# Echantinom: a hand-annotated morphological lexicon of French nouns

Olivier Bonami, Delphine Tribout

# Échantinom: a hand-annotated morphological lexicon of French nouns

**Olivier Bonami**
Université de Paris, LLF, CNRS
olivier.bonami@u-paris.fr

**Delphine Tribout**
Université de Lille, STL, CNRS
delphine.tribout@univ-lille.fr

## Abstract

We present *Échantinom*, a new morphological resource for French nouns based on random sampling of frequent lexemes. The resource documents $5,000$ items in terms of their morphological type at two levels of granularity, as well as, for suffixed nouns, the exact identity of the base and process, and the formal and semantic transparency of the relationship between base and derivative. We outline the motivations for the development of such a resource, the sampling method, main annotation decisions, and provide some preliminary descriptive statistics.

## 1 Motivation

The very existence of the DeriMo workshop series testifies to a renewed interest in the development of large scale resources for derivational morphology. Table 1 lists most of the resources available for French, focussing on freely available machine-readable ressources developed in the last 15 years. This collection of resources provides a very rich view of the French word formation system; and an integration of those resources into a coherent unified database is the main goal of the ongoing Démonext project (Namer et al., 2019).

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

Table 1: Existing resources documenting French word formation

One defining characteristic of that collection of resources is that they were all constructed with a focus on *breadth* rather than *width*. Each resource was designed with the goal of documenting one or more specific word formation processes, and attempted to retrieve as many types as was possible given the practical constraints of the project. As a consequence, the sample of the French lexicon that is documented has strange characteristics. Some vanishingly rare derived lexemes are included in the sample, while very frequent ones are not, because the process they implement happens not to have been the focus of attention. Even for those processes that are documented, samples for different processes have different characteristics. For instance, *Lexeur* or *Dénom* contain many items not documented in dictionaries, because it was relatively straightforward to collect instances from corpus data; by contrast, *Converts* focuses on items documented in a dictionary, in the absence of a good method for extracting conversions semi-automatically from a corpus. An unwelcome consequence of this set of affairs is that there is no obvious way to make meaningful statistical comparisons of different processes by combining resources.

Another characteristic of this collection of resources is the variability of annotation both in terms of quantity and quality. For instance, *Démonette* contrasts with all the other sources in that pairings of derivationally-related words have not systematically been curated manually, leading to an undocumented quantity of false positives.

From these observations it follows that currently available resources do not provide us with a holistic view of the distribution of word formation processes in the lexicon. The goal of the present research is the development of a new resource that fills that gap: we provide a coherent and relatively detailed set of morphological annotations for a carefully sampled set of French nouns.

## 2 Sampling

The sampling procedure was as follows. We started from the *Lexique* (New et al., 2007) and *flexique* (Bonami et al., 2014) databases: *Lexique* provides various types of annotations for words attested in either a French literary corpus or a corpus or subtitles, and *flexique* tabulates all nouns, verbs and adjectives of *Lexique* in inflectional paradigms, and provides manually corrected phonemic transcriptions and grammatical gender information for all forms of the corresponding lexemes. We limited attention to the 13,046 nouns with a summed relative frequency in the two reference corpora higher than 0.3 per million, ensuring that we were focusing on nouns that are relatively frequent, but may be more prevalent either in formal or in informal French.[1]

Sampling was done in two steps. In an initial annotation campaign, we excluded all nouns homophonous with another noun, either with the same orthography but the other gender (there can be inanimates, e.g. LIVRE$_F$ 'pound' vs. LIVRE$_M$ 'book', or animates, e.g. JOURNALISTE$_{F/M}$ 'female/male journalist'), or with different orthographies (e.g. SERRE 'greenhouse' vs. CERF 'deer'). These constitute 8% of the 13,046 nouns we sampled from. This particular sampling strategy was motivated by the needs of a separate study on the phonological and morphological predictability of gender (Bonami et al., 2019). In a second annotation campaign, we first sampled 318 nouns with homophones so as to rebalance the sample; we then sampled more nouns until we reached a total of 5,000 nouns after exclusion of tagging errors. Note that, for purposes of sampling, masculine and feminine variants of common gender nouns such as JOURNALISTE were counted as two separate items; hence for some nouns (e.g. HUMORISTE 'comedian') both the masculine and the feminine variants are present in our sample, while for others either the feminine (e.g. HUMANISTE 'humanist') or the masculine (e.g. EXISTENTIALISTE 'existentialist') is. This is of course a disputable choice (Bonami and Boyé, 2019), but there was no way of avoiding taking a stance on the status of human common gender nouns.

## 3 Manual morphological annotation

The 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. All difficulties were discussed and decisions were made collectively. After guidelines for the annotation were drawn up,[2] the remaining nouns were distributed between the authors. All problems and questions were discussed and solved collectively.

Each noun was annotated for different properties. First, we annotated the broad morphological status of the noun as being either simplex or not; nonsimplex nouns were then classified on the basis of the outermost word formation process involved: prefixation, suffixation, conversion, any nonconcatenative process (`nonconcat` in the tables) or formation from more than one word (`polylexical` in the tables). When there was uncertainty as to what the last process was, we relied on frequency for arbitration. For example, SOUS-ALIMENTATION 'undernourishment' is ambiguous between an outermost prefixation (from ALIMENTATION 'feeding') or suffixation (from SOUS-ALIMENTER 'undernourish'). Because ALIMENTATION has a higher frequency than SOUS-ALIMENTER in *Lexique*'s reference corpora, we considered it to be the base of SOUS-ALIMENTATION and thus coded the last process as prefixation.

[1] The particular threshold of 0.3 per million was motivated by backward compatibility with the previous study of Tribout et al. (2014), although nothing crucial hinges on that choice.

[2] The guidelines are distributed with the resource in the following OSF repository: `https://osf.io/rdxqk/`.

All broad morphological categories except prefixation and suffixation were divided into fine grained sub-categories. Simplex nouns can be native underived nouns (e.g. CAHIER 'notebook'), borrowings (e.g. JAZZ), antonomasia (e.g. POUBELLE 'bin') or onomatopeic nouns (e.g. CLIC 'click'). The nonconcatenative processes found in the database are reduplication (e.g. BABALLE, from BALLE 'ball'), back formations (e.g. NUMISMATE 'numismatist', from NUMISMATIQUE 'numismatics'), slang processes such as verlan (e.g. KEUF from FLIC 'cop') or louchébem (e.g. LARFEUIL, from PORTE-FEUILLE 'wallet') and different types of truncation: mere apocope (e.g. IMPRO, from IMPROVISATION 'improvisation'), apocope with addition of an ending (e.g. VALOCHE, from VALISE 'suitcase') and apheresis (e.g. SCOPE, from MICROSCOPE 'microscope'). Among polylexical processes, we distinguished native compounds (e.g. SÈCHE-CHEVEUX, 'hairdryer', from SÉCHER 'dry' and CHEVEUX 'hair'), neoclassical compounds (e.g. BARYTON, 'baritone'), blends (e.g. FADETTE, from FACTURE 'bill' and DÉTAILLÉE 'detailed'), acronyms (e.g. SIMA, from SILICIUM 'silicon' and MAGNÉSIUM 'magnesium') and frozen word sequences, which we call `agglomerate`s (e.g. ARC-EN-CIEL 'rainbow', litterally 'bow in sky'). The difference between native compounds and agglomerates lies in the nature of elements: a lexeme was classified as an agglomerate if and only if one of the combined expressions is a grammatical word (e.g. *en* in ARC-EN-CIEL) or an inflected form (e.g. *dira* in QU'EN-DIRA-T-ON 'word of mouth', litt. 'what will one say'). Conversions were classified by base part of speech; Table 2 gives an example for each of the documented situations. Note that, following (Tribout, 2012), we distinguish four subcases of conversion from verbs depending on

| POS | Stem type | Base | Translation | DERIVATIVE | Translatiin |
|---|---|---|---|---|---|
| Adjective | — | PEUREUX | 'fearful' | PEUREUX | 'fearful person' |
| Verb | basic stem | RECHERCHER | 'research' | RECHERCHE | 'research' |
| | infinitive | SOUVENIR | 'remember' | SOUVENIR | 'memory' |
| | past participle | ENTRER | 'enter' | ENTRÉE | 'entrance' |
| | learned | CONCEVOIR | 'conceive' | CONCEPT | 'concept' |
| | indeterminate | FAILLIR | 'fail' | FAILLITE | 'bankruptcy' |
| Noun | — | RAVIN | 'ravine' | RAVINE | 'small ravine' |
| Proper name | — | SUISSE | 'Switzerland' | SUISSE | 'Swiss' |
| Adverb | — | DEHORS | 'outside' | DEHORS | 'outside' |
| Pronoun | — | MOI | 'me' | MOI | 'ego' |
| Numeral | — | ONZE | 'eleven' | ONZE | 'the number eleven' |

Table 2: Examples illustrating the diversity of base part of speech in converted nouns.

which stem allomorph is used. Note also that all deverbal nouns that can be analyzed as conversions from past participles are so analyzed, irrespective of whether or not a suffix is involved in the formation of that participle: hence examples such as ENTRÉE 'entrance', ACCALMIE 'lull', VENUE 'arrival', and ENCEINTE 'enclosure' are all treated on a par.

While we neither provide a full account of a lexeme's derivational history nor its relationship to all members of its derivational family, we did use the tabular structure of the database to document more word formation processes involved in a lexeme's formation. For instance, in addition to the outermost process, we also noted in dedicated columns whether other word formation processes (conversion, compounding, prefixation, or suffixation) are involved in the formation of the noun. For instance, the entry for EMBARQUEMENT 'boarding' documents it as formed by suffixation from EMBARQUER 'board', but also notes that it contains the verb-forming prefix *en*.

This annotation is particularly useful in situations where determination of the base-derivative relationship is nonobvious. As a case in point, consider the situation with conversion between nouns and verbs. Following Tribout (2020), we distinguish three situations: where the verb is clearly morphologically complex, it has to be the base of the noun (e.g. RECHERCHER 'research' has to be the base of RECHERCHE 'research' because of the presence of the verb-forming prefix *re-*); where the noun is clearly morphologically complex, it has to be the base (e.g. PARLEMENT 'parliament' is clearly based on PARLER 'speak' by

*-ment* suffixation and hence is the base of PARLEMENTER 'negociate'); in all other cases (e.g. with CLOU 'nail' vs. CLOUER 'to nail'), directionality cannot be established—in particular Tribout (2020) shows that neither etymological information nor semantic intuitions are reliable indicators or directionality. We account for the commonality of all three types of cases by noting in the `conversion` column the existence of a relationship with a verb, but differentiate them by coding the last process as `conversion` for RECHERCHE, `suffixation` for PARLEMENT, and `simplex` for CLOU.

In the same spirit, for compounding, we noted the compound type even if compounding is not the last morphological process. For example, THALASSO is the apocope of THALASSOTHÉRAPIE 'thalassotherapy' which is a neoclassical compound. In this case, the `compound` column indicates `neoclassical` while the last morphological process is noted as `apocope`. When there was hesitation between prefixation or compounding, particularly when the first element corresponds to a preposition such as *sur*, *sous*, *arrière*. . . , we arbitrated in favor of prefixation. Finally, when suffixation is involved in the formation of the noun, we noted the suffix in a specific suffix column, be it the last process (e.g. MINCEUR 'slimness' from MINCE 'slim') or not (e.g. PORTE-CIGARETTES 'cigarette case' from PORTER 'to bear' and CIGARETTE 'cigarette' that itself comes from CIGARE 'cigar'), like we did for the other processes.

Because suffixes are the most frequent derivational processes in our data, in addition to the mention of the suffix we also annotated different kinds of information linked to the suffixation process. First, we annotated suffix identity at two levels of granularity: the `sfx` column indicates the surface orthography of the precise allomorph, while `sfx_broad` lumps together allomorphs and gendered variants. For instance, *-oir* as found in RASOIR 'razor' and *-oire* as found in PASSOIRE 'colander' are distinguished at the fine grained level but grouped together under *-oir* at the coarse grained level. Similarly, the *-able* and *-ible* suffixes found in NOTABLE 'noteworthy' (from NOTER 'to note') and NUISIBLE 'harmful' (from NUIRE 'to harm') are distinguished at the fine grained level and noted as two allomorphs of the suffix *-able* at the coarse grained level. It is important to note that, except for gender variation and basic allomorphy, the identification of the suffixes is only based on the form of the suffixes and the gender they assigned to the nouns: no semantic or syntactic information is taken into account. For example, we distinguished two *-ure* suffixes, one feminine (e.g. BRÛLURE$_{\text{F}}$ 'burn' from BRÛLER 'to burn') and one masculine (e.g. SULFURE$_{\text{M}}$ 'sulphide' from SOUFRE 'sulphur'), but only one suffix *-ier*, be it used to form the name of a tree (e.g. AMANDIER 'almond tree' from AMANDE 'almond'), a person (e.g. BANQUIER 'banker' from BANQUE 'bank') or an artifact (e.g. SUCRIER 'sugar bowl' from SUCRE 'sugar'). As a consequence, we identified only one suffix in cases of homonymous suffixes used in distinct derivational processes if they assign the same gender to the outputs. Therefore the resource contains one suffix *-age*, even if we usually differentiate two suffixation processes: one deverbal *-age* suffixation that forms action nouns (e.g. JARDINAGE 'gardening' from JARDINER 'to garden') and one denominal that forms collective nouns (e.g. OMBRAGE 'shade' from OMBRE 'shadow'). However, the difference between the two *-age* suffixations can still be retrieved through the part of speech of the base that is noted in a dedicated column. In addition to the fine grained and coarse grained suffixes, we noted in dedicated columns the base of suffixation, its part of speech and whether it is autonomous (e.g. MINCEUR 'slimness' from MINCE 'slim') or not (e.g. LACTOSE 'lactose' from LACT- 'milk'). In some cases the identification of the base is tricky. Below, we describe these cases and the decisions we made.

i) There could be a mismatch between the formal and the semantic base of suffixation, as in ROYALISTE 'royalist': it formally derives from the adjective ROYAL by addition of the suffix *-iste*, but it is semantically related to the noun ROI 'king' rather than the adjective ROYAL. In such cases we arbitrated in favor of the formal base.

ii) For all demonym formation processes as well as *-iste* suffixation, which form parallel adjectives and nouns, following (Roché, 2008) we considered a direct suffixation from the base to the inhabitant or supporter noun, without an intermediate adjectival step. For example, the noun PARISIEN 'parisian' is treated as directly derived from PARIS (1a), not from an adjective itself deriving from the city name (1b). However, in order to capture the relation between the noun and the homonymous adjective, the existence of an adjectival counterpart is noted in the `conversion` column.

(1) a. PARIS → PARISIEN$_N$
→ PARISIEN$_A$
b. PARIS → PARISIEN$_A$ → PARISIEN$_N$

The same method was applied to *-isme* and *-iste* nouns: when a shared base exists, both nouns were analyzed as derived from that base. For example, ARRIVISTE 'social climber' and ARRIVISME 'ambition' are annotated as both derived from ARRIVER 'to arrive'.

iii) Sometimes suffixation applies to a bound stem. If this stem appears in at least one other word, it was considered to be a non autonomous base (Corbin, 1987). For example, in DÉLATRICE 'informer' the *-rice* suffix applies to the string *délat-* that is also found in DÉLATION 'informing', so that *délat-* was annotated as the non autonomous base of DÉLATRICE.

iv) When the string the suffix attaches to is not found elsewhere in the lexicon, but the noun belongs to a derivational series—it has the shape and expected meaning of a derivative (Hathout, 2009), the noun was considered to be a suffixed noun having no base. Therefore, we noted `0` in the `base` column. For instance, MAQUETTE 'model' ends with *-ette* while the stem *maqu* does not appear in other words, so that it cannot be a non autonomous base. However, MAQUETTE has the same ending and the same diminutive meaning as suffixed nouns in *-ette* like FILLETTE 'small girl' (from FILLE 'girl') so that it belongs to the derivational series of diminutive nouns suffixed with *-ette*. Therefore MAQUETTE was annotated as a suffixed noun having no base.

## 4 Descriptive statistics

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

Table 3: Type frequency by broad morphological type

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

Table 4: Type frequency of suffixed nouns by base part of speech[3]

We briefly comment on some descriptive statistics. Table 3 reports the breakdown of the dataset in terms of broad morphological types. The striking results are the low prevalence of polylexical units, and the high prevalence of simplex nouns. The latter result is partly due to the presence of 431 borrowings, as well as many items which were morphologically analyzable at some point in the history of French (e.g. MANIÈRE 'manner', historically derived by conversion from a now disappeared adjective MANIER 'to be used with the hand', itself from MAIN 'hand') or were analyzable in Latin (e.g. VICTOIRE 'victory', from Latin VICTORIA which itself was derived from Latin VICTOR 'victor').

Since suffixes make up the bulk of nonsimplex nouns and have been annotated in more detail, we focus on them in the remainder of this paper. 82 broad suffixes are attested in the dataset, with a high diversity of type frequencies, as shown in Figure 1: 8 suffixes account for more than half of the data, and two thirds of the suffixes have a type frequency lower than 10. Table 4 shows that deverbal formations make up almost half of the data, dominating nouns and then adjectives.

Finally, we report in Figure 2 the median token frequency of derivatives by suffix, for those suffixes with 10 or more instances in the dataset. It is striking that suffixes forming abstract feminine nouns

[3]The 'No POS' label corresponds to situations where either there is no identifiable base (while there is an identifiable suffix) or the base is a bound stem.

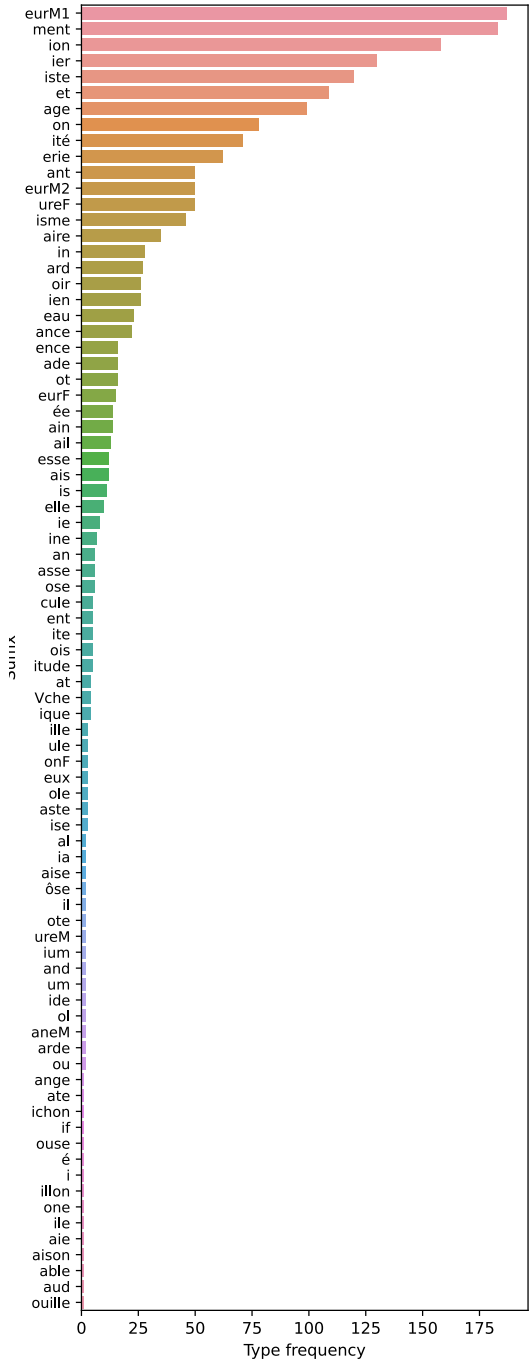


Figure 1: Type frequency of the 82 suffixes

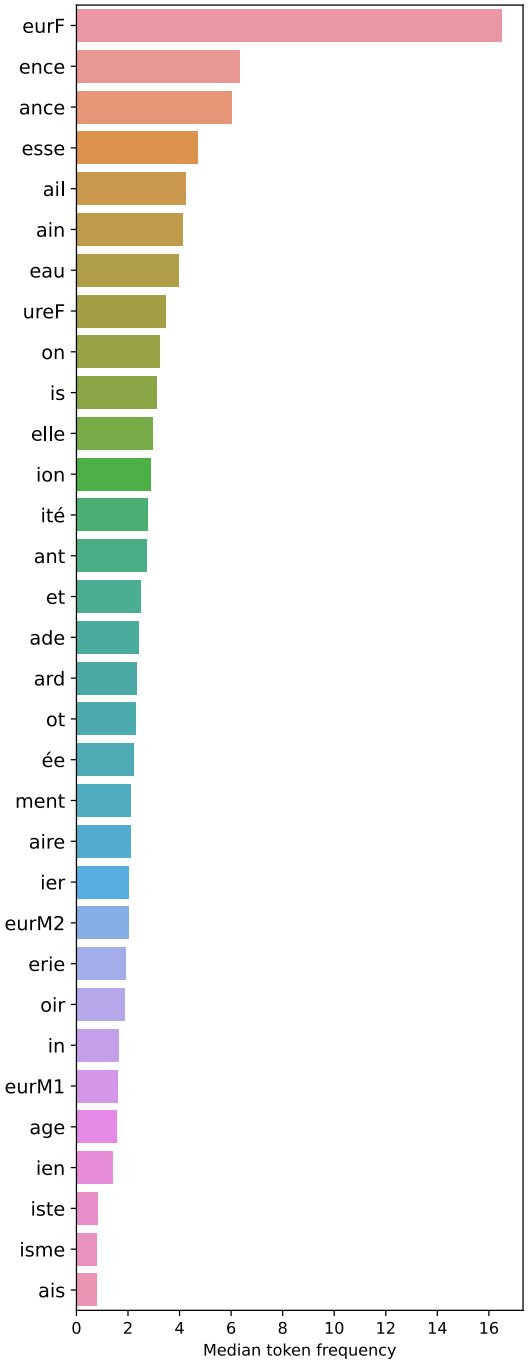


Figure 2: Median token frequency of the 32 most type-frequent suffixes

occupy the bulk of the high frequency range, above those forming individual or event-denoting nouns; and that *-isme* on the other hand has very low median frequency. This paper is not the place to attempt an explanation of these tendencies, but they illustrate the type of study allowed by a balanced annotated sample of a word formation system such as *Échantinom*.

## 5 Transparency

One of our goals with Échantinom was to document the formal and semantic transparency of suffixed derivatives so as to be able to use that information in future modelling efforts. After experimenting with using the raw intuitions of the authors, we concluded that these were unreliable, and that conducting a serious norming experiment over a multi-thousand item lexicon was out of the picture. Hence we report quantitative measures computed from the data.

### 5.1 Formal transparency

We report two measures of formal transparency: edit distance between base stem and derivational stem, and type frequency of patterns of alternation.

To compute the first measure, we first collected from *flexique* phonemic transcriptions for the citation forms of all nouns in the database. The derivational stem of suffixed derivatives was then deduced by simply stripping out the phonology of the appropriate suffix allomorph. Deducing the appropriate base stem was more challenging. First, we collected from *flexique* a reference stem for each lexeme: the singular form of nouns, the feminine singular form of adjectives, and the imperfect indicative 3SG form of verbs, stripped of the final /ɛ/. These are arguably the basic stem for each part of speech (Bonami and Boyé, 2003, 2005), and are definitely the stem allomorph most often relied on by suffixal derivation. Second, we computed an adapted Levenshtein distance between the derivational stem and the candidate base stem, which ignores differences between tense and lax mid-vowels, and between nasal vowels and matching vowel-/n/ sequences. Third, we examined by hand all cases where the resulting edit distance was larger than zero: in some cases this corresponds to genuine lack of formal transparency, in others it was found to be due to regular morphophonology, or the choice of a distinct stem allomorph. In addition, there were a few dozen of cases where the lexeme documented as the base contains a suffix absent from the derivative; e.g. the base for INSOUCIANCE 'carelessness', suffixed in *-ance*, is INSOUCIANT 'careless', itself suffixed in *-ant*. While this is a sensible decision, it leads to an artificially inflated formal distance between the base and derivational stem. In all such cases, the derivational stem was corrected by hand. We report in the resource the edit distance between the derivational stem and this manually corrected base stem. For instance the distance between INFORMATION 'information' and its base INFORMER 'inform' is `0`, that between INTERDICTION 'prohibition' and INTERDIRE 'forbid' is `2`, and that between DESTRUCTION 'destruction' and DÉTRUIRE 'destroy' is `4`.

Whether edit distance is a good measure of formal transparency in derivational morphology is disputable; Strnadová (2014, chap. 4) argues that it is not, and that the type frequency of patterns of alternation between surface forms is a better indication. The idea is that alternations that are judged as opaque are not those that are formally complex but those that are unexpected, and that unexpectedness is a consequence of low type frequency. To provide a rough operationalization of Strnadova's idea, as follows, we used the difflib Python library's SequenceMatcher algorithm to identify patterns relating the citation forms of the base and the derivative, [4] and then report the relative frequency of a pattern among derivatives formed with the same suffix. The higher the frequency of a pattern is, the more transparent the noun is. For example, the alternation pattern between INFORMATION and INFORMER (_∼_asjɔ̃) has the highest relative frequency (around 0.591) among *-ion* derivatives, which indicates that INFORMATION is very transparent. Conversely, the alternation pattern between CONTRADICTION 'contradiction' and CONTREDIRE 'contradict' (_ə_z∼_a_ksjɔ̃) has the lowest frequency (around 0.007), which indicates that the noun is not transparent. Note that the accuracy of our estimation of relative frequencies is highly dependent of the overall frequency of the affix; while we report relative frequencies for all suffixal formations,

[4] The use of SequenceMatcher as a rough but efficient way to classify surface alternations is inspired by Hathout et al. (2020). See Beniamine (2017) for a much more principled approach to the topic.

we advise against using them for suffixes with fewer than 10 types.

### 5.2 Semantic transparency

To operationalize semantic transparency, we rely on distributional semantics (see Boleda 2020 for a recent overview). We rely on a distributional vector space computed from frcow (Schäfer and Bildhauer, 2012) corpus for the purposes of Guzmán Naranjo and Bonami (2021), which provides lexeme-based rather than word-based distributional vectors for French.[5]

Using these vectors, we computed two separate measures of semantic transparency. First, we provide the cosine similarity between the vector for a suffixed noun and the vector for its base. This captures the idea that words that are transparently related occupy adjacent regions of semantic space, which may not be the case if the relationship is not transparent (see e.g. Varvara et al. 2021 for recent discussion). For instance, the cosine similarity between the vector of INITIALISATION 'initialisation' and that of INITIALISER 'initialize' is around 0.785, which indicates a high semantic similarity between the noun and the verb. Conversely, the cosine between the vector of PRESSION 'pressure, stress' and that of PRESSER 'press' is very low (around 0.011), which correlates with the semantic difference between the two words, as the noun is usually used with a psychological meaning while the verb has almost always a physical denotation.

In addition to this, we provide a measure of the predictability of the relationship between base and derivative. To this effect, for all suffixed nouns, we compute the difference (or offset) between the vector for the derivative and the vector for the base: this represents the shift in semantic space from the base semantics to the derived semantics. Such offset vectors tend to be similar for instances of the same derivational processes or even for rival processes (Guzmán Naranjo and Bonami, 2021). However, within a set of pairs of words related by the same process, we expect to find some variation, with transparent formations having very similar vectors while opaque ones will diverge (Bonami and Paperno, 2018). Hence we compute the average of offset vectors for all derivatives formed using the same suffix, and then the cosine similarity between that average and each individual offset vector. This similarity measure, which we call *offset vector typicality*, tells us the extent to which one instance of a derivational process implements a semantic relation that is similar to what happens on average for other instances of that process. For instance, the offset vector typicality of DESTRUCTION is high (around 0.8), in contrast to that of MUNITION 'ammunition' (from MUNIR 'to provide'), which is about 0.4. Just as with formal transparency assessed through the type frequency of patterns, the quality of our evaluation of offset vector typicality is heavily dependent on the number of datapoints going into the average vector. Hence, while we provide numbers for all derivatives, we urge users to proceed with caution, and definitely advaise against using them for suffixes with fewer than 10 types.

### 5.3 Discussion

In both the formal and semantic dimension, we provided two operationalizations of transparency: one based on bare comparison of base and derivative, the other based on an assessment of the typicality of their relationship. In the case of formal transparency, we observe a strong although far from perfect correlation between the two measures (Pearson's $r = -0.62$). The violin plot in Figure 3 confirms that there are very few cases where a nonzero edit distance does not coincide with a very low pattern frequency. This suggests that, despite Strnadová's (2014) principled reservations, in practice, edit distance is not such a bad indicator of the formal regularity of a derivative. In the case of semantic transparency on the other hand, our two measures are not correlated at all (Pearson's $r = -0.02$). The density plot in Figure 4, also suggests no interesting nonlinear relationship between the two variables: this suggests that the two measures indeed capture very different aspects of similarity. This is unsurprising though, for the following reason. While distributional vectors do capture some lexical semantic contrasts, they are also heavily influenced by aspects of distribution that have little to do with the two words corresponding to related concepts. As a case in point, consider the fact that the nouns PATINAGE 'skating' and PATINEUR

[5]The vector space is based on a curated version of the lemmatization provided by the corpus, and was obtained using the gensim (Řehůřek, 2010) implementation of the word2vec algorithm (Mikolov et al., 2013). The vector space is lexeme-based in the sense that each word was replaced by its lemma: thus the vector space documents the distribution of lexemes among other lexemes, rather than inflected forms among other inflected forms.

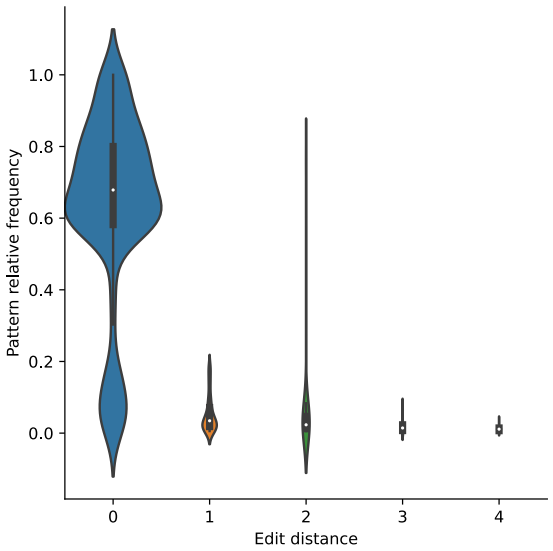


Figure 3: Relative distribution of the two measures of form transparency

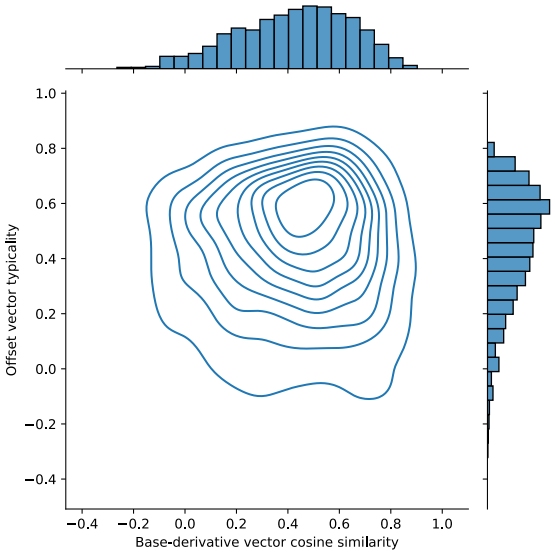


Figure 4: Relative distribution of the two measures of semantic transparency

'skater' are much more similar to one another (cosine similarity 0.72) than either is to the verb PATINER 'skate' (respective cosine similarities 0.20 and 0.28), whereas intuitively the even noun is semantically closer to the verb: clearly what is at play here is the general distributional similarities among nouns and differences between nouns and verbs. On the other hand, offset vector similarity should not be influenced by such factors; as a matter of fact, the similarity between the base and the derivative plays no role here: we do not care how distant they are from one another, but only about the direction in which the difference vector points.

We leave it to future research, or to the attention of future users of the resource, to study how formal and semantic transparency correlate with one another and with other variables of interest.

## Acknowledgments

Part of this work was done in the context of a project on gender assignment in French, in collaboration with Matías Guzmán Naranjo, and which benefited from in internship by Nadège Demanée. We thank them both for their input on relevant aspects of the research. This work was partially supported by the Démonext project (ANR 17-CE23-0005) as well as a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

## References

Sacha Beniamine. 2017. Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques. In *Actes de TALN 2017*. pages 77–85.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1):213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303.

Olivier Bonami and Gilles Boyé. 2003. Supplétion et classes flexionnelles dans la conjugaison du français. *Langages* 152:102–126.

Olivier Bonami and Gilles Boyé. 2005. Construire le paradigme d'un adjectif. *Recherches Linguistiques de Vincennes* 34:77–98.

Olivier Bonami and Gilles Boyé. 2019. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond, and Andrew Hippisley, editors, *Perspectives on morphology: Papers in honour of Greville G. Corbett*, Edinburgh University Press, Edinburgh, pages 171–192.

Olivier Bonami, Gauthier Caron, and Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, and Sophie Prévost, editors, *Actes du quatrième Congrès Mondial de Linguistique Française*. pages 2583–2596.

Olivier Bonami, Matías Guzman Naranjo, and Delphine Tribout. 2019. The role of morphology in gender assignment in French. Presented at the Second International Symposium on Morphology (ISMo 2019).

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2):173–195.

Danielle Corbin. 1987. *Morphologie dérivationnelle et structuration du lexique*. Max Niemeyer Verlag, Tübingen.

Matías Guzmán Naranjo and Olivier Bonami. 2021. Comparing derivational processes with distributional semantics. Presented at the second Paradigmo workshop.

Nabil Hathout. 2009. *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à diriger des recherches. Toulouse 2 - Le Mirail.

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

Nabil Hathout, Franck Sajous, Basilio Calderone, and 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)*. Marseille, pages 3870–3878.

Aurore Koehl. 2012. *La construction morphologique des noms désadjectivaux suffixés en français*. Ph.D. thesis, Université de Lorraine.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout, and Delphine Tribout. 2019. Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de TALN*. Toulouse, France. https://halshs.archives-ouvertes.fr/halshs-02275652/document.

Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28:661–677.

Radim Řehůřek. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pages 45–50.

Michel Roché. 2008. Structuration du lexique et principe d'économie : le cas des ethniques. In *Actes du Congrès Mondial de Linguistique Française 2008*. pages 1571–1585.

Roland Schäfer 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*. pages 486–493.

Jana Strnadová. 2014. *Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français*. Ph.D. thesis, Université Paris Diderot et Univerzita Karlova V Praze.

Delphine Tribout. 2010. *Les conversions de nom à verbe et de verbe à nom en français*. Ph.D. thesis, Université Paris Diderot.

Delphine Tribout. 2012. Verbal stem space and verb to noun conversion in french. *Word Structure* 5(1):109–128.

Delphine Tribout. 2020. Nominalization, verbalization or both? Insights from the directionality of noun-verb conversion in French. *Zeitschrift für Wortbildung / Journal of Word Formation* 2/2020:187–207.

Delphine Tribout, Lucie Barque, Pauline Haas, and Richard Huyghe. 2014. De la simplicité en morphologie. In *Actes du 4$^e$ Congrès Mondial de Linguistique Française (CMLF 2014)*. volume 8 of *SHS Web of Conferences*, pages 1879–1890.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology* https://doi.org/https://doi.org/10.1007/s11525-021-09382-w.

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