

Huyghe, Richard & Wauquier, Marine. 2021. Distributional semantics insights on agentive suffix rivalry in French. *Word Structure* 14:3, 354-391. Pre-print version.

## Distributional semantics insights on agentive suffix rivalry in French

Richard Huyghe  
University of Fribourg

Marine Wauquier  
University of Paris, LLF, CNRS

### Abstract

The formation of French agent nouns (ANs) involves a large variety of morphological constructions, and particularly of suffixes. In this study, we focus on the semantic counterpart of agentive suffix diversity and investigate whether the morphological variety of ANs correlates with different agentive subtypes. We adopt a distributional semantics approach and combine manual, computational and statistical analyses applied to French ANs ending in *-aire*, *-ant*, *-eur*, *-ien*, *-ier* and *-iste*. Our methodology allows for a large-scale study of ANs and involves both top-down and bottom-up procedures. We first characterize agentive suffixes with respect to their morphosemantic and distributional properties, outlining their specificities and similarities. Then we automatically cluster ANs into distributionally relevant subsets and examine their properties. Based on quantitative analysis, our study provides a new perspective on agentive suffix rivalry in French, that both confirms existing claims and sheds light on previously unseen phenomena.

### 1. Introduction

Agent nouns, i.e. nouns that describe performers of actions<sup>1</sup>, are known to be morphologically diverse in many Indo-European languages. In French, agent nouns (ANs) arguably include affixed, converted, compound nouns, as well as morphologically simple or opaque nouns (Author 2020). Furthermore, the

---

<sup>1</sup> The definition of agent has been long debated. Authors notably disagree on the intentionality and animacy of agents (Gruber 1967, Fillmore 1968, Cruse 1973, Jackendoff 1987, Schlesinger 1989, among others). To avoid such controversies, prototypical or scalar approaches to agentivity have been developed, in which some arguments or participants are analyzed as more (prototypically) agentive than others according to different defining criteria (see e.g. Lakoff 1977, Dowty 1991, Grimm 2011). In order not to be a priori restrictive about the lexical items considered agentive, we define here agents as effectors (i.e. entities deploying energy to perform actions) that are prototypically but not necessarily animate and intentional. By *action* we refer to any dynamic eventuality, regardless of the volitive or animate nature of the causer. It will appear that, even based on such a broad definition of agent, most agent nouns restrict their denotation to human beings (see Section 3.2).

regular morphological processes that produce ANs involve a variety of constructions. Words from different parts of speech and with different semantic types can be selected as a base for ANs, and different suffixes are used in AN derivation, including *-eur* (Winther 1975, Anscombre 2001, Fradin & Kerleroux 2003, Roy & Soare 2012, Author & coll. 2015), *-iste* (Roché 2011, Cartoni & Namer 2012), *-ier* (Corbin & Corbin 1991, Roché 2004), *-aire* (Schneidecker & Aleksandrova 2016), among others. In this study, we address the question of whether these morphological differences relate to fine-grained semantic distinctions, and what these distinctions could be. In particular, we investigate the correlation between suffix diversity and the existence of different agentive subtypes, by comparing French ANs ending in *-aire*, *-ant/-ante*<sup>2</sup>, *-eur/-euse/-rice*, *-ien/-ienne*, *-ier/-ière* and *-iste*.

To explore the semantics of ANs, we adopt a distributional semantics approach that will allow us to evaluate the semantic similarity between these nouns. In case agentive suffixes relate to different agentive subtypes, we should observe distributional differences between groups of ANs ending with distinct suffixes—and possibly more or less significant differences between groups, depending on the semantic proximity between the suffixes. More generally, the existence of semantically homogeneous subclasses of ANs can be highlighted by distributional similarities between ANs, and it can be asked whether these subclasses correlate with morphological distinctions between ANs.

The article is organized as follows. In Section 2, we discuss to what extent agentive suffixes in French can be considered morphological rivals. In Section 3, we scrutinize the distributional profile of ANs ending with different suffixes in order to bring out possible semantic specificities. In Section 4, we analyze samples of differently suffixed ANs in a bottom-up approach, and examine their clustering in distributional semantic models. One of the goals of the study is to assess which morphosemantic properties would be the best predictors for distributional similarities between ANs.<sup>3</sup>

## 2. Agentive suffix rivalry

---

<sup>2</sup> ANs ending in *-ant* can be regarded as converted from verbs, *-ant* being the inflection marker of present participle in French. This view may be challenged by the existence of feminine ANs ending in *-ante*, since *-ante* is absent from verbal inflection paradigms. By default, we include in our study both ANs ending in *-ant* and *-ante*, relying on the fact that the ending in *-ant* ensures the direction of derivation from verb to noun—even in the case *-ant* is not considered a true suffix. In contrast, we do not take into account ANs that are possibly converted from verbs if the direction of conversion cannot be decided on a morphological basis, as is the case for *guide* ‘guide’, *juge* ‘judge’, *assassin* ‘murderer’ with respect to *guider* ‘guide’, *juger* ‘judge’, *assassiner* ‘murder’ (see Tribout 2010). For writing convenience, we will group together *-aire*, *-ant/-ante*, *-eur/-euse/-rice*, *-ien/-ienne*, *-ier/-ière* and *-iste* under the label ‘agentive suffixes’.

<sup>3</sup> The complete linguistic and distributional data used in the study will be made available online on publication.

In this section, we briefly address the issue of defining affix rivalry, and examine whether agentive suffixes in French should be considered rivals or not.

### 2.1. Affix rivalry in question

Affix rivalry can be roughly defined as a case of morphological competition<sup>4</sup> in which two or more affixes are associated with equivalent word-formation patterns (modulo phonology). Such a definition needs further clarification.

First, it should be specified whether the equivalence condition applies to morphosyntactic and/or semantic properties of words. In derivational processes, it may concern the part of speech (POS) and/or meaning of the morphological base and/or derived lexeme, which leads to various configurations possibly defined as rival. Second, the granularity with which semantic equivalence is assessed has to be made explicit. The equivalence can be evaluated with coarse-grained or fine-grained semantics. Depending on that granularity, two nominalizing suffixes for instance will be considered rivals if they both form nouns that denote actions, or only a particular type of actions. The criteria used to define the relevant semantic types also have to be explicitly specified. Third, affixes are known to often be polyfunctional, and rivalry could be defined as partial or total depending on whether it occurs between all functions of rival affixes, or only between some of them.

Epistemologically, studying affix rivalry appears to be a fruitful way of approaching general issues such as derivational synonymy, morphological change, overabundance, and the many-to-many relations between form and meaning, as well as a method for refining the analysis of some affixes. In that perspective, affix rivalry is not usually considered in a too strict definition, and cases of loose equivalence are often taken into account. Indeed, affix rivalry in the existing literature is generally considered to occur when two or more affixes, in at least one of their derivation patterns, apply to bases with the same POS, and derive words with the same POS and coarse-grained meaning. Equivalence is not required for input semantics, nor is it for fine-grained output semantics, and partial functional equivalence is seen as a sufficient condition for rivalry.

In studies about affix rivalry, authors investigate differences in the distribution of rival affixes. Distinct explanatory variables are considered, often in combination. They pertain to the phonological selection of morphological bases, syntactic properties of bases or derivatives (e.g. subcategorization and argument structure), semantic properties of bases or derivatives (e.g. ontological types, fine-grained semantic

---

<sup>4</sup> For a theoretical and historical overview of morphological competition, see Gardani et al. (2019).

types), derivational paradigmatic structures, polysemy properties, diachronic distribution, productivity, geographical or social preferences, etc. (see Fábregas 2010, Uth 2010, Arndt Lappe 2014, Koehl & Lignon 2014, Schulte 2015, Varvara 2018, Author & coll. 2018, Bonami & Thuilier 2019, Dressler et al. 2019, Naccarato 2019, among others). Investigations often rely on the assumption of linguistic economy: in a systemic organization of derivational processes, no two affixes should be strictly equivalent. It has to be noted though that economy in word-formation patterns does not exclude local equivalence. On one hand, differences between similar word-formation patterns can be categorical or tendential, i.e. based on actualization frequency. In the latter case, the differences between rival affixes may be neutralized in the formation of some lexemes, and there can be instances of rival word-formation patterns that do not exhibit the specificities of said patterns. For instance, it can be argued that *-age* and *-ment* in French differ with respect to the formation of stative nominalizations, largely favored by *-ment* (e.g. *énervement* ‘irritation’, *découragement* ‘despondency’, *dénouement* ‘deprivation’). Nonetheless, *-ment* also allows for the formation of dynamic nominalizations that are similar to those ending in *-age* (e.g. activity nominalizations in the case of *maniement* ‘manipulation’, *cheminement* ‘progression’, *dandinement* ‘waddle’ and *pilotage* ‘piloting’, *massage* ‘massage’, *boycottage* ‘boycotting’). On the other hand, linguistic economy applies diachronically, and does not exclude synchronic equivalence. For example, the suffixes *-eur* and *-on* were similarly used in Old and Middle French to form ANs (e.g. the deverbal nouns *espion* ‘spy’, *forgeron* ‘blacksmith’, *piéton* ‘pedestrian’ and *danseur* ‘dancer’, *graveur* ‘engraver’, *mangeur* ‘eater’ are first attested in the 13th or 14th century), before *-on* became unproductive. Equivalence can therefore be a transitional stage before specialization or formal reduction: one of the rivals may ultimately specialize or disappear, following Bréal’s law of differentiation (1897). As a consequence, challenging the economy principle needs evidence of stable enduring cases of overabundance.

Affix rivalry does not imply the existence of doublets (i.e. differently affixed lexemes derived from the same base, and sharing the same POS and coarse-grained meaning), even in a hypothetical case of pattern overabundance—bases of the same type could be randomly selected by one affix or its rival with no overlap due to lexical blocking. Actual doublets are nonetheless scrutinized and often considered as revealing of rival affixes specificities (Martin 2010, Rodrigues 2013, Fradin 2016, Dal et al. 2018). And yet conclusions should be cautiously drawn from the observation of doublets. Doublets may undergo lexicalization and specialize accidentally, in which case they do not reflect derivational semantics and cannot be considered as evidence for affix specificities. Conversely, the absence of differences in doublets does not imply the absence of differences in patterns, since it could be an effect of neutralization and local equivalence. Only recurring differences between doublets can be exploited. Finally, the analysis of

doublets cannot fully account for affix rivalry, since fundamental differences between rival affixes could block doublets, and then not be highlighted by the observation of doublets only.

## 2.2. Agentive suffixes in French

In many languages, multiple affixes are used to form ANs. Benveniste (1948) identifies two basic morphological types of ANs in Indo-European languages, and traces back their origin to Vedic, Avestan and Ancient Greek. He relates the existence of two distinct agentive suffixes to a fundamental semantic distinction between occasional agents (i.e. agents in a particular event) and functional agents (i.e. agents with an occupational status, regardless of the actual performing of an action). The direct correspondence between these semantic subtypes and agentive suffixes seems to be rarely observed in contemporary languages<sup>5</sup>, but the existence of a variety of agentive suffixes has been pointed out by many authors. Inventories of agentive suffixes have been proposed for instance in Russian (Andrews 1996), Dutch (Booij 2002), Old Church Slavonic (Luschützky 2011) and English (Bauer et al. 2013). As for French, several suffixes have been described as forming ANs. These include most notably *-eur*, *-iste*, *-aire*, *-ien* and *-ier* (Dubois 1962, Anscombre 2001, Fradin & Kerleroux 2003, Roché 2003, 2011, Lignon 2000, Roy & Soare 2012, Author & coll. 2015, Schnedecker & Aleksandrova 2016, among others). Authors diverge on the particular case of *-ant/-ante*, some of them considering that it does form ANs (Lerat 1984, Rosenberg 2008) while others take the opposite view (Winther 1975, Anscombre 2003, Roy & Soare 2012). It remains that most existing studies about French focus on one suffix, and that French agentive suffixes are rarely extensively compared.

All French agentive suffixes are polysemous (or associated with polysemous word-formation patterns), i.e. they are used to form nouns falling under other semantic types than agent. Moreover, their polysemy capacities are not identical. For instance, *-aire* can form nouns that denote beneficiaries (e.g. *allocataire* ‘beneficiary’, Schnedecker & Aleksandrova 2016); *-ant* can form nouns that denote chemical substances (e.g. *désherbant* ‘weedkiller’, Knittel 2017); *-eur* can form nouns that denote instruments (e.g. *réfrigérateur* ‘fridge’, Fradin & Kerleroux 2003); *-ien* can form nouns that denote inhabitants (e.g. *Malien* ‘Malian’, Lignon 2000); *-ier* can form nouns that denote fruit trees (e.g. *cerisier* ‘cherry tree’, Corbin & Corbin 1991); and *-iste* can form axiological nouns (e.g. *marxiste* ‘Marxist’, Roché 2011). It should be noted that in many cases the suffixes are also used to form adjectives, and that the corresponding nouns can be analyzed as converted from them. Overall, because of differences in suffix versatility (i.e. with respect to

---

<sup>5</sup> Benveniste’s claim has actually been discussed even in the case of Indo-European (Balles 2005).

the semantic diversity of derivatives), the rivalry between *-aire*, *-ant*, *-eur*, *-ien*, *-ier*, *-iste* is only partial, and mainly concerns the agentive interpretation.

Even in their strict agentive interpretation, all French agentive suffixes do not rival each other, since they do not always select bases with the same POS. Possible base POS for agentive suffixes are presented in Table 1.

	Verb	Noun
<i>-aire</i>	+	+
<i>-ant</i>	+	–
<i>-eur</i>	+	+
<i>-ien</i>	–	+
<i>-ier</i>	+	+
<i>-iste</i>	+	+

Tab. 1: Base POS selected by agentive suffixes in French

We identify a possible base POS for an affix  $\alpha$  if at least two  $\alpha$ -affixed words can be analyzed as derived from a word with that POS<sup>6</sup>. French agentive suffixes may or may not select the same range of base POS. For example, *-aire* and *-eur* rival in all constructions, whereas *-eur* and *-ant* rival only when ANs are derived from verbs, and *-ant* and *-ien* do not rival in any case. Furthermore, some suffixes allow for both verbal and nominal bases, but clearly favor one base POS. So is the case for *-iste* and *-ier* which most often select nouns rather than verbs, and consequently weakly rival *-ant* (see Table 2). Therefore agentive suffix rivalry

<sup>6</sup> The morphological bases in Table 1 are identified only for the agentive interpretation of the suffixes, regardless of possible non-agentive constructions. When ANs are semantically derived from multiword expressions, only one word is selected as the morphological base—generally the most semantically discriminant item, as in *feu d’artifice* ‘fireworks’ > *artificier* ‘pyrotechnist’. In such cases, we consider the base POS to be that of the multiword expression head, following Roché (2003) and Tribout & Amiot (2019). We thus reanalyze the apparent adjectival bases in ANs such as *fiscaliste* ‘tax expert’ and *plasticien* ‘visual artist’ as nominal bases, since they stand for *droit fiscal* ‘tax law’ and *arts plastiques* ‘fine arts’.

appears to be both divisible and gradable. Cases of AN doublets can nonetheless be found for different pairs of agentive suffixes, as illustrated in (1)-(2)<sup>7</sup>:

(1) Nominal bases

- a. *-ier / -iste*: anecdotier / anecdotiste ‘anecdotist’
- b. *-iste / -ien*: théologien / théologiste ‘theologian’
- c. *-eur / -iste*: harpeur / harpiste ‘harpist’
- d. *-eur / -ier*: bibelotier / bibeloteur ‘trinket collector’
- e. *-aire / -ier*: boutiquaire / boutiquier ‘storekeeper’
- f. *-aire / -iste*: pamphlétaire / pamphlétiste ‘lampoonist’

(2) Verbal bases

- a. *-eur / -aire*: protestateur / protestataire ‘protestor’
- b. *-eur / -ant*: emprunteur / empruntant ‘borrower’
- c. *-aire / -ant*: exécuteur / exécutant ‘executor’
- d. *-ier / -eur*: boursicoteur / boursicotier ‘speculator’
- e. *-eur / -iste*: arrangeur / arrangiste ‘arranger’

Note that differences in the selection of base POS can to some extent correlate with differences in the semantic construction of agentivity. Whereas verbal bases often involve an agent in their semantic structure, nominal bases often do not, in which case agentivity has to be developed directly in the semantic structure of the derived noun. However, the most important factor may be the semantic type of the base, since action nouns can provide an agentive semantic component, and stative verbs cannot. When derived from action-denoting bases, it is likely that ANs will denote the agent of the action denoted by the base. When ANs are not derived from action-denoting bases, a semantic component of action has to be elaborated through the derivational process. The semantic type of the base might then determine

---

<sup>7</sup> Examples in (1)-(2) are taken from an automatically generated list of possible competing nouns ending in *-aire*, *-ant*, *-eur*, *-ien*, *-ier* and *-iste*. Nominal pairs are extracted from the French *Wikipedia 2018* corpus under the condition that they are tagged as nouns and include the same basic form (independently of POS and semantic specification). Most of the automatically listed nouns are not AN doublets, since they are not necessarily derived nouns (e.g. *bouclier* ‘shield’, as opposed to *boucleur* ‘looper’, is not a derived noun), do not necessarily denote agents (e.g. *dentier* ‘dentures’, as opposed to *dentiste* ‘dentist’, does not denote an agent), or do not necessarily derive from the same lexeme (e.g. *bouquineur* ‘reader (informal)’ can be analyzed as semantically derived from the verb *bouquiner* ‘read (informal)’, whereas *bouquiniste* ‘second-hand book seller’ is semantically related to the noun *bouquin* ‘book (informal)’).

the kind of dynamic predicate involved. For instance, artifact-denoting bases could give rise to predicates of manufacturing (*chocolat* ‘chocolate’ > *chocolatier* ‘chocolate maker’), operating (*grue* ‘crane’ > *grutier* ‘crane driver’), manipulating (*bagage* ‘luggage’ > *bagagiste* ‘luggage handler’), selling (*bijou* ‘jewel’ > *bijoutier* ‘jeweler’), whereas domain-denoting bases could give rise to predicates of intellectual practice (*histoire* ‘history’ > *historien* ‘historian’).

The actual existence of rival agentive suffixes in French raises questions about their possible distinction. Different discriminating factors could determine a non-arbitrary distribution, including phonology, diachrony, syntax and semantics. For instance, Roché (1997) argues that the rejection of the yod repetition in morphosemantically well-formed *\*camionnier* and *\*avionnier* blocks their construction and promotes alternate forms in *-eur* (*camionneur* ‘truck driver’ and *avionneur* ‘aircraft manufacturer’). Similarly, Lignon (2000) claims that in the formation of nouns denoting scientists, *-iste* does not euphonically combine with bases ending with a sibilant, which results in the formation of alternate derivatives in *-ien* (e.g. *mathématicien* vs. *\*mathématiciste* ‘mathematician’). The distribution between rival suffixes can also depend on diachronic evolution and productivity variations. Lignon (2007) shows that the nouns ending in *-ien* that denote specialists and are not morphologically based on words ending with a sibilant (e.g. *grammairien* ‘grammarian’) were formed in a period when *-ien* was more productive than *-iste*. A consultation of the *Neoveille* website<sup>8</sup> that extracts neologisms from French language press, suggests that *-eur* (3a), *-ien* (3b) and *-iste* (3c) could be currently more productive than *-aire*, *-ant* and *-ier*, for which no neologisms are listed.

- (3) a. prestateur ‘provider’, instagrammeur ‘Instagram user’, canyoneur ‘canyoning amateur’
- b. cogniticien ‘cognitics engineer’, parapharmacien ‘parapharmacist’, aromaticien ‘flavorist’
- c. plâtriste ‘plaster cast maker’, neuroscientiste ‘neuroscientist’, parapentiste ‘paraglider’

Some syntactic features might as well account for the distinction between agentive suffixes. For example, Roy and Soare (2012) claim that *-ant* contrasts with *-eur* in being compatible with unaccusative verbal bases (e.g. *arriver* ‘arrive’ > *arrivant* vs. *\*arriveur* ‘arrival’). More generally, morphological differences in AN formation could be related to differences in argument structure, and to the possibility or impossibility of inheriting argument structure from the base.

---

<sup>8</sup> <https://tal.lipn.univ-paris13.fr/neoveille/html/login.php?action=login>



In this study, we will focus on the semantic aspects of agentive suffix rivalry, while keeping in mind that semantics is not the only possible factor to account for agentive suffix distribution. We will investigate whether the existence of multiple agentive suffixes is related to semantic distinctions between agentive constructions and agentive subtypes.

### **3. Distributional analysis of agentive suffixes**

In this section, we present a distributional semantics analysis of the rivalry between agentive suffixes in French. Our aim is to determine (i) whether ANs of each considered construction form a semantically coherent class, and (ii) whether agentive suffixes differ with respect to the semantic properties of the ANs they form. We use distributional semantics because we intend to highlight general tendencies and hierarchize representative and non-representative members of a given class. We first present the theoretical groundings on which we build our distributional setup and the samples of ANs used in the study. Then we outline the different ANs' properties by means of various distributional clues.

#### **3.1. Distributional hypothesis and methodology**

Distributional semantics is based on the distributional hypothesis (Harris 1954, Firth 1957) which states that the meaning of words can be approximated through their distribution. A difference in meaning correlates with a difference in context selection, and the semantic proximity of two words can be evaluated through the similarity of the contexts they appear in (Lenci 2018, Boleda 2020).

Tools such as Word2vec (Mikolov et al. 2013a) build on this assumption to provide a geometric representation of the meaning of the words in a given corpus. The meaning of a word is represented as a vector, also called word embedding, which consists of a set of hundreds of coordinates, computed based on the word distribution in a corpus. These coordinates, or dimensions, situate the vector in a vector space, called Distributional Semantic Model (DSM), in which the spatial proximity of the vectors is an indicator of the semantic similarity of the words they represent. This property allows for the quantification of the distributional and by extension semantic proximity between two words. The proximity in the DSM is evaluated by means of the cosine distance between two word embeddings, and is rendered by a score ranging from 0 for no proximity to 1 for strict equality of the embeddings. DSMs can also be queried to determine the neighbors of given word embeddings, i.e. the words whose embeddings are the nearest to that of some targeted words. As a result of their mathematical nature, word embeddings can moreover be manipulated and combined to approximate semantic compositionality (Mikolov et al. 2013b).

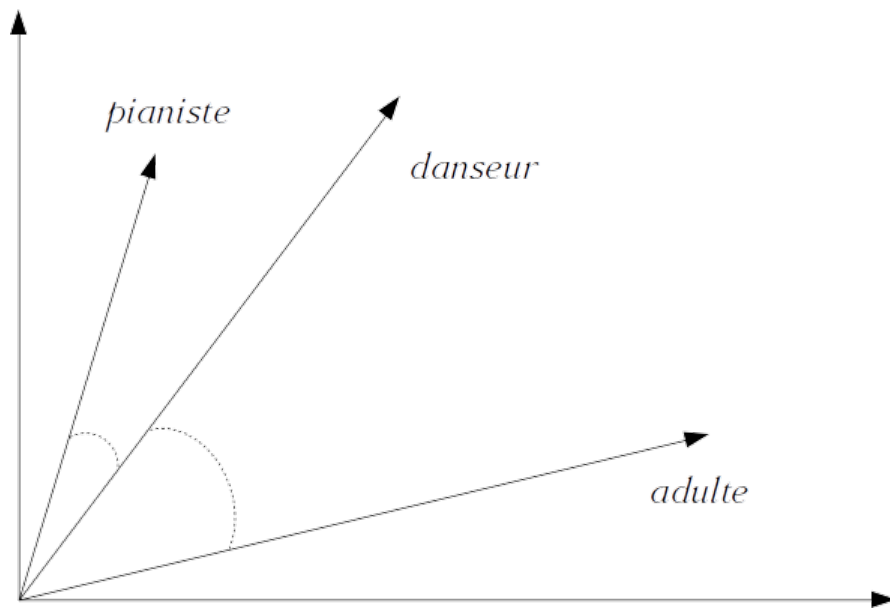


Fig. 1: Fictional two-dimensional vector space

Figure 1 illustrates the proximity relation between word embeddings. It presents a fictional two-dimensional vector space in which the meanings of the words *pianiste* ‘pianist’, *danseur* ‘dancer’ and *adulte* ‘adult’ are represented by three vectors. The proximity between each pair of words is computed as the cosine of the angle formed by their vectors (represented by the dotted lines). For instance, in a DSM computed from the French *Wikipedia 2018* corpus (as described below), the nouns *danseur* and *pianiste* have a proximity score of 0.721, whereas *danseur* and *adulte* have a proximity score of 0.336. The comparison of these scores indicates that *danseur* is semantically closer to *pianiste* than to *adulte*. We can moreover compare the neighbors of each AN: the nearest neighbors of *danseur* are *chorégraphe* ‘choreographer’, *acrobate* ‘acrobat’ and *ballerine* ‘ballerina’, whereas the nearest neighbors of *pianiste* are *violoniste* ‘violinist’, *violoncelliste* ‘cellist’ and *compositeur* ‘composer’. Finally, we could compute the vectors for phrases such as *danseur professionnel* ‘professional dancer’, as opposed to *danseur amateur* ‘amateur dancer’ by combining the vector of the noun *danseur* with the vectors of the adjectives *professionnel* and *amateur*. The properties of both phrases can thus be compared based on their word embeddings.

DSMs can provide an insight on each and every individual AN. Such a procedure does not however give an overall overview of the meaning attached to ANs as a semantic class. To further explore that meaning, we make use of the combinatorial potential of word embeddings. Following Kintsch (2001), Author (2020), and Author & coll. (2020), we hypothesize that the semantic components shared by the members of a

semantic class, such as agentivity in the case of ANs, can be approached through the distributional representation of its members. More precisely, we can compute a single unified distributional representation (called centroid) that aggregates the semantic properties of a subclass of ANs by averaging their word embeddings. We consider the centroid to be representative of the initial target words (also called seeds). In order to investigate the rivalry between agentive suffixes, we build a centroid for each class of ANs ending with a given suffix. We get an insight on their properties by analyzing their 100 nearest neighbors—that number being arbitrarily chosen.

The word embeddings and centroids are computed on DSMs built from the French *Wikipedia 2018* corpus. The corpus was lemmatized and POS-tagged beforehand using the *Talismane* parser (Urieli 2013). The use of a tagged corpus allows us to target only nominal acceptions of forms such as *accusateur*, which can be used as a noun (‘accuser’) or as an adjective (‘accusatory’). Although this method can induce some noise because of POS-tagging errors, it allows us to take into account a larger number of ANs, which would otherwise have to be excluded because of their ambiguity with an adjectival form. On account of the inherent instability of vector spaces (Pierrejean 2020), we average the results over 5 DSMs built with identical default parameters (CBOW architecture, frequency threshold of 5, window size of 5, negative sampling). This is achieved by computing a centroid within each of the 5 DSMs, and selecting the first 100 neighbors shared by the 5 resulting centroids.

### 3.2. Selecting samples of ANs

To perform the distributional analysis, we select samples of French ANs ending in *-aire*, *-ant/-ante*, *-eur/-euse/-rice*, *-ien/-ienne*, *-ier/-ière* and *-iste* (which we will refer to as *-aire*, *-ant*, *-eur*, *-ien*, *-ier*, *-iste* respectively). Two fundamental conditions are required for a noun to feature in one of these samples. First, it has to be a monosemous AN, or at least a monotype AN—nouns with multiple agentive readings being allowed. Monosemy is a crucial condition to avoid ambivalent distributional representations that could be induced by polysemy. Second, since our study focuses on the morphosemantic construction of agentivity, we strictly select ANs for which the agentive meaning can be analytically related to the morphological base. Nouns that can be lexically defined as agentive, but whose agentivity is not constructed through derivation, are dismissed. We define here a derived AN as any X noun that describes an entity as (actually or potentially) performing an action related to what is denoted by the base. Different relations are possible, and can be analyzed as:

- X performs Base;
- X creates/modifies/destroys Base;

- X manipulates/plays with Base;
- X practices/studies Base;
- X performs an action while located in Base;
- etc.

The *-aire*, *-ant*, *-eur*, *-ien*, *-ier*, *-iste* ANs candidates are automatically extracted from the *Lexique* database<sup>9</sup>, with complementary data from the *Lexeur* database<sup>10</sup> in the case of denominal *-eur* ANs. All extracted candidates are manually filtered according to the conditions described above, and to some additional criteria. We exclude from the lists:

- underived nouns (*peur* ‘fear’);
- nouns unanalyzable synchronically as agentively related to their base (*huissier* ‘bailiff’/*huis* ‘front door’);
- nouns whose last formation operation is not suffixation<sup>11</sup> (*télévangéliste* ‘televangelist’);
- nouns in *-euse* for which it cannot be decided whether they are the feminine of nouns ending in *-eur* or in *-eux* (*chatouilleuse* ‘ticklish’);
- nouns with at least one non-agentive sense, be it stative (*maigreur* ‘thinness’), demonymic (*Parisien* ‘Parisian’), phasal (*quadragénaire* ‘quadragenarian’), instrumental (*navigateur* ‘sailor/navigator’), psychological (*croyant* ‘believer’), benefactive (*cessionnaire* ‘transferee’), axiological (*fasciste* ‘fascist’), etc.;
- multifaceted or underspecified nouns with an agentive and non-agentive interpretation (e.g. the agent/member interpretation of *impressionniste* ‘impressionist’).

The methodology presented in Section 3.1 requires that only the nouns with at least 5 occurrences in the *Wikipedia* corpus be selected. The final list consists of 1252 ANs useable as seeds, including 27 *-aire* ANs, 86 *-ant* ANs, 717 *-eur* ANs, 54 *-ien* ANs, 161 *-ier* ANs, and 207 *-iste* ANs<sup>12</sup>. It can be noted that, although no necessary condition on the volitional and animate nature of agents was required (see Note 1), the

---

<sup>9</sup> <http://www.lexique.org/>

<sup>10</sup> The database will soon be online.

<sup>11</sup> We include in our list nouns with the same form as adjectives (e.g. *militant* ‘activist’, *intérimaire* ‘temporary worker’) as long as they can be analyzed as suffixed nouns, i.e. as long as (i) the corresponding suffix unequivocally form ANs (e.g. *assistant* ‘assistant’ from *assister* ‘assist’ and *diamantaire* ‘diamond merchant’ from *diamant* ‘diamond’ are not adjectives), and (ii) there is no morphological evidence that the nouns are deadjectival.

<sup>12</sup> These samples are clearly imbalanced, since all suffixes are not evenly represented. To the best of our knowledge, the impact of sample size has not been studied, nor shown to be detrimental in the context of distributional semantics analyses. We choose not to subsample our data, but favor a maximalist approach by including as many seeds as possible, so as to best cover the semantic spectrum of a category. Such an approach prevents any bias driven by a selection that would not be strictly criterion-based.

sampled ANs mostly denote humans, with the exception of a few nouns that denote animals (*rongeur* ‘rodent’) and a few underspecified nouns with regard to animacy (*catalyseur* ‘catalyst’). Agentive derivation thus appears to be often restricted to animate effectors. For instance, although causer subjects of a verb like *tuer* ‘kill’ are frequently inanimate (e.g. *La pandémie a tué beaucoup de gens* ‘The pandemic has killed many people’), the agentive nominalization *tueur* ‘killer’ (unlike the adjective *tueur* ‘killer’) almost exclusively denotes animate volitional causers<sup>13</sup>.

As a first step to evaluate the rivalry between the sampled ANs, we overview their morphosemantic properties. The POS distribution of their bases according to their suffix is presented in Table 2. The data are consistent with the theoretical selection identified in Table 1. Both verbal and nominal bases are selected by all suffixes, except in the case of *-ant* (strictly deverbal) and *-ien* (strictly denominal). Our data also show that the suffixes selecting both verbal and nominal bases exhibit a strong preference for one base POS: *-ier*, *-iste* and to a lesser extent *-aire* for nominal bases, and *-eur* for verbal bases.

	Verb	Noun
<i>-aire</i>	6 (22.2%)	21 (77.8%)
<i>-ant</i>	86 (100%)	-
<i>-eur</i>	681 (95.0%)	36 (5.0%)
<i>-ien</i>	-	54 (100%)
<i>-ier</i>	6 (3.7%)	155 (96.3%)
<i>-iste</i>	7 (3.4%)	200 (96.6%)

Tab. 2: Distribution of base POS in AN samples. The percentage per suffix (in rows) is given in brackets.

We can get an insight into the semantic specificities associated with each suffix by annotating the meaning of the bases. Six relevant semantic types are considered: action (*chasser* ‘hunt’ > *chasseur* ‘hunter’), object (*machine* ‘machine’ > *machiniste* ‘stagenhand’), domain (*histoire* ‘history’ > *historien* ‘historian’), property (*équilibre* ‘balance’ > *équilibriste* ‘tightrope walker’), institution (*poste* ‘postal service’ > *postier*

<sup>13</sup> As an indication, we annotated randomized samples of 200 tokens of the verb *tuer* ‘kill’ (used with overtly expressed subjects) and of the noun *tueur* ‘killer’ in the FRCOW16A corpus (<https://www.webcorpora.org/>). We found that 139 and 61 subjects of the verb were respectively animate and inanimate, whereas 199 occurrences of the noun denoted humans—the only exception referring to a virus in *ce tueur de sida* ‘this AIDS killer’.

‘postman’), and cognitive object (*pamphlet* ‘lampoon’ > *pamphlétaire* ‘lampoonist’). Verbal bases are analyzed as denoting either actions or properties, according to their dynamic or stative inner aspect. Nominal bases are more diverse, for they actualize the whole range of semantic types. Type assignment is supported by linguistic tests taken from the literature (Godard & Jayez 1996, Flaux & Van de Velde 2000, Author 2015, among others). Action-denoting nouns can be the subject of *avoir lieu* ‘take place’, *se produire* ‘happen’, or the object of *effectuer* ‘carry out’, *accomplir* ‘perform’, *procéder à* ‘proceed to’. Object-denoting nouns can be the subject of *se trouver* ‘be’ followed by a spatial locative. Domain-denoting nouns are compatible with the light verb construction *faire du N* ‘do some N’. Property-denoting nouns can be used in *être d’un grand N* ‘be of great N’, *état de N* ‘state of N’, or as the object of *ressentir* ‘feel’, *éprouver* ‘experience’, *faire preuve de* ‘show’. Institution-denoting nouns can be the subject of *être fondé* ‘be founded’ followed by a temporal locative, or can be used in expressions such as *être nommé à la tête du N* ‘be appointed head of N’. Cognitive-object-denoting nouns can be the object of *écrire* ‘write’, or followed by clauses introduced by *selon lequel* ‘according to which’<sup>14</sup>.

In case base words are polysemous, we annotate the meaning that most closely matches the derived noun, provided it fits an existing pattern of AN formation. For instance, we analyze *farceur* ‘joker’ as derived from the action meaning of *farce* (‘joke’) rather than from its object meaning (‘stuffing’), relying on semantic matching and on the fact that denominal *-eur* ANs can stem from action nouns (e.g. *bienfait* ‘good deed’ > *bienfaiteur* ‘benefactor’).

	Action	Object	Domain	Property	Institution	Cog. Obj.
<b>-aire</b>	<b>19 (70.4%)</b>	6 (22.2%)	-	-	-	2 (7.4%)
<b>-ant</b>	<b>86 (100%)</b>	-	-	-	-	-
<b>-eur</b>	<b>686 (95.7%)</b>	12 (1.7%)	18 (2.5%)	-	-	1 (0.1%)
<b>-ien</b>	4 (7.4%)	2 (3.7%)	<b>46 (85.2%)</b>	1 (1.85%)	1 (1.85%)	-
<b>-ier</b>	21 (13%)	<b>127 (78.9%)</b>	4 (2.5%)	-	3 (1.9%)	6 (3.7%)
<b>-iste</b>	32 (15.5%)	42 (20.3%)	<b>105 (50.7%)</b>	3 (1.4%)	-	25 (12.1%)

<sup>14</sup> The class of cognitive objects includes nouns that have a semantic facet corresponding to the denotation of a physical object (e.g. *roman* ‘novel’), and thus can be the subject of *se trouver* ‘be’ followed by a spatial locative. Such nouns were annotated as denoting cognitive objects rather than just objects, because of the prominence and distinctiveness of the cognitive facet.

Tab. 3: Distribution of base semantic types in AN samples. The percentage per suffix (in rows) is given in brackets. Highest values per suffix (in rows) are indicated in bold.

Table 3 presents for each suffix the semantic type of the selected bases. It shows that *-ant*, *-eur* and to a lesser extent *-aire* strongly privilege action-denoting bases, whereas *-ien* and *-ier* mostly select respectively domain- and object-denoting bases. The *-iste* suffix is more heterogeneous, as it selects bases with a larger variety of semantic types, still with a predilection for domain-denoting bases.

### 3.3. Distributional data

Distributional semantics allows for both individual and global analyses of AN subclasses. We compare agentive suffixes by analyzing the distributional behavior of the different ANs, and various morphosemantic properties of their centroids' neighbors in the DSMs.

#### 3.3.1. Density of ANs in vector space

Figure 2 presents for each suffix the distribution of ANs according to their average proximity to other ANs ending with the same suffix. It gives an overview of the ANs' density in the vector space, which can be interpreted as their semantic cohesion. The more semantically consistent is a morphological subclass of ANs, the closer its members will be in the DSMs.

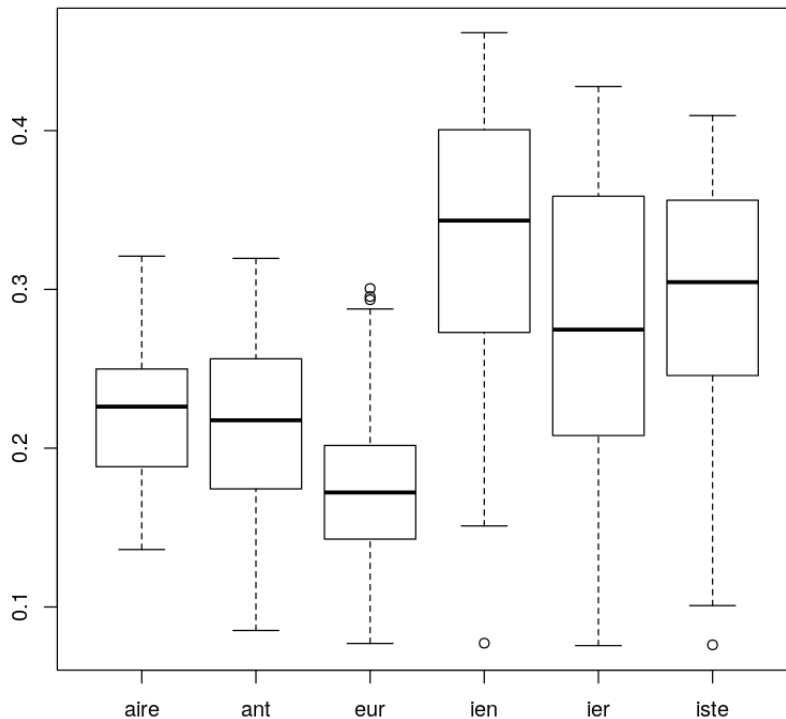


Fig. 2: Distributional density of ANs

It appears that ANs ending in *-ien*, *-ier* and *-iste* tend to be closer to ANs with the same suffix than ANs ending in *-aire*, *-ant* and *-eur*. Median proximity, indicated by bold lines in Figure 2, ranges from 0.275 to 0.343 for the former, and from 0.175 to 0.218 for the latter. The application of the Wilcoxon Mann-Whitney test indicates that the density difference between any two suffixes is statistically significant at  $p < .05$ , except in the case of *-aire* vs. *-ant* ( $p = 0.3869$ ), and of *-ier* vs. *-iste* ( $p = 0.0633$ ). These results suggest that *-aire*, *-ant* and *-eur* derive semantically more diverse ANs than *-ien*, *-ier* and *-iste*. In this respect, *-eur* forms the most heterogeneous category, whereas *-ien* forms the most homogeneous category. It has to be noted though that the deviation shown by *-aire*, *-ant* and *-eur* ANs is lower than that of *-ien*, *-ier* and *-iste*. In the figure, the top and bottom parts of the boxes, corresponding respectively to the 3rd and 1st quartiles, i.e. to the 25% of data above and below the median, are more extended in the case of *-ien*, *-ier*, *-iste* than in the case of *-aire*, *-ant* and *-eur*. This indicates that while being overall closer, ANs ending in *-ien*, *-ier* and *-iste* display more disparities. One explanation may be that semantic subgroups or multiple idiosyncrasies can be found within these three classes.

A final remark with respect to Figure 2 concerns the outliers, represented as dots, i.e. the nouns that differ significantly from the other nouns with the same suffix. Such is the case of *aoûtien* ‘August vacationer’ whose average proximity score to all other *-ien* seeds is 0.077, and *finaliste* ‘finalist’ whose average proximity score to all other *-iste* seeds is 0.076.

### 3.3.2. Centroids’ neighborhood

The distributional analysis of ANs provides information on individual ANs, but does not offer an overall semantic description of the categories they form. In order to capture the semantic specificities of agentive suffixes, we build a unified distributional representation for each of them under the form of a centroid vector, as described in Section 3.1. We compute a centroid per suffix in each DSM, by averaging the ANs’ vectors, and then analyze their 100 nearest neighbors over the 5 models. A sample of the 10 nearest neighbors for each suffix is given in Table 4.

<i>-aire</i>	<i>-ant</i>	<i>-eur</i>	<i>-ien</i>	<i>-ier</i>	<i>-iste</i>
médecin ‘doctor’	détenu ‘prisoner’	plombier ‘plumber’	physicien ‘physicist’	cordonnier ‘shoemaker’	ophtalmologue ‘ophtalmologist’



comptable 'accountant'	citoyen 'citizen'	truand 'crook'	neurologue 'neurologist'	boulangier 'baker'	photographe 'photographer'
collaborateur 'collaborator'	gens 'people'	escroc 'crook'	linguiste 'linguist'	armurier 'gunsmith'	pédagogue 'pedagogue'
bibliothécaire 'librarian'	délinquant 'delinquent'	proxénète 'pimp'	psychiatre 'psychiatrist'	maréchal-ferrant 'farrier'	neurologue 'neurologist'
juriste 'jurist'	personne 'person'	rabatteur 'beater/tout'	mathématicien 'mathematician'	bijoutier 'jeweler'	clarinettiste 'clarinetist'
financier 'financier'	villageois 'villager'	prestidigitateur 'magician'	informaticien 'IT engineer'	serrurier 'locksmith'	pédiatre 'pediatrician'
informateur 'informer'	employé 'employee'	coiffeur 'hairstylist'	statisticien 'statistician'	épiciier 'grocer'	violoniste 'violinist'
avocat 'lawyer'	travailleur 'worker'	garagiste 'mechanic'	biologiste 'biologist'	perruquier 'wigmaker'	mycologue 'mycologist'
savant 'scholar'	wikipédiens 'Wikipedians'	gangster 'gangster'	pédiatre 'pediatrician'	charpentier 'carpenter'	naturopathe 'naturopath'
lobbyiste 'lobbyist'	comploter 'conspirator'	cuisinier 'cook'	anthropologue 'anthropologist'	forgeron 'blacksmith'	ornithologue 'ornithologist'

Tab. 4: 10 nearest neighbors of ANs' centroids

Some seeds used to build the centroids are found among the 100 nearest neighbors of each centroid. Yet the overlap between seeds and neighbors varies depending on the centroids. For instance, 28 of the 100 nearest neighbors of the *-eur* centroid are *-eur* seeds. Likewise, respectively 40 and 37 seeds are included among the 100 nearest neighbors of the *-ier* and *-iste* centroids. However the overlap is lower for *-aire*, *-ant* and *-ien* centroids, with respectively 6, 11 and 13 seeds among the 100 nearest neighbors. These differences tend to confirm that some suffixes form semantically more homogeneous categories, as they seem more likely to cluster with each other in vector space models.

Note that the presence of seeds in a centroid's close neighborhood is not necessarily expected from centroid computing. First, the overlap depends on the targeted suffix. If it was merely due to the

experimental procedure, one could expect the overlap to be roughly similar across suffixes. Its variation in our data suggests that the overlap results from a distributional property of the considered ANs. Second, centroids built from vectors randomly selected in a DSM do not exhibit such an overlap, as shown by Author (2020). The overlap correlates with the semantic homogeneity of the seeds. To find seeds in the neighborhood of a centroid is therefore significant, and contributes to its description. Consequently, when analyzing the neighborhood of a given centroid in order to identify its core semantic properties, it is necessary to take into account the presence of seeds so as not to bias the analysis.

A first observation about the 600 neighbors considered here is that they are almost exclusively nouns<sup>15</sup>, and that they all denote humans—except for *chien* ‘dog’ found in the *-eur* neighborhood at rank 76. These neighbors are quite diverse morphologically. They can be affixed nouns (*voleur* ‘thief’), converted nouns (*émigré* ‘emigrant’), compounds (*maréchal-ferrant* ‘farrier’), morphologically simple nouns (*médecin* ‘doctor’), opaque nouns (*écrivain* ‘writer’), and extra-grammatically formed nouns (e.g. by clipping, *indicateur* > *indic* ‘informer’)<sup>16</sup>. The distribution of the centroids’ neighbors according to their morphological type is given in Table 5.

Neighbors of	Affixed	Converted	Compound	Simple	Complex	Extragram.	Indet.
<b>-aire</b>	57	16	6	12	8	-	1
<b>-ant</b>	48	20	-	14	7	1	10
<b>-eur</b>	66	7	3	10	8	1	5
<b>-ien</b>	81	-	14	2	2	-	1
<b>-ier</b>	78	5	3	1	11	-	2
<b>-iste</b>	87	-	10	-	3	-	-

<sup>15</sup> Two neighbors included in the *-ier* neighborhood, *fermier* ‘farmer’ and *chaudronnier* ‘boilermaker’, are actually tagged as adjectives by the parser, but these may be tagging errors—especially in the case of *chaudronnier*.

<sup>16</sup> Our morphological analysis is based on synchrony, and does not take into account historical word-formation. In case of multiple formation, the annotation only applies to the last morphological operation. For instance, *couturier* ‘fashion designer’ is analyzed as a denominal noun suffixed with *-ier* and based on *couture* ‘fashion design’, even though *couture* is itself derived from *coudre* ‘sew’. An “indeterminate” label is used to account for words that are related to others by means of conversion, in case the direction of conversion cannot be decided on a morphological basis (e.g. for the noun *assassin* ‘murderer’ in relation to the verb *assassiner* ‘murder’, or the noun *complice* ‘accomplice’ in relation to the adjective *complice* ‘conniving’). “Indeterminate” neighbors are thus indeterminate between converted and morphologically simple nouns.

Tab. 5: Morphological type of ANs' neighbors (/100). Highest values per suffix (in rows) are indicated in bold.

ANs' neighbors tend to be affixed nouns. This is particularly the case for *-ien*, *-ier* and *-iste*, for which at least 78 of the 100 nearest neighbors are affixed. The different centroids' neighborhoods do not display the same morphological diversity. The neighbors of *-iste*, *-ien* and to a lesser extent *-ier* fall under fewer morphological types, or are more concentrated as regards morphological type, than the neighbors of *-aire*, *-ant* and *-eur*. Moreover, morphological types are not evenly distributed across suffixes. Compound nouns are mostly found among *-ien* and *-iste* neighbors, whereas simple nouns are mostly present in the *-aire*, *-ant* and *-eur* neighborhood, and converted nouns in the *-aire* and *-ant* neighborhood. Morphologically indeterminate (i.e. possibly converted) items are mostly *-ant* neighbors.

Table 6 indicates the distribution of the derived neighbors' suffixes according to the centroid. Prefixed neighbors are not taken into account because, unlike suffixes, prefixes are not considered to be agent-forming devices. Indeed, prefixed neighbors that can be regarded as ANs are themselves derived from AN bases (e.g. *parapsychologue* 'parapsychologist' is derived from *psychologue* 'psychologist'). Prefixed neighbors are marginal in our data (10 out of 600 neighbors).

Neighbors of	<i>-ain</i>	<i>-aire</i>	<i>-ant</i>	<i>-ard</i>	<i>-eur</i>	<i>-ien</i>	<i>-ier</i>	<i>-iste</i>	<i>-logue</i>	<i>-ois</i>	<i>-on</i>
<i>-aire</i>	-	7 (12.1%)	4 (6.9%)	-	<b>19</b> (32.8%)	7 (12.1%)	8 (13.8%)	13 (22.4%)	-	-	-
<i>-ant</i>	1 (2.1%)	1 (2.1%)	<b>16</b> (34.0%)	-	12 (25.5%)	7 (14.9%)	6 (12.8%)	3 (6.4%)	-	1 (2.1%)	-
<i>-eur</i>	-	1 (1.5%)	-	3 (4.6%)	<b>33</b> (50.8%)	2 (3.1%)	17 (26.1%)	7 (10.8%)	-	-	2 (3.1%)
<i>-ien</i>	-	1 (1.3%)	-	-	3 (3.8%)	17 (21.5%)	-	28 (35.4%)	<b>30</b> (38.0%)	-	-
<i>-ier</i>	-	-	2	-	17	1	<b>52</b>	5	-	-	2

			(2.5%)		(21.6%)	(1.3%)	<b>(65.8%)</b>	(6.3%)			(2.5%)
<b>-iste</b>	-	1 (1.2%)	-	-	4 (4.7%)	10 (11.6%)	-	<b>45 (52.3%)</b>	26 (30.2%)	-	-

Tab. 6: Suffix of ANs' derived neighbors. The percentage per suffix (in rows) is given in brackets. Highest values (in rows) are indicated in bold.

Many suffixes are listed, but they are not equally represented. Suffixes such as *-ain*, *-ois*, *-on* or *-ard* have less than 5 occurrences among the 600 neighbors. These suffixes can infrequently form ANs, such as *-on* in *forgeron* 'blacksmith' and *-ard* in *motard* 'biker', or in no way be agentive and then feature idiosyncratically among neighbors, as in the case of *-ois* in *villageois* 'villager' and *-ain* in *Africains* 'Africans'. Interestingly, a significant portion of nouns ending in *-logue* can be found in both the *-ien* and *-iste* neighborhoods, *-logue* nouns being known to denote specialists (Amiot & Dal 2007, Villoing & Namer 2012, Lasserre & Montermini 2014). In our samples, they represent respectively 38.0% and 30.2% of the suffixed neighbors of the *-ien* and *-iste* centroids.

The distribution of the centroids' derived neighbors with respect to their base POS is presented in Table 7. Derived neighbors are mostly denominal and deverbal. Adjectival bases are marginal, and mostly found in the *-aire*, *-ant* and *-eur* neighborhood. The balance between deverbal and denominal neighbors varies according to the centroid, with some clear predilections. Neighbors of *-ien* and *-iste* are predominantly derived from nouns, whereas the highest proportion of verbal bases is found in the *-ant* neighborhood..

Neighbors of	Verb	Noun	Adjective
<b>-aire</b>	28 (38.4%)	<b>37 (50.7%)</b>	8 (10.9%)
<b>-ant</b>	<b>43 (63.2%)</b>	19 (27.9%)	6 (8.8%)
<b>-eur</b>	<b>38 (52.1%)</b>	30 (41.1%)	5 (6.8%)
<b>-ien</b>	3 (3.7%)	<b>78 (96.3%)</b>	-
<b>-ier</b>	26 (31.3%)	<b>55 (66.3%)</b>	2 (2.4%)
<b>-iste</b>	4 (4.6%)	<b>82 (94.3%)</b>	1 (1.1%)

Tab. 7: Base POS of ANs' derived neighbors. The percentage per suffix (in rows) is given in brackets. Highest values per suffix (in rows) are indicated in bold.

Table 8 indicates the distribution of the centroids' derived neighbors according to the semantic type of their base. Semantic types are analyzed following the criteria presented in Section 3.2. The analysis shows that action, object and domain are prominent as base semantic types. Property, institution and cognitive object are rather marginal—except for property in the *-ant* neighborhood. Important disparities can be observed between the suffixes. Some centroids clearly privilege one base type, as is the case of *-eur* with action, *-ien* and *-iste* with domain, and *-ier* with object. In contrast, the other centroids are more heterogeneous, especially *-aire*. Furthermore, a similar distribution of base semantic types can be observed between *-ien* and *-iste* neighbors, although only the latter ones include words derived from cognitive-object-denoting bases.

Neighbors of	Action	Object	Domain	Property	Institution	Cog. Obj.
<i>-aire</i>	<b>33 (45.2%)</b>	12 (16.4%)	12 (16.4%)	9 (12.3%)	3 (4.1%)	4 (5.5%)
<i>-ant</i>	<b>41 (60.3%)</b>	8 (11.8%)	-	13 (19.1%)	4 (5.9%)	2 (2.9%)
<i>-eur</i>	<b>45 (61.6%)</b>	16 (21.9%)	5 (6.8%)	5 (6.8%)	2 (2.7%)	-
<i>-ien</i>	4 (4.9%)	26 (32.1%)	<b>45 (55.6%)</b>	3 (3.7%)	3 (3.7%)	-
<i>-ier</i>	28 (33.7%)	<b>49 (59.0%)</b>	3 (3.6%)	2 (2.4%)	1 (1.2%)	-
<i>-iste</i>	5 (5.7%)	26 (29.9%)	<b>47 (54.0%)</b>	3 (3.4%)	2 (2.3%)	4 (4.6%)

Tab. 8: Semantic base of ANs' derived neighbors. The percentage per suffix (in rows) is given in brackets. Highest values per suffix (in rows) are indicated in bold.

### 3.3.3. Comparing centroids

Many of our observations suggest that some centroids are more similar than others. To further explore these similarities, we directly compare the centroids by analyzing their neighborhood overlap and vector proximity. The amount of shared neighbors among each centroid's 100 nearest neighbors is given in Table 9.

	<i>-aire</i>	<i>-ant</i>	<i>-eur</i>	<i>-ien</i>	<i>-ier</i>	<i>-iste</i>
<i>-aire</i>	-	<b>16</b>	15	12	12	4
<i>-ant</i>	<b>16</b>	-	5	0	4	0
<i>-eur</i>	15	5	-	4	<b>23</b>	3
<i>-ien</i>	12	0	4	-	1	<b>62</b>
<i>-ier</i>	12	4	<b>23</b>	1	-	1
<i>-iste</i>	4	0	3	<b>62</b>	1	-

Tab. 9: Neighbor overlap. Highest values per suffix (in rows) are indicated in bold.

The different centroids exhibit various amounts of overlapping neighbors. For instance, *-ien* and *-iste* share many neighbors, which reveals a high proximity between the two suffixes. To a lesser extent, the 23 neighbors shared by *-eur* and *-ier* suggest some distributional similarities between these suffixes. On the other hand, the *-aire* centroid's overlapping neighbors distribute rather evenly among other neighborhoods, which indicates that the *-aire* centroid is not specifically close to any of the other centroids.

Table 10 gives the average proximity scores between centroids across the 5 DSMs. These scores confirm the high proximity previously observed between *-ien* and *-iste*, and to a lesser extent between *-eur* and *-ier*. The (quite high) proximity scores displayed by *-aire* also confirm that the *-aire* centroid is more or less equidistant to other centroids. In contrast, the *-ant* centroid appears to be distant from some other centroids, especially from both the *-ien* and *-iste* centroids.

	<i>-aire</i>	<i>-ant</i>	<i>-eur</i>	<i>-ien</i>	<i>-ier</i>	<i>-iste</i>
<i>-aire</i>	-	0.762	<b>0.780</b>	0.732	0.744	0.715
<i>-ant</i>	0.762	-	<b>0.772</b>	0.534	0.656	0.502
<i>-eur</i>	0.780	0.772	-	0.745	<b>0.842</b>	0.773
<i>-ien</i>	0.732	0.534	0.745	-	0.677	<b>0.935</b>
<i>-ier</i>	0.744	0.656	<b>0.842</b>	0.677	-	0.716

<b>-iste</b>	0.715	0.502	0.773	<b>0.935</b>	0.716	-
--------------	-------	-------	-------	--------------	-------	---

Tab. 10: Proximity scores between centroids. Highest values per centroid (in rows) are indicated in bold.

### 3.4. Discussion

The distributional analysis of seeds and centroids and the morphosemantic analysis of seeds and neighbors shed light on some semantic differences between French agentive suffixes. As far as distributional profiles are concerned, two groups of suffixes can be distinguished. In the case of *-ien*, *-ier* and *-iste*, clear correlations can be drawn between suffix distribution and agentive subtypes. In the case of *-aire*, *-ant* and *-eur*, specificities are more complex, although some tendencies emerge from the data. We further detail these observations in this subsection.

First, a strong convergence can be observed between *-iste* and *-ien*, confirming previous analyses (Dubois 1962, Lignon 2007). These two suffixes have similar distributional profiles, as reveal the high proximity between their centroids, the similar distribution of morphological types and base semantic types among their neighbors, and the high number of shared neighbors. Nominal bases prevail in both seeds and derived neighbors, and a distinctive majority of domain-denoting bases can be observed in both seeds and derived neighbors. ANs ending in *-ien* and *-iste* mostly denote domain experts and agents performing intellectual activities, which is confirmed by their proximity to *-logue* nouns. The high internal proximity between ANs ending with each suffix shows their semantic homogeneity and, combined with the distributional similarity between the two suffixes, corroborates their semantic distinctiveness from other agentive suffixes. A closer look at seeds and neighbors shows that *-iste* has the specific capacity to form artist-denoting nouns, many *-iste* seeds and exclusive neighbors being nouns such as those in (4a) and (4b).

(4) a. artiste ‘artist’, violoniste ‘violinist’, clarinettiste ‘clarinetist’, marionnettiste ‘puppeteer’, nouvelliste ‘short-story writer’, soliste ‘soloist’

b. compositeur ‘composer’ (39)<sup>17</sup>, cinéaste ‘moviemaker’ (43), , illustrateur ‘illustrator’ (45), peintre ‘painter’ (65), dessinateur ‘cartoonist’ (75), chorégraphe ‘choreographer’ (85)

<sup>17</sup> Numbers in brackets indicate the rank a noun occupies in the averaged neighborhood of a given suffix’s centroids.

This specificity can also be observed in the translation data presented by Cartoni et al. (2015) in their contrastive study of French and Italian agent nouns. Overall, the *-iste* suffix appears to be slightly more diverse than *-ien*, with more marginal cases (e.g. *automobiliste* ‘car driver’, *gréviste* ‘striker’, *perchiste* ‘pole vaulter’, *projectionniste* ‘projectionist’). In contrast, *-ien* seems more homogeneous, with less semantic variations in base selection, higher distributional density, and exclusive neighbors strictly denoting scientists.

The *-ier* suffix shares a number of properties with *-ien* and *-iste*, including a rather high proximity between seeds and the predilection for nominal bases. However a distinctive predilection for object-denoting bases can be observed in both *-ier* seeds and derived neighbors. ANs ending in *-ier* differ from other ANs in that they mostly denote occupations in crafts or commerce, and correlatively often derive from nouns that denote artifacts, as already noted by Roché (2004). Nouns such as those in (5) are exclusive *-ier* neighbors.

(5) maréchal-ferrant ‘farrier’ (4), tanneur ‘tanner’ (11), aubergiste ‘innkeeper’ (16), tailleur ‘tailor’ (32), tisserand ‘weaver’ (34), palefrenier ‘ostler’ (56), marchand ‘merchant’ (59), orfèvre ‘goldsmith’ (65), apothicaire ‘apothecary’ (68), bûcheron ‘lumberjack’ (79)

The morphosemantic homogeneity of the neighbors, and the high proportion of seeds among the centroid’s nearest neighbors confirm both the consistency and the distinctiveness of the suffix. Some connection to *-eur* can still be observed through centroids’ proximity. It could be due to the fact that some deverbal *-eur* ANs denote manual workers or sellers (e.g. *carreleur* ‘tiler’, *ferrailleur* ‘scrap merchant’, *fraiseur* ‘milling machine operator’, *vendeur* ‘seller’), as noted by Cartoni et al. (2015).

As for *-aire*, *-ant* and *-eur*, they are semantically more versatile than *-iste*, *-ier* and *-ien*. ANs show a greater distributional dispersion, corresponding to more semantic disparities. The centroids’ neighbors are more diverse with respect to their morphological type, as well as to their suffix when derived. In particular, the *-aire*, *-ant* and *-eur* centroids have more simple nouns in their neighborhood than the *-iste*, *-ier* and *-ien* centroids, and simple nouns being morphosemantically unconstrained, they might display a larger range of agent types. Furthermore, derived neighbors stem from more heterogeneous semantic bases in the case of *aire*, *-ant*, *-eur* than in the case of *-iste*, *-ier*, *-ien*, and the bases of seeds and derived neighbors are more dissimilar with respect to POS distribution. For instance, whereas *-eur* seeds are mostly deverbal nouns and *-iste* seeds mostly denominal nouns, the *-eur* centroid’s derived neighbors are based on more diverse POS than the *-iste* ones, which is congruent with greater semantic dispersion.



Another distinctive property of *-aire*, *-ant* and *-eur* is the prominence of action-denoting bases in both seeds and derived neighbors. One could argue that this property is related to agentive versatility, and thus to semantic dispersion. Indeed, action bases allow episodic action features to be semantic constituents of derived words: ANs based on action-denoting words can denote agents in a particular event (i.e. occasional agents), and not exclusively agents with an occupational status (i.e. functional agents).

Despite these common properties, some differences between *-aire*, *-ant* and *-eur* can be observed. The *-ant* centroid is more distant from the *-ien*, *-ier* and *-iste* centroids than are the *-aire* and *-eur* centroids. The *-ant* centroid also shares less neighbors with other centroids than other centroids do. In addition, many ANs ending in *-ant* such as (6a), and many exclusive *-ant* neighbors such as (6b) are occasional ANs.

(6) a. arrivant ‘arrival’, assaillant ‘assailant’, exécutant ‘executor’, intervenant ‘speaker’, plaignant ‘complainant’, poursuivant ‘pursuer’, requérant ‘claimant’, votant ‘voter’

b. client ‘customer’ (18), émeutier ‘rioter’ (27), réfugié ‘refugee’ (41), agresseur ‘attacker’ (78), ravisseur ‘abductor’ (82), gréviste ‘striker’ (83), mutin ‘mutineer’ (89), contributeur ‘contributor’ (91)

The formation of occasional ANs is certainly not the prerogative of *-ant*, since *-eur* derives occasional ANs as well. Furthermore, ANs ending in *-ant* such as *commerçant* ‘shopkeeper’ and *enseignant* ‘teacher’ are functional ANs. Nevertheless, occasional ANs are particularly salient in the distributional analysis of *-ant* ANs. This could be related to the fact that *-ant* ANs are exclusively derived from verbs, which possibly increases the proportion of occasional ANs.

It can be noted that the *-ant* centroid includes more non-ANs in its near neighborhood than other centroids do<sup>18</sup>. For example, *personne* ‘person’ (5), *villageois* ‘villager’ (6), *proche* ‘relative’ (14), *patient* ‘patient’ (30), *jeune* ‘youngster’ (55), *chrétien* ‘Christian’ (71), *enfant* ‘child’ (90), *nécessiteux* ‘needy’ (94), *déporté* ‘deportee’ (97) are exclusive neighbors of *-ant*. This particularity suggests that *-ant* ANs do not denote prototypical agents, which might be correlated with the description of occasional agentive subtypes, as opposed to functional ones—functional ANs being more abundant and more homogeneous in their linguistic distribution than occasional ANs (Author 2020).

The *-eur* suffix appears to be both the most productive and the most versatile agentive suffix. ANs ending in *-eur* are the most heterogeneous ones, *-eur* ANs being more dispersed in DSMs than other ANs. ANs ending in *-eur* can denote occasional (e.g. *narrateur* ‘narrator’, *expéditeur* ‘sender’, *vainqueur* ‘winner’),

<sup>18</sup> The number of words with a clearly non-agentive meaning among the 100 nearest neighbors of each centroid is: 5 for *-aire*, 32 for *-ant*, 2 for *-eur*, 1 for *-ien*, 0 for *-ier*, and 0 for *-iste*.

functional (e.g. *inspecteur* ‘inspector’, *coiffeur* ‘hairdresser’, *bruiteur* ‘sound engineer’) or behavioral agents (e.g. *râleur* ‘complainer’, *bagarreur* ‘fighter’, *séducteur* ‘seducer’). In the latter case, they describe a tendency to act in a certain way (Author & coll. 2015), which can be tested through the compatibility with size adjectives in a non-intersective interpretation (Anscombe 2001, Roy & Soare 2012), as in (7a) vs. (7b).

- (7) a. un gros râleur, un gros bagarreur  
       ‘a big complainer’, ‘a big fighter’ = somebody who complains/fights a lot  
       b. #un gros inspecteur, #un gros narrateur  
       ‘a big inspector’, ‘a big narrator’ ≠ somebody who inspects/narrates a lot

The versatility of *-eur* appears in its centroid’s neighbors, which include shared (e.g. *bijoutier* ‘jeweler’, *dentiste* ‘dentist’) or exclusive (e.g. *coursier* ‘courier’, *pompiste* ‘gas station attendant’) functional ANs, shared (e.g. *criminel* ‘criminal’, *délateur* ‘informer’) or exclusive (e.g. *assassin* ‘murderer’, *racketteur* ‘racketeer’) occasional ANs, and exclusive behavioral ANs such as those in (8).

- (8) voyou ‘rascal’ (18), drogué ‘drug addict’ (36), ivrogne ‘drunkard’ (44), fêtard ‘partygoer’ (50), arriviste ‘upstart’ (62), sadique ‘sadist’ (99)

From all our observations, *-eur* ANs denote behavioral agents more frequently than other suffixed ANs. The *-aire* suffix favors nominal bases, whereas *-ant* selects exclusively verbs, and *-eur* mostly verbs. The bases of the *-aire* derived neighbors are semantically more heterogeneous than those of *-ant* and *-eur* neighbors. Action as a semantic base is dominant, but not as much as in the case of *-ant* and *-eur*. ANs ending in *-aire* are also the smallest group of ANs, and as mentioned earlier, the suffix seems quite unproductive as an agent-forming device. Although some *-aire* ANs are occasional (e.g. *commanditaire* ‘backer’, *contestataire* ‘protester’, *expéditionnaire* ‘forwarding agent’), that property is not salient in the centroid’s neighborhood, unlike what is observed for *-ant*, and to a lesser extent *-eur*. In contrast with *-iste*, *-ien* and *-ier*, *-aire* forms functional ANs without any apparent predilection for a given type of occupation. Therefore the distinctive quality of *-aire* is to tend to non-specialized status characterization, while selecting action nouns as morphological bases.

#### 4. Agentive subtypes

In this section, we investigate agentive suffix rivalry in a bottom-up approach, through the clustering of AN embeddings. The semantic consistency of distributional clusters and the correlation with morphological properties are scrutinized. We first present our methodology, then provide an overview of the clusters, and finally examine how these relate to the morphosemantic properties of ANs.

#### 4.1. Clustering methodology

Our first intention is to identify distributionally relevant clusters of ANs by using clustering methods that do not require any preliminary definition of a number of clusters. Different clustering methods are tested, using hierarchical clustering, tSNE reduction, and HDBSCAN, but none of these methods provide convincing results as regards both the number of clusters derived from the analysis and the consistency of the data. For instance, no clear clusters emerge from tSNE, and hierarchical clustering and HDBSCAN either exclude almost half of the data, or gather most of the ANs into two clusters, one of which including more than 90% of the ANs. Such methods do not allow for any relevant analysis with respect to the semantic organization of the AN class.

Eventually the most fruitful clustering method proves to be the spherical k-means<sup>19</sup>, based on the default hypothesis of a one-to-one relation between suffixes and agentive subtypes—the default hypothesis serving as a basis for evaluating possible discrepancies from an ideal one-to-one relation. Spherical k-means is an unsupervised learning algorithm that aims to partition items into  $k$  clusters so that members of one cluster are more similar to each other than to members of other clusters. The difference between spherical and classical k-means clustering algorithms lies in the metrics used. The former aims to minimize the cosine distance among the members of a given cluster, while the latter minimizes the euclidean distance between the members of a given cluster and its center. Because both Word2vec and the algorithm optimize the cosine similarity, we favor the cosine implementation of the clustering algorithm. Hence, following the default hypothesis, we compute the clustering of our 1252 AN embeddings with  $k$  set to 6.

As stated in Section 3.1, we study the distributional profile of ANs on the basis of 5 DSMs so as to stabilize the results. Because the AN clustering relies on the word embeddings themselves, it may vary from one DSM to another. Consequently, one clustering per DSM is carried out, and the stability of the resulting clusters is evaluated by means of the Rand index (Rand 1971), which quantifies the cluster agreement on a scale from 0 (no agreement) to 1 (full agreement). The average agreement<sup>20</sup> between the 5 models is

---

<sup>19</sup> The clustering was performed using the ‘skmeans’ package in R, with the *pclust* method.

<sup>20</sup> The Rand index was computed using the `rand.index()` function of the R package ‘fossil’.

0.845 ( $\sigma = 0.015$ ), which indicates that the clustering is as stable as can be expected given the inherent instability of DSMs. In other words, ANs are clustered in a relatively similar fashion across the 5 DSMs. We therefore opt to analyze the clusters coming from one randomly selected DSM only, considering them as representative of the other DSMs' clusters.

The 1252 ANs used in the distributional analysis are grouped into 6 clusters whose size ranges from 149 to 263 items (C1: 149, C2: 254, C3: 263, C4: 255, C5: 179, C6: 152). In the next subsections, we briefly investigate which semantic subclasses of ANs are grouped into the different clusters, then elaborate on the relation between the clustering and the morphosemantic properties of ANs.

#### 4.2. Cluster overview

Being based on distributional semantics, AN clusters should coincide with semantic distinctions among ANs. They are expected to correspond, at least to some extent, to agentive subtypes. A detailed examination of the nouns included in the first cluster (C1) reveals that these mainly denote agents involved in non-physical actions, especially in relation to financial operations, as in (9).

(9) acheteur 'buyer', acquéreur 'buyer', assureur 'insurer', bienfaiteur 'benefactor', cotisant 'contributor', donateur 'sponsor', enchérisseur 'bidder', emprunteur 'borrower', épargnant 'saver', gestionnaire 'manager', investisseur 'investor', parieur 'gambler', payeur 'payer', prêteur 'money lender', souscripteur 'subscriber', revendeur 'reseller', spéculateur 'speculator'

More broadly, C1 includes many nouns that denote businessmen (*affairiste* 'wheeler-dealer', *annonceur* 'advertiser', *assureur* 'insurer', *décideur* 'decision-maker', *détaillant* 'retailer', *fournisseur* 'supplier', *exportateur* 'exporter', *promoteur* 'promoter', *repreneur* 'transferee', etc.). It can be noted that some ANs in (9) denote occasional agents, and that this distinctive property extends to other ANs in C1, such as those in (10).

(10) arrivant 'comer', communiant 'communicant', contrevenant 'offender', contributeur 'contributor', demandeur 'applicant', évaluateur 'assessor', expéditeur 'sender', émeutier 'rioter', gagnant 'winner', manifestant 'demonstrator', migrant 'migrant', participant 'participant', postulant 'candidate', requérant 'claimant', signataire 'signatory', visiteur 'visitor', votant 'voter'

Cluster 2 consists mainly of ANs such as those in (11), which denote craftsmen or generally speaking manual professions.

(11) bagagiste ‘baggage handler’, bricoleur ‘handyman’, cardeur ‘carder’, carreleur ‘tiler’, caissier ‘cashier’, couvreur ‘roofer’, cueilleur ‘picker’, déménageur ‘mover’, élagueur ‘trimmer’, ferblantier ‘tinsmith’, fraiseur ‘milling machine operator’, machiniste ‘machinist’, manutentionnaire ‘handler’, réparateur ‘repairer’, rempailleur ‘recaner’

C2 also includes many nouns referring to shopkeepers (12), possibly overlapping with the previous subclass (*chapelier* ‘milliner’, *chocolatier* ‘chocolate maker’, *horloger* ‘clockmaker’, *joaillier* ‘jeweler’).

(12) antiquaire ‘antique dealer’, armurier ‘gunsmith’, barbier ‘barber’, bouquiniste ‘second-hand book seller’, boutiquier ‘shopkeeper’, brocanteur ‘second-hand goods dealer’, cafetier ‘cafe owner’, caviste ‘wine merchant’, droguiste ‘hardware seller’, épicier ‘grocer’, fleuriste ‘florist’, libraire ‘bookseller’, opticien ‘optician’, pharmacien ‘pharmacist’, poissonnier ‘fishmonger’, quincailler ‘ironmonger’, tavernier ‘innkeeper’

The nouns grouped together in Cluster 3 denote agents engaged in intellectual or artistic activities, generically in the case of *chercheur* ‘researcher’, *concepteur* ‘designer’, *inventeur* ‘inventor’, *penseur* ‘thinker’, *artiste* ‘artist’, *créateur* ‘creator’, but more often with a specialization in a given domain. Nouns in C3 denote mostly:

- scientists (*anthropologiste* ‘anthropologist’, *astrophysicien* ‘astrophysicist’, *bactériologiste* ‘bacteriologist’, *biochimiste* ‘biochemist’, *botaniste* ‘botanist’, etc.),
- doctors (*anesthésiste* ‘anaesthetist’, *chirurgien* ‘surgeon’, *chiropraticien* ‘chiropractor’, *clinicien* ‘clinician’, *dermatologiste* ‘dermatologist’, etc.),
- graphical artists (*aquarelliste* ‘watercolorist’, *caricaturiste* ‘caricaturist’, *coloriste* ‘colorist’, *dessinateur* ‘cartoonist’, *graphiste* ‘graphic designer’, etc.),
- musicians (*accordéoniste* ‘accordionist’, *bassiste* ‘bass player’, *clarinettiste* ‘clarinetist’, *clavéciniste* ‘harpsichordist’, *concertiste* ‘concertist’, etc.),
- writers or journalists (*apologiste* ‘apologist’, *chroniqueur* ‘columnist’, *éditorialiste* ‘editorialist’, *essayiste* ‘essayist’, *fabuliste* ‘fabulist’, etc.).

ANs in Cluster 4 denote agents who perform actions that directly affect or involve other human beings. These actions can be oriented towards social relations (e.g. *entremetteur* ‘matchmaker’, *intercesseur* ‘intercessor’, *marieur* ‘matchmaker’, *pacificateur* ‘peacemaker’, *suborneur* ‘surborner’) or psychological purposes (e.g. *amuseur* ‘entertainer’, *consolateur* ‘comforter’, *envoûteur* ‘bewitcher’, *tentateur* ‘tempter’, *trompeur* ‘deceiver’). Many ANs referring to harmful actions are included, such as those in (13).

(13) *abuseur* ‘abuser’, *agresseur* ‘attacker’, *écorcheur* ‘skinner’, *empoisonneur* ‘poisoner’, *éventreur* ‘ripper’, *kidnappeur* ‘kidnapper’, *meurtrier* ‘murderer’, *oppresseur* ‘oppressor’, *racketteur* ‘racketeer’, *ravisser* ‘abductor’, *tortionnaire* ‘torturer’, *tourmenteur* ‘tormentor’, *tueur* ‘killer’, *violeur* ‘rapist’

C4 also comprises ANs denoting agents in communication processes (e.g. *causeur* ‘chatterbox’, *contradictueur* ‘detractor’, *diseur* ‘chatty person’, *écoutant* ‘listener’, *insulteur* ‘insulter’, *parleur* ‘talker’, *prêcheur* ‘preacher’, *questionneur* ‘questioner’, *raconteur* ‘storyteller’), and behavioral ANs such as those in (14), which in some cases are compatible with the description of social interaction.

(14) *arriviste* ‘upstart’, *boudeur* ‘grouch’, *chahuteur* ‘heckler’, *emmerdeur* ‘troublemaker (familiar)’, *farceur* ‘joker’, *gaffeur* ‘blunderer’, *glandeur* ‘lazybones’, *flatteur* ‘flatterer’, *magouilleur* ‘schemer’, *manipulateur* ‘manipulator’, *menteur* ‘liar’, *moqueur* ‘tease’, *noceur* ‘reveler’, *pleurnicheur* ‘weeper’, *radoteur* ‘dotard’, *raisonneur* ‘reasoner’, *râleur* ‘complainer’, *rêveur* ‘dreamer’, *viveur* ‘reveler’

The semantic homogeneity of Clusters 5 and 6 is less clear than in the previous clusters. Both include small identifiable subgroups, but the coherence of the whole clusters is harder to grasp, and the clusters seem more structured by family resemblance than in the previous cases. One of the most distinctive features of C5 is that it includes a high proportion of the marked feminine seeds (42 out of 85, e.g. *allumeuse* ‘tease’, *baroudeuse* ‘adventurer’, *buveuse* ‘drinker’, *charmeuse* ‘charmer’, *cueilleuse* ‘picker’, *dormeuse* ‘sleeper’, *livreuse* ‘delivery woman’, *ouvreuse* ‘usherette’, *promeneuse* ‘stroller’). Many co-clustering masculine ANs are negatively connoted, and relate to seduction (*baratineur* ‘smooth talker’, *cavaleur* ‘tomcat’, *enjôleur* ‘wheedler’), sexual activity (*embrasseur* ‘kisser’, *baiseur* ‘fucker’, *fécondateur* ‘impregnator’, *trousseur* ‘womanizer’, *fouteur* ‘fucker’, *enculeur* ‘fucker’), and sexual harassment (*mateur* ‘voyeur’, *tripoteur* ‘groper’). Similarly to C4, C5 also includes a number of behavioral ANs, many of which are characterized by a feature of dishonesty (e.g. *arnaqueuse* ‘female scammer’, *bluffeur* ‘bluffer’, *carotteur* ‘swindler’, *crâneur* ‘show-off’, *embrouilleur* ‘muddler’, *feinteur* ‘faker’, *frimeur* ‘show-off’,

*lâcheur* ‘quitter’). As for nouns in C6, many of them indicate engagement in competition or confrontation. These nouns denote fighters (15a), military men (15b) and sportsmen (15c), with possible subclass overlaps.

(15) a. *attaquant* ‘assailant’, *bagarreur* ‘fighter’, *cogneur* ‘bruiser’, *combattant* ‘fighter’, *frappeur* ‘striker’, *flagellateur* ‘flagellant’, *joueur* ‘jouster’, *lutteur* ‘wrestler’

b. *archer* ‘archer’, *carabinier* ‘rifleman’, *commandant* ‘commander’, *démineur* ‘bomb disposal expert’, *missilier* ‘missile handler’, *mitrailleur* ‘machine-gunner’, *officier* ‘officer’, *piquier* ‘pikeman’, *tankiste* ‘tank driver’

c. *boxeur* ‘boxer’, *catcheur* ‘wrestler’, *cycliste* ‘cyclist’, *épéiste* ‘swordsman’, *footballeur* ‘football player’, *marathonien* ‘marathoner’, *patineur* ‘ice skater’, *planchiste* ‘windsurfer’, *pugiliste* ‘pugilist’, *skieur* ‘skier’, *surfeur* ‘surfer’

The class of agents denoted in C6 extends to sportsmen characterized by specific technical movements (e.g. *buteur* ‘striker’, *descendeur* ‘downhill racer’, *dribbleur* ‘dribbler’, *plaqueur* ‘tackler’, *pagayeur* ‘paddler’), and more broadly to agents defined by manners of motion (e.g. *coureur* ‘runner’, *fonceur* ‘speeder’, *grimpeur* ‘climber’, *marcheur* ‘walker’, *nageur* ‘swimmer’).

### 4.3. Exploring cluster predictors

A quantitative evaluation of the clusters can be provided by analyzing the morphosemantic properties of the ANs that are grouped together. We first examine the distribution into clusters of agentive suffixes, base POS and base semantic types. Then we use statistical methods to assess the influence of these properties on the clustering of rival agentive suffixes.

#### 4.3.1. Morphosemantic properties of ANs within clusters

Agentive suffixes differ in their distribution between the 6 clusters, as can be seen in Table 12. ANs ending in *-ant*, *-ien*, *-ier*, and *-iste* are mostly grouped into one given cluster, whereas ANs ending in *-aire* and *-eur* distribute more evenly. The majority of ANs ending in *-ien* and *-iste* fall under the same cluster (C3), which confirms the close similarity between these two suffixes. More than half of the *-ant* and *-ier* ANs cluster respectively in C1 and C2. These results suggest a certain semantic homogeneity and specialization of *-ant*, *-ien*, *-ier* and *-iste*. In contrast, the rather even distribution of *-eur* ANs, and to a lesser extent *-aire* ANs, confirms the versatility of both suffixes. It can still be noted that the most significant proportions of

*-aire* and *-eur* ANs are found within the same cluster (C4). The higher proportions of *-eur* ANs in C4 and C5 might be related to the fact that these clusters concentrate the behavioral ANs, and that *-eur* is the suffix that produces most behavioral ANs. Nevertheless, the fact that C5 and C6 are not especially favored by any given suffix shows that suffixes are not sufficient to account for the semantic consistency of these 2 clusters. An additional remark concerns the significant proportion of ANs ending in *-iste* in C2. These *-iste* ANs appear to be non-canonical insofar as they do not denote specialists but manual professions (*métallurgiste* ‘metallurgist’, *garagiste* ‘mechanic’, *chauffagiste* ‘heating engineer’) or shopkeepers (*aubergiste* ‘innkeeper’, *bouquiniste* ‘second-hand book seller’, *caviste* ‘wine merchant’), in line with the overall semantic characterization of C2.

	<i>-aire</i>	<i>-ant</i>	<i>-eur</i>	<i>-ien</i>	<i>-ier</i>	<i>-iste</i>
<b>C1 [Business]<sup>21</sup></b>	6 (22.2%)	<b>49 (57.0%)</b>	77 (10.8%)	2 (3.7%)	9 (5.6%)	6 (2.9%)
<b>C2 [Manual]</b>	5 (18.5%)	6 (7.0%)	115 (16%)	5 (9.3%)	<b>86 (53.4%)</b>	37 (17.9%)
<b>C3 [Intellectual]</b>	5 (18.5%)	2 (2.3%)	84 (11.7%)	<b>34 (62.9%)</b>	11 (6.8%)	<b>127 (61.3%)</b>
<b>C4 [Interpersonal]</b>	<b>7 (26.0%)</b>	17 (19.8%)	<b>198 (27.6%)</b>	6 (11.1%)	18 (11.2%)	9 (4.3%)
<b>C5 [Deceptive]</b>	-	2 (2.3%)	145 (20.2%)	2 (3.7%)	22 (13.7%)	8 (3.9%)
<b>C6 [Confronting]</b>	4 (14.8%)	10 (11.6%)	98 (13.7%)	5 (9.3%)	15 (9.3%)	20 (9.7%)

Tab. 12: Distribution of ANs in clusters according to the suffix. The percentage per suffix (in columns) is given in brackets. Highest values (in columns) are indicated in bold.

Table 13 presents the distribution of ANs within clusters according to the POS of their base. Nominal bases cluster mostly in C2 and C3, whereas verbal bases are more regularly distributed—with a slight concentration in C4. The most important contrasts can be observed within clusters. The distribution weighted by the overall number of verbal vs. nominal bases (respectively 786 and 466 items) exhibits important disparities. C1 and C4 clearly privilege verbal bases (respectively 78.5% and 75.6%) over nominal bases (21.5% and 24.4%), whereas C3 includes mostly ANs derived from nouns (80.1%) rather than from verbs (19.9%). Smaller differences can be observed for C2 (64.6% nouns, 35.4% verbs) and C5

<sup>21</sup> The labels between brackets are used as a reminder of some core semantic features present in each cluster. However, they should not conceal the semantic complexity of the clusters, as detailed in Section 4.2.



(68.0% verbs, 32.0% nouns). C6 appears to be the more balanced cluster with respect to base POS (56.2% verbs, 43.8% nouns).

	Verb	Noun
<b>C1 [Business]</b>	128 (16,3%)	21 (4.5%)
<b>C2 [Manual]</b>	122 (15,5%)	132 (28.3%)
<b>C3 [Intellectual]</b>	78 (9.9%)	185 (39.7%)
<b>C4 [Interpersonal]</b>	214 (27,2%)	41 (8.8%)
<b>C5 [Deceptive]</b>	140 (17,8%)	39 (8.4%)
<b>C6 [Confronting]</b>	104 (13.2%)	48 (10.3%)

Tab. 13: Distribution of ANs in clusters according to base POS. The percentage per base POS (in columns) is given in brackets.

The clustering of ANs according to the semantic type of the morphological base is presented in Table 14. Some ANs favor specific clusters depending on the semantic type of their base. This is especially the case for ANs that derive from object-denoting bases (mostly grouped in C2), and for ANs that derive from domain- and cognitive-object-denoting bases (mostly grouped in C3). The co-clustering of the latter relates to the fact that C3 groups together ANs that denote specialists and artists. In contrast, action-denoting bases distribute rather evenly across clusters, even though a significant proportion of them can be found in C4. As for property- and institution-denoting bases, they are too few to draw any conclusion about their distribution. It can also be noted that some clusters favor one base type, as is the case for C1 and C4, which gather mostly ANs derived from action-denoting bases.

	Action	Object	Domain	Property	Institution	Cog. Obj.
<b>C1 [Business]</b>	139 (16.4%)	8 (4.2%)	1 (0.6%)	-	1 (25%)	-
<b>C2 [Manual]</b>	129 (15.2%)	<b>104 (55.1%)</b>	17 (9.8%)	-	<b>2 (50%)</b>	2 (5.9%)
<b>C3 [Intellectual]</b>	92 (10.9%)	20 (10.6%)	<b>121 (70.0%)</b>	2 (50%)	-	<b>28 (82.4%)</b>

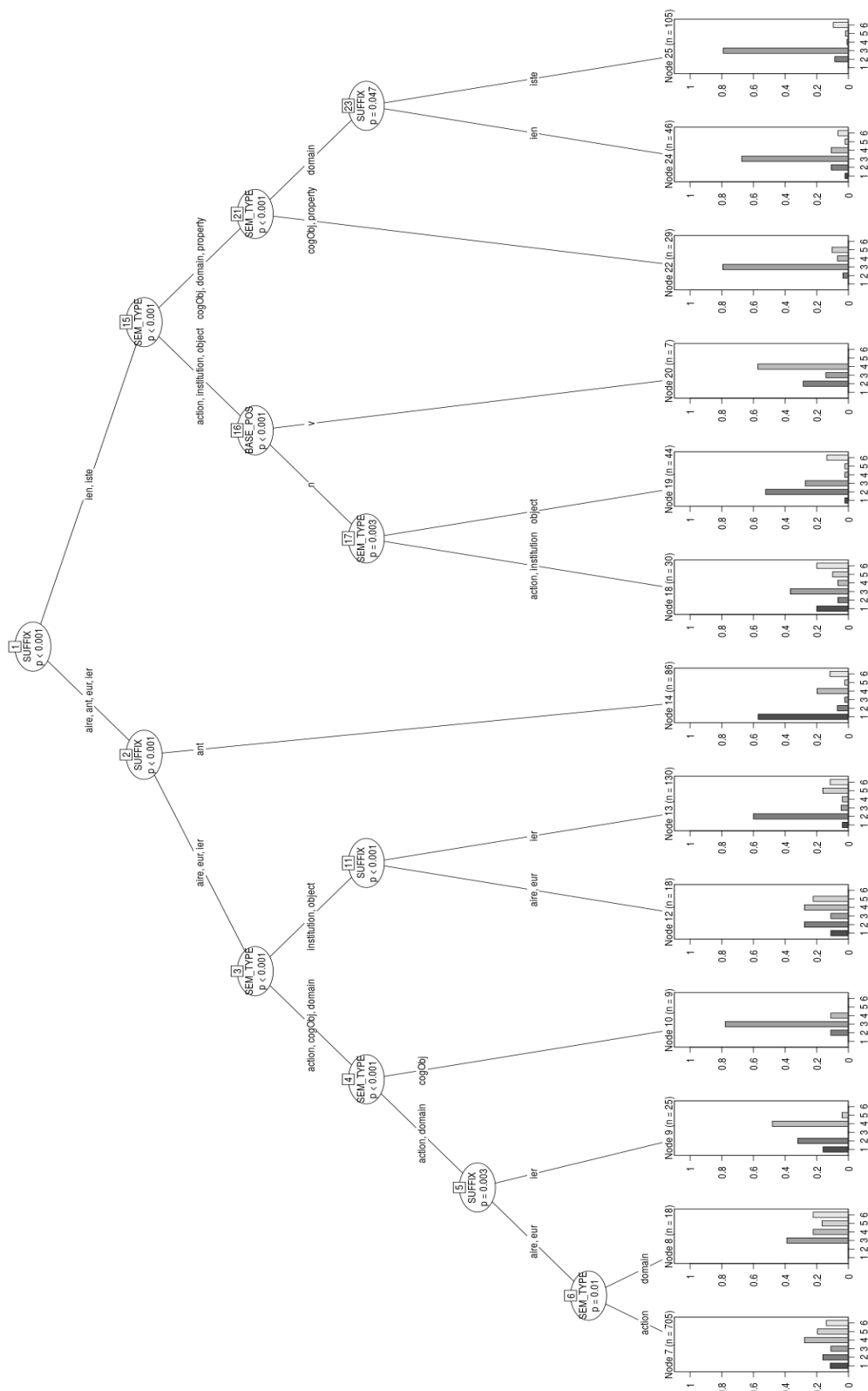
<b>C4 [Interpersonal]</b>	<b>230 (27.1%)</b>	11 (5.8%)	11 (6.3%)	-	-	3 (8.8%)
<b>C5 [Deceptive]</b>	144 (17.0%)	25 (13.2%)	6 (3.5%)	2 (50%)	1 (25%)	1 (2.9%)
<b>C6 [Confronting]</b>	114 (13.4%)	21 (11.1%)	17 (9.8%)	-	-	-

Tab. 14: Distribution of ANs in clusters according to base semantic type. The percentage per type (in columns) is given in brackets. Highest values (in columns) are indicated in bold.

#### 4.3.2. Statistical analysis

Some of our observations concerning the distribution of AN suffixes, base POS and base semantic types among clusters are congruent. For instance, the denotation of occasional agents, the concentration of *-ant* ANs, and the selection of verbal dynamic bases in C1 are morphosemantically coherent. So are the denotation of intellectual activities, the concentration of *-ien* and *-iste* ANs, and the predilection for domain-denoting nominal bases in C3. To further examine how these properties interact in the semantic clustering of ANs, we fitted conditional inference tree and random forest models to our data set<sup>22</sup>. These models aim to predict the most probable outcome on the basis of given factors, and to identify which ones have the largest independent effects on a given response variable. Conditional inference trees recursively split the data into two significant subsets, and stop when the predictors no longer allow for significant data partitioning. The conditional inference tree presented in Figure 2 was trained with suffix, base POS and base semantic type as predictors, and clustering as the response variable. Each node represents a splitting point: the predicting factor is specified in the oval and the branches indicate the relevant variables. The final bar plots display the distribution of the response variable in each terminal node.

<sup>22</sup> The analysis was performed using the `cforest()` and `ctree()` functions from the R package ‘party’.



The conditional inference tree confirms the importance of some morphosemantic properties of ANs. As far as suffixes are concerned, the proximity between *-ien* and *-iste* on one hand, and *-aire* and *-eur* on the other, seems central insofar as both pairs of suffixes are analyzed as joint predictors of the clusters. *-ien* and *-iste* are separated from other suffixes from Node 1, and only split on Node 23, when derived from domain-denoting bases, and with a rather high p-value of 0.047. *-Aire* and *-eur* are distinguished from other suffixes on Nodes 5 and 11, but not from one another, and are co-predictors of final Nodes 7, 8 and 12. It can be noted that ultimately, ANs ending in *-aire* and *-eur* are not separated according to the suffix, but to the semantic type of the base (Nodes 3 and 6). The *-ant* suffix is remarkable in that it is isolated from Node 2 and independently gives the best prediction for C1. As for *-ier*, it is also analyzed as a predictor, but only when combined with other properties (Nodes 5 and 11). The semantic type of the base appears to be an important predictor as well. For instance, object-denoting bases are separated on Nodes 3 and 17, domain-denoting bases on Nodes 6 and 21, and cognitive-object-denoting bases on Nodes 4 and 21. In contrast, the base POS criterion is only used once, when clustering *-ien* and *-iste* ANs derived from action-, institution- or object-denoting bases (Node 16).

The two variables of suffix and base semantic type often combine to best predict some clusters. For instance, the best predictors for C2 are *-ier* ANs derived from institution- or object-denoting bases (Node 13), and the best predictors for C3 are *-ien* and *-iste* ANs derived from domain- and cognitive-object-denoting bases (Nodes 22, 24 and 25). C4 cannot be as strongly predicted as C1-C3. It is rather weakly related to *-aire* and *-eur* ANs derived from action-denoting bases (Node 7), and only predicted with high probability by small classes such as *-ier* ANs derived from action- and domain-denoting bases (Node 9), and deverbal *-ien* and *-iste* ANs (Node 20). Interestingly, the conditional inference tree does not make any high prediction for C5 and C6, i.e. none of our variables is strongly correlated with the semantic distinctions that characterize these two clusters. Although the morphosemantic properties of ANs can to some extent predict their linguistic distribution, they cannot fully account for the distributionally relevant medium-grained distinctions that apply to them.

To precisely determine what are the most significant predictors of the clustering, we ran a random forest model, which computed a large number of conditional inference trees in order to estimate measures of variable importance<sup>23</sup>. These measures quantify the impact of the different predictors considered.

---

<sup>23</sup> The measures were computed using the `varimp()` function from the R package.

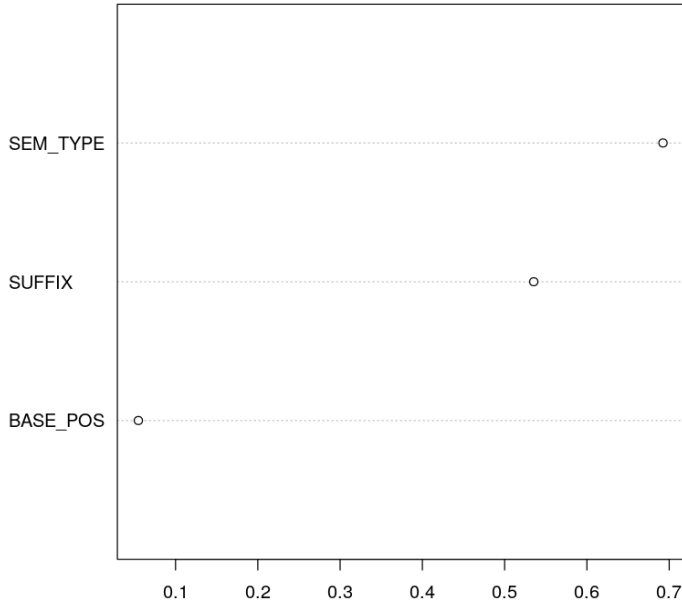


Fig. 3: Conditional importance of variables

Figure 3 presents the relative influence of the three variables. It confirms that base semantic type and suffix are the two main variables used by the algorithm to predict the clustering, the former being an even stronger predictor than the latter. It could be extrapolated from these results that ANs deriving from bases with the same semantic type are more likely to be semantically close than ANs deriving from bases with the same POS—hence they are more likely to be rivals, if one considers that the more two affixes rival, the semantically closer are the words they form. For instance, *-eur* and *-aire* are distributionally closer than *-ier* and *-iste*. They differ from *-ier* and *-iste* in that they do not mostly select bases with the same POS, but with the same semantic type. If one understands affix rivalry as fundamentally dependent on output semantics, one could hypothesize that agentive suffix rivalry in French is more dependent on the semantic type of the base than on its POS.

## 5. Conclusion

In this article, we have investigated the rivalry between 6 agentive suffixes used to form ANs in French, in order to determine whether these suffixes are related to different agentive morphosemantic constructions, and to the denotation of different subclasses of agents. We combined manual, computational and statistical analyses to provide some answers to these questions.

Some suffix specificities already identified in the literature have been confirmed by converging observations (e.g. the propensity to denote specialists in the case of *-ien* and *-iste*, or craftsmen and

shopkeepers in the case of *-ier*), while others have emerged (e.g. the versatility of *-aire*, the distinctive construction of behavioral ANs in the case of *-eur*, the preference for occasional ANs in the case of *-ant*). Our methodological approach made it possible first to evaluate these specificities as tendencies rather than categorical distinctions, second to relate these specificities to different agentive morphosemantic constructions. In particular, the semantics of the base appears to play an important role in determining both how agentivity is semantically developed in the derivational process, and which specific type of agent is denoted.

The diversity of agentive suffixes is correlated to semantic distinctions among ANs, but these correlations are not systematic, and many cases of overlapping construction patterns can be observed. Our analysis shows that the many-to-many relations between form and meaning extend to medium- or fine-grained semantics—e.g. craftsmen and shopkeepers can be denoted not only by *-ier* ANs, but also by *-eur* and *-iste* ANs, whereas *-ier* ANs do not only denote craftsmen and shopkeepers, but also in some cases writers or military men. Suffix distribution and agentive distinctions coincide to various degrees, according to the suffix and agentive subtype, which in turn determines different configurations and degrees of rivalry. Some suffixes are closer rivals than others: *-ien* and *-iste*, *-eur* and *-aire* are close rivals, but *-eur* also rivals *-ier* to some extent, while *-ant* appears to be the more isolated agentive suffix. There are obvious implications between the POS and semantic types of ANs' bases, but the degree of rivalry seems more determined by the semantic type than by the POS. It may be asked whether affix rivalry should be defined by the identity of base semantic type rather than base POS.

The distributional analysis, especially in a bottom-up approach, can support a semantic subclassification of ANs. The distributionally relevant AN subclasses that emerge are characterized by two factors: ontological distinctions between action types (depending on the kind of activity involved, or on the relation to other participants), and structural distinctions between aspectual or predicate level types (as regards the denotation of functional, occasional or behavioral agents). The morphosemantic construction of ANs is clearly related to these distinctions, but cannot account for all of them.

Our study focused on the semantic aspects of the rivalry between some agentive suffixes, but further investigation could expand the scope of research. The possibility for agentive affixes to denote other semantic types than agent should be investigated, and could shed light on the semantic specificities of each agentive derivational pattern, if one assumes a non-arbitrary relation between the different interpretations associated with a given suffix. More broadly, considering that semantics is probably not the only factor that determines the rivalry between agentive suffixes, differences in phonology, syntax,

diachrony, productivity and lexicalization should be taken into account to provide a complete picture of agentive suffix rivalry.

## References

- Amiot, D. & Dal, G. (2007). Integrating Combining Forms into a Lexeme-Based Morphology. In G. Booij, B. Fradin, A. Ralli & S. Scalise (eds.), *Online Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*, 323-336.
- Andrews, S. (1996). *The Semantics of Suffixation. Agentive substantival Suffixes in Contemporary Standard Russian*. München/Newcastle: LINCOM EUROPA.
- Anscombre, J.-C. (2001). A propos des mécanismes sémantiques de formation de certains noms d'agent en français et en espagnol. *Langages*, 143, 28–48.
- Anscombre, J.-C. (2003) L'agent ne fait pas le bonheur : agentivité et aspectualité dans certains noms d'agent en espagnol et en français. *Thélème, Revista Complutense de Estudios Franceses*, 11–27.
- Arndt-Lappe, S. (2014). Analogy in suffix rivalry: The case of English *-ity* and *-ness*. *English Language & Linguistics*, 18(3), 497-548.
- Balles, I. (2005). Indogermanische Nomina agentis: Probleme und Lösungsansätze. In I. Balles, & R. Lühr (eds.), *Indogermanische Nomina agentis (Linguistische Arbeitsberichte, 83, p.7-70)*. Leipzig: Institut für Linguistik.
- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.
- Benveniste, E. (1948). *Noms d'agent et noms d'action en indo-européen*. Paris: Maisonneuve.
- Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6, 213-234
- Bonami, O. & Thuilier, J. (2018). A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure*, 11(2), 4-41.
- Booij, G. (2002). *The Morphology of Dutch*. Oxford: Oxford University Press.
- Bréal, M. (1897) [2005]. *Essai de sémantique*. Limoges: Lambert-Lucas.
- Cartoni B., Namer F. (2012). Linguistique contrastive et morphologie : les noms en *-iste* dans une approche onomasiologique. *Actes du CMLF 2012*, 1245-1260.
- Cartoni, B., Namer, F. & Lignon, S (2015). A cross-linguistic insight on agentive noun formation in Italian and French. *Selected Papers from the 8th Décembrettes: Morphology in Bordeaux, Carnets de Grammaire*, 22, 81–98.
- Corbin, D., & Corbin, P. (1991). Un traitement unifié du suffixe *-ier(e)*. *Lexique*, (10), 61-145.

- Dal, G., Hathout, N., Lignon, S., Namer, F., & Tanguy, L. (2018). Toile versus dictionnaires: Les nominalisations du français en *-age* et en *-ment*. In *SHS Web of Conferences* (Vol. 46, 08003). EDP Sciences.
- Dressler, W. U., Barbaresi, L. M., Schwaiger, S., Ransmayr, J., Sommer-Lolei, S., & Korecky-Kröll, K. (2019). Rivalry and lack of blocking among Italian and German diminutives in adult and child language. In *Competition in Inflection and Word-Formation* (p. 123-143). Springer, Cham.
- Dubois, J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain: essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Paris: Larousse.
- Fábregas, A. (2010). A syntactic account of affix rivalry in Spanish nominalizations. In A. Alexiadou & M. Rathert (eds.), *The Semantics of Nominalizations across Languages and Frameworks* (p. 67-92). Berlin/New York: Mouton de Gruyter.
- Firth, J.R. (1957). A synopsis of linguistic theory, 1930-1955. In J.R. Firth (ed.), *Studies in Linguistic Analysis* (p. 1-32). Oxford: Basil Blackwell
- Flaux, N. & Van de Velde, D. (2000). *Les noms en français: esquisse de classement*. Paris: Editions Ophrys.
- Fradin, B. (2016). Competition in derivation: what can we learn from doublets?. In *International Morphology Meeting*, Vienna, Austria).
- Fradin, B. (2016). L'interprétation des nominalisations en N-age et N-ment en français. In E. Buchi, J.-P. Chauveau, J.-M. Pierrel (éds), *Actes du XXVIIe congrès international de linguistique et philologie romane* (53-66). Strasbourg : Société de Linguistique Romane / Eliphi.
- Fradin, B. & Kerleroux, F. (2003). Troubles with lexemes. In: G. Booij, J. Cesaris, S. Scalise & A. Ralli (eds.) *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting* (p. 177–196). IULA-Universitat Pompeu Fabra,
- Gardani, F., Rainer, F., & Luschützky, H. C. (2019). Competition in Morphology: A Historical Outline. In *Competition in Inflection and Word-Formation* (pp. 3-36). Springer.
- Godard, D. & Jayez, J. (1996). Types nominaux et anaphores : le cas des objets et des événements. In W. De Mulder, L. Tasmowski-De Ryck & C. Vetters (eds.), *Anaphores temporelles et (in-)coherence, Cahiers Chronos* (Vol.1, p.41-58). Amsterdam: Rodopi.
- Harris, Z.S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.
- Kintsch, W. (2001). Predication. *Cognitive Science*, 25(2), 173–201.
- Knittel, M.L. (2017). French derived nominals in *-ant*: semantic properties. Talk given at ISMO, Lille, December 2017.



- Koehl, A., & Lignon, S. (2014). Property nouns with -ité and -itude: formal alternation and morphopragmatics or the sad-itude of the AitéN. *Morphology*, 24(4), 351-376.
- Lasserre, M., & Montermini, F. (2014). Pour une typologie des lexèmes construits: entre composition, composition néoclassique et affixation. In *SHS Web of Conferences* (Vol. 8, p. 1797-1812). EDP Sciences.
- Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, 4, 151–171.
- Lerat, P. (1984). Grammaire des noms d'agent en -ant en français contemporain. *Cahiers de Lexicologie*, 24, 23–29.
- Lignon, S. (2000). *La suffixation en -ien: aspects sémantiques et phonologiques* (Doctoral dissertation, University Toulouse Jean Jaurès).
- Lignon, S. (2007). Les noms de spécialistes en -iste et en -ien : le chimiste perturbé ou comment le physicien se réajuste. In B. Vaxélaire, R. Sock, G. Kleiber & F. Marsac (eds.), *Perturbations et Réajustements. Langue et langage* (p. 287-295). University Marc Bloch - Strasbourg 2.
- Luschützky, H. C. (2011). Agent-noun polysemy in Slavic: some examples. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 64(1), 75-97.
- Martin, F. (2010). The semantics of eventive suffixes in French. In A. Alexiadou & M. Rathert (eds.), *The Semantics of Nominalizations across Languages and Frameworks* (p. 109-139). Berlin/New York: Mouton de Gruyter.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*
- Mikolov, T., Yih, W. & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 13, 746–751.
- Naccarato, C. (2019). Agentive (para) synthetic compounds in Russian: a quantitative study of rival constructions. *Morphology*, 29(1), 1-30.
- Pierrejean, B (2020). *Qualitative Evaluation of Word Embeddings: Investigating the Instability in Neural-Based Models*. (Doctoral dissertation, University Toulouse Jean Jaurès)
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846-850.
- Roché, M. (1997). Briard, bougeoir et camionneur: dérivés aberrants, dérivés possibles. In D. Corbin, B. Fradin, B. Habert, F. Kerleroux & M. Plénat (éds.), *Mots possibles et mots existants. 1res rencontres du forum de morphologie* (Vol. 1, pp. 241-250).

- Roché, M. (2003). Catégorisation et recatégorisation en morphologie dérivationnelle: le cas de la dérivation en-ier (e). In G. Coll & J.P. Régis (dir.), *Morphosyntaxe du lexique. Catégorisation et mise en discours, Actes du Colloque de Tours, 7-8 juin 2002, Travaux Linguistique du CerLiCO*, (16, 75-92).
- Roché, M. (2004). Mot construit ? mot non construit ? quelques réflexions à partir des dérivés en -ier (e). *Verbum*, 26(2), 459–480
- Roché, M. (2011). Quel traitement unifié pour les dérivation en -isme et en -iste ? In M. Roché, G. Boyé, N. Hathout, S. Lignon, & M. Plénat (eds.), *Des unités morphologiques au lexique* (p. 69-143). Paris: Hermès.
- Rodrigues, A. S. (2013). Semantic affix rivalry: the case of Portuguese nominalisers. In *9th Mediterranean Morphology Meeting* (p. 134-137). Universidade de Coimbra, Centro de Estudos e Linguística Geral e Aplicada.
- Rosenberg, M. (2008). *La formation agentive en français. les composés [VN/A/Adv/P]N/A et les dérivés V-ant, V-eur et V-oir (e)*. (Doctoral dissertation, University of Stockholm)
- Roy, I. & Soare, E. (2012). L'enquêteur, le surveillant et le détenu : les noms déverbaux de participants aux événements, lectures événementielles et structure argumentale. *Lexique*, 20, 207–231.
- Schnedecker, C. & Aleksandrova, A. (2016). Les noms d'humains en -aire : essai de classification. In F. Neveu, G. Bergounioux, M.H. Côté, J.M. Fournier, L. Hriba and S. Prévost (eds.), *Congrès Mondial de Linguistique Française 2016* (27, 12001). Paris: Institut de Linguistique Française.
- Schulte, M. (2015). Polysemy and synonymy in derivational affixation—a case study of the English suffixes-age and-ery. *Morphology*, 25(4), 371-390.
- Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. (Doctoral dissertation, University of Paris Diderot)
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. (Doctoral dissertation, University Toulouse Jean Jaurès)
- Uth, M. (2010). The rivalry of French -ment and -age from a diachronic perspective. In A. Alexiadou & M. Rathert (eds.), *The Semantics of Nominalizations across Languages and Frameworks* (p. 215-244). Berlin/New York: Mouton de Gruyter.
- Varvara, R. (2018). *Verbs as nouns: empirical investigations on event-denoting nominalizations*. (Doctoral dissertation, Università Degli Studi di Trento)
- Villoing, F., & Namer, F. (2012). Composition néoclassique en-logue et-logiste: les noms en-logue sont-ils encore des noms de spécialistes. *Verbum*, (2), 213-231.
- Winther, A. (1975). Note sur les formations déverbaux en -eur et -ant. *Cahiers de lexicologie*, 26, 35-54.