A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*

Olivier Bonami

Université Paris Diderot & LLF

olivier.bonami@linguist.univ-paris-diderot.fr

Juliette Thuilier

Université Toulouse Jean Jaurès & CLLE-ERSS

juliette.thuilier@univ-tlse2.fr

Abstract

Rivalry in lexeme formation refers to a situation where multiple, rival lexeme formation processes may be used to fill a gap in a morphological family. In this paper we study one such situation, the rivalry between the suffixes *-iser* and *-ifier* in French to derive verbs from nouns and/or adjectives. We propose a statistical approach to the problem, and use multivariate logistic regression applied to a large dataset derived from existing ressources to establish that phonological, morphological, and semantic properties of the morphological family all contribute independently to predicting preference for one or the other suffix. One main result of this study is that rivalry can not be studied in terms of the relationship of a single base and a derived lexeme, as multiple members of the morphological family play a role in jointly predicting the choice of a suffix.

1 Introduction

A notable property of systems of lexeme formation is their many-to-many nature:[1] more often than not, a single process may be used to construct derived lexemes of multiple syntactic and semantic types (1 form - many contents), while multiple processes may be used to construct derived lexemes holding the same syntactic and semantic relationship to their base (many forms - 1 content). This is illustrated in Table 1 on the basis of a sample of French deverbal nouns: the table illustrates various formal operations in columns, and various syntactic/semantic types of deverbal nouns in rows. As the table shows, most processes are associated with multiple contents and most derivational relations can be realized by multiple processes. However, this does not mean that the various processes lead to the exact same set of possible derivations: there are gaps in the table, suggesting that grammar still constrains the choice of a particular process to express a particular meaning.[2] The question is then to determine the nature of such constraints.

	-oir	*-age*	*-eur*	*-ette*
Patient	*tirer* > *tiroir* ‘draw’ ‘drawer’	—	—	*sucer* > *sucette* ‘suck’ ‘lollypop’
Instrument	*hacher* > *hachoir* ‘chop’ ‘chopper’	*maquiller* > *maquillage* ‘make up’ ‘makeup’	*tracter* > *tracteur* ‘tow’ ‘tractor’	*laver* > *lavette* ‘wash’ ‘dishcloth’
Agent	—	—	*nager* > *nageur* ‘swim’ ‘swimmer’	*coudre* > *cousette* ‘sew’ ‘seamstress’
Location	*laver* > *lavoir* ‘wash’ ‘washhouse’	*garer* > *garage* ‘park’ ‘garage’	—	*boire* > *buvette* ‘drink’ ‘small bar’
Event	—	*guider* > *guidage* ‘guide’ ‘guidance’	—	*bronzer* > *bronzette* ‘tan’ ‘sunbath’

Tab. 1: Sample of French deverbal nouns

A cursory look at the examples will lead an investigator with some knowledge of French towards the conclusion that non-categorical constraints are at play here. For instance, if we focus on formal processes and ask what meaning they associate with, it is clear that *-oir* nouns are much more commonly used to form Instrument or Location nouns than Patient nouns; all *-age* nouns are at least compatible with an Event interpretation, although only some of them lexicalize another type of content; *-eur* nouns seem just as likely to denote Agents or Instruments. If we examine the table horizontally and ask what formal means are used to express each meaning, *-oir* and *-eur* are the preferred affixes for coining Instrument nouns, although there are other possibilities; on the other hand, Location nouns in *-age* are rather uncommon.

The goal of this paper is to take the gradient, many-to-many nature of lexeme formation systems at face value, and to explore how statistical modelling applied to large lexical databases can shed light on the constraints at play in shaping the systems. Although the topic of rivalry between lexeme formation processes has been extensively studied (see among many others Aronoff 1976; Anshen and Aronoff 1981; Giegerich 1999; Plag 1999; Lindsay and Aronoff 2013 for English, Corbin 1987; Ferret et al. 2010; Lignon 2013; Strnadová 2014b; Tribout and Villoing 2014 for French), little literature is devoted to the quantitative study of its gradient nature. A notable exception is Arndt-Lappe (2014), who shows that an analogical model trained on phonological and syntactic properties of bases predicts with reasonable accuracy whether English denominal adjectives use the suffix *-ity* or *-ness*. While this result is definitely relevant, an analogical model mostly informs us on the existence of relevant correlations, without directly allowing for examination of which correlations hold. In this paper, we use a more conservative modelling tool, multivariate logistic regression, in an attempt to tease apart which properties of a derived lexeme's morphological family may influence the use of a particular formal process.

Process	Adjectival base	Nominal base
a-	*grand* > *agrandir*	*ligne* > *aligner*
	'large' 'enlarge'	'line' 'align'
é-	*large* > *élargir*	*fil* > *effiler*
	'wide' 'widen'	'thread' 'fray'
en-	*beau* > *embellir*	*flamme* > *enflammer*
	'beautiful' 'embellish'	'flame' 'set fire to'
conversion	*mûr* > *mûrir*	*liasse* > *liasser*
	'ripe' 'ripen'	'bundle' 'bundle'
-iser	*banal* > *banaliser*	*canal* > *canaliser*
	'banal' 'trivialize'	'canal' 'channel'
-ifier	*dense* > *densifier*	*momie* > *momifier*
	'dense' 'densify'	'mummy' 'mummify'

Tab. 2: Sample of French denominal verbs

The empirical domain to be explored is that of French verbs derived from nouns and adjectives. As Table 2 illustrates, the same six main processes are used in both cases: prefixation by *a-*, *é-* or *en-*, conversion, and suffixation by *-iser* and *-ifier*.[3] For practical reasons, we focus here on rivalry between *-iser* and *-ifier*, building heavily on previous descriptive work by Namer (2009) and Lignon (2013). Namer (2009) provides strong evidence that the same type of many-to-many relation discussed above in the case of deverbal noun also holds here: just like their English cognates (Plag, 1999; Lieber, 2004), *-iser* and *-ifier* are heavily polysemous, and give rise to the same range of meanings, although the likelihood of finding one or the other suffix is conditioned in part by the meaning to be expressed. Table 3 illustrates this situation: with the exception of the two small classes of 'instrumental' (the base noun denotes an instrument used to perform the action denoted by the verb) and 'production' (the base noun denotes the product resulting from the action denoted by the verb) verbs, all types of meanings are found with both *-iser* and *-ifier*; and the two exceptional classes are small enough that the absence of attestation of both suffixes might be due to chance.

While Namer provides a very useful classification on the basis of extensive data from dictionaries, the book only reports impressionistically on the relative frequency of the two suffixes in association with different meanings, and there is no distributed annotated dataset that one may draw from for statistical analysis. In fact, a robust, systematic morphosemantic annotation of

morphosemantic type	*-iser* examples	*-ifier* examples
causative	*banal* > *banaliser*	*rigide* > *rigidifier*
	‘trivial’ ‘trivialize’	‘rigid’ ‘rigidify’
ornative	*alcool* > *alcooliser*	*gaz* > *gazéifier*
	‘alcohol’ ‘alcoholize’	‘gas’ ‘gasify’
locative	*hôpital* > *hospitaliser*	*plastique* > *plastifier*
	‘hospital’ ‘hospitalize’	‘plastic’ ‘laminate’
resultative	*atome* > *atomiser*	*momie* > *momifier*
	‘atom’ ‘atomize’	‘mummy’ ‘mummify’
inchoative	*caramel* > *caraméliser*	*ours* > *oursifier*
	‘caramel’ ‘caramelize’	‘bear’ ‘become isolated’
instrumental	*cautère* > *cautériser*	—
	‘cautery’ ‘cauterize’	
production	—	*fruit* > *fructifier*
		‘fruit’ ‘yield a profit’
similative	*nomade* > *nomadiser*	*pontife* > *pontifier*
	‘nomad’ ‘behave like nomads’	‘pontiff’ ‘pontificate’

Tab. 3: Sample lexemes illustrating the parallel polysemy if *-iser* and *-ifier* (adapted from Namer 2009; labels for meanings taken from Plag 1999 where possible)

derived verbs would be a considerable endeavour in itself, that neither previous work nor the present paper has pursued.

Unlike Namer, Lignon (2013) adopts an explicitly quantitative approach, and provides descriptive statistics on the influence of morphological and phonological factors on the use of *-iser* and *-ifier* to derive verbs from adjectives. The present paper builds heavily on Lignon's results and attempts to improve on them in two directions. First, descriptive statistics do not allow one to determine whether the tendencies observed in a sample are robust enough that one can exlude their being due to chance; neither do they allow one to conclude on the relative role of highly correlated properties of bases, such as phonological and morphological characterizations of their shape. In both cases, statistical modeling is the analytic tool of choice to confirm or disprove Lignon's findings. Second, Lignon focuses on those verbs that may reasonably be taken to derive from an adjective. However, as we will discuss in detail below in section 2.2.2, it is often difficult, if not impossible, to decide whether a given verb derives form a noun or an adjective, when the morphological family under consideration contains both (Namer, 2013). As a consequence, it is not clear that the data chosen for consideration by Lignon forms a coherent sample, and that the method of data selection does not bias the results. In the present study we avoid this problem by considering denominal and deadjectival formations within the same models, and by making as few assumptions as possible on the exact identity of the base when it is not self-evident.

The structure of the paper is as follows. In section 2 we present how we compiled a dataset of 791 derived verbs and annotated them for various historical, phonological, morphological and semantic properties. Section 3 discusses the value of different properties of a derived lexeme's morphological family that seem to have an incidence on preference for *-iser* or *-ifier*. In section 4, we use multivariate logistic regression to tease apart the predictive value of different grammatical distinctions. Section 5 discusses the methodological and theoretical lessons we may draw from the statistical study.

2 Extraction of the data and annotation

2.1 Data selection

Any statistical study of the lexicon will strive to use a dataset that is as large as possible but on which rich and reliable annotation is readily available. For the present study, we decided to start from a combination of two lexical resources that provide much relevant information. First, we used GLÀFF (Hathout et al., 2014), a large scale inflectional lexicon of French derived from the French version of Wiktionary. GLÀFF gives easy access to morphosyntactic descriptions and phonological transcriptions present in Wiktionary as of May 2013, making it by far the largest machine-readable inflected lexicon for French that has also been validated by speakers (since a lexeme will be documented in GLÀFF only if a speaker of the language decided to create a description of that lexeme on Wiktionary). Second, we intersected the set of verbs in *-iser* and *-ifier* from GLÀFF with those that are attested as unigrams in the French Google Ngrams dataset (Michel et al., 2010).[4] The Google Ngrams dataset documents all n-grams (sequences of n orthographic words) identified through Optical Character Recognition in books from the Google Books collection.[5] This has two advantages. First, we have attestations of all the words under investigation at hand, confirming that they are in actual use. Second, the Google Ngrams dataset provides for each word a time series of number of attestations by year of publication of the books. While the datations are definitely noisy (due for instance to citation of one book by another, or multiple editions of the same book), this allows one to take into account evolution in the usage of a lexeme; in particular it allows for a rough estimate of its age.

This selection strategy gives us an initial set of 1263 verbs. The list was then filtered by hand so as to eliminate any false positives (verbs that happen to have an infinitive in *-iser* or *-ifier* without being suffixed, e.g. *miser* 'bet'). In addition, we left out all verbs beginning with a prefix, since in the vast majority of those cases prefixation is posterior to deverbal suffixation (e.g. *décoloniser* 'decolonize'), and it was preferable to eliminate the few cases where verbs were clearly derived from prefixed adjectives (e.g. *insonoriser* 'soundproof'), rather than to have to take a disputable decision in cases of apparent parasynthesis (e.g. *amenuiser* 'reduce' from *menu* 'slight'). Finally, we filtered out all verbs based on a learned Latin stem distinct from the corresponding noun or adjective stem (e.g. *pacifier* ‘pacify’, cf. *paix* ‘peace’, *paisible* ‘peaceful’; *sanctifier* ‘sanctify’, cf. *saint* ‘saint, holy’) and all verbs clearly borrowed from English (e.g. *pressuriser* ‘pressurize’, cf. English *pressure* vs. French *pression*; *randomiser* ‘randomize’, cf. English *random* vs. French *aléatoire*). These various filtering operations reduced our dataset to a final size of 791 lexemes.[6]

2.2 Annotation

2.2.1 Age of the lexeme

Any study of rivalry in lexeme formation must address the fact that constraints on the relative productivity of processes varies over time. As a result, observation of the extant lexicon at some point in time may lead to an inaccurate estimation of such constraints, given that the lexemes under examination often have been coined over centuries.

One obvious solution to this problem is to focus on lexemes coined during a relatively limited time span, e.g. the 20th century (see e.g. Plag, 1999). While this is certainly a wiser choice than not paying attention to the effects of time depth, it has a number of drawbacks, on both a practical and a conceptual level. First, for many processes, a shorter time span will lead to data sparseness, making it impossible for apparent generalizations to reach statistical significance. As a case in point, Plag’s discussion of *-ize* and *-ify* verbs in English is based on a dataset of 20th century neologisms from the OED which only contains 30 *-ify* verbs; this makes it very hard to reach any conclusion as to constraints on the use of that pattern.[7] Second, choosing a time span presupposes that one knows productivity not to have varied during that time span; but assumptions on the diachronic variability of productivity is precisely what one set out to avoid. Third and finally, any choice of a particular cutting point amounts to discretizing what is certainly an evolving situation: productivity usually does not suddenly change at particular points in time; when it does, this can only be established by detailed philological work.[8]

We thus submit that a more promising approach to addressing diachronic variation in productivity is to explicitly take into account the usage history of lexemes in our statistical models. This will allow us both to observe directly whether a change in productivity occurred (rather than presuming that it could have), and to assess whether some constraints hold irrespective of the time at which a lexeme was coined. To assess the age of lexemes, we rely again on the Google Ngrams dataset: we note as the operational date of birth of a lexeme the middle of the first sequence of 10 years during which 10 attestations are found in the corpus—i.e., the middle of the first sequence of 10 years with 1 attestation per year on average.[9] This is the only practical way of obtaining an estimation of date of coinage for our dataset, which contains many verbs not discussed in etymological dictionaries. In addition, it presents the advantage that the dating procedure is very uniform, since it is obtained automatically from a fixed diachronic corpus.

2.2.2 Identifying bases

As we already alluded to in the introduction, a main challenge posed by the study of competition between *-iser* and *-ifier* is the fact that both affixes combine with both adjectival and nominal bases. Since a single morphological family may contain closely related nouns and adjectives, there is in many cases uncertainty as to which lexeme should be considered the base of the verb. Table 4 illustrates the various situations that occur.

	Noun	Adjective	Derived verb
(i)	—	*banal*	*banaliser*
		'trivial'	'trivialize'
(ii)	*aval*	—	*avaliser*
	'approval'		'approve'
(iii)	*île*	*insulaire*	*insulariser*
	'island'	'insular'	'make insular'
(iv)	*république*	*républicain*	*républicaniser*
	'republic'	'republican'	'make republican'
(v)	*Staline*	*stalinien*	*staliniser*
	'Stalin'	'stalinist'	'make stalinist'
(vi)	*municipalité*	*municipal*	*municipaliser*
	'municipality'	'municipal'	'make municipal'
(vii)	*fédération*	*fédéral*	*fédéraliser*
	'federation'	'federal'	'federalize'
(viii)	*morale*	*moral*	*moraliser*
	'morality'	'moral'	'make ethical'
(ix)	*cardinal*	*cardinal*	*cardinaliser*
	'cardinal$_N$'	'cardinal$_A$'	'name cardinal$_N$/ render cardinal$_A$'
(x)	*hôpital*	*hospitalier*	*hospitaliser*
	'hospital'	'of hospital'	'hospitalize'
(xi)	*salaire*	*salarial*	*salariser*
	'salary'	'of salary'	'provide the status of salaried employee'

Tab. 4: Derived verbs in *-iser* and their relation to nouns and adjectives

The obvious way of deciding whether the noun or adjective is the base of a derived verb is to examine which of the two has a stem which matches the stem to which *-iser* is suffixed. This simple heuristic allows us to reach an easy decision in cases where there is no noun (i), there is no adjective (ii), the noun and the adjective stand in a suppletive relation (iii), one of the two is derived from the other by suffixation (iv-vi), or, even if the noun and the adjective do not stand in a direct derivation relation, the stem of one of them matches the stem to which *-iser* is attached (vii). Such a simple heuristic however proves unhelpful in a number of situations. First, there is a sizable number of cases where the noun and adjective stand in a conversion relation, so that they are equally good candidates to serve as a base (viii-ix). Conversely, there are many cases where the noun and adjective are equally bad candidates to be a base: in (x) and (xi), the stem alternant of the adjective is used in the verb, but without the adjectival suffix.

	Noun	Adjective	Derived verb
(i)	*algèbre* 'algebra'	*algébrique* 'algebraic'	*algébriser* 'make algebraic'
(ii)	*ethnie* 'ethnic group'	*ethnique* 'ethnic'	*ethniciser* 'give an ethnic character'
(iii)	—	*pathétique* 'pathetic'	*pathétiser* 'make pathetic'
(iv)	*drame* 'drama'	*dramatique* 'dramatic'	*dramatiser* 'make dramatic'
(v)	*politique* 'politics'	*politique* 'political'	*politiser* 'politicize'

Tab. 5: The special case of bases in *-ique*

More complications surface when we consider the situation of verbs with an adjective in *-ique* in their morphological family, as illustrated in Table 5. The simple form-based heuristic works well in cases such as (i) and (ii). However, most adjectives in *-ique* pair with a derived verb in *-iser* not including the suffixe *-ique* in any form, even when there is no corresponding noun (iii), the noun uses a different stem allomorph (iv), or the noun and adjective are homophonous (v). This strongly suggests that *-ique* is normally, though not obligatorily (ii), truncated before suffixation. As satisfactory as that suggestion may be from a theoretical perspective, in pratice it entails that our form-based heuristic is useless when it comes to morphological families containing an adjective in *-ique*, except in those rare instances where *-iser* indeed attaches to a stem in *-ic-* (ii).

Let us finally note that there are many cases where there is more than one noun or adjective to choose from as a possible base. The largest class of such cases is that of morphological families containing a place name and a matching demonym adjective, which in French can always also be used as a demonym noun. In effect, we thus systematically have two nouns and one adjective in the morphological family, with one of the two nouns homophonous to the adjective. As Table 6 illustrates, the formal relationship between the place name and the demonym is variable: suffixed demonym (i, ii), suffixed place name (iii), independent suffixes on a common stem (iv), identity (v), suffixed demonym on a suppletive (vi) or allomorphic (vii) stem. The stem used in the derived verb can match that of the place name (ii), the demonym (i, iii), both (v), or neither (iv, vi, vii). Thus in this large family of cases, we do not even know which of the two nouns in the

	place name	demonym A	demonym N	Derived verb
(i)	*Afrique*	*africain*	*Africain*	*africaniser*
	‘Africa’	‘African’	‘African’	‘africanize’
(ii)	*Japon*	*japonais*	*Japonais*	*japoniser*
	‘Japan’	‘Japanese’	‘Japanese’	‘make Japanese’
(iii)	*Turquie*	*turc*	*Turc*	*turciser*
	‘Turky’	‘Turkish’	‘Turk’	‘make Turkish’
(iv)	*Albanie*	*albanais*	*Albanais*	*albaniser*
	‘Albany’	‘Albanian’	‘Albanian’	‘make Albanian’
(v)	*Corse*	*corse*	*Corse*	*corsiser*
	‘Corsica’	‘Corsican’	‘Corsican’	‘make Corsican’
(vi)	*Grande Bretagne*	*britannique*	*Britannique*	*britanniser*
	‘Great Britain’	‘British’	‘Brit’	‘make British’
(vii)	*Asie*	*asiatique*	*Asiatique*	*asiatiser*
	‘Asia’	‘Asian’	‘Asian’	‘make Asian’

Tab. 6: Verbs derived from locative morphological families

morphological family we should take into consideration to decide whether the base is a noun or an adjective.

In the absence of a workable way of deciding what the base of a derived verb is on the basis of form, we may turn to semantics for help. This however proves even more problematic. First, an operational procedure that can be deployed on a large scale to decide whether a verb is semantically based on a noun or adjective is currently out of reach. A quick examination of the data discussed so far should convince the reader that in most cases, it is easy to come up with natural paraphrases for the verb that are based on the meaning of either the noun or the adjective (e.g. *moraliser* ‘make compatible with morality’ vs. ‘make moral’). Second, where there is a strong intuition going in one direction, the semantics-based heuristic does not necessarily match the form-based heuristic (see Namer 2013 for extensive discussion). As a case in point, consider the verb *staliniser* from Table 4: the meaning connects more closely to that of the adjective (‘turn into something stalinist’, not ‘turn into Stalin’), but the form does not. Third and finally, there are many cases where the derived verb is actually ambiguous between two meanings connecting to the noun and the adjective respectively—witness the case of *cardinaliser* in Table 4.[10]

We thus conclude that there is no operational way of deciding, in general, whether the base of an *-iser* or *-ifier* verb is a noun or adjective, when the derivational family contains both a noun and an adjective; in many cases, there are even strong arguments to the effect that such a decision does not make sense.

2.2.3 Properties of the morphological family

Whereas we cannot decide what the base of a derived verb is, we still want to collect information on its morphological family that may affect preferences for one or the other suffix. We thus classified verbs in three groups according to the makeup of their morphological family:

N: Verbs whose morphological family contains at least one noun that is a fitting possible formal base, but no such adjective.

A: Verbs whose morphological family contains at least one adjective that is a fitting possible formal base, but no such noun.

both: Verbs whose morphological family contains at least one noun and one adjective that are fitting possible formal bases.

Decisions on inclusion were based on the nomenclature of French Wiktionary as accessible through *GLAWI* (Hathout and Sajous, 2016). Note that the classification makes no attempt to document which of the noun or adjective is the more likely base, just whether there is such a candidate noun or adjective. Table 7 provides counts and examples.

		Example	
Class	Counts	Noun	Adjective
N	60	*aval*	—
A	114	—	*banal*
both	617	*république*	*républicain*

Tab. 7: Classification by profile of the morphological family

Where there is both a noun and an adjective in the morphological family, we found it useful to collect more information on the relationship between the noun and adjective. For this we relied mostly on the *Dénom* database (Strnadová, 2014a). On the formal side, we coded whether the noun and adjective were in a direct derivational relation, formed a conversion pair, or were in no direct relation.[11] In the subcase of N>A suffixation, we further classified the pairs by affix, as indicated in Table 8.

Because of the very high prevalence of suffixed verbs that contain both a noun and an adjective in their morphological family, it is relevant to explore how different classes of relationships between the noun and the adjective influence preference for *-iser* or *-ifier*. Besides examination of the formal process relating the two, an obvious parameter to take into consideration is whether the adjective qualifies as a relational adjective (Bally, 1944). Bally's basic intuition is that in some attributive uses, an adjective plays a syntactic and semantic role equivalent to that of a PP

Class	Counts	Example Noun	 Adjective
Conversion	192	*politique* ‘politics’	*politique* ‘political’
A>N suffixation	5	*municipalité* ‘municipality’	*municipal* ‘municipal’
No direct relation	4	*fédération* ‘fedaration’	*fédéral* ‘federal’
N>A suffixation	416		
-ique	172	*volcan* ‘volcano’	*volcanique* ‘volcanic’
-al	72	*globe* ‘globe’	*global* ‘global’
-el	49	*forme* ‘form’	*formel* ‘formal’
-aire	33	*banque* ‘bank’	*bancaire* ‘banking’
-ien/-éen	23	*Brésil* ‘Brazil’	*brésilien* ‘Brasilian’
-eux	15	*os* ‘bone’	*osseux* ‘bony’
-ais	11	*Japon* ‘Japan’	*japonais* ‘Japanese’
-in	10	*cristal* ‘crystal’	*cristallin* ‘crystalline’
-ain	10	*république* ‘republic’	*républicain* ‘republican’
other	21	*hôpital* ‘hospital’	*hospitalier* ‘of hospital’

Tab. 8: Classification of morphological families with N and A by derivational relation

complement; in particular it may saturate an argument position of the head noun it combines with. This characterization purportedly explains the felicity of paraphrases relying on a PP, as illustrated in the attested example in (1-a) from Canadian parliamentary debates; as well as the reluctance of such adjectives to be used predicatively (1-b) or to be subject to degree modification (1-c).

(1) a. Les allusions à la mauvaise **gestion budgétaire** sont à mon avis un véritable canular. Personne ici ne verrait d’un oeil favorable la mauvaise **gestion des budgets** destinés aux enfants.[12]
‘This idea about **fiscal mismanagement** [litt. (bad) budgetary management] I think is really a red herring. None of us around this table would view in any favourable terms the **mismanagement of dollars** [litt. (bad) management of the budgets] intended to go to children.’

b. #Sa gestion est budgétaire.
(litt.) ‘Its management is budgetary.’

c. #une gestion très budgétaire
(litt.) ‘a very budgetary management’

In this, relational adjectives contrast with what Bally calls ‘true adjectives’ (also known as *qualifying adjectives* in the French tradition), which may or may not be morphologically related to a noun. For instance *brutal* ‘brutal’ derives from the noun *brute* ‘brute’ but has none of the relevant properties.

(2) a. une augmentation brutale $\neq$ une augmentation de brute
'a sudden rise' 'a rise of a brute'
b. L'augmentation a été brutale.
'The rise was sudden.'
c. une augmentation très brutale
'a very sudden rise'

The notion of a relational adjective is both omnipresent in Romance grammars and remarkably elusive. First, it is well known that many adjectives give rise to both relational and non-relational readings, cf e.g. *la pression populaire* 'public pressure' vs. *une chanson populaire* 'a popular song'. Second, in some contexts, relational readings arise where they are not supposed to according to the criteria above (McNally and Boleda, 2004), as shown by the following attested examples.

(3) a. [...] la CAF [...] pratique une gestion très budgétaire de la question [...][13]
'the child benefit office manages the issue in stricly budgetary terms' (litterally: makes use of a highly budgetary management of the issue)
b. La gestion est budgétaire et mesure les moyens indépendamment des résultats que l'on peut obtenir dans nos productions.[14]
'The management is budget-based' (litterally: is budgetary) 'and measures means independently of the results we may get in our productions.'

Third, the traditional criteria for relational adjectives do not really allow for a partition of adjectives (or adjective readings) in two sets: rather, subclasses of adjectives satisfy various subsets of criteria (Fradin, 2007). Fourth, the ideas that relational adjectives are 'semantically vacuous' or 'equivalent to a noun' do not resist careful scrutiny (Rainer, 2013; Strnadová, 2014a).

Be that as it may, it remains that some adjectives give rise to relational readings and others do not. This suggests that, as elusive as it may be, there exists some lexical semantic partition of adjectives that accounts for the ability of some to possess such readings. In addition, the close semantic relationship between nouns and relational adjectives suggests that choosing a noun or the related adjective as base has no consequence beside phonology. As a matter of fact, in his study of derivatives in *-isme* and *-iste*, Roché (2011) suggests that the derivational stem often coincides with the stem of a relational adjective, although the derived lexemes clearly relate morphosemantically to the corresponding noun. If that is so, we do expect that the inclusion of a relational adjective in a morphological family may have an incidence on how that family is extended with a derived verb.

We thus tentatively annotated our data points for this property. For this annotation, we considered all 617 morphological families containing both a noun and an adjective, irrespective of the nature of the morphological relation between them. For each of these, we attempted to determine whether the adjective possessed at least one relational reading. In this we were guided by the definitions in French Wiktionary (as accessible through *GLAWI*): a definition of the form *relatif à N* 'relative to N', *en rapport avec N* 'in a relation with N', *de N* 'of N', etc., was taken as a hint that we were dealing with a candidate. We then used our own intuition on the felicity of predicative use and degree modification, and, in some difficult cases, a quick exploration of the *FrWac* web corpus (Baroni et al., 2009), to confirm whether a relational reading was found.

This annotation was done in a matter of hours by a single annotator (the first alphabetic author), and is hence probably quite noisy. However, it still allowed us to identify 367 examples of adjectives admitting relational readings. Interestingly, relational readings arise irrespective of the morphological relation between the noun and the adjective. Table 9 gives some examples of what was found, as well as counts and proportions of types admitting a relational reading for each class of derivational relation.

			Example	
Class	Count	Proportion	Noun	Adjective
Conversion	67	0.35	*politique* 'politics'	*politique* 'political'
A>N suffixation	2	0.40	*municipalité* 'municipality'	*municipal* 'municipal'
No direct relation	2	0.5	*fédération* 'fedaration'	*fédéral* 'federal'
N>A suffixation	296	0.71		
-ique	115	0.67	*volcan* 'volcano'	*volcanique* 'volcanic'
-al	55	0.76	*nation* 'nation'	*national* 'national'
-el	33	0.67	*culture* 'culture'	*culturel* 'cultural'
-aire	26	0.79	*banque* 'bank'	*bancaire* 'banking'
-Ven	22	0.96	*Brésil* 'Brazil'	*brésilien* 'Brasilian'
-eux	4	0.29	*os* 'bone'	*osseux* 'bony'
-ais	11	1.00	*Japon* 'Japan'	*japonais* 'Japanese'
-in	7	0.70	*femme* 'woman'	*féminin* 'feminine'
-ain	10	1.00	*république* 'republic'	*républicain* 'republican'
other	13	0.62	*hôpital* 'hospital'	*hospitalier* 'of hospital'

Tab. 9: Adjectives admitting a relational reading, sorted by derivational relation

2.2.4 Phonology

It is well known that affix rivalry is often conditioned by phonological properties of the base (see e.g. Plag 1999; Lignon 2013). We thus want to be able to include information on those phonological properties in our statistical study. For this we rely on the phonemic transcriptions from Wiktionary as accessible through the *GLÀFF* lexicon.[15]

Taking into account the phonology of the base raises an immediate challenge in the case at hand: as we discussed above, there is often more than one candidate for the status of base in a lexeme's morphological family. When that happens, the two potential bases typically differ both in length and in the makeup of the final syllable of their stem, which are the two aspects of base phonology that are most likely to have an impact. Table 10 shows relevant examples, that also illustrate the fact that the verbalizing suffix may attach to a stem that coincides with that of one of the two potential bases, or differ from both in some fashion, in the case of *salariser*.

	Potential bases			
Derived verb	Orthography	Stem phonology	Length	Final σ
centraliser	centre	sɑ̃tʁ	1	sɑ̃tʁ
'centralize'	central	sɑ̃tʁal	2	tʁal
franciser	France	fʁɑ̃s	1	fʁɑ̃s
'make French'	français	fʁɑ̃sɛ	2	sɛ
salariser	salaire	salɛʁ	2	lɛʁ
'make salaried'	salarial	salaʁjal	3	ʁjal

Tab. 10: Illustration of the uncertainty of base phonology

Given this situation, there seem to be three ways of taking into account phonological information. The first, more ambitious way would be to include in our models phonological information on all relevant members of the morphological family. This is pertinent inasmuch as both phonological material that is shared by the two potential bases (e.g. /sɑ̃tʁ/, /fʁɑ̃s/, /sal_ʁ/ in the examples above) and material that differentiates them (e.g. the suffix /-al/ or /-ɛ/ and the /ɛ∼a/ alternation above) may be relevant. While this presents no particular challenge in terms of annotation, it is unclear how to operationalize that information in terms of an uniform set of independent predictors, given that the size of the relevant subfamily and the morphological relation among its members varies from case to case. A second, less ambitious possibility is to rely on the material that is shared among potential bases—that is, in simple cases, the stem of the underived member of the pair. While this is clearly feasible, in many cases (see *centraliser* in Table 10) the verbalizing affix attaches to the stem of the derived member of the pair; hence the influence of the phonotactics of the adjacent part of the stem would not be taken into account.

The third solution, which we will adopt, is to only code the phonology of what we call the DERIVATIONAL STEM, the stem that the verbalizing affix happens to attach to. Table 11 indicates the derivational stems for the examples from Table 10, as well as the three properties of that stem we will use in our modelling: length in syllables, quality of the last vowel, makeup of the final consonant cluster if any is present.

Clearly, this solution is less than fully satisfactory: there is relevant phonological information that it does not take into account. However, we submit that it has the advantage of being easy

	Derivational stem			
Derived verb	Transcription	Length	V	Cluster
centraliser	sɑ̃tʁal	2	a	l
franciser	fʁɑ̃s	1	ɑ̃	s
salariser	salaʁ	2	a	ʁ

Tab. 11: Illustration of Derivational stems

to operationalize and conceptually clear. Coining a new verb involves two separate (but not necessarily independent) decisions: which potential base to rely on for choosing a stem, and which suffix to attach to that stem. By using exclusively the derivational stem as a predictor, we explicitly decide to focus on modelling the second of these two decisions when the first one has already been taken. In doing so, we are leaving aside two issues. First, it may be the case that the choice of the stem and the choice of the suffix are interdependent, something that could only be established by modelling both binary choices at the same time. Second, it may be that the phonology of the ultimate stem (e.g. /sɑ̃tʁ/ in the case of *centraliser*) also has an influence on the choice of verbalizing suffix. We leave the investigation of these two issues for future work. What should be stressed for the purposes of this paper is that identifying what we call the derivational stem does not commit us to the fact that a morphological family member sharing that stem plays the privileged role of a base in the derivation. In that sense the concept of a derivational stem is independent of the concept of a derivational base.

3 Descriptive statistics

In this section we use descriptive statistics to assess the relationship between different possible predictor variables and the choice of the *-iser* and *-ifier* suffix. Our selection of predictors is guided by previous studies of affix rivalry, and most prominently Lignon (2013), which discusses preference for *-iser* or *-ifier* in verbs derived from adjectives, on the basis of both dictionary data[16] and data collected semi-automatically on the web.[17] Remember that, unlike us, Lignon focused on cases where the verb can reasonably be assumed to be derived from an adjective: her data hence includes examples that we would have classified as having either just an adjective, or both a noun and an adjective, in their morphological family. Lignon made no distinction between these two situations.[18] Table 12 compares our dataset with Lignon's.

Our data table contains 791 lexemes, with 88.4% of *-iser* verbs. Thus our dataset is notably smaller than Lignon's, and we also observe a lower proportion of *-ifier* suffixation. These differences can be partly explained by the fact that our filtering method is more selective.[19] Finally, our data comprise 9 doublets, which are listed below. This represents around 2% of the data, a

	-iser	*-ifier*	Total	Doublets
Present dataset	699 (88.4%)	92 (11.6%)	791 (100%)	9
TLFi (Lignon, 2013)	789 (83.9%)	151 (16.1%)	940 (100%)	12
Lignon (2013)'s web data	1231 (78.6%)	336 (21.4%)	1567 (100%)	258

Tab. 12: Proportions of *-iser* and *-ifier* verbs in different datasets

proportion similar to Lignon's dictionary data.

(4) *turquifier/turquiser* 'turkify' ; *électrifier / électriser* 'electrify'; *estérifier / estériser* 'esterify' ; *étanchéifier/étanchéiser* 'make watertight' ; *éthérifier / éthériser* 'etherify' ; *fluidifier / fluidiser* 'liquefy'; *nanifier / naniser* 'dwarf' ; *terrorifier / terroriser* 'terrorize' ; *typifier / typiser* 'typify'.

3.1 The date of birth of the derived verb

Both *-iser* and *-ifier* are learned suffixes, which became productive in the late middle ages on the model of borrowed Latin verbs in *-izicare* (e.g. Lat. *baptizare* > Fr. *baptiser* 'baptize') and *-ificare* (e.g. Lat. *clarificare*> Fr. *clarifier* 'clarify').[20] Etymological dictionaries document French formations from the 14th century onwards (e.g. *humaniser* 'humanize', 1554; *exemplifier* 'exemplify', 1365), which coexist with continued learned borrowing. Hence the origin of the two suffixes in itself does not lead one to expect variation in their relative productivity. However, in a study of the relative productivity of *-ize* and *-ify* in English, Lindsay (2012) documents a growing number of *-iser* verbs between the 16th and the 20th century, together with a relatively stable number of newly attested *-ifier* forms. Since the history of English and French learned vocabulary is intimately tied, it is worth considering whether the same tendency is found in French—if so, this would contribute to explaining the variation in the use of the two suffixes.

In order to assess whether Lindsay's (2012) observation extend to French, we use an estimated "date of birth" for each constructed verb, calculated on the basis of its first attestations in *Google ngrams* (cf. section 2.2.1). The estimated dates range from 1580 to the 2000s. Examples of attested words for the temporal boundaries of this period are provided in (5).

(5) a. 1580 : *favoriser* 'favor', *fortifier* 'fortify', *justifier* 'justify', *réaliser* 'realize', *sacrifier* 'to sacrifice', *spécifier* 'specify'

b. 2000s : *stendhaliser* 'write like stendhal', *galliciser* 'to gallicize', *chimériser* 'turn into a chimara', *métropoliser* 'turn into a metropole'

Figure 1 displays the number of newly attested verbs for 5 periods of 50 years and a larger period ranging from 1580 to 1749.

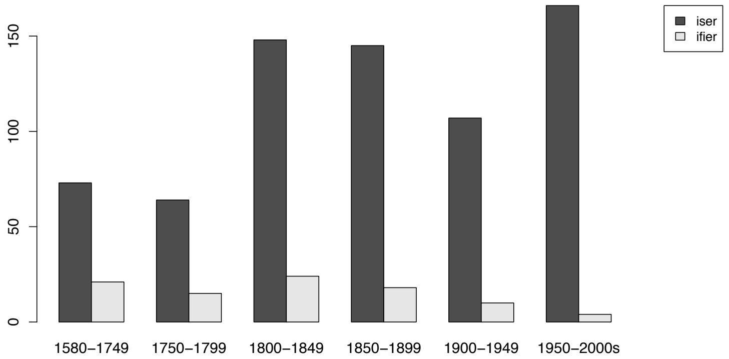

Fig. 1: Raw number of *-ifier* and *-iser* suffixations according to the estimated date of first attestation.

We observe that, as in English, *-iser* was the more productive of the two from the beginning. However, it is difficult to interpret the raw numbers displayed in Figure 1 because we do not know whether the number of *-iser* verbs estimated for the oldest periods are smaller either on account of the lack of data or on account of the lower productivity of *-iser* at that time. To compensate for this, we calculated the adjusted number of *-ifier* and *-iser* suffixation for each period by dividing the number of verbs by the estimated number of neologisms in the period.[21] Figure 2 indicates that the adjusted number of *-ifier* suffixation is relatively stable over time, while the adjusted number of newly attested *-iser* forms increases regularly until the middle of the 20th century and shows a major increase in the most recent period (1950-2000s). More strinkingly, Figure 2 reveals that the proportion of *-ifier* suffixation decreases over time: while between the end of the 16th century and the first half of the 18th century more than 22% of the newly attested forms occur with *-ifier*, the proportion falls to 2% for verbs coined since 1950. The correlation between the date and the number of *-ifier/-iser* suffixations is statistically significant: the regression coefficient combined with the estimated date of first attestation is significantly different from 0 (p-value $< .00001$), in a model evaluating the relationship between the attested suffix and the date of first attestation.[22] Thus the productivity of *-ifier* decreases over time in favor of *-iser*. This is in line with Lindsay's (2012) observations for English. The data presented in this section show that the rivalry between *-ifier* and *-iser* is linked to the productivity of each suffix in each period. Taking into account the age of the derived verb allows to partly explain the competition between the two suffixes: over time, the use of *-ifier* to coin new verbs becomes less and less likely.

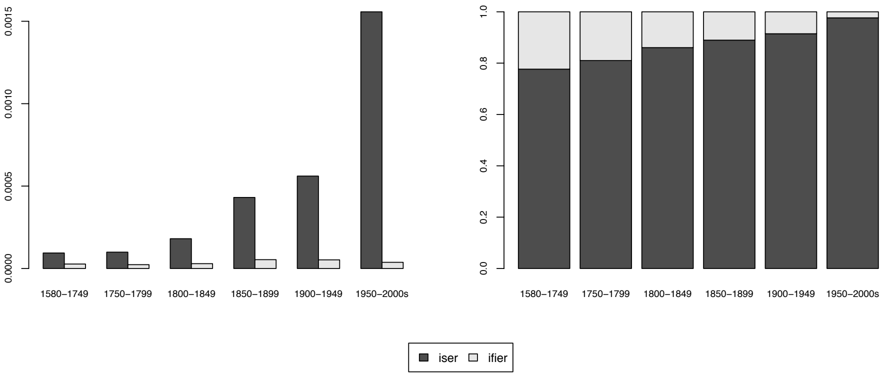

Fig. 2: Adjusted number (left) and proportion (right) of *-ifier* and *-iser* suffixations according to the estimated date of first attestation.

3.2 Phonological properties

Lindsay and Aronoff (2013) observe that *-ify* suffixation strongly correlates with the length of the base in English. In their data, almost 80% of the *-ify* forms are constructed from monosyllabic bases, while only 3% of *-ize* suffixations occur with non polysyllabic bases (cf. Table 13).

	monosyllabic	polysyllabic	Total
-ize	68 (3.1%)	2127 (96.9%)	2195 (100%)
-ify	322 (78.3%)	89 (21.6%)	411 (100%)

Tab. 13: Distribution of *-ize* and *-ify* suffixation by number of syllables in stem in Lindsay and Aronoff (2013).

A similar but weaker trend has been observed by Lignon (2013, 116) in French: "*More than half of the TLFi [dictionary] Xifier are derived from monosyllabic bases, while the percentage of monosyllabic Xiser is lower than 4%*". Although we measured the length of the derivational stem, rather than that of the stem of the adjectival member of a morphological family like Lignon (2013), the distribution of *-ifier* and *-iser* suffixations with monosyllabic stems in our data is similar, with 65% of *-ifier* and 5% of *-iser*. As Table 14 illustrates, there is an inverse correlation between the length of the stem and the prevalence of *-ifier*. The correlation is statistically significant: as for the date of attestation, the regression coefficient combined with the length measured in syllables is significantly different from 0 (p-value $< .00001$), in a model evaluating the relationship between the attested suffix and the length of the stem.[23] The preference of monosyllabic stems for *-ifier* over *-iser* is likely due to the difference in length between the two suffixes (Lignon, 2013, 116): although they are usually cited in combination with the infinitive ending, the derivational affixes in themselves are /iz/ (one syllable) and /i.fi/ (two syllables). When a vocalic suffix follows, semi-vocalization of /i.fi/ leads to the same syllable count, but in the common situation where no suffix (present and subjunctive singular and third plural) or a consonant-initial suffix (future and conditional) follows, inflected forms do indeed contrast in length.

	1 syll	2 syll	more than 2	Total
-iser	36 (5.1%)	350 (50.1%)	313 (44.8%)	699 (100%)
-ifier	60 (65.2%)	31 (33.7%)	1 (1.1%)	92 (100%)

Tab. 14: Distribution of *-iser* and *-ifier* suffixation by number of syllables in stem

French morphophonology is partly driven by dissimilative constraints that tend to avoid having phonemes sharing articulatory features at the end of the base and in the suffix. For instance, Plénat (2000) showed that dissimilative constraints push nominal bases whose last syllable contains phonemes similar to those making up the French derivational suffix *-esque* into being shortened when an *-esque* adjective is derived. More precisely, when the base is long enough, the addition of *-esque* triggers the truncation of the final rime of the base, if this rime holds identical or similar phonemes to the ones comprised in the suffix, i.e. mid front vowel (6-a), sibilant (6-b), or velar plosive consonant (6-c) (examples taken from Plénat 2000, 32-34).

(6) a. *Crédit Lyonnais* > *Crédit Lyonnesque* ‘Crédit Lyonnais like’
b. *show-bizness* > *show-biznesque* ‘show-business like’
c. *kremlinologue* > *kremelinolesque* ‘kremlinologist like’

In the case of rivalry between two suffixes, dissimilative constraints could influence the choice of the suffix in such a way as to avoid the proximity of similar phonemes at the boundary between the stem and the suffix. Concerning *-iser vs. -ifier* competition, Lignon (2013) notes that in her dictionary data, *-ifier* suffixation is over 8 times more frequent than *-iser* suffixation when the preceding consonant is a sibilant. Moreover, still in her data, a large proportion of *-ifier* suffixation occurs with bases ending with alveolar plosives. This means that the *-iser* suffix combines reluctantly with a base whose last syllable contains an alveolar obstruent. Hence, there is a dissimilative tendency that favors *-ifier* over *-iser*. Looking at the final consonant of stems in our data in Table 15, we observe, as expected, that the proportion of *-ifier* suffixations increases when the last consonant of the stem is an alveolar obstruent (/s/, /z/, /t/ or /d/).

Lignon (2013) also pointed out that stems ending with /l/, /ʁ/ or /n/ are more frequently combined with *-iser*. Our data display a similar trend: sonorant consonants are predominantly followed by the *-iser* suffix, as shown in Table 15. According to Lignon (2013), this is due to the high proportion of denominal adjectives used as bases to derive the *-iser* verbs. The French adjective-forming suffixes end predominantly with a sonorant consonant. This will be considered further in section 3.3. The effect of alveolar obstruent and sonorant consonants on the distribution of *-ifier* and *-iser* is statistically significant ($\chi^2 = 42.35$; df $= 2$; p-value $< .0001$).

	Alveolar Obstruent	Sonorant	Other
-iser	131 (76,2%)	518 (93,2%)	50 (79.4%)
-ifier	41 (23,8%)	38 (6,8%)	13 (20.6%)
Total	172 (100%)	556 (100%)	63 (100%)

Tab. 15: Distribution of *-iser* and *-ifier* suffixation by final consonant of the stem

Lignon (2013) did not take a look at the vowel of the last syllable of the base. However given the observations of Plénat (2000) concerning the *-esque* suffix, one could expect that the presence of two closed front vowels in *-ifier* has an impact on the choice of the suffix. A dissimilative constraint would favor *-iser* over *-ifier* if the last syllable contains a high vowel. Our data exhibits the opposite situation: there is a large proportion of *-ifier* suffixations that occurs with stems ending with /i/, /y/ or /u/, as Table 16 shows. The effect of high vowels over the distribution of the suffixes is statistically significant ($\chi^2 = 9.8$; df $= 1$; p-value < 0.002). Thus it seems that vowels follow a slight assimilative tendency in the case of *-iser vs. -ifier* rivalry.[24]

	High	Other
-iser	121 (80,7%)	578 (90,2%)
-ifier	29 (19.3%)	63 (9.8%)
Total	150 (100%)	641 (100%)

Tab. 16: Distribution of *-iser* and *-ifier* suffixation by final vowel of the stem

As we saw above, the nature of the phonemes in the last syllable of the stem seems to affect the choice between the two verbal suffixes. Our data also indicates that the presence of a consonant cluster at the boundary between the stem and the suffix correlates with a higher prevalence of *-ifier* suffixation. More precisely, we coded stems ending in a cluster, irrespective of whether this consists of a branching onset (7-a) or a coda-onset sequence (7-b).[25] Table 17 shows that the suffix *-ifier* is less disfavored when the stem ends in a cluster. The correlation between the selected suffix and the presence of a consonant cluster is statistically significant ($\chi^2 = 35.3$; df $= 1$; p-value < 0.0001).

(7) a. *am.**pl**ifier* ‘amplify’, *sa.**cr**ifier* ‘sacrifice’

b. *giscar**r.d**iser* ‘act like Giscard’, *mo**r.t**ifier* ‘mortify’

3.3 Morphological properties

As discussed in section 2.2.3, we classified derived verbs in three categories, depending on whether the morphological family contains a candidate nominal base, a candidate adjectival base,

	No Cons. Cluster	Consonant Cluster
-iser	656 (90.5%)	43 (65.2%)
-ifier	69 (9.5%)	23 (34.8%)
Total	725 (100%)	66 (100%)

Tab. 17: Distribution of *-iser* and *-ifier* suffixation by consonant cluster at the end of the stem

or both. This variable we call the ASCENDING MORPHOLOGICAL FAMILY (AMF), as it is concerned with possible ancestors of the derived verb in the morphological family, and ignores its possible descendants.

Table 18 shows that when the AMF contains only one possible base (noun or adjective), the distribution of *-iser* and *-ifier* is quite similar (light grey cells in the table), and there is no significant difference. However, if we compare the distribution of the verbal suffixes according to the distinction between AMF with one (light grey cells) or two (darker grey cells) candidate bases, we observe that a single possible base favors *-ifier*. The correlation is statistically significant ($\chi^2 = 7.52$; df $= 1$; p-value < 0.006). Our observations thus suggests skepticism towards Lignon (2013)'s statement that adjectival bases favor *-iser* over *-ifier*.[26] It seems more satisfactory to picture the *-iser vs. -ifier* rivalry in terms of the makeup of the AMF than in terms of a single base, and to assume that the configuration of the morphological family affects the choice of the suffix.

	N	Adj	Both
-iser	94 (82,5%)	47 (78,3%)	558 (90,4%)
-ifier	20 (17,5%)	13 (21,7%)	59 (9,6%)
Total	114 (100%)	60 (100%)	617(100%)

Tab. 18: Distribution of *-iser* and *-ifier* suffixation by type of AMF

In order to better understand the AMF configuration effect on the suffix choice, we explore the relationship between the two candidate bases for the 617 verbs under consideration. First, we focus on the formal relationship. Based on the annotation presented in section 2.2.3, we classified our data into the nine following categories:

1. Denominal suffix *-aire*
2. Denominal suffix *-al*
3. Denominal suffix *-el*
4. Denominal suffix *-en*
5. Denominal suffix *-ique*
6. Other : other denominal suffix
7. Conversion
8. A>N suffixation
9. No direct relation

Table 19 displays the distribution of *-iser* and *-ifier* according to these nine categories.

	-ique	*-aire*	*-al*	*-el*	*-en*	other	conver	A>N	no relation
-iser	166 96.5%	30 90.9%	71 98.6%	49 100%	21 91.3%	52 77.6%	162 84.4%	5 100%	2 50%
-ifier	6 3.5%	3 9.1%	1 1.3%	0 0%	2 8.7%	15 22.4%	30 15.6%	0 0%	2 50%
Total	172 100%	33 100%	72 100%	49 100%	23 100%	67 100%	192 100%	5 100%	4 100%

Tab. 19: Distribution of *-iser* and *-ifier* suffixation by derivational relation

It should first be noted that the number of verbs having 'A>N suffixation' or 'no direct relation' values is too small to show a significant trend.[27] Second, our data confirm Lignon's observation that "*one can suggest that* -iser *preferentially selects denominal adjectives*", with a twist. If the AMF contains a denominal adjective in *-aire, -al, -el, -en* or *-ique*, the proportion of *-iser* suffixations ranges from 91% to 100%, with a mean of 96.5% (dark grey cells in Table 19). Note that, at least on the surface, it is the presence of a denominal adjective that favors the *-iser* suffixation, not the fact that verb obviously derives from an advectival base. The case of morphological families containing an *-ique* denominal is particularly clear: in most cases, the derivational stem of the verb does not contain the *-ique* denominal suffix, but coincides with the stem of the noun, as shown in examples (8). In our data, there are only two exceptions to this observation, presented in (9): the derivational stem contains the /is/ allomorph of the /ik/ denominal suffix.

(8) a. *fanatique* 'fanatical' – *fanatiser* 'to make fanatical'
b. *folklorique* 'folkloric' – *folkloriser* 'to make folkloric'
c. *dogmatique* 'dogmatic' – *dogmatiser* 'to be dogmatic'

(9) a. *historique* 'historical' – *historiciser* 'to make historical'
b. *ethnique* 'ethnic'– *ethniciser* 'to make ethnic'

Lignon describes such situations as involving an adjectival base with truncation of *-ique*. We contend that in the present context, where we have established pervasive uncertainty as to the existence and identity of a base for derived verbs, it is hard to support such a view. The only surface-true generalisation not involving unsupported asumptions about base identity holds that the overwhelming proportion of *-iser* suffixation for verbs having *-ique* denominals in their AMF can be linked to the structure of the AMF, but not directly to the fact that their base is a suffixed denominal adjective.[28]

Finally, Table 19 shows that a conversion relation between the basis (light grey cells) slightly favors *-ifier* verbs.

We mentioned in the previous section that there is a correlation between stems ending with a sonorant and the proportion of *-iser* suffixation. Taking into account morphological information, we observe that among the 518 derivational stems displaying a sonorant in their last syllable and combining with the *-iser* suffix (cf. Table 15), 160 end with a denominal suffix[29], as in examples (10). This means that the combination of the morphological parameter (denominal suffix in the base) with the phonological parameter (sonorant in the last syllable of the stem) leads to a very strong preference for *-iser*.

(10) a. *brésilien* 'brazilian' – *brésilianiser* 'to make brazilian'
b. *nasal* 'nasal' – *nasaliser*
c. *spectaculaire* – *spectaculariser*

However, if we look at the distribution of *-iser* leaving aside the derivational stems ending with a sonorant and a denominal suffix, as well as the derivational stems ending with an alveolar obstruent and favoring *-ifier* suffix, the prevalence of *-iser* is still important after a final sonorant in the stem, as shown in Table 20 (light grey cells). The correlation between the presence of a final sonorant and the preference for *-iser* is still significant ($\chi^2 = 5.635$; df $= 1$; p-value < 0.02).

	Derivational stems ending with a sonorant and a denominal suffix	Derivational stems ending with a sonorant	Derivational stems not ending with a sonorant nor an alveolar obstruent	Derivational stems ending with an alveolar obstruent
-iser	160 (100%)	358 (90.4%)	50 (79.4%)	131 (76.2%)
-ifier	0 (0%)	38 (9.6 %)	13 (20.6%)	41 (23.8%)
Total	160 (100%)	396 (100%)	63 (100%)	172 (100%)

Tab. 20: Distribution of *-iser* and *-ifier* suffixation by the presence of a sonorant in the last syllable of the derivational stem

Thus, our data indicate that sonorants favor the choice of *-iser*, beyond the cases of derivational stems ending with denominal suffixes displaying a sonorant. This disproves Lignon's claim that the impact of final sonorants is a side-effect of the morphological structure of the base. Of course, one may still conjecture that this effect emerged diachronically thanks to the high prevalence of denominal suffixes ending in a sonorant, hence in effect leading to the reanalysis of a morphological preference into phonology. Still, synchronically, the phonological effect is present even in the absence of a morphological trigger.

Finally, concerning the verbs displaying both a noun and an adjective in their AMF, we observe that the presence of an adjective admitting at least one relational reading (cf. section 2.2.3) favours *-iser*. Table 21 shows that more than 94% of the relational adjectives give rise to *-iser* verbs, while there is only 84% of *-iser* verbs when the adjective does not qualify as a relational adjective (light grey cells). The correlation between the selected suffix and the presence of a relational adjective in the AMF is statistically significant ($\chi^2 = 18.91$; df $= 1$; p-value < 0.00001). This suggests that, besides the expected effect of formal complexity of the stem on the rivalry, the semantic aspect of the relationship plays a part.

	Rel	No Rel	na
-iser	348 (94.8 %)	210 (84%)	141 (81%)
-ifier	19 (5.2%)	40 (16%)	33 (19%)
Total	367 (100%)	250 (100%)	174 (100%)

Tab. 21: Distribution of *-iser* and *-ifier* suffixation by semantic relationship in the AMF

Now that we described the content of our data table, we present a statistical modelling of *-iser* and *-ifier* rivalry.

4 Multifactorial statistical analysis

In the previous section, we showed that competition between *-iser* and *-ifier* suffixation is driven by non-deterministic factors relating to phonology, morphology, semantics, and diachronic evolution of the system. We now attempt to understand how those factors work jointly and how each of them affects competition when the other factors are taken into account. To this end, we use logistic regression, a multifactorial statistical tool suited to examining the relationship between a dependent categorical variable and several predictor variables.[30] In our case, the dependent variable is a binary variable whose value corresponds to the suffix *-ifier* or *-iser*. The predictors are the factors described in section 3. Formally, logistic regression involves estimating the probability, noted of $P(Y = \text{IFIER} \mid X)$, that the *-ifier* suffix be chosen, given a set of predictor variables X. The model is defined as follows, where X refers to the set of predictors and β to the estimated parameters combined with each predictor variable.

$$P(Y = \text{IFIER} \mid X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

4.1 Methodology

The predictor variables used for the statistical modelling are the following:

- Scaled DATE of birth of the derived verb: continuous variable ranging from 0 to 1.45 (mean $= 0.95$).

- Scaled LENGTH of the stem: discrete variable ranging from 0.39 to 1.98 (mean $= 0.94$).
- Last CONSONANT of the stem: Alveolar obstruent or Sonorant or Other;
- Last VOWEL of the stem: High or Other.
- Consonant CLUSTER $=$ True or False.
- AMF: A, N or Both;
- Morphological class of the adjective (MCA) : *-ique* denominal, other denominal or conversion.
- RELATIONAL adjective : True or False.

We scaled the DATE and LENGTH variables so that both predictors have standard deviation 1. The DATE variable has also been reduced: we subtracted 1580 from each value of the variable, so that the date and the length have the same minimum, in addition to sharing their scale. This operation does not affect the result of the modelling, but facilitates the interpretation and the comparison of the coefficients. Given that the last two variables give information about the formal and semantic relationship between the noun and the adjective in the AMF, these variables concern only the portion of the data displaying a 'both' value for AMF. In addition, given that we need a reasonable number of observations for each level of a predictor value for statistical modelling, we grouped different morphological classes showing a similar distribution (cf. section 3). Hence the morphological variable, named MCA, has only three different levels, among which the *-ique* denominal level includes verbs whose AMF contains an *-ique* denominal adjective, the denominal level corresponds to derived verbs displaying a non-*ique* denominal adjective in their AMF, and the conversion level refers to the cases where the noun and the adjective stand in a conversion relationship.[31]

In the following sections, we will present two different models. The first one is built on the entire data table and aims to understand the role of age, phonological properties and general makeup of the morphological family (AMF). The second focuses on the verbs having both a noun and an adjective in their AMF and tries to better understand how the relationship between the two potential bases affects the choice of the suffix. This model will give us insight on the effect of both formal (MCA) and semantic (RELATIONAL) aspects of the relationship.

4.2 AMF model

The AMF model is a logistic regression model computed with a set of 6 predictors: DATE, CONSONANT, LENGTH, VOWEL, CLUSTER and AMF. The probability of constructing an *-ifier* verb is calculated as a function of these six variables. The predictive power of each variable is evaluated by comparing a model with a given predictor and another without this predictor. If the added predictor does not significantly improve the explanatory power of the model, the predictor is ruled out.[32] This operation led to shrinking the model, excluding the variables VOWEL and CLUSTER. This means that taking into account all the predictor variables, the effects of the consonant cluster and the high vowel observed in section 3.3 are negligible. The regression model is displayed in Table 22.

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.6290	0.8355	6.737	1.61e-11
LENGTH	-7.0933	0.7080	-10.018	< 2e-16
CONSONANT==AlvObs	0.2324	0.5009	0.464	0.64267
CONSONANT==Son	-0.8537	0.4926	-1.733	0.08308
DATE	-1.1360	0.4173	-2.722	0.00649
AMF==N	-0.7734	0.5400	-1.432	0.15211
AMF==both	-1.2199	0.4506	-2.707	0.00679

Tab. 22: AMF Model

The predictive power of logistic regression models is used to assess their accuracy. In particular, we use the value of the Area Under the ROC (Receiver Operating Characteristic) Curve, henceforth AUC, that estimates how well the model discriminates between true positives and true negatives: a value of $AUC = 0.5$ indicates that the discrimination accuracy is not better than chance ; a value of $AUC = 1$ indicates that the predictions are perfect. According to this measure, the AMF model makes very accurate predictions: $AUC = 0.918$.

One advantage of logistic models is that the coefficients combined with the predictors can be interpreted as follows: a positively estimated coefficient means that the value of the predictor increases the probability $P(Y = \text{IFIER}|X)$, i.e. favors the choice of *-ifier*, while a negative coefficient indicates that the predictor decreases $P(Y = \text{IFIER}|X)$, i.e. favors the choice of *-iser*. For instance, according to the AMF model, the presence of an alveolar obstruent at the end of the stem (positively estimated coefficient) comes out in favor of *-ifier* whereas sonorant consonants (negatively estimated coefficient) vote for *-iser*. In order to better understand the effect of each factor, Figure 3 displays the values of the predictors by log odds, an alternate way of expressing probabilities as values between $-\infty$ and $+\infty$.[33] First we note that the increase of the derivational stem length clearly reduces the odds of having the *-ifier* suffix. The date of birth of the word goes in the same direction, but the effect is weaker. The effects of categorical predictors, CONSONANT and AMF, are congruent with the observations based on descriptive statistics and their amplitude is in the same range as the one of DATE effect. The final consonant of the stem affects the choice of the verbal suffix: regarding the *Other* value as the benchmark, sonorants lower the probability $P(Y = \text{IFIER}|X)$, i.e. favor *-iser*, and alveolar obstruents drive the probability up. As for the AMF predictor, if we compare the effect of the presence of a single candidate base (either A or N) to cases where both A and N are candidate bases, Figure 3 shows that the existence of two candidate bases give advantage to *-iser* verbs. Moreover there is a slight numerical difference between nominal and adjectival candidate bases, although the estimation of that effect does not reach statistical significance.[34]

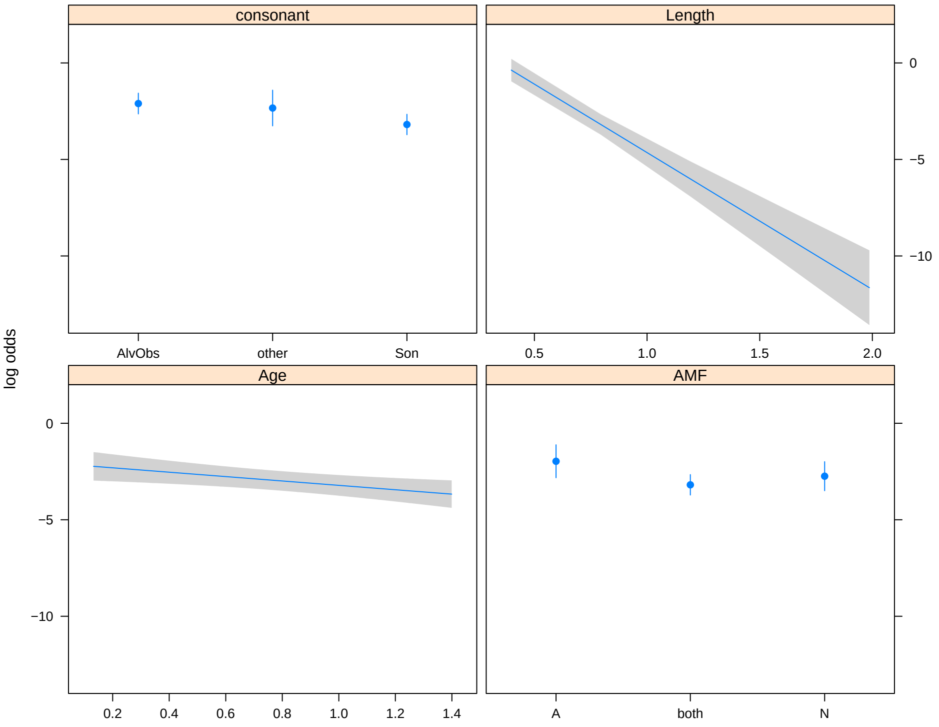

Fig. 3: Effect of the predictors in the AMF model.

4.3 Complex AMF model

The Complex AMF model is a logistic regression model applied only to the data points with complex AMFs, and relying on 5 predictors: DATE, CONSONANT,[35], LENGTH, MCA and RELATIONAL. No variable has been ruled out by the shrinkage method, which means that all variables contribute to explaining the choice between the verbal suffixes *-iser* and *-ifier*. Table 23 shows the coefficients of the model.

The AUC measure demonstrates that the predictions of the Complex AMF model are very ac-

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.5029	0.8535	6.447	1.14e-10
LENGTH	-7.8332	0.9512	-8.235	< 2e-16
MCA==ique	-1.2240	0.5540	-2.210	0.027134
MCA==denom	1.0299	0.4797	2.147	0.031810
RELATIONAL==True	-1.4755	0.4390	-3.361	0.000777
CONSONANT==son	-1.2054	0.4174	-2.888	0.003878
DATE	-0.8419	0.5905	-1.426	0.153923

Tab. 23: Complex AMF Model

curate, even more than those of the AMF model : $AUC = 0.943$. Taking into account information regarding the relationships in the morphological family of the verb appears to improve the quality of the modelling.

Figure 4 displays the effects of the predictors. The effects of the three variables shared by the two models (CONSONANT, DATE, LENGTH) go in the same direction : increased stem length and earlier date of first attestation both lower the probability $P(\text{IFIER})$; stems ending with a sonorant also favor *-iser* suffixation. The MCA predictor shows that, compared to conversion, the presence of an *-ique* denominal adjective in the AMF favors *-iser* suffixation, whereas other denominal adjectives drive up the probability of *-ifier* suffixation. Finally, relational adjectives are more likely to be combined with *-iser* than non-relational ones.

The p-value of the DATE predictor is higher than 0.05 (p-value $= 0.15$), which means that the estimated coefficient is not significantly different from 0. In other words, taking into account the makeup of the morphological family greatly weakens the effect of the DATE variable. We tested different nested models (without RELATIONAL, without MCA and without both RELATIONAL and MCA) and it appears clearly that removing the RELATIONAL variable from the Complex AMF model gives back its explanatory power to the DATE variable. This means that it is the presence of the RELATIONAL variable that seems to inhibit the effect of diachrony. This observation leads to the following comments. At first glance, one could suppose that the weakness of DATE is the result of a strong correlation between DATE and RELATIONAL variables. But the relatively low values of correlation coefficients (Spearman's $\rho = 0.17$; Rank Biserial Correlation $\rho = 0.27$) indicate that there is only a slight correlation, where verbs having a relational adjective in their AMF tend to be relatively more recent coinages.[36] Second, without being strongly correlated, these two predictors affect in the same manner the dependent variable, $P(Y = \text{IFIER}|X)$. Both variables lower this probability: the younger the constructed verb, the higher the chances of the *-iser* being used; while at the same time, having a relational adjective in the morphological family increases the chances of coining a new verb with *-iser*.

Given this, one could speculate that the large proportion of verbs recently coined with *-iser* be the result of an incidental correlation between the date of first attestation of the verb and the fact that its morphological family comprises a relational adjective. Under this view, the apparent effect of DATE would thus be spurious, and a side effect of the RELATIONAL variable. However, one can not rule out the possibility that the two variables affect the choice of the verbal suffix independently from one another, and that the lack of statistical significance for DATE in the present model is due to the size of the dataset.[37] As things stand, it is actually difficult to disentangle the respective explanatory power of those two variables.

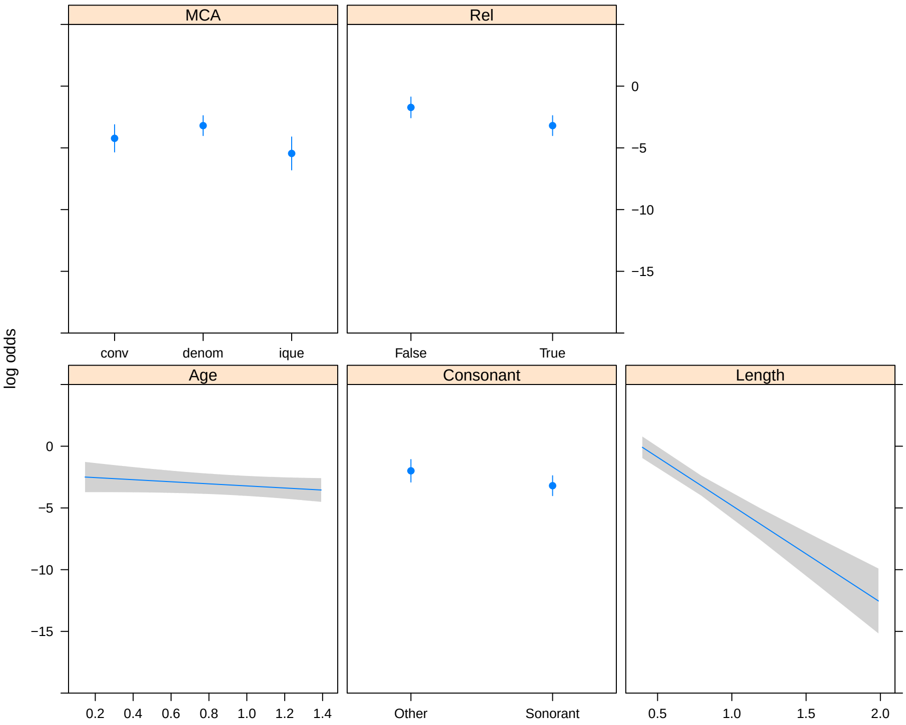

Fig. 4: Effect of the predictors in the Complex AMF model.

5 Discussion

As stated in the introduction, this paper has a dual goal: showcasing the usefulness of statistical analysis in the study of rivalry in lexeme formation, and providing a concrete analysis of one phenomenon of rivalry, between verbs in *-iser* and *-ifier* in French. Although our analysis has mainly confirmed and reinforced previously established generalisations by Lignon (2013) (and related observations by Lindsay (2012) on English), three main new insights should be highlighted.

We first note that the present study highlights the usefulness of statistical hypothesis testing when studying rivalry in lexeme formation. There is a general consensus that lexeme formation is subject to noncategorical constraints, and that studying rivalry amounts to uncovering and evaluating these constraints. In such a context, any generalisation takes a quantitative character, and falsifiability must rely on statistical tests. As a case in point, consider the discussion in sections 3.2 and 3.3 of phonological and morphological preferences. As we saw, stems ending in a sonorant have a stronger preference for *-iser* than other stems. Lignon (2013) claims that this preference is due to the high prevalence of denominal adjectives formed with a suffix ending in a sonorant (*-aire*, *-al*, *-el*, *en*) as bases for derived verbs, since the existence of a denominal adjective in the morphological family also favors *-iser*. While this reduction of an apparently phonological to a morphological constraint is appealing, it is an empirical claim that can and should be tested as such. As it happens, we showed it to be false: if we limit our attention to stems ending in a sonorant that is not the final consonant of a denominal suffix, these still have a significantly higher preference for *-iser*. Of course, the use of a statistical test is crucial here to arguing that the correlation is not spurious.

Second, descriptive statistics and hypothesis testing are useful but inherently limited in the study of a multifactorial phenomenon such as lexeme formation. Since constraints from various linguistic dimensions (phonology, morphology, lexical semantics, diachrony, possibly syntax) are thought to be at play in rivalry, and since there are undisputably correlations between features in these various dimensions, an adequate modelling strategy should be able to establish that each of the relevant features produces an effect even when the effects of other relevant features have been taken into account. In this paper we relied on the simplest and most well-established statistical modelling framework, namely multivariate logistic regression. While this is by no means the only possibility (basically, any classification method could be put to use in this context), it has the advantage of being well-understood and producing results that are relatively easy to interpret. Specifically, our final models allowed us to establish firmly the complementary role of phonological (length of the stem, final consonant) and morphological (makeup of the morphological family, nature of the relation between noun and adjective within the family) in explaining preference for *-iser* or *-ifier*. The models also establish that preferences varied over time, with *-iser* becoming more prevalent in newer formations, although it is not possible to disentangle whether this is an inherent effect of time or it is due to the distribution of other features also changing over time. Overall, an important conclusion is that rivalry in lexeme formation is inherently gradient and multidimensional: there is an irreducible multiplicity of reasons to prefer one suffix over the other, and these include at least phonological and morphological considerations.

The third and final lesson we want to draw is more theoretical than methodological. The most surprising result from this study is the role that is played by the overall makeup of the morphological family in shaping decisions as to how this family can be extended with a new verb. The crucial morphological predictor for the choice between *-iser* and *-ifier* is not whether the base is a noun or an adjective, but whether there are BOTH a noun and a related adjective in the morphological family. This suggests that in general, derivation should be seen as a process whose input is not one lexeme, the base, but a whole structured morphological family, with the presence or absence of various members jointly shaping the new member. This suggestion is congruent with psycholinguistic observations on the role of morphological families in the processing of individual lexemes (see among many others Schreuder and Baayen 1997; del Prado Martín et al. 2005), and lends support to so-called paradigmatic views of lexeme formation (see e.g. the classical discussions in van Marle 1984; Becker 1993; Bochner 1993 and renewed interest in studies such as Namer 2013; Lignon et al. 2014; Štekauer 2014; Strnadová 2014a). In particular, the effects observed here in lexeme formation are reminiscent of the phenomenon of joint predictiveness in inflectional paradigms, where knowledge of multiple forms of a lexeme helps predict the shape of an unknown form (Stump and Finkel, 2013; Bonami and Beniamine, 2016). Bonami and Strnadová (Submitted) explicitly explores that parallelism on the basis of different data (the relationship between verbs, action nouns and agentn nouns in French) and reaches conclusions parallel to those of the present paper.

6 Conclusion

The present study had a limited scope, mostly due to limitations in the availability of data and annotation. In conclusion, it is worth considering how these affect our results, and what steps could be taken to progress.

First, the size and diversity of the dataset is inherently limited, as we relied on the intersection of French *Wiktionary* and the *Google ngrams* unigram collection. Size in itself should not be an issue: it should be kept in mind that, with close to 1000 derived verbs, we are dealing with a sample that is arguably comparable in size to the overall exposition of a speaker to derived verbs over a lifetime; and our statistical study has shown this to be sufficient to establishing significant effects of almost all variables under consideration. On the other hand, limiting ourselves to lexemes attested in a dictionary and a collection of edited and published texts, as massive as these are, affects the sociolinguistic and stylistic diversity of the dataset, and is expected to henceforth limit the diversity of observed data, as Hathout et al. (2003) forcefully showed when studying properties of French *-able* adjectives attested on the web. Although there is no particular reason to believe that it would affect our overall conclusions, carrying out a similar study on non-edited usage would definitely be beneficial.

Second, we decided at the outset that we would focus on two of the six processes forming verbs from nouns and adjectives in French. By doing so, instead of addressing directly the question in (11-a), we address the slightly artificial question in (11-b).

(11) a. Given the rest of the morphological family, what should be the derived verb in that family?

b. Given the rest of the morphological family, if the verb is in either *-iser* or *-ifier*, which of the two should it be?

Ignoring prefixation by *a-*, *é-* and *en-* should not be much of an issue, as these have limited productivity. But ignoring conversion clearly forces us to leave out a highly prevalent option and hence may lead us to not assess correctly which factors are at stake. Unfortunately, this is a problem that is harder to address than one may think. As Tribout (2010) shows at length, for the vast majority of verb-noun conversion pairs in French, it is not possible to decide which of the two lexemes is the base and which is the derivative. As a consequence, we do not have a simple way of deciding which morphological families should be considered to contain a verb derived by conversion. Progress in the study of conversion is hence a necessary precondition to being able to address question (11-a) directly.

Third, one main limitation of the present study is the fact that we could only use information on morphological families that is coarse-grained, limited in scope, and in need of systematic validation. If, as we argued in section 5, constraints on lexeme formation arise from the whole morphological family rather than a single base, the availability of resources that document in detail the size, structure and properties of families is a precondition to descriptive and theoretical progress. Large scale semantic annotation would be particularly welcome in this respect, as it would allow one to assess whether different morphosemantic types of derived verbs have different preferences.

Finally, our goal in this study was limited to establishing firmly the existence of correlations between properties of a morphological family, the date of coining of a lexeme, and the choice of a particular affix. An obvious next step is to ask why the correlations we established hold. Although at this point we can only speculate, there are interesting relevant ideas that could be tested. As we saw, the proportion of new *-ifier* verbs has always been low, and decreases over time. This may be related to the fact that *-ifier* is more likely to attach to monosyllabic stems: at any point in the history of the language, the number of monosyllabic stems is small; good candidates for *-ifier* suffixation are hence likely to be fewer in number, and to decrease over time, as the stock of monosyllabic stems is not renewed quickly (note that new formations in French, even by clipping, tend to be bisyllabic). Establishing whether such an explanation is valid would necessitate a large scale quantitative study of the evolution of the lexicon that is clearly beyond the scope of this paper. Another question worth asking is why we find the tendencies we find rather than others. For instance, why do final sonorants, or morphological families containing a relational adjective, favor *-iser*, rather than the other way around? Firm conclusions are out of reach here, but one plausible line of explanation relies on Guzman Naranjo's (2017) proposals on the role of analogy in grammar. On the basis of evidence from a wealth of different morphological systems, Guzman Naranjo shows that lexemes that are similar in terms of stem phonology and/or semantics tend to exhibit the same morphological behavior, both in inflection and in derivation. If that is so, we do expect that whatever statistical distribution of properties was present in the initial stages of the system will, all other things being equal, tend to be constant over time as new lexemes are coined. Hence it is not surprising that *some* preferences are found: on the contrary, true random use of the two rival suffixes would be surprising. On the other hand, *which* preferences are found may follow from a chance statistical distribution in some early stage of the language. We leave it to specialists in the history of the lexicon to assess whether such a line of explanation can be confirmed.

Notes

[1] This work started as a pedagogical exercise prepared by the authors for a group of French morphologists interested in applying statistical modelling to the study of derivational morphology, and was first presented in Toulouse on December 5, 2016, in a workshop peripheral to the *Décembrettes 9* morphology conference. The results turned out to be interesting enough to warrant publication. We thank all participants in that event, and in particular Gilles Boyé, Georgette Dal, Bernard Fradin, Nabil Hathout, Stéphanie Lignon, Fiammetta Namer and Fabio Montermini, for their remarks and support. Stéphanie Lignon, Franck Sajous, Jana Strnadová and Delphine Tribout provided help at various stages in the form of data, comments, criticisms, and suggestions. Finally, we thank for their work two anonymous reviewers and our handling editor, Gregory Stump. All remaining errors are obviously the sole responsibility of the authors. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

[2] See Riehemann (1998); Bonami and Boyé (2006); Desmets and Villoing (2009); Tribout (2010); Bonami and Crysmann (2016) for a formal approach to lexeme formation that takes the many-to-many situation at face value. This approach however is strictly categorical, and hence complementary to that developed here.

[3] All three prefixations and conversion produce either first or second conjugation verbs (Tribout, 2012, 113–122). Both suffixations produce exclusively first conjugation verbs. We leave aside minor, unproductive processes such as *-oyer* suffixation.

[4] We only looked for infinitives in the Google Ngrams dataset. While that certainly somewhat diminishes the size of our dataset, the number of false positives would be very high if we included e.g. indicative present or participial forms,

which commonly have homographic nouns and/or adjectives.

[5] Only n-grams with more than 40 occurrences in the corpus are included. The version of the dataset used here was compiled in July 2012, and hence reflects the state of the Google Books collection at that time.

[6] By comparison, there are 1015 relevant verbs in the *Trésor de la Langue Française*. Note that both the GLÀFF and the Google Ngrams datasets contain many more verbs than their intersection (respectively 3212 and 2660), suggesting that larger datasets could be used under less stringent documentation requirements.

[7] Our dataset likewise contains only 332 verbs in *-iser* and 26 verbs in *-ifier* attested only since 1900 or later.

[8] See Koehl (2012) for a detailed discussion of brutal changes in the use of the suffix *-itude* in French.

[9] We take this to be preferable to looking at just the first attestation for two reasons. First, there are lots of OCR errors in the Google Books collection; while a single occurrence has a high risk of being a false positive due to poor OCR of a noisy book page, the risk is much lower for the first 10 occurrences to all be false positives. Second, a hapax in an arcane book runs the risk of not being noticed by many speakers, which is of course less true of a set of 10 occurrences.

[10] Note that in French the relational adjective corresponding to the noun *cardinal* is *cardinalice*. The adjective *cardinal* is not used with such a meaning.

[11] In the case of locative families such as those illustrated in Table 6, which all contain one adjective and two nouns, there is often no single answer to this question: conversion is always an option since demonyms play double duty as nouns and adjectives, but the relationship between the place name and the demonym is variable. We arbitrated in favor of N>A suffixation when the demonym derives from the place name, and in favor of conversion when the place name derives from the demonym. This is an arbitrary decision, which however will have no incidence on our results, since cases of conversion and A>N suffixation will be grouped together in the forthcoming analyses.

[12] `http://www.parl.gc.ca/HousePublications/Publication.aspx?Ses=1&Parl=41&DocId=5546931&Mode=1&Language=F`. Translation from the original parallel text.

[13] `http://www.correze.fr/fileadmin/user_upload/Correze_et_institution/Collectivite/Comptes_Rendus_Seances/Seance_18122015/CD_18122015_debats.pdf`.

[14] `http://www.davidautissier.com/telechargement/articles/AUTISSIER_THESE_TOME1-1997.pdf`.

[15] See Hathout et al. (2014) for a discussion of the reliability of these transcriptions. While we did encounter some errors in our manual examinations, those are not numerous enough to affect the results of a statistical study.

[16] From the *Trésor de la langue française*: `http://atilf.atilf.fr`.

[17] Candidates generated from adjectives in the BRULEX database (accessible through `www.lexique.org`) and attested on the web through the *Yahoo* search engine in November, 2010 using the WaLiM validation tool (Namer, 2003).

[18] Lignon (personal communication), December 2015.

[19] In particular, we excluded all prefixed verbs (cf. section 2), which ruled out 297 lexemes. With these prefixed verbs, the proportion of *-ifier* suffixation is 12%.

[20] The native descendant of *-izicare*, *-oyer*, has been supplanted by its learned rivals, and survives only in a few dozen verbs such as *nettoyer* ‘clean up’ (from *net* ‘clean’).

[21] We estimated the number of neologisms on the basis of Google Ngrams dataset, using the dating method presented in 2.2.1: A word counts as a neologism at its estimated date of first attestation.

[22] Logistic regression with the choice of the suffix as the dependent variable, and the date of attestation as continuous predictor variable. All calculations and modelling were done with R software (R Core Team, 2013).

[23] Logistic regression with the choice of the suffix as the dependent variable, and the length of the stem as continuous predictor variable.

[24] Another interesting observation arising in the data is the large proportion of stems having a nasal vowel in their last

syllable and followed by the *-ifier* suffix. Nevertheless the low amount of relevant data (only 27 derived verbs) calls for caution. Further research into a potential effect of nasality on the choice of the suffix needs to rely on a more extensive dataset.

[25] Treating the two types of consonant clusters as distinct categories led to no significant difference.

[26] Nyrop (1936) states that *-ifier* tends to attach to nouns and *-iser* tends to attach to adjectives for verbs coined in French, and his wording suggests that this may not hold for learned borrowings from Latin. Testing whether this is true in our dataset would necessitate arbitrating systematically which verbs were borrowed from Latin, a task we are in no position to complete.

[27] However, all 'A>N' verbs use the *-iser* suffix, suggesting that the relevant feature explaining the distribution of *-iser* and *-ifier* might be not only the presence of a derived adjective in the AMF, but also the presence of any derived potential base in AMF. In order to answer this question, we would need more data with desadjectival nouns in the AMF.

[28] As an anonymous reviewer points out, this observation is quite mysterious at this point. In section 5 we will suggest that it contributes to a wealth of evidence that coining of new lexemes is influenced by the whole preexisting morphological family rather than a single, distinguished base within that family.

[29] Note that the figures slightly differ between table 19 and table 20, because in the former table, the figures take into consideration the adjective candidate base, while the second table refers to the derivational stem. For instance, if we add up the number of candidate bases having a denominal suffix ending with a sonorant (dark grey cells in Table 19, minus -ique suffix), we end up with a figure (177) different from the one presented in Table 20 (160). This means that 17 coined verbs have a denominal adjective ending with a sonorant in their AMF, that is not in their derivational stem.

[30] Note that, unlike many studies that use logistic regression to model alternations (e.g Bresnan et al., 2007; Boleda et al., 2012; Faghiri and Samvelian, 2014; Thuilier, 2014), we do not model the probability of using an alternant in an utterance, but the probability of inclusion of a possible lexeme in the lexicon. For previous uses of logistic regression in the prediction of types of items in the lexicon, see e.g. Baayen and Moscoso del Prado Martín (2005) and Henri and Bonami (inpress).

[31] Note that the five verbs having a deadjectival noun in the AMF were included in the conversion level, because of the sparseness of the data. The 4 verbs having both a noun and an adjective in the AMF that are in no direct relation were ruled out, in order to work with homogeneous data.

[32] We used the `step` function (`stats` package) of R.

[33] The log odds of the probability $P(Y = \textsc{ifier}|X)$ are calculated using the logit function : $\text{logit}(x) = \log \frac{x}{1-x}$.

[34] In the statistical model presented in Table 22, there is not a coefficient corresponding to the A value of AMF, because this coefficient is incorporated into the intercept.

[35] The CONSONANT variable has been simplified by grouping the levels 'Alveolar Obstruent' and 'Other', because the 'Alveolar Obstruent' level is not significant in the Complex AMF model.

[36] More generally, we estimated the overall collinearity of the all Complex AMF model variables, using the condition number κ. The low value of κ for the Complex AMF model ($\kappa = 11.2$) indicates that collinearity is not much an issue in the model. The condition number has been calculated with the `collin.fnc` function (Baayen, 2008). According to Baayen (2008, 182), "*When the condition number is between 0 and 6, there is no collinearity to speak of. Medium collinearity is indicated by condition numbers around 15, and condition numbers of 30 or more indicate potentially harmful collinearity.*"

[37] Note that the p-value combined with the DATE predictor is higher than 0.05, but the variable is not ruled out by the `step` function of R software. This indicates that, while we find no evidence that DATE is a significant predictor on its own, a model that includes it is still significantly more performant than a model that does not.

References

Anshen, F. and Aronoff, M. (1981). ‘Morphological productivity and phonological transparency’. *Canadian Journal of Linguistics*, 26:63–72.

Arndt-Lappe, S. (2014). ‘Analogy in suffix rivalry: the case of English *-ity* and *-ness*’. *English Language and Linguistics*, 18:497–548.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge: MIT Press.

Baayen, H. and Moscoso del Prado Martín, F. (2005). ‘Semantic density and past-tense formation in three Germanic languages’. *Language*, 81:666–698.

Baayen, R. H. (2008). *Analyzing Linguistic Data : A A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Bally, C. (1944). *Linguistique générale et linguistique française*. Berne: Francke.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). ‘The wacky wide web: A collection of very large linguistically processed web-crawled corpora’. In *Language Resources and Evaluation*, vol. 43. 209–226.

Becker, T. (1993). ‘Back-formation, cross-formation, and ‘bracketing paradoxes’ in paradigmatic morphology’. In G. Booij and J. van Marle (eds.), *Yearbook of Morphology 1993*. Dordrecht: Kluwer, 1–25.

Bochner, H. (1993). *Simplicity in Generative Morphology*. Berlin: Mouton de Gruyter.

Boleda, G., Evert, S., Gehrke, B., and McNally, L. (2012). ‘Adjectives as saturators vs. modifiers: Statistical evidence’. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, and M. Westera (eds.), *Logic, Language and Meaning - 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised Selected Papers*. Dordrecht: Springer, 112–121.

Bonami, O. and Beniamine, S. (2016). ‘Joint predictiveness in inflectional paradigms’. *Word Structure*, 9:156–182.

Bonami, O. and Boyé, G. (2006). ‘Deriving inflectional irregularity’. In *Proceedings of the 13th International Conference on HPSG*. Stanford: CSLI Publications, 39–59.

Bonami, O. and Crysmann, B. (2016). ‘The role of morphology in constraint-based lexicalist grammars’. In A. Hippisley and G. T. Stump (eds.), *Cambridge Handbook of Morphology*. Cambridge: Cambridge University Press, 609–656.

Bonami, O. and Strnadová, J. (Submitted). 'Paradigm structure and predictability in derivational morphology'. *Morphology*.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). 'Predicting the dative alternation'. In G. Bouma, I. Kramer, and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Sciences, 69–94.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Tübingen: Max Niemeyer Verlag.

del Prado Martín, F. M., Deutsch, A., Frost, R., Schreuder, R., Jong, N. H. D., and Baayen, R. H. (2005). 'Changing places: A cross-language perspective on frequency and family size in dutch and hebrew'. *Journal of Memory and Language*, 53:496–512.

Desmets, M. and Villoing, F. (2009). 'French VN lexemes: morphological compounding in HPSG'. In *Proceedings of the HPSG 2009 Conference*. Stanford: CSLI Publications, 89–109.

Faghiri, P. and Samvelian, P. (2014). 'Constituent Ordering in Persian and the Weight Factor'. In C. Piñón (ed.), *Empirical Issues in Syntax and Semantics 10*. CNRS, 215–232.

Ferret, K., Soare, E., and Villoing, F. (2010). 'Rivalry between french *-age* and *-ée* : the role of grammatical aspect in nominalizations.' In M. Aloni, H. Bastiaanse, T. de Jager, and K. Schulz (eds.), *Logic, Language and Meaning*. Berlin, Heidelberg: Springer, 284–294.

Fradin, B. (2007). 'Three puzzles about denominal adjectives in *-eux*'. *Acta Linguistica Hungarica*, 54:3–32.

Giegerich, H. (1999). *Lexical strata in English: Morphological causes, phonological effects*. Cambridge: Cambridge University Press.

Guzman Naranjo, M. (2017). *Analogy in Formal Grammar*. Ph.D. thesis, Universität Leipzig.

Hathout, N., Plénat, M., and Tanguy, L. (2003). 'Enquête sur les dérivés en *-able*'. *Cahiers de grammaire*, 28:49–90.

Hathout, N. and Sajous, F. (2016). 'Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary'. In N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Hathout, N., Sajous, F., and Calderone, B. (2014). ‘GLÀFF, a large versatile French lexicon’. In *Proceedings of LREC 2014*.

Henri, F. and Bonami, O. (inpress). ‘Prédire l’agglutination de l’article en mauricien’. *Faits de langue*.

Koehl, A. (2012). ‘Altitude, négritude, bravitude ou la résurgence d’une suffixation’. In *Actes du troisième Congrès Mondial de Linguistique Française*. 1307–1323.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Lignon, S. (2013). ‘*-ISER* and *-IFIER* suffixation in French: Verifying data to ‘verize’ hypotheses’. In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse, Selected prodeedings of Décembrettes 7*. Munich: Lincom Europa, 119–132.

Lignon, S., Namer, F., and Villoing, F. (2014). ‘De l’agglutination à la triangulation ou comment expliquer certaines séries morphologiques’. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, and S. Prévost (eds.), *Actes du quatrième Congrès Mondial de Linguistique Française*. 1813–1835.

Lindsay, M. (2012). ‘Rival suffixes: synonymy, competition, and the emergence of productivity’. In A. Ralli, G. Booij, S. Scalise, and A. Karasimos (eds.), *On-Line Proceedings of the 8^{th} Mediterranean Morphology Meeting*. Rio, Greece: University of Patras, 193–203.

Lindsay, M. and Aronoff, M. (2013). ‘Natural selection in self-organizing morphological systems’. In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse, Selected prodeedings of Décembrettes 7*. Munich: Lincom Europa, 133–153.

McNally, L. and Boleda, G. (2004). ‘Relational adjectives as properties of kinds’. In O. Bonami and P. C. Hofherr (eds.), *Empirical issues in formal syntax and semantics*, vol. 5. 179–196.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., andSteven Pinker, J. O., Nowak, M. A., and Aiden, E. L. (2010). ‘Quantitative analysis of culture using millions of digitized books’. *Science*, 14:176–182.

Namer, F. (2003). ‘WaliM: valider les unités morphologiques par le Web’. In B. Fradin, G. Dal, N. Hathout, F. Kerleroux, M. Plénat, and M. Roché (eds.), *Silexicales 3: Les unités morphologiques*. Villeneuve d’Ascq: Université Lille 3, 142–150.

——— (2009). *Morphologie, Lexique et Traitement Automatique des Langues*. London: Hermès Science Publishing.

——— (2013). ‘Adjectival bases of French *-aliser* and *-ariser* verbs: syncretism or underspecification?’ In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse, Selected prodeedings of Décembrettes 7*. Munich: Lincom Europa, 185–210.

Nyrop, K. R. (1936). *Grammaire historique de la langue française, Tome troisième formation des mots*. Copenhague: Nordisk Forlag.

Plag, I. (1999). *Morphological productivity*. Berlin: Mouton de Gruyter.

Plénat, M. (2000). ‘Quelques thèmes de recherche actuels en morphophonologie française’. *Cahiers de lexicologie*, 77:27–62.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rainer, F. (2013). ‘Can relational adjectives really express any relation? an onomasiological perspective’. *SKASE Journal of Theoretical Linguistics*, 10:12–40.

Riehemann, S. (1998). ‘Type-based derivational morphology’. *Journal of Comparative Germanic Linguistics*, 2:49–77.

Roché, M. (2011). ‘Quel traitement unifié pour les dérivations en *-isme* et en *-iste*?’ In M. Roché, G. Boyé, N. Hathout, S. Lignon, and M. Plénat (eds.), *Des unités morphologiques au lexique*. Hermès Lavoisier.

Schreuder, R. and Baayen, R. H. (1997). ‘How complex simple words can be’. *Journal of Memory and Language*, 37:118–139.

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

——— (2014b). ‘Multiple derivation in French denominal adjectives’. In S. Augendre, G. Couasnon-Torlois, D. Lebon, C. Michard, G. Boyé, and F. Montermini (eds.), *Proceedings of the 8th Décembrettes*, no. 22 in Carnets de grammaire. Toulouse: CLLE-ERSS, 327–346.

Stump, G. T. and Finkel, R. (2013). *Morphological Typology: From Word to Paradigm*. Cambridge: Cambridge University Press.

Thuilier, J. (2014). ‘An experimental approach to French attributive adjective syntax’. In C. Piñon (ed.), *Empirical issues in formal syntax and semantics, vol. 10*. 287–304.

Tribout, D. (2010). ‘How many conversions from verb to noun are there in French?’ In *Proceedings of the HPSG 2010 conference*. Stanford: CSLI Publications, 341–357.

——— (2012). ‘Verbal stem space and verb to noun conversion in French’. *Word Structure*, 5:109–128.

Tribout, D. and Villoing, F. (2014). ‘La composition vn et la conversion v>n en français: un nouveau cas de concurrence morphologique?’ In F. Villoing, S. David, and S. Leroy (eds.), *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Nanterre: Presses Universitaires de Paris Ouest, 75–108.

van Marle, J. (1984). *On the Paradigmatic Dimension of Morphological Creativity*. Dordrecht: Foris.

Štekauer, P. (2014). ‘Derivational paradigms’. In R. Lieber and P. Štekauer (eds.), *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press, 354–369.