Échantinom: a hand-annotated morphological lexicon of French nouns

Olivier Bonami[1] Delphine Tribout[2]

[1]Université de Paris, LLF, CNRS
[2]Université de Lille, STL, CNRS

Derimo 2021 – Nancy/online – September 10, 2021

Why bother?

- Rich set of existing resources on word formation in French

Resource	Publication	Processes
Démonette	Hathout and Namer (2014)	Agent/Instrument deverbal nouns, Event nominalizations, *-if* adjectives, …
Lexeur	Wauquier, Fabre, and Hathout (2020)	Agent/Instrument deverbal nouns, Event nominalizations
Dénom	Strnadová (2014)	All derived adjectives
Mordan	Koehl (2012)	Deadjectival nouns
Converts	Tribout (2010)	Verb<>Noun conversions
…	…	…

- The Démonext project (Namer et al., 2019) aims to combine and streamline these resources into a coherent whole.

Problems

- While this is an exciting development, this mode of data collection has drawbacks for some applications:
  - No uniform sampling procedure: data collected from dictionaries vs. corpora vs. web crawls.
  - Focus on depth rather than breadth: many obscure words are included, while some word formation processes are not documented at all.
  - Depth and quality of annotation is variable from source to source.
  - Annotation decisions tend to be poorly documented.
- As a result, these resources are an imperfect starting point for statistical studies of the word formation system.

Our agenda

- Our goal: build a resource that is
  - carefully sampled
  - fully manually curated
  - fully documented.
- Because this is very expensive, we focused on nouns and on a smaller sample size (5000).
- Already available: `https://osf.io/rdxqk`

Sampling

- We start frop the Lexique database (New et al., 2007) and other resources derived from it.
- Nouns with a lemma frequency of at least 0.15 tokens per million, averaging over the two reference corpora (post-1950 French literature, subtitles)
  - 13,046 nouns
- We randomly sample items from this set until we had 5000 confirmed nouns (after correction of tagging errors).

  NB: for purposes of sampling, human masculine and feminine nouns (e.g. BANQUIER, BANQUIÈRE) were counted as distinct, even when they have the same form (e.g. JOURNALISTE).
  - This is a disputable compromise (Bonami and Boyé, 2019), but at least it is coherent.

Morphological annotation

- The morphological annotation of the dataset was made by two annotators, both authors of the paper.
- In a first step, each one annotated about 850 nouns that were checked by the other annotator afterwards.
- In a second step, the remaining nouns were distributed between the authors.
- All problems and questions were discussed and solved collectively.
- Each noun was annotated for different properties.

Outermost word formation process

- We annotated the **broad outermost word formation process**:
  - prefixation
  - suffixation
  - conversion
  - non concatenative process (`nonconcat`)
  - formation from more than one word (`polylexical`)
  - simplex for underived nouns

- When the last process was ambiguous, we relied on frequency

  e.g. SOUS-ALIMENTATION 'undernourishment' can derive from
  - ALIMENTATION 'feeding' (last process = prefixation)
  - SOUS-ALIMENTER 'undernourish' (last process = suffixation)

  ☞ ALIMENTATION has a higher frequency than SOUS-ALIMENTER in *Lexique*'s reference corpora ☞ last process = prefixation

Outermost word formation process

Each category was divided into **fine grained sub-categories**

- Simplex:
  - native simplex (CAHIER 'notebook')
  - borrowings (JAZZ)
  - antonomasia (POUBELLE 'bin')
  - onomatopeic nouns (CLIC 'click')

- Non concatenative processes:
  - reduplication (BABALLE ← BALLE 'ball')
  - back formations (NUMISMATE 'numismatist' ← NUMISMATIQUE 'numismatics')
  - slang processes: verlan (KEUF ← FLIC 'cop') or louchébem (LARFEUIL ← PORTE-FEUILLE 'wallet')
  - truncation: apocope (IMPRO ← IMPROVISATION), apocope with addition of an ending (VALOCHE ← VALISE 'suitcase') and apheresis (SCOPE ← MICROSCOPE 'microscope').

Outermost word formation process

- Conversion:
  - one type for each base POS (adjective, adverb, pronoun, etc.)
  - 5 different types of verb→noun conversions

- Polylexical processes:
  - native compounds (SÈCHE-CHEVEUX, 'hairdryer' ← SÉCHER 'dry' and CHEVEUX 'hair')
  - neoclassical compounds (BARYTON, 'baritone'),
  - blends (FADETTE ← FACTURE 'bill' and DÉTAILLÉE 'detailed')
  - acronyms (SIMA ← SILICIUM 'silicon' and MAGNÉSIUM 'magnesium')
  - frozen word sequences (ARC-EN-CIEL 'rainbow').

  - Difference between native compounds and frozen sequences:
    if one of the element is a grammatical word (*en* in ARC-EN-CIEL)
    ☞ coded as a frozen sequence (`agglomerate`)

Annotation of main word formation processes

In addition to the last process, 4 columns for the **main word formation processes**: prefixation, suffixation, compounding, conversion

- ☞ These columns allow to specify the prefix/suffix used and the type of compound/conversion

lexeme	last_process	prefix	compound	conversion	suffix
EX-FEMME ‘ex-wife’	prefix	ex	0	0	0
GRANDEUR ‘size’	suffix	0	0	0	eurF
OUVRE-BOITE ‘tin opener’	native compound	0	VERB-NOUN	0	0
AVEUGLE ‘blind person’	conversion-A	0	0	A	0

Annotation of main word formation processes

We do not provide a full account of each lexeme's derivational history.

- However, the 4 dedicated columns allow to indicate wether another process is involved in the formation of the lexeme.

lexeme	last_process	prefix	compound	conversion	suffix
EMBARQUEMENT 'boarding'	suffix	en	0	0	ment
COMMERCIAL 'salesman'	conversion-A	0	0	A	al
BIOLOGISTE 'biologist'	suffix	0	neoclassical	0	iste
CLOU 'nail'	simplex	0	0	V	0
MARCHE 'walk'	simplex	0	0	V	0

☞ This is particularly useful when the directionality of the derivation is nonobvious (e.g. conversion).

The case of suffixation

Because suffixation is the most frequent process in our data, we included:

- 2 levels of granularity for the suffix: i) the surface form of the suffix, ii) a form that neutralizes gender variation and allomorphy
- additional columns for the base of suffixation, its POS, wether it is autonomous or not

lexeme	suffix	suffix broad	sfx_base	sfx_base POS	autonomous base
PASSOIRE 'colander'	oire	oir	PASSER	V	TRUE
RASOIR 'razor'	oir	oir	RASER	V	TRUE
NOTABLE 'noteworthy'	able	able	NOTER	V	TRUE
NUISIBLE 'harmful'	ible	able	NUIRE	V	TRUE

The case of suffixation: a few decisions

- We did not differentiate suffixes according to fine semantic distinction
  - e.g. one suffix *-ier* for AMANDIER 'almond tree' (tree) BANQUIER 'banker' (person) and SUCRIER 'sugar bowl' (artifact)

- We did not differentiate homonymous suffixes
  - e.g. one suffix *-age* for JARDINAGE 'gardening' (deverbal noun) and OMBRAGE 'shade' (denominal collective nouns)
    - ☞ the information is available in the `sfx_base_POS` column

- If the formal and semantic bases are different, the formal base is indicated
  - e.g. ROYALISTE 'royalist' formally derives from ROYAL 'royal' and semantically from ROI 'king' ☞ ROYAL is noted as the base

- We distinguished cases where the base of suffixation is a bound stem:
  - e.g. COMPÉTITRICE 'rival' derives from the bound stem COMPÉTIT- also found in COMPÉTITION 'competition'

  and cases where there is no base:
  - e.g. MAQUETTE 'model' belongs to the derivational series of *-ette* diminutive nouns but has no base (*MAQU-)

Other data provided

- Frequency data from Lexique (New et al., 2007) and FrCoW (Schäfer and Bildhauer, 2012)
- Phonemic transcriptions from flexique (Bonami, Caron, and Plancq, 2014)
- For suffixed nouns:
  - Measures of formal transparency derived from the transcriptions
  - Measures of semantic transparency derived from a distributional vector space based on FrCoW

Descriptive statistics I

- Striking prevalence of simplex nouns
- Striking prevalence of deverbal suffixations

	Count	Proportion
Simplex	2064	41%
Suffix	1865	37%
Conversion	564	11%
Polylexical	298	6%
Nonconcat	125	2%
Prefix	84	2%

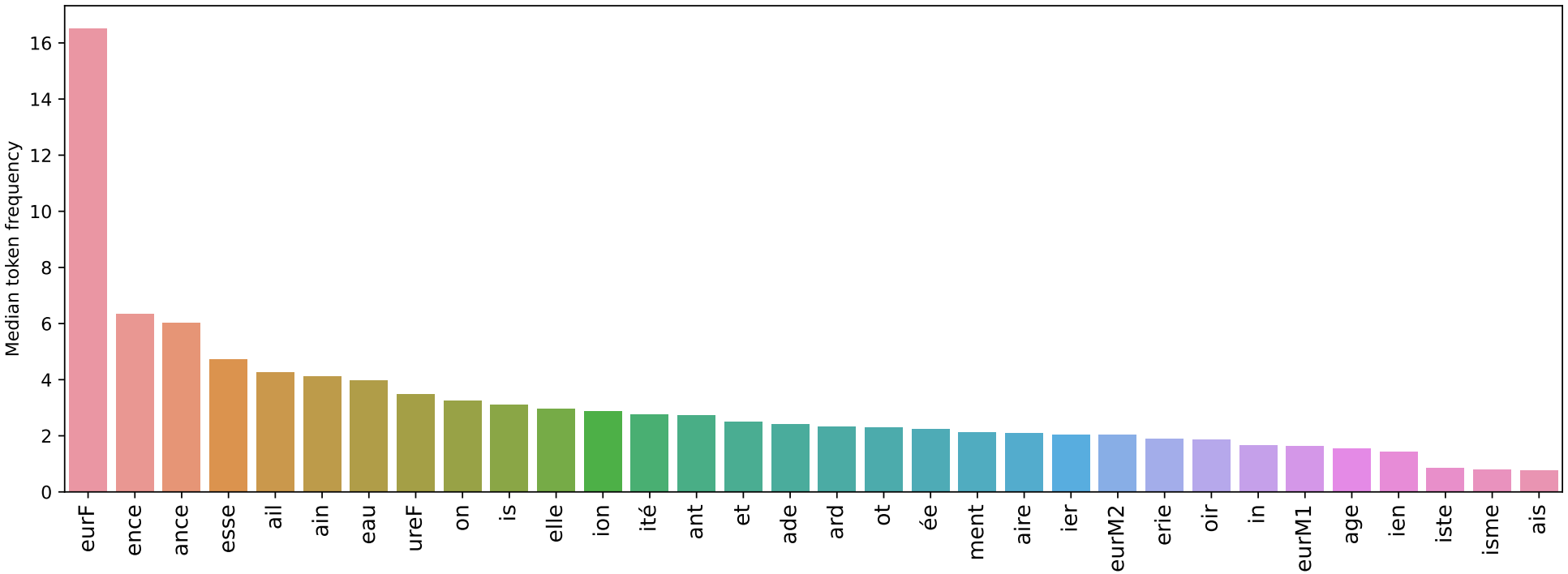
Broad types of last process

	Count	Proportion
Verb	887	48%
Noun	603	32%
Adjective	179	10%
No POS	101	5%
Name	83	4%
Numeral	11	1%
Adverb	1	0%

Base POS of suffixed nouns

Descriptive statistics II

- Interesting distribution of token frequency by affix:



Median token frequency of nouns based on the same suffix

- Striking high token frequency of abstract feminine nouns (*-eur*$_F$, *-ence*, *-ance*, *-esse*, *ité*, *-erie*)

Two approaches to formal transparency I

- We provide two separate measures of formal transparency (for suffixed nouns):
  1. The edit distance between the closest stem of the base and the derivational stem, e.g.
     - MENSUEL>MENSUALITÉ:
     1. Derivational stem: mãsɥalite $\ominus$ -ite = mãsɥal
     2. Closest stem of the base: mãsɥɛl
     3. ED(mãsɥɛl, mãsɥal) = 1
  2. The relative frequency of a surface pattern of alternation between citation forms, e.g.

$$\text{pat}(\text{mãsɥɛl}, \text{mãsɥalite}) = \text{\_ɛ\_}\sim\text{\_a\_ite}$$

$$\text{PRF}(\textsc{mensualité}) = \frac{|\text{suffixed in } \textit{-ité} \wedge \text{pattern is \_ɛ\_}\sim\text{\_a\_ite}|}{|\text{suffixed in } \textit{ité}|} = \frac{10}{55} \approx 0.18$$

Two approaches to formal transparency II

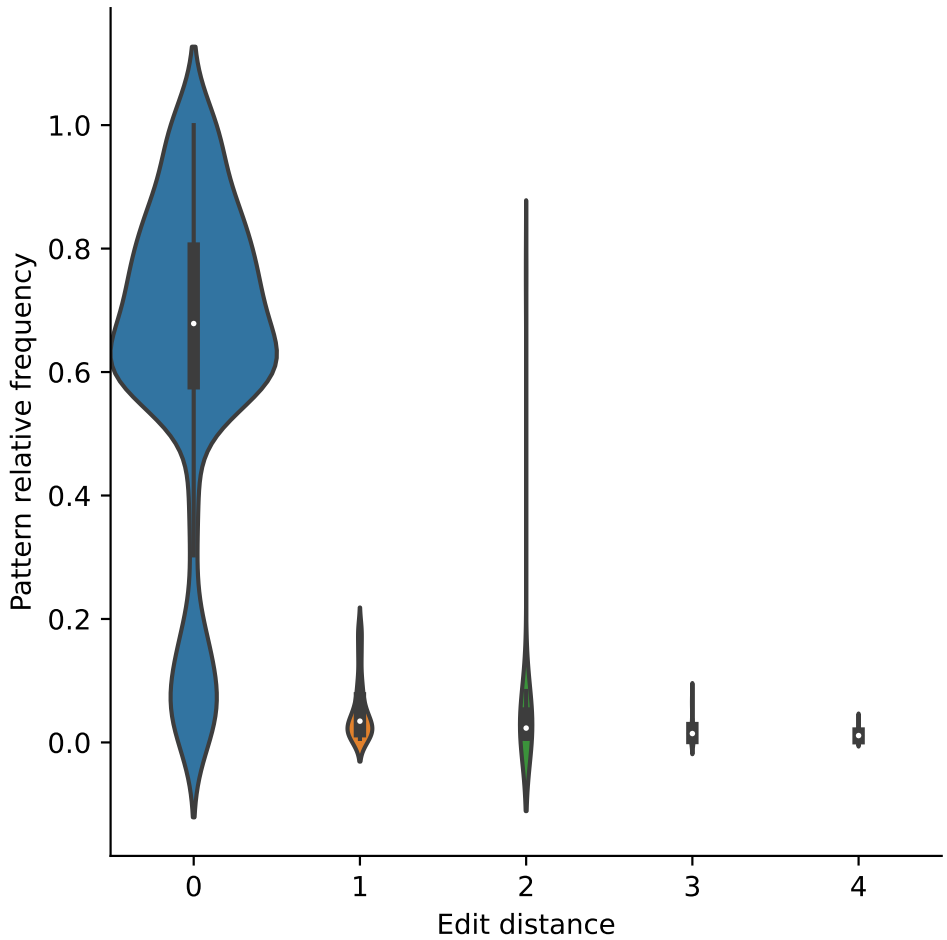
- The two measures are related but different:

Noun	Base	ED	PRF
TIMIDITÉ	TIMIDE	0	0.58
MENSUALITÉ	MENSUEL	1	0.18
SINGULARITÉ	SINGULIER	2	0.02

- Strnadová (2014) argues that pattern relative frequency is a better correlate of perceived regularity than edit distance between stems.
  - E.g. DISPERSER∼DISPERSION is less expected than AGITER∼AGITATION.
- How do the two measures compare in our dataset?

Two approaches to formal transparency III

- Interestingly, in our data:
  - Strong correlation between the two measures ($r = -0.62$)
  - In most cases the edit distance is 0, so that pattern relative frequency gives us more granularity.



Conclusions

- Hopefully Échantinom can be used:
  - To make statistically meaningful comparisons between word formation processes
  - As a training set for machine learning
  - As a test set for (semi-)automatically derived resources
- Please use it!

`https://osf.io/rdxqk`

References I

Bonami, Olivier and Gilles Boyé (2019). "Paradigm uniformity and the French gender system." In: *Perspectives on morphology: Papers in honour of Greville G. Corbett*. Ed. by Matthew Baerman, Oliver Bond, and Andrew Hippisley. Edinburgh: Edinburgh University Press, pp. 171–192 (cit. on p. 5).

Bonami, Olivier, Gauthier Caron, and Clément Plancq (2014). "Construction d'un lexique flexionnel phonétisé libre du français." In: *Actes du quatrième Congrès Mondial de Linguistique Française*. Ed. by Franck Neveu et al., pp. 2583–2596 (cit. on p. 14).

Hathout, Nabil and Fiammetta Namer (2014). "Démonette, a French derivational morpho-semantic network." In: *Linguistic Issues in Language Technology* 11.5, pp. 125–168 (cit. on p. 2).

Koehl, Aurore (2012). "La construction morphologique des noms désadjectivaux suffixés en français." PhD thesis. Université de Lorraine (cit. on p. 2).

Namer, Fiammetta et al. (2019). "Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats." In: *Actes de TALN*. Toulouse, France. URL: https://halshs.archives-ouvertes.fr/halshs-02275652/document (cit. on p. 2).

New, Boris et al. (2007). "The use of film subtitles to estimate word frequencies." In: *Applied Psycholinguistics* 28, pp. 661–677 (cit. on pp. 5, 14).

Schäfer, Roland and Felix Bildhauer (2012). "Building Large Corpora from the Web Using a New Efficient Tool Chain." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 486–493 (cit. on p. 14).

Strnadová, Jana (2014). "Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français." PhD thesis. Université Paris Diderot et Univerzita Karlova V Praze (cit. on pp. 2, 18).

References II

Tribout, Delphine (2010). “Les conversions de nom à verbe et de verbe à nom en français.” PhD thesis. Université Paris Diderot (cit. on p. 2).

Wauquier, Marine, Cécile Fabre, and Nabil Hathout (2020). “Semantic discrimination of technicality in French nominalizations.” In: *Zeitschrift für Wortbildung / Journal of Word Formation* 2/2020, pp. 100–121 (cit. on p. 2).