Affix rivalry in French demonym formation: The role of linguistic and non-linguistic parameters

Juliette Thuilier, Delphine Tribout, Marine Wauquier

Affixal rivalry in French demonym formation: The role of linguistic and non-linguistic parameters

Juliette Thuilier, Delphine Tribout & Marine Wauquier

Abstract

Affix rivalry is defined as the phenomenon of morphological competition where affixes and meaning are in a many-to-many relationship. Because of their poor semantic content, demonyms are perfect candidates for the investigation of selectional constraints in such a context. Indeed the morphological processes they originate from are characterized by their shared, straightfoward semantic relation, as they denote inhabitants linked to the toponym they derive from, which allows for the apparently simplified scrutinization of non-semantic properties. Investigations however prove to be not as straightforward.

The present study provides a quantitative and statistical investigation of the rivalry between French *-ois*, *-ais*, *-ien* and *-éen* suffixes. It notably relies on phonological and morphological features. Its contribution pertains to the use a statistical modeling to provide a quantitative description, and to the integration of extralinguistic features in the nature of geographical proximity in a quantitative approach. The study shows that while the model can't predict with a good accuracy the suffix of a given demonym based on these features, it still draws on the main tendencies underlying the French demonym affix rivalry.

Keywords: demonym, affix rivalry, statistical modeling, random forest, French, geographical features

1 Introduction

Affix rivalry is defined as the morphological competition between several affixes which associated word-formation patterns are equivalent on all levels except the phonological one (Gardani et al., 2019). This long-studied phenomenon has been approached by means of various explanatory factors aiming at identifying the condition of selection of affixes. Numerous studies build on phonological properties of the base word, such as the length of the base or the last phoneme (Arndt-Lappe, 2014; Bonami and Thuilier, 2019). Syntactic properties such as argument structure (Fábregas, 2010; Martin, 2010) and morphological properties such as the morphological type of the base (Missud and Villoing, 2020) are also investigated. Authors as well delve into factors such as telicity, speciality domains or fine-grained semantic types as far as semantics is concerned

(Dubois, 1962; Martin, 2010; Huyghe and Wauquier, 2021). Some authors even rely on diachronic distribution as a complementary insight on affix rivalry (Uth, 2010; Bonami and Thuilier, 2019).

Overall, the study of affix rivalry highlights the many-to-many relationship between form and meaning. This multiplicity of relations usually makes it difficult to compare morphological processes. Because of their weak semantic instruction, demonyms stand out as a particular case of rival morphological derivatives. They are nouns denoting inhabitants. When derived, their base denotes a location to which the inhabitants are linked. The base toponym can refer to varied geographical entities, ranging from neighborhood, through cities and countries, to larger areas such as continents. In that regard, all demonyms retain a similar semantic link with respect to their base, regardless of said bases. Demonyms are therefore in a simplified many-to-one relationship, where all affixes point toward one single meaning. This specificity of demonyms facilitates the comparison of morphological processes in that it dismisses the semantic factor.

The literature with respect to the study of affix rivalry in the context of demonyms formation is to the best of our knowledge pretty scarce. One of the most discussed language in that respect is Spanish, for which Mexican and South American demonyms have been addressed by various authors, among which García Sánchez (2005), Brizuela (2017) and Chesnokova et al. (2021). Few studies focus on other languages, such as Danner (2016) and Roberts (2017) for English, and Faust (2017) for Hebrew.

As far as French is concerned, demonym affix rivarly has notably been addressed by Eggert (2002), Plénat (2008b) as well as Roché and Plénat (2016). All three works focus on the four main demonym suffixes: *-ois*, *-ais*, *-ien* and *-éen*. These studies notably hightlight the impact of various dissimilative constraints. For instance, authors suggest that the suffix *-ois* tends to disfavor bases ending with a back vowel, while *-ais* actually disfavors bases ending with a front vowel. As for the suffix *-ien*, it is said to prefer fricative-ending bases, at the expense of nasal-ending bases.

The present study aims at shedding light on the constraints at stake for the formation of French demonyms through a large scale statistical modeling. This modeling relies on features related to phonological, morphological and geographical properties, both individually and in combination. This quantitative and statistical assessment confirms empirically the previous observations from the literature and provides new insights to be explored for future work. Overall, it shows that while there are no strict constraints on the selection of either *-ois*, *-ais*, *-ien* or *-éen*, strong tendencies still emerge.

The article is organized as followed. Section 2 is dedicated to the presentation of the data and the methodology of our study. Section 3 focuses on the description of morphological properties, while Section 4 deals with that of phonological properties. Geographical insights on affix rivalry

are investigated in Section 5. We assess in Section 6 the association of all features in random forests to give an overview of the overall tendencies. We briefly conclude in Section 7.

2 Data and methodology

The present section is dedicated to presenting the data selection and the overall methodology we used in order to investigate French demonym formation and evaluate the impact of linguistic and non-linguistic features of the base toponym on affix rivalry. The presentation of the features themselves is provided in Sections 3, 4 and 5.

Our study of French demonyms is based on the Prolex Database (Tran and Maurel, 2006). Prolex is a multilingual dictionary of proper names that also contains information about relational names and adjectives associated to proper nouns. We extracted from Prolex a list of 10,213 pairs of French toponyms with their demonyms, be they adjectives or nouns. We found no case where the derived adjective differs from the derived noun, so that we kept only one form as the demonym, regardless of its category. When the same toponym can have different forms as demonym, like *Bologne* → *bolonais* ([bolonɛ])*/bolognais* ([bolɔɲɛ]), we duplicated the line in order to always have one demonym associated to each toponym. However, we kept the information that both demonyms are linked to the same toponym thanks to a numeric identifier associated to each toponym.

Because our goal is to determine whether some properties of the toponym may have influenced the choice of the suffix, we annotated each pair with different kinds of properties: morphological, phonological and geographical. The phonological transcription of toponyms and demonyms has been retrieved from the French Wiktionary, through Glawi (Sajous and Hathout, 2015). In order to analyze whether phonological properties may impact the choice of the suffix, we only kept pairs for which we had at least the transcription of the toponym. That choice reduced the dataset to 2,218 pairs of toponym-demonym suffixed with *-ais*, *-éen*, *-ien* or *-ois*, of which few examples are given in (1)[1].

(1) *-ais* Antilles→antillais, Népal→népalais, Bagneux→bagnolais, New York→new-yorkais

-éen Guadeloupe→Guadeloupéen, Guinée→guinéen, Foix→fuxéen, Noisy-le-Sec→noiséen

-ien Nanterre→nanterrien, Sochaux→sochalien, Saint-Maurice→saint-mauricien

-ois Belleville→bellevillois, Meaux→meldois, Le Blanc-Mesnil→blanc-mesnilois

The distribution of the four suffixes is presented in Table 1. The figures show that *-ois* is the

most frequent suffix in our dataset with 38.5% of the data, and that *-éen* is very uncommon (only 7.5%). Note that *-en* [ɛ̃] was considered as an allomorph of *-éen* when the toponym ends in [e] or [ɛ] (*e.g. Vendée* [vɑ̃de] → *vendéen* [vɑ̃deɛ̃]) and an allomorph of *-ien* when the toponym ends in [i] (*e.g. Algérie* [alʒeʁi] → *algérien* [alʒeʁjɛ̃]).

Suffix	#	%
-ois	854	38.5
-ien	644	29.0
-ais	555	25.0
-éen	165	7.5
total	2218	100.0

Tab. 1: Distribution of the four suffixes under study

Our study relies on statistical modeling in order to better understand the distribution of the four suffixes among demonyms in the existing lexicon. We aim at modeling the demonym system from a synchronic point of view, without any claim on the formation of neologisms. However, the synchronic description of the system can give insights into the formation of new demonyms if the situation arises. As an example, the experiment described in Akin (2006) shows that speakers consciously rely on their knowledge of the existing lexicon when they are asked to form new demonyms.

In the next three sections, we will provide a descriptive assessment of linguistic and non-linguistic features, based on the raw figures and a correlation test for each feature to be described, in order to investigate whether the distribution is random or not. We use chi-square test for discrete variables, and Kruskal-Wallis test for continuous variables. Pairwise comparisons between the suffixes are conducted using Wilcoxon rank sum.

In Section 6, we provide a multifactorial modeling to see how the combination of all features contributes to predicting the choice of the suffix, and to assess the importance of each feature. We make use of random forests of conditional inference trees because it is a non-parametric method, which allows for unbalanced data with small number of observations for some features, as it will be explained at the beginning of Section 6.

3 Morphological properties

We coded three different morphological properties: i) whether the toponym is a polylexical unit; ii) whether the toponym is an opaque compound; and iii) whether the formation of the demonym implies a form variation.

3.1 Polylexical toponyms

Polylexical toponyms are those composed of more than one word, whether these words are separated with a space like *New York*, a hyphen like *Saint-Tropez* or both like *Le Blanc-Mesnil*. 611 of our toponyms, that is 27.5% of the dataset, are polylexical. We observe different cases in the formation of demonyms out of polyexical toponyms:

- 271 derive from the first element of the toponym (*e.g.* **Dives**-sur-Mer→**div**ais)
- 114 derive from the whole toponym (*e.g.* **Lot-et-Garonne**→**lot-et-garonn**ais)
- 96 derive from the last element of toponym (*e.g.* Saint-**Tropez**→**tropéz**ien)
- The remainder (130 demonyms) are divided into many different complexe cases (*e.g.* **Cinq-Mars**-La-Pile→**cinq-mars**ien, Saint-**Rémy**-des-**Monts**→**rémy-mont**ais)

The issue of what part(s) of a polylexical toponym is/are chosen to derive its demonym has not been addressed yet. In the remainder of the paper, particularly when it comes to the analysis of phonological properties of toponyms, we only take into account what is used as radical in the demonym. For example, in the case of *Saint-Rémy-des-Monts*, we only analyzed *Rémy-Monts* because it is the segment of the toponym that is used to form the demonym *rémy-montais*.

Whether the toponym is polylexical or not does not seem to have an influence on the choice of the suffix. As can be seen in Table 2, the distribution of the suffixes in each category (polylexical or non-polylexical toponym) is similar. This observation is statistically confirmed: there is no significant correlation between the choice of the suffix and the polylexicality of the toponym ($\chi^2(3, N = 2218) = 4.4$, $p = .22$).

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
Polylexical	140	22.9	51	8.4	192	31.4	228	37.3	611
Non polylexical	415	25.8	114	7.1	452	28.1	626	39.0	1607
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 2: Distribution of suffixes according to the polylexical status of the toponym

3.2 Opaque compounds

A few toponyms (145, *i.e.* 6.5%) are noted as opaque compounds when they show no hyphen nor space, but are formed with words such as *bourg* ‘market town’, *court* ‘court’, *fort* ‘fort’, *mont* ‘mount’, *terre* ‘land’ or *ville* ‘city’, as illustrated in (2).

(2) a. Cabourg, Fribourg, Luxembourg, Strasbourg

b. Élancourt, Guyancourt, Quièvrecourt, Rocquencourt

c. Beaufort, Rocquefort

d. Aiglemont, Boisemont, Chaumont, Ermont, Rumont

e. Angleterre, Nanterre

f. Belleville, Franconville, Sartrouville, Trouville

In very few cases, the toponym can still be perceived by speakers as a compound, like *Belleville*, that comes from the words *belle* 'beautiful' and *ville* 'city' and literally means 'beautiful city'. In some cases it is likely that the first part of the compound is not recognized anymore as a word, like *Angle* in *Angleterre*, that refers to an ancient tribe, the toponym literally meaning 'land of the Angles'. In most cases the first part does not seem to be a word, like *nan* in *Nanterre* or *sartrou* in *Sartrouville*. Most of the time only the endings *bourg*, *court*, *fort*, *mont*, *terre* and *ville* are recognized, that is why we coded these toponyms as opaque compounds.

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
bourg	1	7.7	0	0	1	7.7	11	84.6	13
court	1	5.0	0	0	0	0	19	95.0	20
fort	2	66.7	0	0	0	0	1	33.3	3
mont	7	28.0	0	0	0	0	18	72.0	25
terre	1	33.3	0	0	2	66.7	0	0	3
ville	17	21.0	0	0	0	0	64	79.0	81
Total	29	20.0	0	0	3	2.1	113	**77.9**	145

Tab. 3: Distribution of the suffixes with respect to the different types of opaque compound bases

As the figures in Table 3 show, when the base is an opaque compound, the suffix *-ois* is clearly favored (77.9%). This preference is in line with the phonological properties of the word ending the compound base. Indeed, the final consonant (be it latent or not) is either a plosive [t], an approximant [l, ʁ] or a non alveolar fricative [ʒ], and each favors *-ois*, as we will see in Section 4 (cf. Table 8). From the statistical point of view, there is a significant correlation between the choice of the suffix and the fact that the base is an opaque compound or not ($\chi^2(3, N = 2218) = 114.6$, $p < .00001$).

3.3 Formal variation

The last morphological property we annotated is the formal variation of the toponym in the derivation of the demonym. We looked at coarse level whether the presence or absence of any kind of a formal variation of the base toponym can be linked to one suffix. We settled for this approach because we faced numerous problems in trying to take into account more fine-grained annotations, due to the extreme variability of toponym bases and to the special status of demonym derivation. In the end, the presence or absence of formal variation is the only objective and reliable criteria that we have found. We detail this issue below.

All kinds of variation grouped together, 1193 demonyms involve a variation of the base toponym, that is, more than half of the data (53.8%). As can be seen in Table 4, the variation of the toponym seems to be more tightly linked to suffixes *-ien* and *-ais*. In the whole dataset 29% of the demonyms are suffixed with *-ien*, while they are 32.4% when variation is involved. Similarly, *-ais* demonyms are 25% of the whole dataset, but 29.4% when there is a variation. Conversely, the *-ois* suffix seems to be more favored by the absence of variation: demonyms suffixed with *-ois* correspond to 38.5% of the whole dataset, but 46.3% of the data without variation. This can be related to the observation that *-ois* is favored when the toponym ends with a consonant, while *-ais* is favored when the toponym ends with a nasal vowel and *-ien* when it ends with a front oral vowel, as will be seen in Section 4.2. Final vowels are indeed prone to variation before suffixes beginning with a vowel, while final consonants present no special difficulty for such suffixes. These tendencies are statistically significant : there is a statistical correlation between the suffix and the presence or absence of formal variation ($\chi^2(3, N = 2218) = 63.54$, $p < .00001$).

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
Variation	351	**29.4**	77	6.4	386	**32.4**	379	31.8	1193
No variation	204	19.9	88	8.6	258	25.2	475	**46.3**	1025
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 4: Distribution of the suffixes with respect to variation of the base

We are aware that a mere distinction between presence and absence of any kind of variation is problematical because the presence of variation covers many different cases, from regular and predictable alternations like final latent consonants, to suppletion. For instance, the formal variation of the base toponym can be the mere realization of the final latent consonant (3a) or the denazalization of a final nasal vowel (3b). But it can also be the insertion of an interfix (3c), an epenthesis (3d) or, in the opposite, the deletion of the last vowel (3e) or the truncation of the final segment (3f). It can be a regular consonantic (3g) or vocalic (3h) alternation pattern, but

also be the use of a suppletive base (3i).

(3) a. Lorient [lɔʁjɑ̃] → lorientais [lɔʁjɑ̃t-ɛ], Paris [paʁi] → parisien [paʁiz-jɛ̃]

b. Berlin [bɛʁlɛ̃] → berlinois [bɛʁlin-wa], Japon [ʒapɔ̃] → japonais [ʒapon-ɛ]

c. Carthage [kaʁtaʒ] → carthaginois [kaʁtaʒ-in-wa], Quercy [kɛʁsi] → quercinois [kɛʁsi-n-wa]

d. Bray-Sur-Seine [bʁɛ] → braytois [bʁɛt-wa], Jura [ʒyʁa] → jurassien [ʒyʁas-jɛ̃]

e. Angola [ɑ̃gola] → angolais [ɑ̃gol-ɛ], Palaiseau [palɛzo] → palaisien [palɛz-jɛ̃]

f. Angleterre [ɑ̃glətɛʁ] → anglais [ɑ̃gl-ɛ], Gennevilliers [ʒœnvilje] → gennevillois [ʒœnvil-wa]

g. Cognac [kɔɲak] → cognaçais [kɔɲas-ɛ], Salonique [salonik] → salonicien [salonis-jɛ̃]

h. Fontaine [fɔ̃tɛn] → fontanien [fɔ̃tan-jɛ̃], Martel [maʁtɛl] → martelais [maʁtəl-ɛ]

i. Ahun [aœ̃] → acitodunois [asitodyn-wa], Houilles [uj] → ovillois [ovil-wa]

The manual annotation of these 10 categories of formal variation has been blurred by numerous cases of combinations, such as vocalic alternation + epenthesis (4a), vocalic alternation + consonant alternation (4b), vocalic alternation + latent consonant (4c), latent consonant + interfix (4d), etc. Overall we ended up with 49 different cases of variations, among which 36 apply to less than 10 items.

(4) a. Bagneux [baɲø] → bagnolais [baɲol-ɛ], Sochaux [soʃo] → sochalien [soʃal-jɛ̃]

b. Champagne [ʃɑ̃paɲ] → champenois [ʃɑ̃pən-wa], Cardroc [kaʁdʁɔk] → cardreucien [kaʁdʁøs-jɛ̃]

c. Arthies [aʁti] → arthésien [aʁtez-jɛ̃], Bourg-en-Bresse [buʁ] → burgien [byʁʒ-jɛ̃]

d. Gars [gaʁ] → garcinois [gaʁsin-wa], Saint-Amand [amɑ̃] → amandinois [amɑ̃din-wa]

Given the explosion in the number of the toponym variations, we considered reducing formal variation to the distinction between regular allomorphy and suppletion, but we abandonned this approach for two reasons. First, as Boyé (2006) showed, allomorphy and suppletion form a continuum of form variation and, if both ends of the continuum are clear, such as cases in (3a) and in (3i), intermediate cases are not always easy to classify. For instance, the [k]-[s] alternation can be considered as regular in the case of *Salonique→salonicien* because the [ik]-[is] alternation is very common in the lexicon (see for instance *logique* [lɔʒik] 'logic'→*logicien* [lɔʒisjɛ̃] 'logician', *mathématiques* [matematik] 'mathematics'→*mathématicien* [matematisjɛ̃] 'mathematician', etc.). But we can wonder if the same [k]-[s] alternation is also regular in the case of *Cognac→cognaçais* when

compared to examples in (5): with the same final segment [k] *Aurillac* does not give rise to any variation in its demonym, while *Balzac* and *Blagnac* show other variations (respectively [k]-[t] and [k]-[d]).

(5) a. Aurillac [oʁijak] → aurillacois [oʁijak-wa]

b. Balzac [balzak] → balzatois [balzat-wa]

c. Blagnac [blaɲak] → blagnadais [blaɲad-ɛ]

Second, the two annotators sometimes relied on different criteria for the distinction between allomorphic and suppletive bases. For instance the alternation between *château* and *castel* has been considered as allomorphic by annotator 1 in (6a) because this alternation appears several times in the lexicon, whereas annotator 2 considered it as suppletive in (6b) because the two forms are formally too distant (only 2 segments remain identical: [a] and [t]). This is a clear illustration of the criticism made by Boyé (2006) towards the various criteria proposed in the litterature in order to distinguish between allomorphy and suppletion: different criteria may apply (here the frequency of the variation and the formal distance between the two forms), leading to different results.

(6) a. Château-Chalon → castelchalonais

b. Châteaubriant → castelbriantais

Finally, following Roché and Plénat (2016), we also tried to account for latent consonants and final nasal vowels, which seem to be entirely predictable from orthography, thanks to the notion of *thème B* (B stem). The B stem is defined by the authors as a derivational stem of nouns that is only used in derivation (see Bonami and Boyé, 2003 and Roché, 2010 for details on the description of stem space). However, the identification of the B stem raises other problems. Within inflection, stems are identified by their use in the formation of inflected forms. For French, Bonami and Boyé (2003) have shown that we need to postulate 12 stems in the stem space of verbs in order to account for the whole conjugation. In derivation, the formation of one lexeme's derivatives mostly rely on the various stems used to inflect that lexeme. For example, according to Bonami et al. (2009), deverbal derivation applies to the verbs stems 1 and 3. The authors have also shown that there is one special verb stem that is not used for inflection but only for derivation. This special derivational stem has been identified because all derivatives of one given verb rely on this stem, which differs from all other inflectional stems. Bonami, Boyé and Kerleroux called this special stem the *hidden stem* (because it is hidden to inflection). Building on that proposal,

Plénat (2008a) postulates that nouns also have different derivational stems. The stem under study has been named the *B stem* by Roché (2010). Like nouns, toponyms do not inflect, so that we cannot identify stems with the help of inflection. But unlike nouns, which can have many derivatives, toponyms usually give rise to only one derivative, which is the demonym. Therefore, the identification of a derivational stem for toponyms only relies on one lexeme. This is problematical, because it does not allow to make a distinction between cases of form variation that are due to the identity of the toponym (for example its etymology) and cases where the variation can be caused by the suffix or by phonological rules that will add or delete a phoneme in order to avoid hiatus between the base and the suffix.

In the cases of final latent consonant and final nasal vowel, the orthography of the toponym usually helps predict the form of the demonym and therefore its B stem: as shown in (7), the orthography displays a final segment (*t, s, p, c*) that is phonetically realized in the demonym. However, there are cases like examples in (8) where the form of the demonym is not what could be expected if we rely on orthography. In these cases, like in (7), the toponym ends with a nasal vowel but the orthography shows an unpronounced final consonant. We could expect this final written consonant to be realized in the demonym, as in examples (7), but it is not what happens. These examples show that relying on orthography in order to identify the B stem is not straightforward, even in the case of final nasal vowel and final latent consonant. In this respect the comparison between *Le Blanc* (7d) and *Montblanc* (8c) is particularly striking.

(7) a. Belmont [bɛlmɔ̃] → belmontais [bɛlmɔ̃t-ɛ]

b. Nyons [njɔ̃] → nyonsais [njɔ̃s-ɛ]

c. Beauchamp [boʃɑ̃] → beauchampois [boʃɑ̃p-wa]

d. Le Blanc [blɑ̃] → blancois [blɑ̃k-wa]

(8) a. Montpont-en-Bresse [mɔ̃pɔ̃] → montponnais [mɔ̃pon-ɛ] (expected [mɔ̃pɔ̃t-ɛ])

b. Louhans [luɑ̃] → louhannais [luan-ɛ] (expected [luɑ̃s-ɛ])

c. Montblanc [mɔ̃blɑ̃] → montblanais [mɔ̃blan-ɛ] (expected [mɔ̃blɑ̃k-ɛ] or [mɔ̃blɑ̃ʃ-ɛ])

d. Provins [pʁovɛ̃] → provinois [pʁovin-wa] (expected [pʁovɛ̃s-wa])

To sum up, all our attempts to have a better account of the variety of form alternations led us to unreliable annotations. This is why we only kept the notion of presence/absence of any kind of form variation.

4 Phonological properties

Different phonological properties were annotated, among which: the number of syllables in the toponym, the last segment of the toponym, a backness score of the vowels in the toponym. It is reminded that in the case of polylexical toponyms only the part of the toponym that is used to form the demonym has been taken into account.

4.1 Number of syllables

The distribution of the four suffixes with respect to the length of the toponym is presented in Table 5. As can be seen in the table, half of the toponyms are dissyllabic. The distribution of the suffixes for dissyllabic toponyms is similar to their general distribution in the whole dataset (compare with the Total line in the table). In the other cases we found no clear effect. Nevertheless, two tendancies can be identified: monosyllabic toponyms like those presented in (9) favor *-ois*, while long toponyms, *i.e.* toponyms with four syllables or more, like examples in (10), favor *-ien*. Indeed, demonyms suffixed with *-ois* represents 38.5% of the whole dataset, while they are 61.6% in the set of monosyllabic base toponyms. As for demonyms suffixed with *-ien*, they are 29% in the whole dataset but 38.1% in the set of toponyms of four syllables or more. Considering that length has four levels, as presented in Table 5, there is a significant correlation between the suffix and the length of the base based on chi-square test ($\chi^2(9, N = 2218) = 112.49$, $p < .00001$).

(9) Lille [lil] → lillois, Cannes [kan] → cannois

(10) Mésopotamie [me.zo.po.ta.mi] → mésopotamien, Saint-Léonard [sɛ̃.le.o.naʁ] → saint-leonardien

	-ais		*-éen*		*-ien*		*-ois*		
Syllables	#	%	#	%	#	%	#	%	Total
1	63	16.4	14	3.6	71	18.4	237	**61.6**	385
2	312	26.8	104	8.9	349	30.0	399	34.3	1164
3	148	26.9	37	6.7	179	32.5	187	33.9	551
4+	32	27.1	10	8.5	45	**38.1**	31	26.3	118
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 5: Distribution of the suffixes with respect to the length of the toponym

4.2 Last segment of the toponym

Different aspects of the last segment of the toponym have been analyzed. At a broad level we only distinguished between final vowels and consonants. At a fine-grained level we looked at the phonological properties of the last segment. Just like the number of syllables, we found no clear

effect of the last segment on the choice of the suffix, but we observed several tendencies that are described below.

The distribution of the suffixes with respect to final vowels or consonants is presented in Table 6. As shown in the table, final consonants are more frequently associated to *-ois* (52.8% of the data, compared to 38.5% in the whole dataset). Few examples are given in (11).

Last	*-ais*		*-éen*		*-ien*		*-ois*		
segment	#	%	#	%	#	%	#	%	Total
C	259	23.5	14	1.3	247	22.4	581	**52.8**	1101
V	296	26.5	151	**13.5**	397	**35.5**	273	24.5	1117
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 6: Distribution of the suffixes with respect to final vowels and consonants

(11) Aurillac [oʁijak] → aurillacois, Cannes [kan] → cannois, Lille [lil] → lillois

Conversely, final vowels seem to favor *-ien* and *-éen*. Indeed, 29% of the whole demonyms are suffixed with *-ien*, but they are 35.5% when the toponym ends with a vowel. Similarly, only 7.5% of the whole dataset are suffixed with *-éen*, but they are 13.5% when the toponym ends with a vowel. This tendency is statistically significant: the choice of the suffix is significantly correlated to the type of final segment of the toponym ($\chi^2(3, N = 2218) = 262.14$, $p < .00001$).

This high score of *-ien* and *-éen* when the toponym ends with a vowel seems to be particularly linked to front vowels ([i,e,y]), as can be seen in Table 7. Indeed, the proportion of *-éen* demonyms when the toponym ends with a front vowel is more than twice its proportion in the whole dataset (7.5%). As for *-ien* demonyms, their number increases by 50% when the toponym ends with a front vowel. Few examples are given in (12). The correlation between the choice of the suffix and the fact that the final segment is a front vowel or not is statistically significant ($\chi^2(3, N = 2218) = 453.68$, $p < .00001$).

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
Front V	69	9.0	136	**17.7**	358	**46.5**	206	26.8	769
Other	486	33.5	29	2.0	286	19.8	648	44.7	1449
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 7: Distribution of the suffixes with respect to final front vowel/other segment

(12) a. Ivry [ivʁi] → ivrien, Vertus [vɛʁty] → vertusien

b. Nancy [nɑ̃si] → nancéen, Vendée [vɑ̃de] → vendéen

As for final consonants, they reveal other tendencies that are presented in Table 8. As can be seen in the table, plosive (13a) and approximant (13b) consonants favor *-ois*. As already observed

by Roché and Plénat (2016), alveolar fricatives (14) favor *-ien*, in order to avoid the succession of two alveolar fricatives in the feminine forms of the demonyms. Other fricatives (15) favor *-ois*. There is a statistically significant correlation between the two variables ($\chi^2(12, N = 2218) = 265.67$, $p < .00001$).

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
plosive	56	31.8	7	4.0	30	17.0	83	**47.2**	176
approximant	108	21.3	5	1.0	104	20.5	290	**57.2**	507
alveolar fricative	16	12.0	1	0.8	66	**49.6**	50	37.6	133
other fricative	6	6.3	1	1.1	22	23.2	66	**69.5**	95
other segment	369	28.2	151	11.6	422	32.3	365	27.9	1307
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 8: Distribution of the suffixes according to final consonants

(13) a. Dunkerque [dœ̃kɛʁk] → dunkerquois, Étampes [etɑ̃p] → étampois, Sète [sɛt] → sétois

b. Lille [lil]→ lillois, Quimper [kɛ̃pɛʁ] → quimpérois

(14) Alsace [alzas] → alsacien, Mulhouse [myluz]→ mulhousien

(15) Loches [lɔʃ] → lochois, Orange [ɔʁɑ̃ʒ] → orangeois

We also observed dissimilative constraints between the last segment of the toponym and the suffix used to form the demonym. A toponym ending with a final nasal segment, be it a vowel or a consonant, is very unlikely to combine with a suffix containing a nasal vowel like *-ien* and *-éen*, as can be seen in Table 9 . The figures in the table show that while the proportion of *-ois* demonyms when the toponym ends with a nasal segment is the same as in the whole dataset, that of *-ien* and *-éen* demonyms is much lower than in the general dataset (respectively 29% and 7.5%). Conversely, nasal segments, and particularly nasal vowels, largely favor *-ais* (16). Moreover, among nasal vowels, the back nasal vowels [ɑ̃] (16a) and [ɔ̃] (16b) are particularly frequent with the suffix *-ais*. They cover 206 cases over the 218 nasal vowels observed with *-ais*. From the statistical point of view, there is a significant correlation between the choice of the suffix and the nasality of the last segment ($\chi^2(3, N = 2218) = 394.18$, $p < .00001$).

(16) a. Dourdan [durdɑ̃] → dourdannais, Orléans [ɔʁleɑ̃] → orléanais

b. Avignon [aviɲɔ̃] → avignonnais, Meudon [mødɔ̃] → meudonnais

c. Gaume [gom] → gaumais, Rennes [ʁɛn] → rennais, Valogne [valɔɲ] → valognais

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
nasal	291	53.5	3	**0.6**	41	**7.5**	209	38.4	544
vow	218	**61.6**	3	0.9	16	4.5	117	33.0	354
cons	73	38.4	0	0	25	13.2	92	48.4	190
other	264	15.8	162	9.7	603	36.0	645	38.5	1674
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 9: Distribution of the suffixes with respect to nasal segments

4.3 Vowel backness

As seen above about the last segment of the toponym, front vowel can have an impact on the choice of the suffix (cf. Table 7). In order to determine the possible role of all vowels of a toponym in the choice of the suffix, we also calculated a backness score based on all vowels of the toponym. The methodology to calculate the score is adapted from Lohmann (2017). Each front vowel were coded as 1, while back vowels were coded as 3. Note that we considered only front and back vowels, schwa being treated as a back vowel. The backness score was calculated by adding the score associated to all vowels of the toponym. The result was then divided by the total number of vowels in the toponym. *Bougival* [buʒival], for example, has three vowels: two front vowels and a back one. The calculation of the score is then $(3 + 1 + 1)/3 = 1.7$. The distribution of the score is presented in Figure 1.

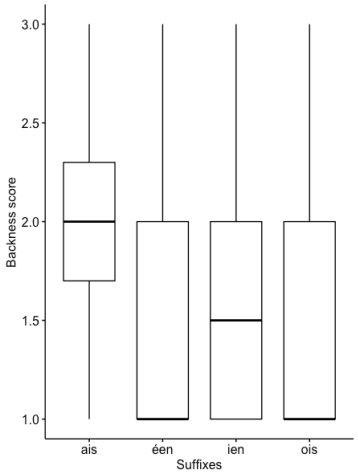

Fig. 1: Distribution of the suffixes as a function of backness score

The boxplot shows that demonyms coined with *-ais* tend to have a base with a higher backness vowel mean. Kruskal-Wallis test confirms that the suffixes behave differently ($H(3) = 163.8$, $p < 0.00001$, $N = 2218$) and pairwise comparisons using Wilcoxon rank shows that only *-ais* suffix is different from the other suffixes according to backness score (*-ais vs. -ois* $p < .00001$; *-ais vs. -ien* $p < .00001$; *-ais vs. -éen* $p < .00001$). Other pairwise comparisons are not significant

(*-ois vs. -ien* $p = 1$; *-ois vs. -éen* $p = .94$; *-ien vs. -éen* $p = .75$).

5 Geographical properties

While affix rivalry studies benefit from the scrutinizing of various linguistic properties, demonyms differ from most morphological rivals because of the lack of semantic variability and the specificities of their base referents, being geographical entities. Such a characterization entails distinctive properties, notably geographical, that have been but briefly addressed on very few occasions in the literature. Among the criteria that could be addressed, the size of the city or country (when relevant). It can indeed be hypothesized that the bigger a city is, the more impactful it is bound to be, and thus the suffix of its demonym to be known and reuse by other cities. Size could here be evaluated in terms of surface area, but also on the basis of demographic data. Diachronic aspects could also be taken into account, such as the affixes productivity and the date of creation of cities, or the date of first attestation of a demonym, as all three might be correlated. However such perspectives are left out from the scope of this study for future work.

This section provides a preliminary quantitative investigation on two non-linguistic features: the type of location denoted by the base toponym (Section 5.1) and its geographical location (Section and 5.2).

5.1 Type of locations

While we postulate that demonyms do not display any significant semantic variation from one another, they nevertheless display some kind of specialization linked to the nature of the denoted geographical entity. This specialization particularly emerges when expliciting the link of the inhabitant to the location denoted by the base toponym. For instance, if a *parisien* is a person who lives in the city of Paris, a *méditerranéen* is a person that lives by the mediterranean sea, while *européen* designates a citizen or inhabitant of the European continent or the European Union. The relationship between the base toponym and the demonym is not strictly equivalent depending on whether the toponym denotes a country, a city, or a natural entity, and one might hypothesize that suffixes may specialize with respect to the type of entity denoted by the base.

To test this hypothesis, we annotated the demonyms of our dataset with respect to the kind of entities denoted by their base toponym. The annotation was performed according to 3 labels: country, city and area. Countries and capital cities were automatically identified based on established lists.[2] Area-denoting toponyms include various configurations such as continent (*Eurasie*), natural areas or elements (*Amazonie*, *Méditerranée*), administrative and geopolitical areas that are

not defined as a country or a city (*Californie*, *Angleterre*, *Languedoc*, *Hollywood*).

Table 10 presents a quantitative description of the annotation results. As can be seen, a large majority of our toponyms refers to cities (almost 85% of the dataset). The others categories are very poorly represented, with about 6% for the countries, 8% for the areas. As far as the suffixes are concerned, we observe that the suffix *-ien* is favored by toponyms that do not designate a city. On the one hand, there are 26.6% (171/644) demonyms in *-ien* which base toponym does not refer to a city, *vs.* 16.4% for *-ais* (91/555) , 14.5% for *-éen* (24/165) and 5.6% for *-ois* (48/854) . On the other hand, 46.2 and 57.1% of area- and country-denoting toponyms have demonyms in *-ien*. We also observe a strong association between *-ois* and toponyms denoting cities, as 42.8% of *-ois* demonyms are linked to cities, *vs.* 19.8% and 8.2% with areas and countries. Overall, the correlation between distribution of suffixes and the type of location denoted by the base toponyms is statistically significant based on a Kruskal-Wallis test ($H(6) = 132.66$, $p < .00001$,$N = 2218$)

	-ais		*-éen*		*-ien*		*-ois*		
	#	%	#	%	#	%	#	%	Total
City	464	24.6	141	7.5	473	25.1	806	**42.8**	1884
Area	51	27.3	13	6.9	87	**46.5**	36	19.3	187
Country	40	27.2	11	7.5	84	**57.1**	12	8.2	147
Total	555	25.0	165	7.5	644	29.0	854	38.5	2218

Tab. 10: Distribution of suffix according to the geographical types of base toponyms

5.2 Local influence of series effect

Though not extensively, the geographic location of base toponym referents as a potential explanatory factor in demonym affix rivalry has been evoked. García Sánchez (2005), quoted by Brizuela (2017), discusses the geographical position of cities as one factor among others in Mexican demonym affix rivalry. García Sánchez (2005) and Brizuela (2017) support the hypothesis of an analogical effect in affix selection for a given demonym, based on the demonyms of nearby (major) cities. They show that suffixes distribute to some extent non-randomly with respect to the considered geographical area. The geographical distribution of rival suffixes has also been discussed for French and notably with respect to *-ais*, *-éen*, *-ien* and *-ois* suffixes (Eggert, 2002; Plénat, 2008b). Authors show that while there are no clear exclusive distribution of suffixes across the country, some suffixes tend to be favored on a regional scale, as is *-ais* along the Atlantic coast, or *-éen* in Brittany and Pays de la Loire. Our data display similar patterns, as shown in the map in Figure 2. In order to have as many examples as possible, the map was drawn based on 5007 demonyms suffixed with *-ais*, *-éen*, *-ien* and *-ois* found in Prolex and for which we had the GPS coordinates, even if we did not have the phonetic transcription of their base toponym. On

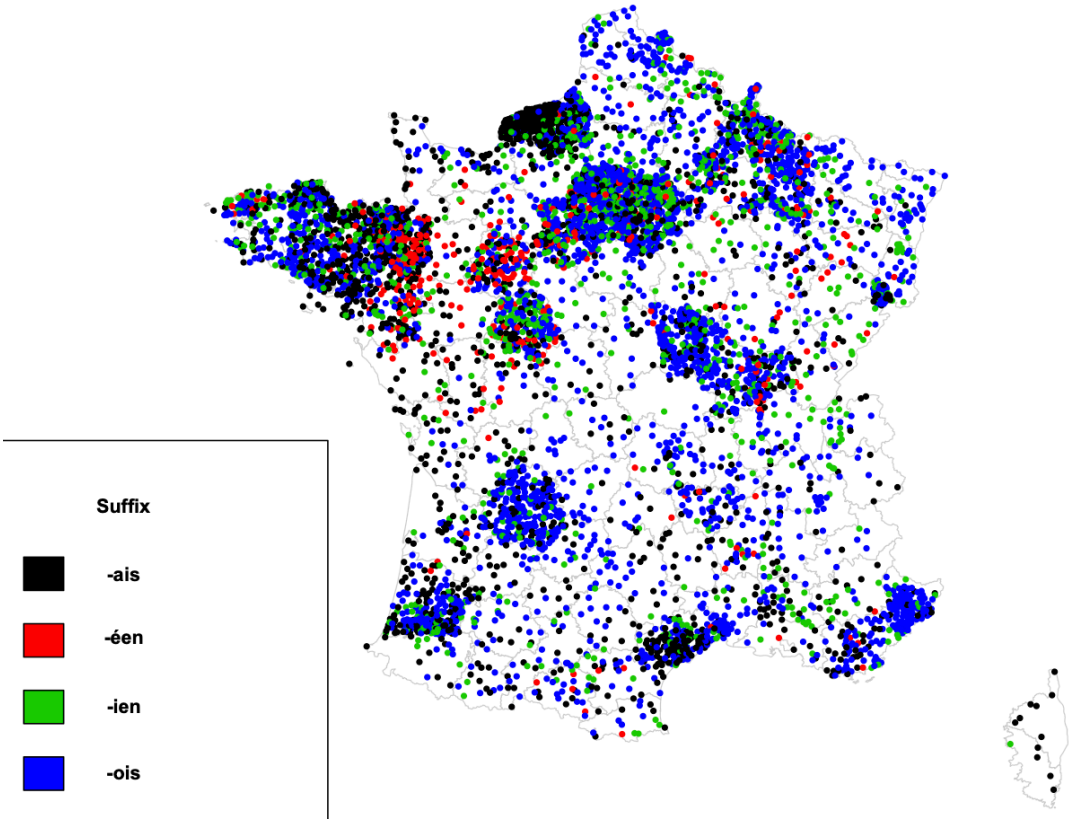

Fig. 2: Distribution of the four demonym suffixes on French metropolitan territory

the map, areas with few dots may correspond either to less populated regions or to areas where denomyms are coined with other suffixes.

While these tendencies can be explained partly based on regional phonological properties (Plénat, 2008b; Brizuela, 2017), the phenomenon has not been clearly quantified, and the strength of this factor has not yet been properly evaluated. In this section, we explore whether the selection of a suffix is a consequence of the geographical proximity between cities referred to by the corresponding toponyms. We hypothesize that the closer to each other two cities are, the more likely they are to be formed by the same affix, by analogy.

To assess the impact of geographical localization on affix selection, we consider the geographical proximity of the city associated to a given demonym to other cities, not their cardinal localization as such. To this end, we compute the crow distance[3] between two cities, *i.e.* the euclidian distance between them (modulo the earth curve) based on their GPS coordinates[4]. Because it relies on GPS coordinates, our analysis only takes into account French cities from our dataset for which we have their GPS coordinates. The final sample contains 1435 French cities, whose distribution is displayed in Table 11. As the figures show, *-ois* demonyms are more numerous in this sample than in the general corpus of 2218 demonyms, while *-ien* demonyms are fewer. This

is in line with the results of Section 5.1: *-ois* is favored when the toponyms refer to cities, while *-ien* is more frequent with non-cities toponyms (see Table 10).

Suffix	#	%
-ois	639	44.5
-ien	336	23.4
-ais	354	24.7
-éen	106	7.4
total	1435	100.0

Tab. 11: Distribution of the four suffixes in the narrow sample

For all 1435 demonyms, we compute the distance of its corresponding city to the 1434 other cities. We thus end up with a total of 1,028,895 unique pairs of cities and their corresponding measures. For each pair of cities, we define whether their demonyms are coined by the same suffix (TRUE) or not (FALSE), regardless of the affixes. Table 12 gives the distribution of our pairs according to their suffixes. Pairs of demonyms whose suffixes are identical (TRUE condition) are highlighted in gray. As shown by table 12, data are unbalanced, both in terms of conditions and suffixes. Pairs with distinctive suffixes are twice as numerous as pairs with identical suffixes (700,728 *vs.* 328,167), and TRUE pairs in *-ois* are 36 times more numerous than TRUE pairs suffixed in *-éen*. Interestingly, while the average distance is somewhat similar between the two conditions (376.4km for TRUE pairs *vs.* 397 for FALSE pairs), the distribution of distance is nevertheless significant based on a Kruskal-Wallis test ($H(630003) = 683553$, $p < 0.00001$, $N = 1028895$).

	-ais	*-éen*	*-ien*	*-ois*
-ais	62481	37524	118944	226206
-éen		5565	35616	67734
-ien			56280	214704
-ois				203841

Tab. 12: Distribution of pairs of demonyms according to their suffixes

However, the crow distance indicates to what extent two cities are close to each other, but it's not informative as to whether they are the closest cities to each other. Even though two cities have a low distance, closest cities might have a stronger impact with respect to the affix selection. To account for that gap, we approach geographical proximity in terms of series effect. This effect is quantified by means of the proportion of *-ois*, *-ais*, *-ien* and *-éen* suffixes in the close neighborhood, *i.e.* the closest cities of a given target city. We arbitrarily set the close neighborhood to 50. We identify for each of these 50 closest cities the suffix used to coin their associated demonym, and we compute 4 scores corresponding to the percentage of each suffix among these 50 neighbors

demonyms. If geographical proximity is indeed a factor in the suffix selection, we expect the proportion of the targeted suffix to be higher than on average for a given demonym.

The boxplots in Figure 3 display the distribution of the suffixes according to the proportion of each suffix among the demonyms of the 50 closest cities. For each measure, we observe that the average proportion, represented by the line in the box, is slightly higher when it comes to the suffix represented by the proportion. In other words, the average proportion of *-ien* is higher for *-ien* suffix than for other suffixes, the average proportion of *-ois* is higher for *-ois* suffix than for other suffixes, and so on. This is in line with what we expected, but some differences are very slim: for instance the mean rate of proportion of *-ais* is 0.2866 for *-ais* and 0.2734 for *-éen*. The main substantial observations concern the demonyms in *-éen*, which on average tend to have higher rate of demonyms in *-éen* in their proximity than other demonyms do (0.1542 *vs.* 0.08746, 0.08244, 0.06188), and the lowest rate of demonyms in *-ois* (0.3345 *vs.* 0.4112,0.4376, 0.4699). Kruskall-Wallis tests applied to all 4 measures show that there is a significant correlation between the suffixes and the proportion of demonyms in *-ais* ($H(3) = 74.7$, $p < 0.00001$, $N = 1435$), in *-éen* ($H(3) = 63.77$, $p < 0.00001$, $N = 1435$), *-ien* ($H(3) = 26.51$, $p < 0.0001$, $N = 1435$) and *-ois* ($H(3) = 84.67$, $p < 0.00001$, $N = 1435$) among the 50 closest cities.

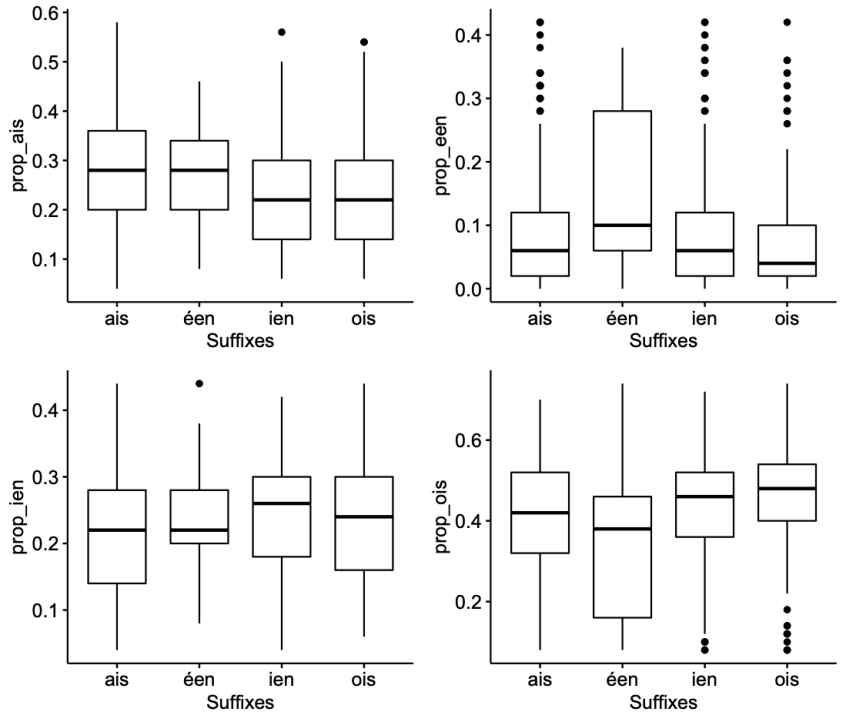

Fig. 3: Distribution of each suffix as a function of the proportion of *-ais*, *-éen*, *-ien* and *-ois* in the 50 nearest demonyms

6 Statistical modeling of demonym affix rivalry

Most of the features presented so far are significantly correlated to the suffixes. One can hypothesize that the distribution of *-ais*, *-éen*, *-ien* and *-ois* suffixes in the formation of French demonyms is a multifactorial phenomenon. At this stage, two questions remain to be investigated. First, to what extent does the combination of the morphological and phonological features presented in Sections 3 and 4 accurately account for such distribution? Second, are geographical features relevant for the modeling of demonym affix rivalry? We provide insights on these two aspects in Sections 6.1 and 6.2, after the presentation of the methodology.

To model the contribution of the features, we compute random forests based on conditional inference trees. Conditional inference trees are non-parametric decision trees that aim at predicting the most probable outcome for a response variable on the basis of given factors. They recursively partition the data into two significant subsets as long as the factors allow it. Conditional inference trees differ from traditional decision trees by the use of a significance test, instead of an information measure, to decide whether each split significantly improves the classification of the data. Random forest algorithm builds on conditional inference trees as it averages predictions from a large number of conditional inference trees, thus allowing for the inference of the variable importance, i.e. the most valuable features. More precisely, the variable importance measures how much the accuracy of the model decreases when a given predictor is removed from the model. The more important the decrease compared to the other variables, the higher the importance of the variable (quantified on a normalized scale from 0 to 100). Such algorithms prove to be particularly relevant when dealing with high dimensional data that display correlated features, and to provide a more robust classification.

In what follows, models are fitted with the train() function from the R package 'caret' (Kuhn, 2008). We use the *rf* method, the *accuracy* metric, and 10-fold cross-validation as our resampling method. The random forest algorithm is trained with the following parameters: the number of tree *ntree* is arbitrarily set to 1000, and the maximum number of terminal nodes *maxnodes* to 35. The number of random variables selected for each tree *mtry* is computed as the square root of the number of predictors. The computation of variables importance is performed with the varImp() function of the 'caret' package, based on the mean decrease of accuracy. For reproductibility purposes, we set the random seed at 42.

6.1 Linguistic features

In order to understand which linguistic features contribute to the selection of the demonym suffix, we first compute a random forest based on linguistic features only (henceforth MODEL1), *i.e.* with the suffix as the response variable, and 12 dependent variables (which involves setting *mtry* to 3), among which 3 are morphological features: polylexicality (*polylex*), opacity (*opaque*), and formal variation (*formal_var*). We decided to keep polylexicality as a feature even though it bears no significant correlation with the suffixes to investigate whether there is any joint effect with other features. The 9 remaining phonological features are the length of the base, *i.e.* its number of syllables (*length*), the backness score (*backness_score*), and the presence or absence of a nasal vowel (*nas_vow*), a front oral vowel (*front_oral_vow*), an approximant consonant (*approx_cons*), a plosive consonant (*plos_cons*), a nasal consonant (*nas_cons*), an alveolar fricative consonant (*alv_fric_cons*), and any other fricative consonant (*oth_fric_cons*) in the final segment of the base.[5] All but *backness_score* and *length* predictors are 2-level factor variables. The resulting importance of the variables are provided in Figure 4.

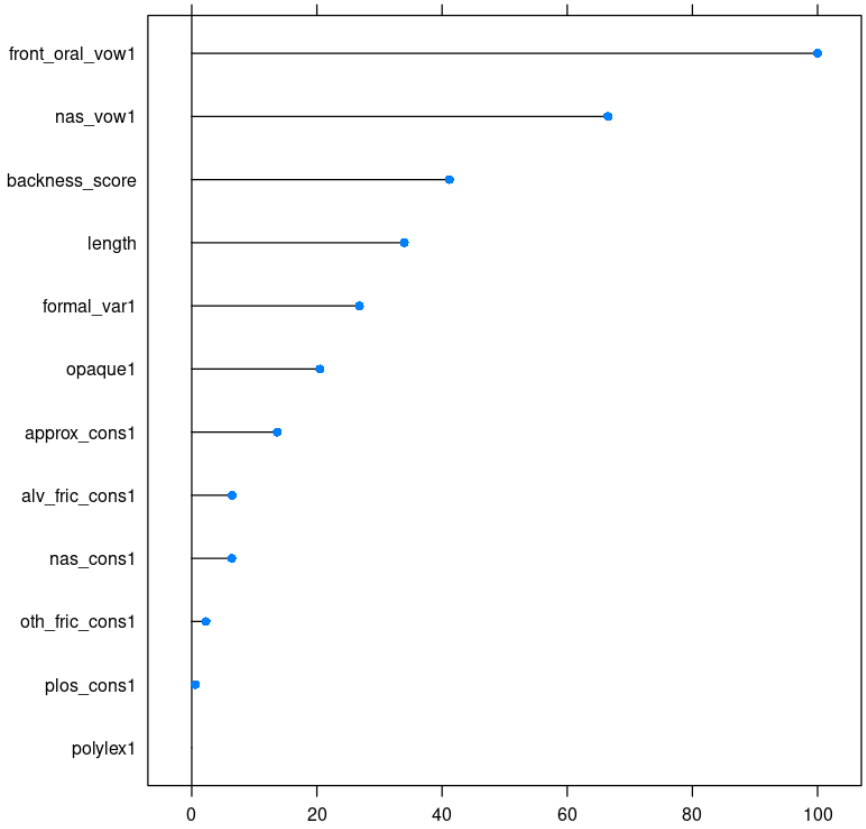

Fig. 4: Variable importance for the MODEL1 random forest with the suffix as the response variable, the linguistic features as dependent variables, and the mean decrease in accuracy as the measure

We can see that the four most important variables are phonological ones: the final front oral vowel which favors *-ien* and *-éen* suffixes (cf. Section 4.2); the final nasal vowel and the backness score, which favors *-ais* (cf. Sections 4.2 and 4.3); the length of the base whose contribution varies along the number of syllables (cf. Section 4.1). Two morphological variables appear to also contribute to the selection of the suffix: the formal variation and the opacity. The least important variables are the polylexicality, the presence of a final plosive consonant or a non-alveolar fricative consonant, which is in line with what we saw in Sections 3.1 and 4.2.

To assess the explanatory power of these features, we provide two measures of accuracy of the MODEL1 model. The first, that we call classification accuracy, is the mean proportion of suffixes that the model correctly classify, in a 10-fold cross-validation[6]. The second is the area under the ROC (Receiver Operating Characteristic) curve (henceforth AUC).[7] It 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. The current model achieves a classification accuracy of 0.57 (with a confidence interval ranging from 0.55 to 0.59) and an AUC value of 0.77 (confidence interval ranging from 0.76 to 0.78). These two measures indicate that the model clearly has a predictive power, without making very accurate predictions. It confirms that the phonological and morphological features participate in the selection of the demonym suffix.

While the model performs relatively well in so far as our response variable has four levels (*-ais*, *-éen*, *-ien*, *-ois*), all levels are not equally well classified. Table 13 presents the confusion matrix for the final model of the MODEL1 random forest (*i.e.* the optimized model obtained after 10-fold cross-validation of random forest). It shows that the suffix *-éen*, and *-ais* to some extent, presents a higher classification error, *i.e.* a lower accuracy, than the suffixes *-ois* and *-ien*. Around 56% and 79% of *-ais* and *-éen* demonyms are not classified as such, while only 29% and 35% of *-ien* and *-ois* demonyms were wrongly classified.

		Reference				
		-ais	*-éen*	*-ien*	*-ois*	Total
Prediction	*-ais*	242	4	33	83	362
	-éen	0	35	17	0	52
	-ien	109	114	457	218	898
	-ois	204	12	137	553	906
Total		555	165	644	854	2218
Class. error		0.564	0.788	0.290	0.352	

Tab. 13: Confusion matrix from the final model from the MODEL1 random forest based on linguistic features. True positives are highlighted in gray.

More specifically, two major sources of error can be identified. First, *-éen* demonyms tend to

be classified as *-ien* demonyms (the reverse not being true). Second, *-ais* demonyms are largely classified as *-ois* demonyms (the reverse still not being true). Such errors can be explained by the similar distribution of these suffixes with respect to some of their morphological and phonological properties described in Sections 3 and 4, such as the presence of a final nasal or front oral vowel for *-éen* and *-ien*.

6.2 Overall modeling

As presented in Section 5, we implemented 5 geographical features: the type of location (*loc_type*), and the proportion of each suffix in the demonyms of the 50 nearest cities (*prop_Ais*, *prop_Een*, *prop_Ien* and *prop_Ois*). However, not all of them can be added to our current model because of their coverage of the data. The four proportion measures were only computed for French cities, while the type of location allows for the distinction of cities and other entities.

To get a preliminary insight on the contribution of geographical features on demonym affix rivalry, we first assess the impact of the one geographical feature available for the whole dataset, namely the type of location. Thus we compute a random forest (henceforth MODEL1b) with the suffix as the response variable, and the association of the 12 previous linguistic predictors and the type of location, according to the same parameters (*mtry* 3, *maxnodes* 35, *ntree* 1000, 10-fold cross-validation).

The resulting MODEL1b model shows little change from the previous MODEL1 model, with a similar classification accuracy (0.58, within the 0.56-0.60 confidence interval) and AUC (0.77, within the 0.75-0.79 confidence interval), and a similar distribution of variables in terms of importance as that of Figure 4. While significant individually with respect to the suffix distribution (see Section 5.1), the addition of the type of location as a feature does not lead to a neat improvement of the classification in a quantitative perspective. However the computation of the variables importance for the MODEL1b random forest shows the relevance of the *loc_type* feature. While both MODEL1 and MODEL1b models display a similar distribution of their variables with respect to their importance, the variable *loc_type* appears among the most important feature of the MODEL1b random forest, in 4th position, after *front_vow*, *nas_vow* and *backness_score* features, and before *length*. This suggest that the *loc_type* feature does contribute to the classification of demonyms, even though it does not help the accurate classification of more demonyms.

To evaluate the contribution of the 4 other geographical features, we compute a third random forest (henceforth MODEL2) which takes into account in addition to the 12 linguistic variables the 4 measures of suffix proportion *prop_Ais*, *prop_Een*, *prop_Ien* and *prop_Ois*. Because these fea-

tures rely on GPS coordinates that were not available at the time of constitution of the data, the random forest is trained on the 1435 demonyms for which these measures were computed (see Section 5.2). Moreover, Because all but one demonym are formed on city-denoting toponyms, the integration of the 5th geographical feature, the type of location, was not deemed relevant. The same parameters are used, except for *mtry* which is set to 4. The variable importance of the MODEL2 random forest model is presented in Figure 5.

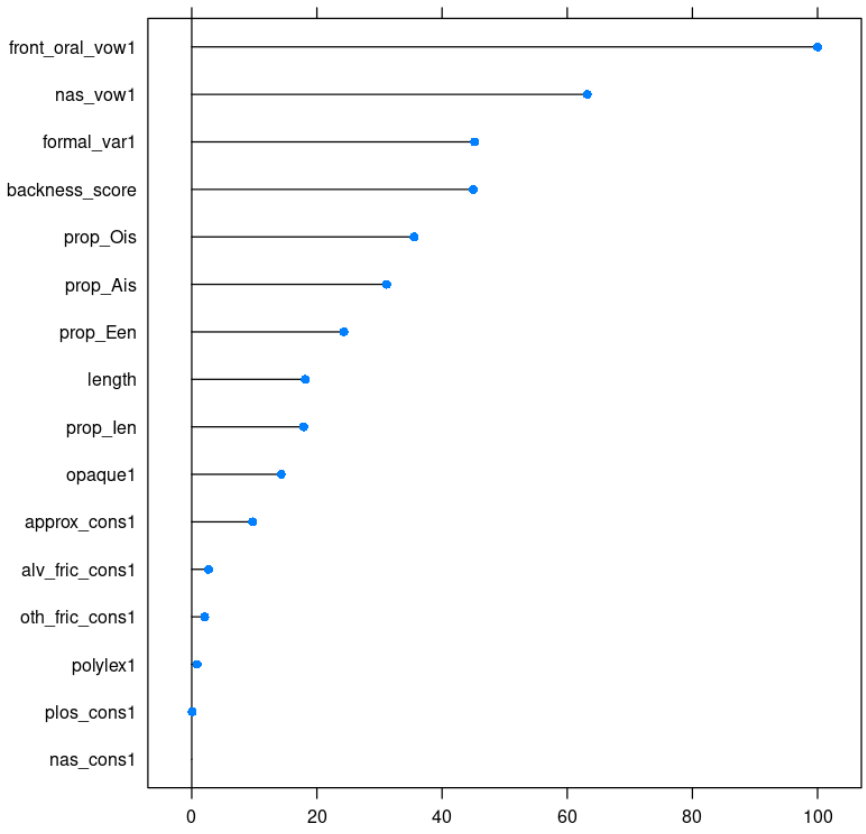

Fig. 5: Overall variable importance from the MODEL2 random forest based on 1435 demonyms, with the suffix as the response variable, the linguistics and non linguistics features as dependant variables, and the mean decrease in accuracy as the measure

The variable importance of MODEL2 in Figure 5 presents an overall similar distribution to that of the MODEL1 random forest in Figure 4. The most important features are once more the presence or absence of a front oral vowel or a nasal vowel in the last segment of the base, and the backness score. Both models differ with respect to the presence or absence of a formal variation in the base, which seems to be more important in MODEL2 than in MODEL1. As for the geographical features included in the MODEL2, 3 out of 4 appear to be rather important features, following backness score (respectively *prop_Ois*, *prop_Ais* and *prop_Een*). By contrast, *prop_Ien* does not seem as important, even though it can be noted that its importance is not null.

Both MODEL1 and MODEL2 models also differ with respect to their performance. The MODEL2 model gets a classification accuracy of 0.62 (with a 0.593-0.643 confidence interval) and an AUC of 0.82 (with a 0.81-0.84 confidence interval), which turns out to be a real improvement compared both to the MODEL1 and MODEL1b models (0.57 and 0.77, 0.58 and 0.77 respectively). This strongly suggests that these four geographical features contribute significantly to the discrimination of the suffixes, contrarily to the type of location feature which proved to be important, but had no additional effect on the classification.

To guarantee that the improvement is indeed due to the geographical features themselves and not the consequence of a bias induced by the resampling of the data, we train a fourth random forest (henceforth MODEL2b) on the reduced sample of 1435 items, with the 12 linguistic dependant variables only. This corresponds to the training of a similar random forest as that of MODEL1 but on the reduced sample. The resulting MODEL2b random forest performs better than MODEL1, with a classification accuracy of 0.61 (confidence interval ranging from 0.583 to 0.634) and an AUC of 0.79 (confidence interval ranging from 0.76 to 0.82), yet not as good as the previous geographically-enhanced MODEL2 random forest. Moreover, the analysis of the variables importance for MODEL2b shows that the variation of some features observed for the MODEL2 model (and notably the increase of the importance of the formal variation feature) is already instantiated in MODEL2b, and thus a consequence of the resampling. These results confirm that while part of the variations and improvement can be attributed to the resampling, the geographical features do have a significant effect on the overall model.

In the remainder of this Section, we investigate more thoroughly the properties of the most complete model, MODEL2 with respect to each suffix. The observed improvement for MODEL2 does not indeed affect similarly all four suffixes. This much emerges from comparing the confusion matrix for the final model of the MODEL2 random forest, presented in Table 14, with that of the MODEL1 model (Table 13). It appears that the improvement benefits to *-ois*, which classification error decreases from 0.352 to 0.280, *-ais* (from 0.564 to 0.458) and *-éen* (from 0.788 to 0.642). By contrast, the geographical features are detrimental to *-ien*, which classification error rate increased from 0.290 to 0.432. More specifically, they increase the confusion between *-ien* and *-ois*, when the classification of *-ien* demonyms is concerned.

Figure 6 presents the variable importance from MODEL2 for each suffix. Despite similar trends as these exhibited in Figure 5, the variable importance presents local variations according to the suffixes. For instance, the presence or absence of a nasal vowel in the last segment of the base turns out to be the most importance feature for *-ais*, and more specifically mainly associated to *-ais*, as it displays a far lower importance for the other three suffixes. A similar analysis can

		Reference				
		-ais	*-éen*	*-ien*	*-ois*	Total
Prediction	*-ais*	192	6	22	63	283
	-éen	0	38	8	1	47
	-ien	31	52	191	115	389
	-ois	131	10	115	460	716
Total		354	106	336	639	1435
Class. error		0.458	0.642	0.432	0.280	

Tab. 14: Confusion matrix from MODEL2 random forest based on linguistic and non linguistic features features. True positives are highlighted in gray.

be made for the backness score, which is among the most important feature for *-ais*, but almost among the least important for the other three suffixes. On the other hand, the presence or absence of an alveolar fricative consonant in the last segment of the base is of no importance for *-ois*, and mainly important for *-ien*. These observations are in line with the descriptions made in Sections 4.2 and 4.3.

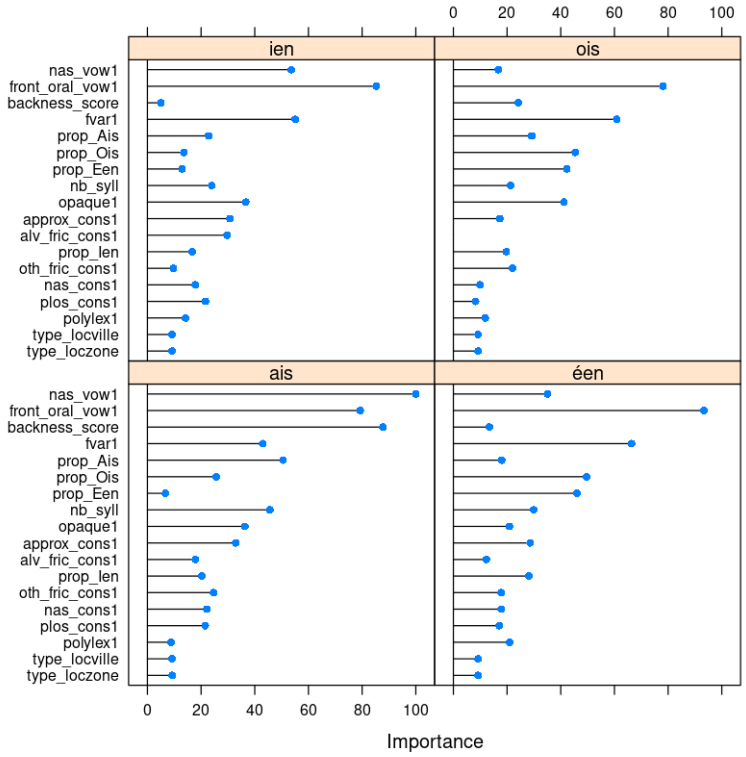

Fig. 6: Variable importance for each suffix from the MODEL2 random forest based 1435 demonyms, with the suffix as the response variable, the linguistics and non linguistics features as dependent variables, and the mean decrease in accuracy as the measure

With respect to the geographical features, Figure 6 highlights the respective importance of *prop_Een* for *-éen* demonyms, *prop_Ais* for *-ais* demonyms, and *prop_Ois* for *-ois* demonyms. By contrast, the proportion of each suffix in the demonyms of the 50 nearest cities does not appear

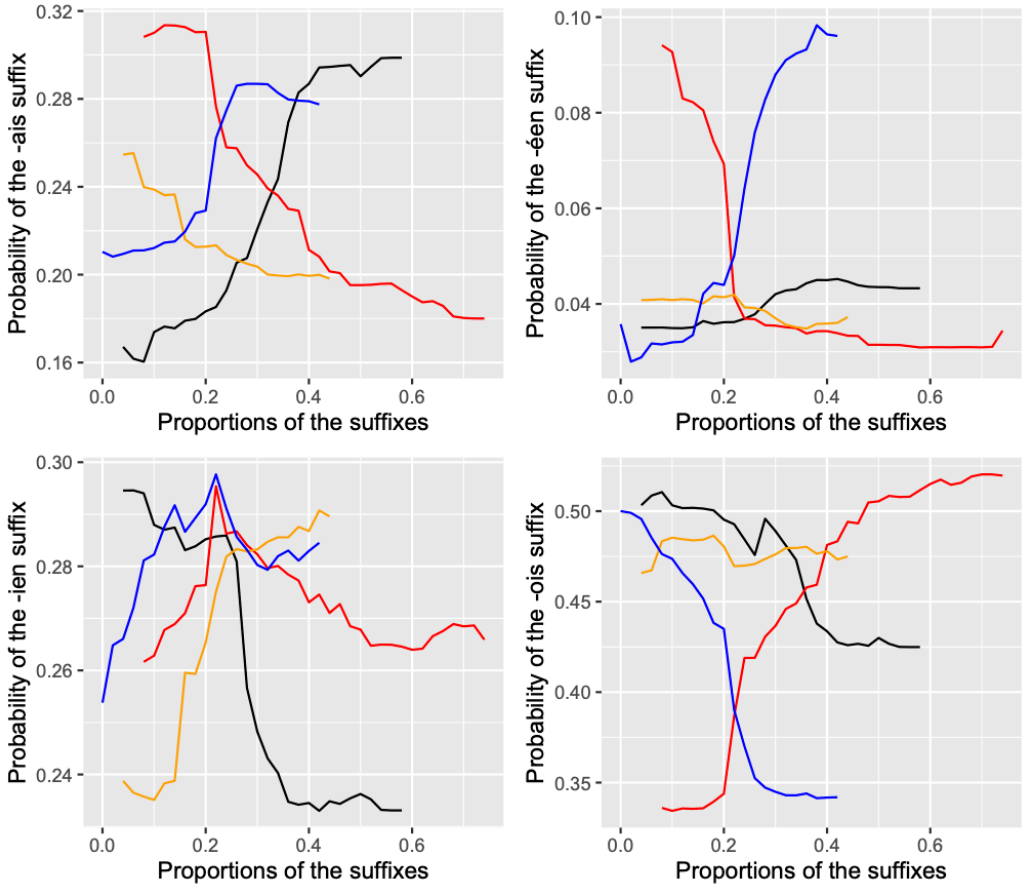

Fig. 7: Partial effects of the geographical features in MODEL2. Black lines represent the proportion of *-ais*, blue lines the proportion of *-éen*, orange lines the proportion of *-ien*, and red lines the proportion of *-ois*.

quite as important in the context of *-ien* demonyms, all four features display a rather low and similar importance.

We compute partial dependence scores[8] in order to better understand the direction that the values of the geographical variables have on the probability of each suffix (Greenwell, 2017). Each plot in Figure 7 represents the probability of a suffix as a function of the four geographical predictors (*i.e.* the proportions of *-ais*, *-éen*, *-ien*, or *-ois* among the 50 closest cities), given that the other predictors are neutralized by averaging their effect. Note that the y axis scale is not the same in all plots: for instance, the probability of *-éen* ranges from 0.029 to 0.105, while the probability of *-ois* ranges from 0.332 to 0.521.

The main observation is that in each plot, the line corresponding to the proportion of the suffix concerned by the probability rises as the proportion increases (the black line in the upper left plot; the blue line in the upper right plot; the orange line in the lower left plot; the red line in the lower right plot). This is consistent with our series effect hypothesis: the higher the

proportion of a suffix around a city, the higher the probability that the demonym is coined with that suffix.

The use of the suffixes seems also to be disfavored by the presence of some other suffixes in their geographical neighborhood: for instance, the higher the proportion of *-ois* (red line), the lower the probability of *-ais* (upper left plot), of *-éen* (upper right plot), and in a lesser extent, of *-ien* (lower left plot). Along the same lines, the probability of *-ien* (lower left plot) and *-ois* (lower right plot) decreases as the proportion of *-ais* increases (black line). We also observe that the higher the proportion of *-éen* (blue line) the lower the probability of *-ois* (lower left plot). However, the effect of the proportion of *-éen* on the probability of *-ais* (upper left plot) and *-ien* (lower left plot) is less clear. Finally, the size of the effect of the presence of *-ien* among the 50 closest cities seems to be weak compared the other proportions: it slightly disfavors the use of *-ais* (decreasing orange line in upper left plot), but it does not seem to affect the probability of *éen* or *-ois* (relatively constant orange line in upper right and lower right plots).

Overall, these observations reinforce our point: the choice of a demonym suffix is affected by the other demonyms in the geographical area. Not only is a specific suffix favored by its own high proportion among the 50 closest cities, but it tends to be disadvantaged by the high proportions of other suffixes.

All these observations are in line with our initial hypothesis. It does suggest the existence of an effect from referential properties, mostly in terms of geographical proximity, in affix selection for demonym formation. Results on such small sample are promising, and complementary data should allow for a better modeling of the joint effect of linguistic and geographical features.

7 Conclusion

This study focused on explaining the selectional constraints at stake in affix rivalry as far as French demonyms are concerned. Based on statistical modeling, we showed that various phonological, morphological and geographical features played a role. Features such as the length of the base, the nature of the final segment of the base, or the overall backness of the base all are linguistic properties that discriminate to some extent one or another suffix. While these observations are globally in line with previous studies of this particular affix rivalry, the present study models more precisely their interplay. A specific contribution of this work is to provide a new insight through the use of quantitative geographical properties. Our models showed that the overall distribution of other suffixes on a local scale had a substantial role in the matter, allowing for a finer-grained discrimination when combined with phonological and morphological features.

Despite difficulties with respect to some suffixes, our model performs a surprisingly good classification when considering that we are dealing with four-level response variable. While preliminary to some extent, this study shows the benefits of using geographical properties as predictive features. Future work should include the pursuit of this lead, by taking into account a larger sample. Metrics should also be revised, such as the 50 threshold of nearest cities, and other metrics could be explored, as discussed in Section 5.

Beyond geographical considerations, other factors could also be investigated. Similarly to Uth (2010) and Bonami and Thuilier (2019), diachronic information could be used: the correlation between the diachronic productivity of all 4 affixes and the date of first attestation of the demonyms could be investigated. Axiological criteria could also be studied, such as the perception by the speakers of various connotations associated to the suffixes (see Akin 2006; García Sánchez 2005; Chesnokova et al. 2021).

Notes

[1]Other suffixes forming demonyms are, among others: *-al* (Provence→provençal), *-at* (Auvergne→auvergnat), *-i* (Rabat→rabati), *-ite* (Yémen→yéménite), *-ot* (Sologne→solognot), etc.

[2]Countries and capital cities were extracted from the official French site https://www.data.gouv.fr/fr/datasets/etats-et-capitales-du-monde/. French areas and cities were retrieved from the web sites www.regions-et-departements.fr and https://sql.sh/736-base-donnees-villes-francaises.

[3]While a good initial approximation, the crow distance builds on the idea on a linear assessment of distance. Yet if we consider the diffusion of demonyms as the consequence of human activity, the actual distance between two cities can quite heavily vary, depending on the geographic relief between them. Two cities might be close by crow distance, and yet necessitate a longer travel time. A preliminary evaluation of this gap using OpenStreetMap suggests a significant difference between crow distance and what we'll call travel distance, but that remains to be explore in depth, notably to assess the impact of this parameter on our current observations.

[4]We use the distance() function from the *GeoPy* library, which computes by default the geodesic distance between two points based on their lattitude and longitude. Because of the scale, we set the unit of measurement in kilometer.

[5]Following the discussion relative to B stem in Section 3.3, we still tried to include it in one of our models. In this case, the B stem has been identified as: i) the base with the pronounced final consonant in the case of final latent consonant; ii) the base with an oral wovel followed by a nasal consonant in the case of final nasal wovel; iii) the base in all other cases. However, our model based on the B stem gave no better results than the model based on the surface form of the toponym. Given the problems raised by the annotation of B stem (cf. Section 3.3), we chose to keep the toponym as the base.

[6]Classification accuracy is computed by means of the train() function from the 'caret' package.

[7]AUC is computed with the multiclass.roc() function from the 'pROC' package.

[8]We used partial() function from the R pdp package.

References

Akin, S. (2006). 'Comment dériver un gentilé à partir d'un toponyme? Les potentialités signifiantes de Seine-Maritime'. *Cahiers de Sociolinguistique*, 11:63–80.

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

Bonami, O. and Boyé, G. (2003). 'Supplétion et classes flexionnelles'. *Langages*:102–126.

Bonami, O., Boyé, G., and Kerleroux, F. (2009). 'L'allomorphie radicale et la relation flexion-construction'. In B. Fradin, F. Kerleroux, and M. Plénat (eds.), *Aperçus de morphologie du français*. Saint-Denis: Presses Universitaires de Vincennes, 103–125.

Bonami, O. and Thuilier, J. (2019). 'A statistical approach to rivalry in lexeme formation: French-iser and-ifier'. *Word structure*, 12:4–41.

Boyé, G. (2006). 'Suppletion'. In K. Brown (ed.), *Encyclopedia of Language and Linguistics, t. 12*. Oxford: Elsevier, 297–299.

Brizuela, S. L. E. (2017). 'Sufijos gentilicios en el español de méxico. rivalidad y restricciones de aplicabilidad'. *Anuario de letras. Lingüística y filología*, 5:67–90.

Chesnokova, O. S., Radović, M., and Kotenyatkina, I. B. (2021). 'Spanish South American and Brazilian Demonyms: Morphosyntactic Structure and Axiological Values'. *RUDN Journal of Language Studies, Semiotics and Semantics*, 12:576–596.

Danner, S. G. (2016). 'Selectional Effects in Allomorph Competition'. In *Proceedings of the Annual Meetings on Phonology*, vol. 2.

Dubois, J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain: essai d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Paris: Larousse.

Eggert, E. (2002). *La dérivation toponymes-gentilés en français: mise en évidence des régularités utilisables dans le cadre d'un traitement automatique*. Ph.D. thesis, Tours.

Fábregas, A. (2010). 'A syntactic account of affix rivalry in spanish nominalizations'. In *The syntax of nominalizations across languages and frameworks*. De Gruyter Mouton, 67–92.

Faust, N. (2017). 'The effect of a null pronoun on morpho-phonology: Israeli Hebrew denomyns'. *IATL 2017 Proceedings*:23–38.

García Sánchez, J. J. (2005). 'Irradiación analógica en la formación de gentilicios'. *Vox romanica*, 64:160–170.

Gardani, F., Rainer, F., and Luschützky, H. C. (2019). 'Competition in morphology: A historical outline'. In *Competition in inflection and word-formation*. Springer, 3–36.

Greenwell, B. M. (2017). 'pdp: An R Package for Constructing Partial Dependence Plots'. *The R Journal*, 9:421–436.

Huyghe, R. and Wauquier, M. (2021). 'Distributional semantics insights on agentive suffix rivalry in French'. *Word Structure*, 14:354–391.

Kuhn, M. (2008). 'Building predictive models in r using the caret package'. *Journal of statistical software*, 28:1–26.

Lohmann, A. (2017). 'Phonological properties of word classes and directionality in conversion'. *Word Structure*, 10:204–234.

Martin, F. (2010). *The semantics of eventive suffixes in French*. De Gruyter Mouton.

Missud, A. and Villoing, F. (2020). 'The morphology of rival -ion, -age and -ment selected verbal bases'. *Lexique*, 26:29–52.

Plénat, M. (2008a). 'Le thème L de l'adjectif et du nom'. In J. Durand, B. Habert, and B. Laks (eds.), *Actes du Congrès Mondial de Linguistique Française 2008*. Paris: Institut de Linguistique Française, 1613–1626.

——— (2008b). 'Quelques considérations sur la formation des gentilés'. In B. Fradin (ed.), *La raison morphologique. Hommage à la mémoire de Danièle Corbin*. Amsterdam: John Benjamins, 155–174.

Roberts, M. (2017). 'The semantics of demonyms in English'. In Z. Ye (ed.), *The Semantics of Nouns*. Oxford Scholarship Online, 205–220.

Roché, M. (2010). 'Base, thème, radical'. *Recherches linguistiques de Vincennes*:95–134.

Roché, M. and Plénat, M. (2016). 'De l'harmonie dans la construction des mots français.' In *Actes de CMLF 2016*.

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

Tran, M. and Maurel, D. (2006). ‘Prolexbase : un dictionnaire relationnel multilingue de noms propres’. *Traitement Automatique des Langues*, 47:115–139.

Uth, M. (2010). ‘The rivalry of French-ment and-age from a diachronic perspective’. In *The semantics of nominalizations across languages and frameworks*. De Gruyter Mouton, 215–244.