabil. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the coling workshop textgraphs-3*, 1–8. Manchester: ACL.

Hathout, Nabil. 2011. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio* 2011(2). 243–262.

Hathout, Nabil & Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168.

Hathout, Nabil & Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a workable French machine-readable dictionary. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*, Portorož, Slovenia.

Références

Hathout, Nabil, Franck Sajous, Basilio Calderone & Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the twelfth international conference on language resources and evaluation (LREC 2020)*, 3870–3878. Marseille.

Langlais, Philippe & François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the 22nd international conference on computational linguistics (coling 2008)*, 51–54. Manchester.

Lepage, Yves. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 36th annual meeting of the association for computational linguistics and of the 17th international conference on computational linguistics*, vol. 2, 728–735. Montréal.

Lepage, Yves. 2004. Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on computational linguistics (COLING-2004)*, 736–742. Genève.

Références

Lignon, Stéphanie & Michel Roché. 2011. Entre histoire et morphophonologie, quelle distribution pour *-éen* vs *-ien*. In *Des unités morphologiques au lexique*, 191–250. Paris: Hermès Science-Lavoisier.

Martin, Robert. 1992. *Pour une logique du sens* Linguistique nouvelle. Paris: Presses universitaires de France.

Namer, Fiammetta. 2009. *Morphologie, lexique et traitement automatique des langues : L'analyseur dérif*. Paris: Hermès Science-Lavoisier.

Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 - une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26^{e} conférence annuelle sur le traitement automatique des langues naturelles (taln 2019)*, 233–243. Toulouse.

Roché, Michel. 2004. Mot construit ? Mots non construits ? Quelques réflexions à partir des dérivés en *-ier(e)*. *Verbum* 26. 459–480.

Roché, Michel. 2006. La dérivation en *-ier(e)* en ancien français. *Lexique* 17. 55–96. Claude Buridant (éd.), La morphologie dérivationnelle dans l'ancienne langue française et occitane.

Références

Roché, Michel. 2008a. Quelques exemples de morphologie non conventionnelle dans les formations construites à partir d'un mot en *-ouille(r)*. In Bernard Fradin (ed.), *La raison morphologique. hommage à la mémoire de danièle corbin*, 215–238. Amsterdam: John Benjamins.

Roché, Michel. 2008b. Structuration du lexique et principe d'économie : Le cas des ethniques. In Jacques Durand, Benoît Habert & Bernard Laks (eds.), *Actes du congrès mondial de linguistique française (cmlf-2008)*, 1571–1585. Paris: ILF.

Roché, Michel, Gilles Boyé, Nabil Hathout, Stéphanie Lignon & Marc Plénat. 2011. *Des unités morphologiques au lexique*. Paris: Hermès Science-Lavoisier.

Roché, Michel & Marc Plénat. 2014. Le jeu des contraintes dans la sélection du thème présuffixal. In *Actes du congrès mondial de linguistique française (cmlf-2014)*, 1863–1878. ILF.

Références

Sajous, Franck & Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, 405–426. Herstmonceux, England.

Strnadová, Jana. 2014. *Les réseaux adjectivaux : sur la grammaire des adjectifs dénominaux en français*: Université Paris Diderot / Univerzita Karlova V Praze Thèse de doctorat.

Stroppa, Nicolas & François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th conference on computational natural language learning (conll-2005)*, 120–127. Ann Arbor, MI: ACL.

Tribout, Delphine. 2010. *Les conversions de nom à verbe et de verbe à nom en français*: Université Paris 7 Thèse de doctorat.

Wauquier, Marine, Cécile Fabre & Nabil Hathout. 2020. Semantic discrimination of technicality in french nominalizations. *Zeitschrift für Wortbildung / Journal of Word Formation* 2020(2). 100–121.