Investigating the distributional properties of rival -age suffixation and verb to noun conversion in French

Alice Missud, Florence Villoing

Verbum XXXVIII, 2016, no

INVESTIGATING THE DISTRIBUTIONAL PROPERTIES OF RIVAL *-AGE* SUFFIXATION AND VERB TO NOUN CONVERSION IN FRENCH[1]

Alice Missud
UMR 7114 MoDyCo, CNRS & Université Paris Nanterre
UMR 8094 LaTTiCe, CNRS, E.N.S. & Université Paris 3 Sorbonne-Nouvelle
missud.a@parisnanterre.fr

Florence Villoing
UMR 7114 MoDyCo, CNRS & Université Paris Nanterre
villoing@parisnanterre.fr

RESUME

Les travaux sur la compétition en morphologie dérivationnelle en français se sont principalement focalisés sur l'identification des propriétés sémantiques des suffixations rivales en -age, -ment *et* -ion *construisant des noms d'événements à partir de verbes. Cependant, un autre schéma rival jusque-là non pris en compte est la conversion de verbe à nom, qui dérive également des noms événementiels. Cet article présente une étude de la rivalité entre la conversion de verbe à nom et la suffixation en* -age *en s'attardant sur les propriétés distributionnelles de leurs dérivés. En utilisant des méthodes quantitatives et des modèles de sémantique distributionnelle (DSM), nous montrons que le degré de dispersion et spécificité sémantique des nominalisations diffère d'un schéma à l'autre.*

[1] We are grateful to the participants of ISMo 2019 (Second International Symposium of Morphology), *12th Mediterranean Morphology Meeting* (2019) and the Workshop "Concurrence & Polysémie" at Université Sorbonne Paris Cité, (2019) where we presented preliminary results. We also thank the reviewers of the first version of this paper for their valuable comments.

ABSTRACT

Work on competition in French word-formation has mostly focused on the semantic properties of the rival -age, -ment *and* -ion *suffixations that construct deverbal event nouns. However, another important rival schema that has been neglected so far is verb to noun conversion as it also derives a significant number of eventive nominalizations. This paper presents a study of the rivalry between verb to noun conversion and* -age *suffixation by investigating the distributional properties of their lexemes. By comparing word vectors using Distributional Semantic Models (DSMs), we show that the degree of semantic dispersion and specificity of nominalizations differs from one schema to another and that different types of converted nouns can be discriminated.*

0. INTRODUCTION

Research in derivational morphology has long sought to explain the reasons for the coexistence of rival morphological schemas that select the same bases to construct similar meanings. In French, the rivalry between *-age*, *-ment* and *-ion* suffixations that derive deverbal event nouns has received the most attention (see, for example, Dubois 1962; Martin 2010; Uth 2010; Fradin 2014, 2019; Dal *et al.* 2018; Wauquier *et al.* 2019). However, another important rival in this competition that has not been considered as such is verb to noun conversion (henceforth "V to N conversion"), which has been shown to derive a significant proportion of event nouns from verbs (Tribout 2010) as in (1).

(1) *survoler* 'to fly over' → *survol* 'hovering',
baisser 'to drop' → *baisse* 'drop',
secouer 'to shake' → *secousse* 'shake'
venir 'to come' → *venue* 'coming'

This neglect is mostly due to its belated recognition as a derivational morphology schema (see section 1). Our work is rooted in the theoretical background of lexemic morphology (Aronoff 1994), for which verb to noun conversion is considered a lexeme-formation process similar to affixal ones, except for the phonological part of the derivation, since conversion is characterized by a phonological identity between the two lexemes (or, more precisely, between a stem of the base verb and the stem of the derived lexeme).

There are many factors (phonological, morphological, syntactic, semantic, pragmatic, etc.) that may shed light on the competition between two morphological processes. The literature devoted to the morphological rivalry that constructs deverbal event nouns has mostly focused on the semantic features that can differentiate them by investigating the base verbs and the aspectual and argumental properties of the derived nouns. For example, with regard to the case of rivalry we are interested in, some deverbal converted nouns ending in ée (*arrivée* 'arrival', *plongée* 'diving', *traversée* 'crossing')

have already been compared with -age suffixed nouns in order to find distinctive syntactic and aspectual properties (Ferret *et al.* 2010; Ferret & Villoing 2012). Our research goes further and takes all types of event deverbal converted nouns into account and not only those ending in *ée*. In addition, we are studying semantic issues that have not yet been addressed. Thus, for the purpose of introducing V to N conversion as a rival in the competition that opposes deverbal event nouns in French, this paper investigates the semantics of -age suffixation and V to N conversion as a first step by looking at the distributional properties of the lexemes they derive.

In this paper, we explore the distributional properties of derived nouns using Distributional Semantics Models (DSMs). Such models, based on the distributional hypothesis (Harris 1954; Firth 1957), allow for quantitative analyses of the semantics of words by converting nouns into word vectors that represent their distribution in a corpus. Word vectors can then be used to calculate semantic similarities between words by measuring the cosine distance between vectors. Advanced word vectors generated by Word2Vec neural network-based models (Mikolov *et al.* 2013) have recently been used to discriminate French rival nominalization schemas. Notably, Wauquier *et al.* (2019) provided evidence that *-age*, *-ment* and *-ion* suffixations could be semantically discriminated based on the semantic similarity between morphologically related nouns. By computing word similarities, they showed that suffixed nouns attract an overwhelming majority of nouns that are derived from the same schema. We aim to investigate if this holds for *-age* suffixation versus V to N conversion, as well as for discriminating converted nouns based on their stem.

Apart from distinctive semantic properties between rival schemas, word vectors can also be used to measure the semantic dispersion of lexemes in order to get a sense of the extent to which they semantically cluster together depending on the schema they derive from. In the case of morphological rivalry, Lindsay and Aronoff (2013) argued that two rival schemas can be discriminated by their degree of specialization and versatility. This suggests that schemas will therefore self-organize in order to coexist while remaining productive: one will be more specialized by investing a specific niche while the other will be more versatile. Consequently, we measured the degree of specialization of *-age* suffixation and V to N conversion by investigating the semantic niches that lexemes might occupy and the degree of semantic relatedness they maintain with their pairs using word vectors.

We believe that the results obtained by manipulating word vectors need to be verified scrutinizing them from various angles The series of experiments that are presented in this paper form part of this perspective as their results intersect and complement each other. First, we present the data and the DSM we used as well as the frequency distribution of the data according to the schema they derive from. The second section of this paper presents results

showing that *-age* suffixation and V to N conversion (and its different types) differ in terms of semantic relatedness with their pairs by computing word similarities. The following section proposes a method for predicting the semantic dispersion of lexemes depending on their schema. The results are then used in Section 4 to categorize the different semantic niches that N-age and V to N converted nouns occupy.

1. DATA

1.1. Conversion, a particular case of lexical derivation

Our work is rooted in the perspective that conversion is a derivational morphological process (see, among others, Plag 2003; Don 2004; Bauer *et al*. 2013; and, for French, Corbin 1987; Kerleroux 1999; Tribout 2010, 2012, 2015). The arguments for this position are based on properties such as: conversion involves substitution of a new inflectional paradigm, new syntactic properties (new word-class), and new semantic properties. Thus, according to these properties, the same form is interpreted as a different lexeme and conversion is part of lexical derivation (see Valera 2014 for a summary of these questions). The specificity of conversion with regard to affixation, for example, lies in the formal identity between the base and the derivative (which poses specific problems in determining the directionality of the process). By postulating that conversion is a word-formation process that results in unmarked (by affixes) word-class change, we reject other approaches that attribute another status to conversion, in particular: (i) the one that views conversion as "zero-derivation" or "zero-affixation"; (ii) the one that considers conversion as a non-derivational lexical creation that consists of a second introduction of an existing word within a different category in the lexicon (as for Lieber 2005) and (iii) the one that denies the derivational or lexical process because it considers that word-class change does not exist: according to this view, lexical items are unspecified as regards word class and may be specified as members of different categories according to the context (see Distributed Morphology, Marantz 1997, for example).

1.2. French N to V and V to N conversion: the issue of phonological identity

The confusion surrounding the recognition of V to N French conversion as a derivational process dates back to Darmesteter (1877) and was reproduced by Nyrop (1936) which have long served as a reference. This is due in particular to the difficulty in recognizing a phonological identity between verb and noun (see Corbin 1987; Kerleroux 1999). While formal identity between the base and the derived lexeme is an important condition for conversion, it

may fail to apply in many languages where formal changes appear nevertheless: for example, stress shift between English nouns and verbs (*torm´ent*$_V$ – *t´orment*$_N$; *constr´uct*$_V$ – *c´onstruct*$_N$) (Plag 2003); or formal differences between the base verbal lexeme and the converted noun lexeme in German (*Antworten*$_V$ – *Antwort*$_N$; *Fragen*$_V$ – *Frage*$_N$) (Valera 2014) or in Italian (*caminare*$_V$ 'to walk' – *cammino*$_N$ 'walking'; *sostare*$_V$ 'to stop, to rest' – *sosta*$_N$ 'stopping, rest') (Marzo 2013). French V to N conversion also presents the same kind of phonological difference: because of a noticeable stem allomorphy of French verbs, and the decision to represent the verbal lexeme by the infinitive form, the base verbal lexeme is not phonologically identical to the converted noun lexeme (see examples in (2)). To account for it, we follow Tribout's analysis (2010, 2012) based on Aronoff (1994) and Bonami and Boyé (2003)'s treatment of allomorphy for which each verb has a list of indexed morphemic stems. In this perspective, inflectional and derivational morphological formations select one of these stems to construct either a word form or a lexeme. Tribout showed how stem spaces can be used in derivation to account for verb to noun conversion. She demonstrated that the French verbal stem space contains fourteen stems, and that "each of them is potentially available to be the input of deverbal lexeme-formation processes" (Tribout 2012: 122). Her work proposed that three sorts of stems are available to derive deverbal converted nouns: stem 0 (2a), (also used to inflect, for example, the present singular forms of 1st conjugation verbs), stem 12 (2b) (also used to inflect the past participle forms of verbs), and stem 13 (2c) (hidden to inflection and only used in derivation for deverbal *-if, -eur*/*-rice* and *-ion* suffixations) (see more examples in Tribout 2012).

(2) a. *marcher*$_V$ 'to walk' /maʁʃ/ → *marche*$_N$ 'walk' /maʁʃ/
b. *sortir*$_V$ 'to go out' /sɔʁti/ → *sortie*$_N$ 'exit/outing' /sɔʁti/
c. *défendre*$_V$ 'to defend' /defɑ̃s/ → *défense*$_N$ 'defence' /defɑ̃s/

Thus, French V to N conversion can be characterized by a phonological identity between the base and the derivative where the variety of stems involved is considered instead of the phonological form of the lexeme. Stem selection in V to N conversion is crucial for the present research as we aim to investigate the hypothetical semantic distributional differences between converted nouns depending on the stem they select.

1.3. Word vectors

Following Firth's intuition that "You should know a word by the company it keeps" (1957), the distributional hypothesis stipulates that words that have similar meanings share similar contexts. The idea that the meaning of a word can be inferred by knowing the words that surround it has been widely used

in the field of natural language processing. Distributional semantics models (DSMs) such as Latent Semantic Analysis (Landauer *et al.* 1998) operationalize this principle by representing words with word vectors. These vectors are the result of a transformation from textual (actual word) to numerical (vector) that captures the contextual-meaning usage of words, where each dimension represents the frequency of cooccurrence of a target word with others in a given corpus. In derivational morphology, DSMs have proven successful in predicting the directionality of verb to noun conversion in English (Kisselew *et al.* 2016) or the lack of semantic regularity in derivation as opposed to inflection in French (Bonami & Paperno 2018).

More recently, predictive models generated by neural network-based tools such as Word2Vec (Mikolov *et al.* 2013) have been popularized because of their high performance in computing word similarities or analogies. While they rely on the same principle, these models generate dense vectors (or word embeddings) through unsupervised machine learning techniques. Although they can provide results that are close to accurate semantic intuitions, the opacity of word embeddings' dimensions makes the interpretation tricky and results need to be complemented by qualitative analyses when investigating linguistic phenomena.

This study was undertaken from this perspective with the idea that morphologically related lexemes that have the same distribution might cluster because they share at least one specific morpho-semantic property. We used a CBOW model from Word2Vec trained on a concatenation of three massive French corpora extracted from the web: frCOW (Schäfer & Bildhauer 2012; Schäfer 2015: 9 billion words in 2016), frWaC (Baroni *et al.* 2009: 1.9 billion words in 2009) and frWiki (178 million words, a dump of Wikipedia encyclopedic pages from 2007). The model was trained on words with a frequency of at least 5 using default parameters (window of 5; negative sampling, 5 items).

1.4. Collecting the lexicons

Data collection was subject to several constraints in order to find a balance between quantity and quality while manipulating word vectors.

A first constraint is that the quality of word vectors depends on the number of examples that is found in the corpus that the model is trained on. Consequently, highly frequent words are better represented as they appear in many contexts. Zipf's law states that the frequency of words in a corpus is inversely proportional to their rank. In other words, we expect to find few high-frequency words and many low-frequency ones. Therefore, the number of frequent derivatives that we can select to ensure the quality of the vectors is rather limited.

Another constraint lies in the extraction of converted nouns. Converted lexemes pose four major challenges for computational linguistics. First, as no

affix is involved, their identification in corpora cannot rely on spotting an additional and specific phonological sequence. Secondly, inflected nouns and verbs are sometimes phonologically identical (*marche*$_{V}$ / *marche*$_{N}$, *avancée*$_{V}$ / *avancée*$_{N}$). Their identification therefore relies on the quality of morphosyntactic tagging, which is usually subject to many errors when it comes to participles in French. Another major problem is the direction of derivation (verb to noun or noun to verb), which is sometimes impossible to predict. As stated by Tribout (2015) for French, the orientation of conversion is always questionable since no information can be truly reliable to assess which lexeme appeared first (dating, phonetics, semantic interpretation, semantic range, frequency of occurrence). Lastly, converted nouns are not necessarily eventive and can have multiple semantic interpretations such as instrument, agent or location (Tribout 2010, 2015). Instead of extracting converted nouns automatically, we had to rely on existing annotated datasets.

As a result, we collected a total of 300 nouns that have at least one eventive interpretation (some polysemous nouns can also denote results). 150 of them are *-age* suffixed nouns extracted from VerNom (Missud *et al.* 2020), a lexical database consisting of 25 857 verb-noun pairs acquired from frCOW (Schäfer & Bildhauer 2012; Schäfer 2015), a massive corpus from the French web that dates back to 2016 and that covers a wide variety of texts (press, forums, encylopedias...). The other 150 are converted nouns taken from Tribout's lexicographic database (2010). Tribout's data come from two online dictionaries: *Trésor de la Langue Française informatisé* and *Le Petit Robert Electronique*. The data were given an annotation for the stem (0, 12 or 13), the semantic interpretation (action, result, agent, location...), as well as the orientation of conversion when it has been retrieved (verb to noun, noun to verb or unknown). Only nouns that were annotated as deverbal and eventive were selected. Among them, 71 are derived from stem 12 of their base verb (*sortie* 'exit', *complainte* 'lamentation', *huée* 'booing'), 67 from stem 0 (*attaque* 'attack', *mépris* 'contempt', *retard* 'delay') and 12 from stem 13 (*course* 'race', *défense* 'defence', *plagiat* 'plagiarism'). Although we tried to balance out the number of converted nouns of each type, this distribution reflects the one that is found in Tribout's dataset. The derivatives we selected are the most frequent event nouns we could find in frCOW (see appendix for the word frequency list). We excluded doublets (in our case, situations where two morphological processes select the same verbal base to derive at least three different event nouns, as in *porter* 'to carry, to convey' → *port* 'carrying', *portée* 'scope', *portage* 'portage') to focus on the prototypical base selection behavior of each schema that distinguishes both of them by making sure that the verbal bases were exclusively selected by one schema to construct event nouns. Each derivative was then assigned a 100-dimensional vector from the Word2Vec model. Note that only a small portion of converted nouns have low frequency (*découple* 'decoupling', 2 occurrences) and were kept in our data as we needed the same amount of N-age and converted nouns.

1.5. Token frequency of lexemes according to their schema

In order to get a sense of the frequency distribution of our data, we compared the frequencies of the derivatives according to the schema they derive from and the stem selected for conversion. The token frequency of each noun was extracted from frCOW's word frequency list.

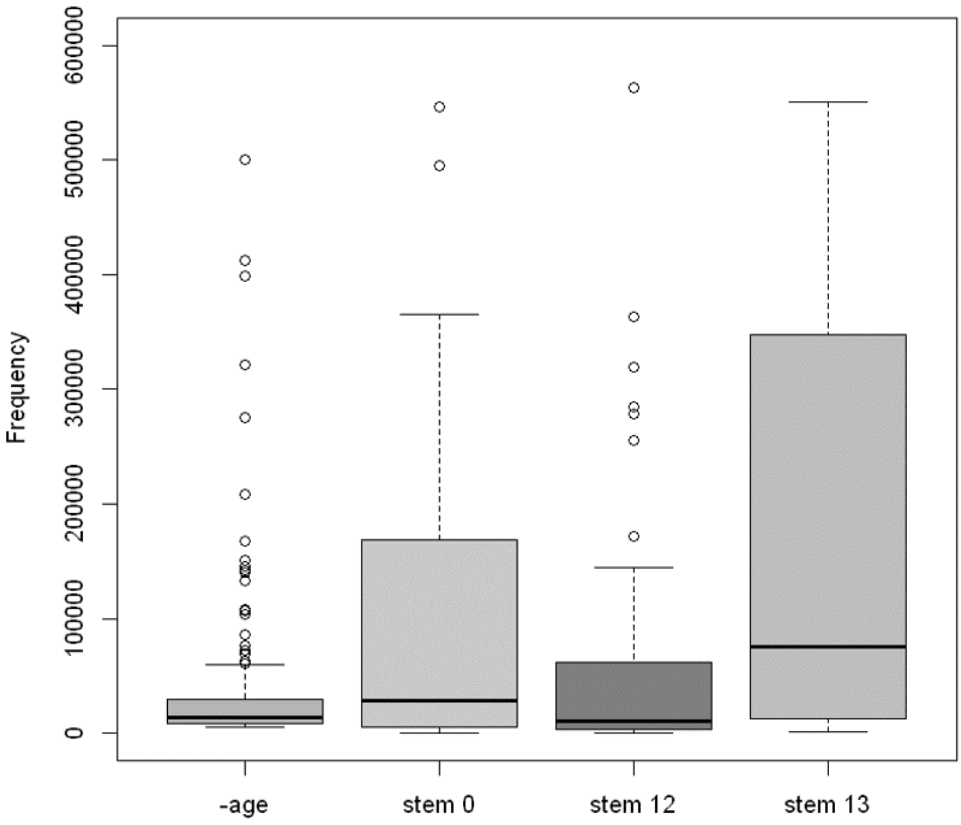

Figure 1. – Token frequency of the nouns depending on the schema

Figure 1 shows a boxplot[2] that represents the distribution of the frequencies of nouns; the Y-axis shows the frequency and the boxes in the X-axis each

[2] A boxplot is a graphical tool for visualizing the distribution of data through their quartiles. The colored box, delimited by the first (lower bound) and third (upper bound) quartiles, indicates where 50% of the data are concentrated. The bold line inside the box corresponds to the median that separates the data into two halves. The lines extending from the boxes each indicate the concentration of 25% of the data, and their endpoints show the lowest and

indicate a schema (*-age* suffixation '-age', conversion on stem 0 'stem 0', conversion on stem 12 'stem 12' and conversion on stem 13 'stem 13'). The larger the box, the higher the frequency for a specific type of noun.

The results show that *-age* suffixation and stem 12 conversion comprise the least frequent derivatives. As shown by the median situated at 15 000, *-age* suffixed nouns have the lowest frequencies. Half of them have a token frequency that ranges from 10 000 to 30 000. Stem 12 converted nouns follow the same trend: 50% have a frequency that goes from 5 000 to 60 000, with a median of 10 000. Stem 13 and stem 0 converted nouns are the ones that have the highest frequencies in comparison. 50% of stem 13 converted nouns (the most frequent in our data) have a token frequency that ranges from 15 000 to 370 000 and half of them have frequencies that go beyond 75 000 as shown by the median. Stem 0 converted nouns are the second most frequent: 50% of them have a frequency that ranges from 15 000 to 180 000. Overall, the most frequent converted nouns that we found in Tribout's data are more frequent than the most frequent *-age* suffixed nouns that are found in frCOW.

As proposed by Resnik (1995), the token frequency of a word can correlate with its informational content: as highly frequent words are more likely to appear in various contexts than low-frequency ones, they are likely to lose semantic specificity and be more generic and polysemous. Considering this hypothesis and the distribution observed in Figure 1, we would expect converted nouns (especially those derived from stems 0 and 13) to be more polysemous and generic than *-age* suffixed nouns. Additionally, as pointed out by Baayen (1992), the productivity of a schema can have an effect on the token frequency of its lexemes: highly productive schemas will construct many nonce-formations with compositional meanings and low frequency, while unproductive schemas will concentrate frequencies around a small portion of highly frequent derivatives. While *-age* suffixation is the second most productive deverbal suffixation that derives event nouns according to the data available on the French web (Missud *et al.* 2020), there is no evidence for the productivity of conversion. Although this paper does not aim to measure the productivity of conversion, we would expect *-age* suffixation to be more productive than V to N conversion when deriving event nouns since converted nouns display higher frequencies overall. In the following sections, we investigate the hypothesis of a lack of homogeneity among converted nouns compared to *-age* derivatives by manipulating word vectors.

largest data points. Points that are shown outside the box and the lines are outliers, i.e. marginal datapoints that lie outside the pattern of the distribution.

2. SEMANTIC DISCRIMINATION OF MORPHOLOGICAL SCHEMAS

Our first task was to investigate the semantic relatedness between derivatives according to their morphological schema. Because we are manipulating word embeddings, semantic relatedness between lexemes can be measured by computing the cosine similarity between two word vectors. This measure gives a similarity score that ranges from -1 (diametrical opposition) to 1 (strict similarity) to a pair of vectors. As an example, the cosine similarity of *nettoyage* 'cleaning' and *lavage* 'washing' is 0.85, which means that *nettoyage* and *lavage* are remarkably similar in our model. In contrast, the cosine similarity of *nettoyage* and *fricassée* 'fricassee' is 0.4, meaning that their semantic similarity is low in comparison.

Cosine similarity can be used to determine how semantically related the lexemes that belong to a certain morphological group are on average. Consequently, we searched the corpus for word similarities in order to elucidate the semantic homogeneity of derivatives of each type. For example, if *-age* suffixed nouns are more semantically similar to their pairs than to V to N converted nouns, there could be a proper semantic identity that differentiates N-age from other eventive nouns. Similarly, if V to N converted nouns are closer to each other than they are to N-age, converted nouns might present specific semantic properties related to the schema they derive from. Differences between *-age* suffixation and V to N conversion could be found in the degree of semantic relatedness their derivatives share with their pairs.

Moreover, the different types of V to N converted nouns can be investigated as well. Although Tribout (2010) showed that the semantic interpretation of converted nouns globally remained the same regardless of the stem selected for derivation, DSMs could help find discrepancies in the distributional properties of converted nouns that have not yet been examined. By discriminating V to N converted lexemes based on the stem they are derived from, we explore the hypothesis that there might be a semantic differentiation between stem 0, stem 12 and stem 13 V to N conversion and aim to address several questions: do V to N converted nouns have higher semantic similarity scores with converted nouns that are derived from the same stem? Do some stems group more semantically homogeneous V to N converted nouns than others; for example: are stem 12 converted nouns closer to their counterparts than stem 0 converted nouns are?

First, we present a study of the semantic attraction of derivatives based on their closest semantic neighbors. Then, we complement this study by examining the degree of attraction for each schema.

2.1. *-age* suffixation vs. V to N conversion

The *n*-closest neighbors of a given word are the n words that have the highest cosine similarity scores with the word, ranked from the most similar

to the least similar (among the *n* most similar). For example, the 10 closest neighbors of *nettoyage* 'cleaning' are shown in (3).

(3) 1. *lavage* 'washing': 0.85 ; 2. *séchage* 'drying': 0.77 ; 3. *démontage* 'disassembling': 0.75 ; 4. *ramassage* 'pick-up': 0.73 ; 5. *remplissage* 'filling': 0.73 ; 6. *broyage* 'crushing': 0.72 ; 7. *rinçage* 'rinsing': 0.71 ; 8. *drainage* 'drainage': 0.71 ; 9. *désherbage* 'weeding': 0.71; 10. *polissage* 'polishing': 0.7

Here, *lavage* 'washing' is the closest neighbor of *nettoyage* 'cleaning', with a cosine similarity score of 0.85. The last neighbor, *polissage* 'polishing' is the one that has the lowest similarity score (0.7) among the 10 closest neighbors. For the purpose of this study, we computed the 10 closest neighbors of each derivative (except for stem 13 converted nouns that were too few in number to be considered for this experiment). In order to calculate the morphosemantic attraction between the lexemes, we counted the number of neighbors that belonged to the same morphological category as the tested derivative (*-age* suffixation, V to N conversion, stem 0 V to N conversion or stem 12 V to N conversion). For example, with (3), as all the neighbors are also N-age, *nettoyage* has an attraction score of 10/10, because all ten of its closest neighbors are derived from the same morphological schema. Each derivative was given an attraction score based on its neighbors. To make sure that the tested stem 0 and stem 12 converted nouns (for which fewer examples were collected compared to N-age) had equal chances of finding N-age or converted nouns derived from different stems in their neighbors, we downsampled the number of other derivatives. For example, when 71 stem 12 V to N converted nouns were tested, the cosine similarity scores were calculated on a sample of 71 stem 12 V to N converted nouns and 71 randomly chosen derivatives of other types.

We compared the distribution of attraction scores for *-age* suffixation, V to N conversion, stem 0 and stem 12 V to N conversion. If attraction scores are high for a great majority of lexemes of a certain type, semantic unity could be attributed to the schema they derive from.

Table 1 shows the attraction scores of the derivatives depending on their schema. The first row ("Range") shows the range of scores that lexemes derived from a schema can have. IQR ("interquartile range") shows the scores that are situated between the first and third quartile of the distribution (i.e. middle 50% of the lexemes' scores). The median is the value that separates the data into two halves: a median at 5 for a schema indicates that 50% of the scores are below 5 and that 50% are above 5. The last row indicates the average score (arithmetic mean) for each schema.

	N-age	V to N converted N	Stem 0 converted N	Stem 12 Converted N
Range	**8-10**	0-10	1-10	5-10
IQR	**9-10**	5-10	5-8	8-10
Median	**10**	**10**	6	9
Mean	**91%**	77%	63%	84%

Table 1. – Distribution of attraction scores

The results in Table 1 show that N-age are the derivatives that attract their pairs the most. The total N-age in our data can have 8 to 10 N-age in their closest neighbors while 50% have 9 to 10 (IQR). The median of 10 indicates that in most cases, no V to N converted noun can be found among the closest neighbors of N-age. On average, 91% of their neighbors are *-age* suffixed nouns as well.

Additionally (not shown in Table 1), we calculated the average proportion of stem 0, stem 12 and stem 13 V to N converted nouns that are found in *-age* suffixation's closest neighbors. On average, we found 6.4% of stem 0, 1.8% of stem 13 and 0.5% of stem 12 V to N converted nouns. Stem 0 converted nouns are by far the most frequent converted lexemes among the 10 closest neighbors of N-age, while stem 12 V to N converted nouns only appear marginally. Note that stem 13 converted nouns are underrepresented in our data (12 items) and are therefore less likely to appear among the neighbors or any derivative than N-age, stem 0 and stem 12 converted nouns. Their striking representativity among the closest neighbors of *-age* derivatives is thus more significant than that of stem 12 V to N converted nouns. These results indicate that stem 12 V to N converted nouns might not share *-age* derivatives' distributional properties, while stem 0 and stem 13 V to N converted nouns could have some semantic properties in common.

Although attraction scores are not as high for V to N conversion as a whole, the results in Table 1 show that a majority of V to N converted nouns are found among their 10 closest neighbors as well. While the number of V to N converted nouns among V to N converted nouns' neighbors ranges from 0 to 10, 50% of converted nouns have 5 to 10 converted nouns among their neighbors. With a median of 10, most converted nouns only attract converted nouns. On average, 77% of their neighbors are V to N converted nouns. Overall, V to N converted nouns attract their pairs, but not as strikingly as N-age do. Compared to V to N conversion, *-age* suffixation appears as more semantically homogeneous, a feature that has already been observed when comparing N-age with N-ment and N-ion (Wauquier *et al.* 2019, forthcoming).

Nonetheless, the results for stem 0 and stem 12 V to N converted nouns indicate that the semantic homogeneity varies greatly from one stem to the

other. Stem 0 V to N converted nouns can only be slightly discriminated from the other derivatives in our data. The number of stem 0 V to N converted nouns among the neighbors ranges from 1 to 10. As the median shows, 50% of stem 0 converted nouns have more than 6 stem 0 converted nouns among their neighbors. With an average of 63%, stem 0 V to N converted nouns do not attract their pairs as significantly as V to N converted nouns in the general case. Stem 12 V to N converted nouns have much higher scores in comparison. The number of stem 12 converted nouns found in the 10 closest neighbors ranges from 5 to 10, with 50% of them having 8 to 10 stem 12 V to N converted nouns among their neighbors. The median shows that half of stem 12 V to N converted nouns attract more than 9 of their pairs. On average, 84% of their neighbors are stem 12 converted nouns. Stem 12 conversion attraction scores are closer to those of *-age* suffixation than to those of V to N conversion (without discriminating stems).

While Tribout (2010, 2015) showed that the semantic classes to which converted nouns belonged were not strikingly distinguishable depending on the stem selected for derivation, these preliminary results reinforce our hypothesis that V to N converted nouns can in fact be discriminated. In the following section, we investigate the degree of attraction that lexemes maintain with each other depending on the schema they derive from.

2.2. Degrees of semantic attraction between the closest lexemes

As we are manipulating a small amount of data it is usual to come across closest neighbors that have low similarity scores with the tested derivative. For example, in the 10 closest neighbors of *regain* 'revival' (stem 0), we find *renouveau* 'renewal' (stem 0), with a satisfactory cosine similarity score of 0.63, but also *dégoût* 'disgust' (stem 0) with a score of 0.42, which would not intuitively be considered as semantically close to *regain* as *renouveau*. To get a better sense of the actual semantic relatedness of the neighbors of a derivative, the average cosine similarity scores of the neighbors of each derivative must be measured and compared according to the schema it derives from. This would help gain clearer insight into whether N-age' neighbors have higher similarity scores with their pairs at each position than V to N converted nouns' neighbors, thereby supporting the intuition that *-age* suffixation derives nouns that are more semantically homogeneous than conversion.

For each schema, we computed the average cosine similarity score at each position for all the neighbors that derive from the same schema. Neighbors are ranked from the closest (position 1) to the farthest among the 10 closest (position 10). For example, with *-age* suffixation, the cosine similarity scores of all N-age that appeared among N-age' neighbors were averaged at each position (1 to 10). When no N-age could be found for a position, the position was given a score of 0. This gave us a list of 10 scores for *-age* suffixation that

we could compare with the same lists of scores obtained for V to N conversion, stem 0, stem 12 and stem 13 conversion. This enables several questions to be addressed: on average, does the closest neighbor of an N-age have a higher similarity score than that of a stem 12 V to N converted noun? Do the closest neighbors of stem 12 converted nouns have higher similarity scores than stem 0 converted nouns? Although there are few of them, what are stem 13 converted nouns' similarity scores with their pairs?

Figure 2 shows the difference between schemas according to the average similarity scores of their neighbors.

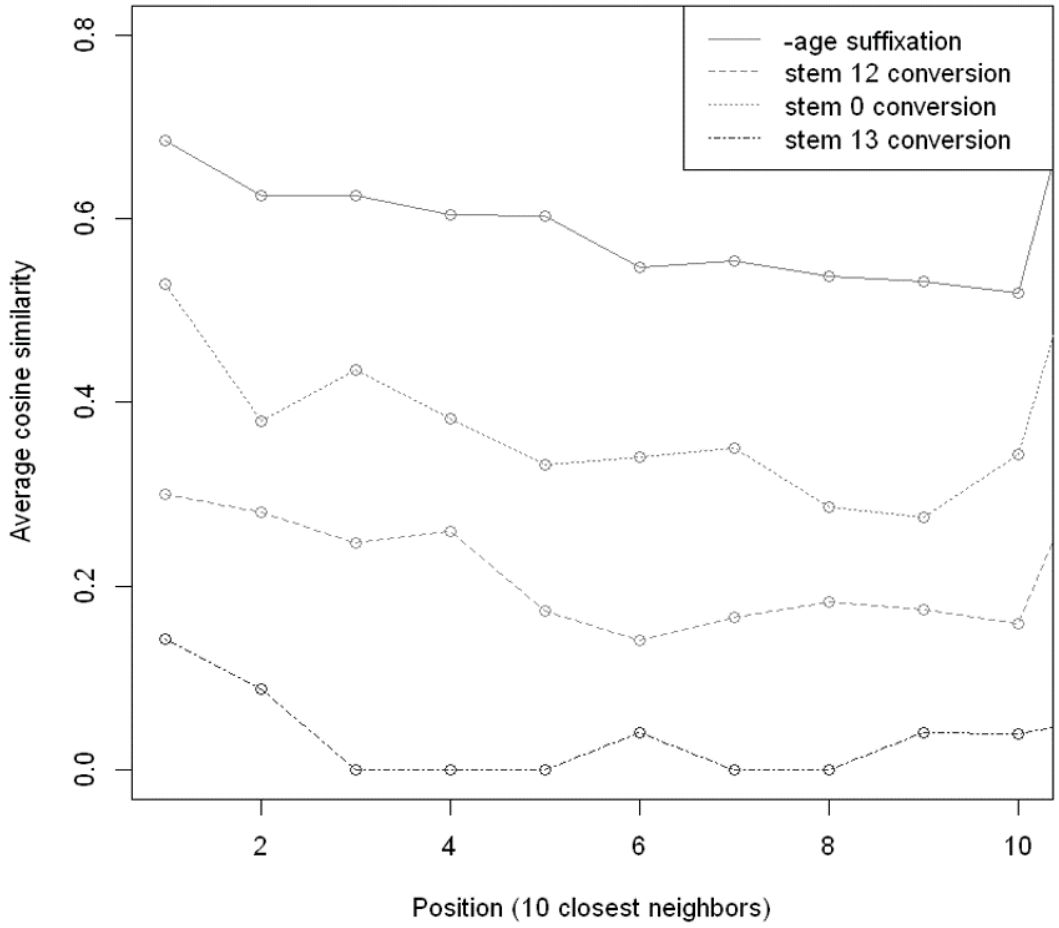

Figure 2. – Average cosine similarity scores of the neighbors according to their position

The X-axis indicates the position of the neighbors (from 1 to 10) and the Y-axis the average cosine similarity of the neighbors that appear at a certain position. Each schema is represented by a curve. A higher curve at each position means that the average cosine similarity of the neighbors derived from the same morphological schema is greater for a schema, therefore the semantic relatedness between nouns derived with the same morphological schema is higher. When the curve declines at a position before rising (as in position 2 for stem 12 V to N converted nouns), it indicates that few derivatives of the same schema are found at one position among the neighbors.

The results show that N-age are by far the ones that maintain the highest similarity scores with their *-age* suffixed neighbors regardless of the position. Their closest neighbors (in position 1) have an average score of approximately 0.7. The average cosine similarity scores do not fall below 0.55 (position 10). The other curves indicate the scores of converted nouns according to the stem they select. Among them, stem 12 V to N converted nouns are the ones that have the highest similarity scores with their neighbors at each position, the highest similarity score being at position 1 with a little more than 0.5 and the lowest score below 0.4 at position 9. Comparatively, stem 0 converted nouns have lower similarity scores at each position, with their highest score reaching 0.3 at position 1. Their scores can fall below 0.2 at position 6. Finally, stem 13 V to N converted lexemes are the ones with the lowest similarity scores. Their highest average similarity score reaches 0.2 at position 1. Unsurprisingly, as few stem 13 V to N converted nouns are taken into account, half of the positions (3, 4, 5, 7 and 8) have a score of 0 which means that no stem 13 converted noun could be found in the neighbors at these positions.

These results correlate with the ones presented in the previous section (Table 1). On average, N-age derivatives have the highest semantic proximity with their *-age* suffixed neighbors: they attract them more than the other schemas, and the ones they attract have high similarity scores. Converted nouns maintain a lower semantic proximity with their pairs in comparison. However, stem 12 V to N converted nouns stand out as they are more likely to attract their pairs with which they have the highest similarity scores. The semantic consistency observed among N-age and stem 12 converted nouns led us to hypothesize that the two schemas might occupy distinct semantic niches. In order to consolidate our intuition that the distributional properties of N-age and stem 12 V to N converted nouns might not overlap, we also looked at the farthest neighbors of N-age with the expectation that we would find a majority of stem 12 converted nouns.

2.3. The farthest neighbors of N-age

The 10 farthest neighbors of *-age* suffixed nouns include the derivatives that have the lowest cosine similarity scores, i.e. the nouns with which N-age

have the least in common. For example, the 10 farthest neighbors of *nettoyage* are given in (4). According to our data, *pensée* is the derivative that has the lowest similarity score (-0.17) with *nettoyage*.

(4) 1. *pensée* 'thinking': -0.17 ; 2. *vocalise* 'singing exercise': -0.16 ; 3. *revue* 'magazine': -0.15 ; 4. *huée* 'booing': -0.15 ; 5. *feinte* 'feint': -0.13 ; 6. *promesse* 'promise': -0.13 ; 7. *défaite* 'defeat': -0.13 ; 8. *dictée* ; dictation': -0.12 ; 9. *embrouille* 'confusion': -0.11 ; 10. *venue* 'arrival': -0.1

We calculated the number of stem 0, stem 12 and stem 13 V to N converted nouns that were found among the 10 farthest neighbors of all N-age in our data. Table 2 shows the proportion of derivatives according to the schema they derive from.

	N-age	Stem 0 V to N conversion	Stem 12 V to N conversion	Stem 13 V to N conversion
Range	0-2	0-7	**2-10**	0-2
IQR	0-1	2-4	**5-8**	0-1
Median	0	3	**6**	0
Mean	3%	30%	**61%**	4%

Table 2. – Proportion of derivatives among the 10 farthest neighbors of *-age* suffixed nouns

Unsurprisingly, the proportion of N-age does not exceed 2 and is closer to 0 as shown by the median. Of all V to N converted nouns, those that are derived from stem 13 are the ones that are the hardest to find. As shown in Table 2, there are as many N-age as stem 13 V to N converted nouns in the farthest neighbors (0 to 2); however, this may be due to their poor level of representativeness. Among stem 0 V to N converted nouns, between 0 and 7 appear in the farthest neighbors. In 50% of cases, we find 2 to 4 stem 0 V to N converted nouns (with a median of 3). Finally, stem 12 V to N converted nouns are the most numerous: 61% of the least similar derivatives of N-age in our data are stem 12 converted nouns. With a median of 6, the number of stem 12 V to N converted nouns found in the farthest neighbors ranges from 2 to 10. 50% of N-age have between 5 and 8 stem 12 V to N converted nouns among their farthest neighbors. Stem 12 V to N conversion is the type of conversion that has the least in common with *-age* suffixation.

2.4. Overview

Our results provided evidence in favor of *-age* suffixation's semantic homogeneity: N-age predominantly attract their pairs and maintain the highest

semantic similarity with them. Although V to N converted nouns are not as semantically homogeneous, stem 12 V to N converted nouns stand out as more likely to attract other stem 12 converted nouns with which they have the highest scores compared to other types of conversion. The results in Table 2 show that stem 12 V to N converted nouns are the derivatives that appear the most among N-age' farthest neighbors, which indicates that the two can be discriminated. These results are consistent with the hypotheses that were made in 1.4 regarding the potential homogeneity of less frequent N-age compared to highly frequent converted nouns. One reason could be that the two schemas occupy distinct semantic niches that do not overlap while stem 0 and stem 13 V to N conversion are more semantically versatile. To explore this idea further, we investigated how derivatives scatter or cluster depending on the schema they derive from.

3. SEMANTIC DISPERSION

In this section, we investigate the semantic dispersion of derivatives in order to explore whether schemas can be discriminated based on the degree of scattering and clustering of the lexemes that they construct. This experiment aims to confirm the quantitative observations given in Section 2. Considering previous results, N-age and stem 12 V to N converted nouns were expected to cluster the most with their pairs, while stem 0 and stem 13 converted nouns were expected to show a wider spread.

In order to quantitatively assess the semantic distribution of our lexemes, we built a classifier that discriminates schemas based on the dispersion of their lexemes using the cosine similarity scores that were presented in Section 2. If our intuitions are correct, we expect that a classifier based solely on attraction scores will confuse N-age with stem 12 V to N converted nouns as both behave in a similar fashion, and stem 0 converted nouns with stem 13 converted nouns as they are more semantically dispersed. We also expect that the classifier will be able to easily discriminate N-age and stem 12 V to N converted nouns from stem 0 and stem 13.

The classifier takes lists of scores as input vectors and predicts the schema of a noun by comparing its vector with centroid vectors of the same dimensions. For each derivative, we constructed an 11-dimensional vector: the first dimension corresponds to the proportion of nouns derived from the same morphological schema in the 10 closest neighbors of the derivative (as computed in 2.1.). For example, the first dimension of *nettoyage* 'cleaning' would be 1.0 as 10 out of its 10 neighbors are N-age (cf. 'prop.' in Table 3). The other 10 dimensions of the vector correspond to the cosine similarity scores of the 10 closest neighbors of the derivatives ranked from closest to

farthest among the closest (n1, ..., n10 in Table 3). Neighbors that are not derived from the same schema are assigned a score of 0.0.

	prop.	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
nettoyage	**1.0**	**0.85**	**0.77**	**0.75**	**0.73**	**0.73**	**0.72**	**0.71**	**0.71**	**0.71**	**0.7**

Table 3. – Example of an input vector (*nettoyage* 'cleaning')

To predict a schema for a noun, the classifier computes the Euclidean distance (2-norm distance) between its input vector (as in Table 3) and 4 centroid vectors representing each schema. In our case, a centroid vector is the average of all input vectors of nouns derived from the same schema, except for the input vector of the noun that is tested[3]. For example, stem 12 conversion's centroid vector will be the average of all stem 12 converted nouns' input vectors. The prediction is based on the smallest Euclidean distance between the input and the centroid vectors. Thus, if an input vector is closer to *-age* suffixation's centroid than to stem 0, stem 12 or stem 13 conversion's centroids, the noun corresponding to the input vector will be classified as an *-age* derivative.

Table 4 shows the results of the classifier through a confusion matrix. The rows indicate the schema predicted by the classifier and the columns the actual schema. Values in bold are the true positives (nouns that were correctly predicted by the classifier).

	-age	stem 0	stem 12	stem 13	total
-age	**127**	5	17	0	149
stem 0	5	**34**	11	1	51
stem 12	18	10	**38**	0	66
stem 13	0	18	5	**11**	34
total	150	67	71	12	300

Table 4. – Confusion matrix of the classifier's prediction

3.1. *-age* suffixation results

As shown in Table 4, the classifier predicted that 149 derived nouns were N-age as their input vectors were closer to *-age* suffixation's centroid vector. Among them, 127 are indeed N-age, while 5 are actually stem 0 V to N

[3] We used cross-validation to compute centroids that did not include the tested input vector.

converted nouns and 17 are stem 12 converted nouns in our data. None of them are stem 13 converted nouns.

Additionally, we calculated the precision, recall and F-1 scores using the results in Table 4. Precision is a measure that evaluates the sensitivity of a classification for a class (in our case, a schema) by dividing the number of correctly predicted instances by the number of predicted instances. Recall, on the other hand, measures the specificity; it is the fraction of the total amount of correctly predicted instances. Here, *-age* suffixation has a precision score of 85.2% and a recall score of 84.6%. The F1-score (or F-measure) is the weighted harmonic mean of precision and recall, which gives 84.8% for *-age* suffixation, meaning that the classifier is able to predict N-age correctly.

3.2. Stem 0 conversion results

51 stem 0 V to N converted nouns were predicted by the classifier. Among them, 34 are indeed stem 0 converted nouns, 5 are N-age and 11 are stem 12 V to N converted nouns while only one was confounded with stem 13 conversion. The classifier obtains a precision of 66.6%, a recall of 50.7% and an F1-score of 57.5% for this class, which is only slightly better than random guessing (50%).

3.3. Stem 12 conversion results

In total, 66 nouns were classified as stem 12 V to N converted nouns. 38 were correctly predicted, while 18 were N-age and 10 were stem 0 V to N converted nouns. As for *-age* suffixation, none were confused stem 13 converted nouns. The scores are slightly lower than those of stem 0 conversion: with a precision of 57.5%, a recall of 53.5% and an F1-score of 55.4%, the classification is not significantly better than random guessing.

3.4. Stem 13 conversion results

Lastly, the classifier predicted 34 stem 13 V to N converted nouns. Among them, 11 stem 13 converted nouns in our data were correctly predicted. 5 were actually stem 12 V to N converted nouns and 18 were stem 0 converted nouns. No stem 13 converted noun was mistaken for an N-age. Stem 13 conversion obtains a precision of 32.3%, a recall of 91.6% and an F1-score of 47.7%. Stem 13 converted nouns are the ones that the classifier is the least able to predict.

3.5. Conversion results

The performance of the classifier on converted nouns as a whole was measured by combining stem 0, stem 12 and stem 13 conversion results. The

precision score for conversion is 84.7%, while the recall is 85.3%. As for *-age* suffixation, conversion obtains an F1-score of 84.9% which shows that the classifier can easily discriminate the two schemas.

For the most part, our classifier based on closest neighbors' attraction scores successfully predicted the schema of a noun when the noun is highly clustered with its pairs and when it is highly dispersed. Consequently, *-age* suffixation (the most homogeneous) and stem 13 conversion (the most scattered) are the schemas for which the classifier exhibits the highest recall scores. As expected, the classifier has trouble discriminating between *-age* suffixed nouns and stem 12 converted nouns, which is not surprising as both attract a great proportion of their pairs with high similarity scores. Stem 12 converted nouns that the classifier mistakes for *-age* suffixed nouns are the ones that cluster the most (*traversée* 'crossing', *randonnée* 'hike', *découverte* 'discovery', *fricassée* 'fricassee', *gelée* 'frost', *étuvée* 'steaming'). The *-age* suffixed nouns that are confounded with converted nouns (mostly stem 0 and stem 13) are those that are less clustered such as *blocage* 'blocking', *apprentissage* 'learning', *vernissage* 'coating, vernissage' or *témoignage* 'testimony'. Unsurprisingly, stem 0 converted nouns that are scattered are predicted as stem 13 converted nouns. Stem 0 converted nouns can sometimes be mistaken with stem 12 converted nouns when they are clustered (*dégonfle* 'letting down', *démerde* '(action of) getting by', *débrouille* '(action of') dealing with it'). Again, the results obtained with the classifier show that the least frequent (N-age and stem 12 converted nouns) and the most frequent kinds of derivatives (stem 0 and stem 13 converted nouns), as shown in 1.4., are the ones that are the most frequently confounded.

Overall, the results show that the dispersion of lexemes differs significantly depending on the schema (although not always depending on the stem selected for conversion, as stated by Tribout 2010), thus confirming our intuitions that the semantic behavior of lexemes can be quantitatively discriminated. In the following section, we present a qualitative analysis of the semantic niches that were found for each schema.

4. SEMANTIC NICHES

We hypothesize that the groupings observed in the distribution of derivatives are motivated by the sharing of common semantic values related to the morphological schema the lexemes are derived from. Here, we define a semantic niche as an ensemble of lexemes derived from the same schema that gather semantically (in terms of cosine similarity) and that share a common semantic property that we identified.

4.1. Semantic niches of *-age* suffixation

The distribution of N-age led us to identify a semantic property characteristic of these nominalizations: these event nouns seem to

systematically imply either a concrete object for the realization of the process denoted by the base verb, or the linking of objects. The object concerned can be either an instrument (5) or an object (which can be a place) used as a container or a storage place (6). The linking of objects mainly concerns humans who are linked to one another via what the N-age denotes (7).

(5) *freinage* 'braking', *étiquetage* 'labelling', *patinage* 'skating, polishing', *broyage* 'shredding', *raffinage* 'refining', *forage* 'drilling'

(6) *remplissage* 'filling', *compostage* 'composting', *stockage* 'storage', *entreposage* 'storing', *garage* 'garage', *archivage* 'archiving', *recyclage* 'recycling'

(7) *mariage* 'wedding', *concubinage* 'cohabitation', *parrainage* 'patronage', *jumelage* 'twinning'

For N-age involving an instrument, several semantic clusters can be observed:

- Cluster referring to body care (8)

(8) *rasage* 'shaving', *gommage* 'exfoliation', *massage* 'massage', *modelage* 'body massage', *bronzage* 'suntan'

- Cluster referring to household activities (9)

(9) *lavage* 'washing', *rinçage* 'rinsing', *nettoyage* 'cleaning', *séchage* 'drying'

- Cluster referring to waste management (10)

(10) *ramassage* 'collection', *broyage* 'shredding', *désherbage* 'weeding', *compostage* 'composting', *recyclage* 'recycling'

- Cluster referring to cultivation, gardening (11)

(11) *arrosage* 'watering', *élagage* 'pruning', *épandage* 'spreading', *abattage* 'felling', *hivernage* 'wintering'

- Cluster referring to vehicles (12)

(12) *décollage* 'takeoff', *démarrage* 'start-up', *atterrissage* 'landing', *mouillage* 'anchorage', *allumage* 'ignition'

- Cluster referring to the organization/disorganization of data (13)

(13) *décryptage* 'decoding', *décodage* 'decoding', *paramétrage* 'configuration', *filtrage* 'filtering', *brouillage* 'jamming'

- Cluster referring to the manufacture of films, movies (14)

(14) *doublage* 'dubbing', *visionnage* 'viewing', *mixage* 'mixing', *montage* 'editing', *bruitage* 'sound effects', *coloriage* 'coloring', *reportage* 'report'

- Cluster referring to malicious acts (15)

(15) *cambriolage* ‘burglary’, *braquage* ‘hold-up’, *sabotage* ‘sabotage’, *pillage* ‘pillaging’

The *-age* suffixed nouns in our corpus comprise very few pure event nouns that do not imply an object. This result is in line with Wauquier *et al.* (forthcoming) that studies the semantic distinction between *-age* and *-ion* deverbals in French (also using DSMs) and that highlights the technical nature of N-age, which are more related to the fields of industry, agriculture or crafts[4]. This is also consistent with Fradin (2014)'s results showing that N-age, compared to N-ment, combine preferentially with complements denoting concrete objects.

4.2. Semantic properties of V to N conversion

The examination of the distribution of deverbal converted nouns showed the opposite: these are usually generic event nouns. However, they can be grouped into semantic clusters, especially depending on the stem of the base verb on which they are built.

4.2.1. Deverbal converted nouns deriving from stem 12

The conversion selecting stem 12 of the base verb provides the most clustering next to pure event nouns (16).

- Pure event nouns

(16) *venue* ‘arrival’, *plainte* ‘complaint’, *poussée* ‘thrust’, *avancée* ‘advance’, *battue* ‘beat’, *pensée* ‘thought’, *percée* ‘opening’, *remontée* ‘increase’, *astreinte* ‘constraint’.

- Clusters referring to meteorological events (17)

(17) *éclaircie* ‘sunny spell’, *accalmie* ‘lull’, *crue* ‘flood’

- Cluster referring to moving (18)

(18) *chevauchée* ‘horse ride’, *virée* ‘trip’, *tournée* ‘round’, *ruée* ‘rush’

4 Uth (2010) suggests that this semantic affinity between N-age and things of technical nature is linked to the multiplication of the suffix in the 19th century, during the industrial revolution.

- Cluster referring to activities in which a body part is involved (without this being considered as an instrument) (19)

(19) *étreinte* ‘hug’, *tétée* ‘breast feeding’, *enjambée* ‘stride’, *fessée* ‘smack’, *plumée* ‘pluck’, *suée* ‘sweating’

Note that converted nouns deriving from stem 12 have a cluster that involves instruments in the realization of the process, as N-age do (however, these nouns are most likely interpreted as result nouns):

- Cluster linked to culinary vocabulary referring to preparations, cooking methods, etc. (20)

(20) *fondue* ‘fondue’, *poêlée* ‘stir fry’, *fricassée* ‘fricassee’, *étuvée* ‘stew’, *gelée* ‘jelly’, *rôti* ‘roast’

4.2.2. Deverbal converted nouns deriving from stem 0

The conversion selecting stem 0 of the base verb forms pure event nouns, even if they never appear isolated, but grouped together with (i) either N-age (21) or (ii) converted nouns deriving from stem 12 (22).

(21) *retour* ‘return’, *réveil* ‘waking up’, *rappel* ‘recall’, *renvoi* ‘expulsion’, *rejet* ‘rejection’, *afflux* ‘influx’, *envol* ‘flight’, *aide* ‘help’

(22) *recherche* ‘search’, *repousse* ‘regrowth’, *rééquilibre* ‘rebalancing’, *décroît* ‘decrease’, *grogne* ‘discontent’, *surchauffe* ‘overheating’, *attaque* ‘raid’, *baisse* ‘drop’, *chute* ‘fall’, *dérive* ‘drift’

Only two semantic clusters appear, depending on whether the converted nouns are grouped with N-age or converted nouns deriving from stem 12: the semantic cluster of nouns referring to psychological acts when grouped with N-age (23) and the semantic cluster referring to the price of a purchase when associated with converted nouns deriving from stem 12 (24).

(23) *dégoût* ‘distaste’, *mépris* ‘contempt’, *rejet* ‘rejection’, *repli* ‘withdrawal’, *aveu* ‘confession’

(24) *enchère* ‘bid’, *détaxe* ‘tax reduction’, *rabais* ‘discount’

4.2.3 Deverbal converted nouns deriving from stem 13

The converted nouns that select stem 13 of the base verb also form pure event nouns, although they are very rare and always grouped either with N-age (25) or with converted nouns deriving from stem 12 (26):

(25) *suspense* ‘suspense’, plagiat ‘plagiarism’

(26) *défense* ‘defence’, *réponse* ‘answer’, *promesse* ‘promise’, *secousse* ‘tremor’, *course* ‘race’.

However, some of them, grouped together with some N-age, form the cluster of humans who are linked to one another via what the derivative denotes (27).

(27) *concordat* ‘concordat’, *attentat* ‘attack’, *assassinat* ‘assassination’

Investigating the distributional properties of converted nouns using word vectors has highlighted their genericity compared to N-age, a distinctive property that the analysis of the semantic outputs of V to N conversion and *-age* suffixation has not been able to show (Tribout 2015). Such genericity allows converted nouns to cluster with more specific N-age when some semantic properties of converted nouns resemble the prototypical ones of an N-age.

4.3. Overview

So far, our observations show that the schemas that are the most dispersed and the least semantically specific are also the schemas that group the most frequent derivatives (stem 13 and stem 0 conversion), while the ones that are the most specific and clustered comprise the least frequent lexemes (*-age* suffixation and stem 12 conversion). Although this correlation could be due to the inherent properties of each schema (including the different stems that are selected for conversion), another reason could be that some converted nouns are more lexicalized than others, as proposed by Resnik (1995). Such a hypothesis could explain why stem 12 converted nouns, the least frequent converted nouns in our data, cluster together away from other more frequent converted nouns. On the other hand, the potential polysemy of stem 13 and stem 0 converted nouns, which may be the consequence of their advanced lexicalization reflected by high frequencies, could explain why they can be found among *-age* suffixed nouns and stem 12 converted noun clusters: some of their semantic interpretations, far from the original (compositional) ones, might have something in common with nouns that are found in *-age* suffixation and stem 12 conversion clusters. These results lead us to consider that a discrimination based on distributional properties between converted nouns depending on the stem selected for derivation might depend on the degree of lexicalization of the converted nouns rather than on their semantics.

5. CONCLUSION

In this paper, we have proposed quantitative and qualitative analyses of the distributional properties of *-age* suffixed nouns and verb to noun converted nouns by manipulating word embeddings. Our methods, applied on a small set of data and based on word similarities, allowed us to measure the semantic

dispersion and relatedness of these lexemes while ensuring the interpretability of the results. We provide evidence that *-age* suffixation and verb to noun conversion have different semantic properties and distributional behaviors, and that converted nouns can be discriminated by the stem they derive from when looking at their distributional properties regardless of their semantic outputs. First, we showed that *-age* suffixation is the most semantically homogeneous schema as *-age* suffixed nouns mostly attract their pairs and have high semantic similarity scores with them. This is consistent with the frequency distribution of our data that shows that N-age are generally less frequent than converted nouns, which implies that *-age* suffixation might be more productive and therefore derive nouns that have a compositional meaning that is less influenced by the consequences of lexicalization compared to converted nouns. We also demonstrated that *-age* suffixed nouns cluster together and occupy semantic niches that we categorized. A semantic value that is specific to N-age is that the event nouns they denote imply, in one way or another, a concrete object. In comparison, V to N converted nouns appeared less homogeneous and more scattered at first glance. However, when differentiating them based on the stem they select, some regularities were found. V to N converted nouns that derive from stem 12 actually behave in a similar fashion to N-age: they attract a majority of their counterparts and cluster together to occupy distinct semantic niches that do not overlap with the ones invested by *-age* suffixation. Unlike N-age, stem 12 converted nouns denote event or result nouns that do not imply the use of a concrete object, a feature that is also found in the other converted nouns in our data. The other types of converted nouns, based on stem 0 and stem 13, lack semantic homogeneity and show a much wider spread. While their versatility could be consistent with the idea that rival schemas adjust their degree of specialization in order to coexist (Lindsay & Aronoff 2013) and could be an inherent property of both schemas, we hypothesize that another reason for such dispersion is that stem 0 and stem 13 converted nouns are more lexicalized because of their high frequencies and generic semantic interpretation. So far, although V to N converted nouns can be discriminated based on their stems, it appears more likely that they constitute a unique morphological schema that is preferentially used by selecting stem 12 when deriving event nouns, while stem 0 and stem 13 selection could have become obsolete. To evaluate this hypothesis, the productivity of V to N conversion and the different stems that are selected must be thoroughly investigated. As for the rivalry between *-age* suffixation and V to N conversion, their semantic relatedness and dispersion features will need to be compared with those of other rival nominalizations such as *-ion*, *-ment*, *-ance*, *-ade*, *-ure* and *-aison* and be matched against morphological, phonological and syntactical properties in order to clarify the distribution that allows them to coexist.

REFERENCES

ARONOFF M. (1994). *Morphology by Itself*. Cambridge: The MIT Press.

BAAYEN H. (1992). Quantitative aspects of morphological productivity. *Yearbook of morphology 1991*, 109-149.

BARONI M., BERNARDINI S., FERRARESI A., & ZANCHETTA E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 2009, vol. 43, no 3, 209-226.

BAUER L., LIEBER R. & PLAG I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

BONAMI O., BOYE G. (2003). Supplétion et classes flexionnelles dans la conjugaison du français. *Langages 152*, 102-126.

BONAMI O. & PAPERNO D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio*, 2018, vol. 17, no 2, 173-196.

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

DAL G., HATHOUT N., LIGNON S., NAMER F. & TANGUY L. (2018). Toile versus dictionnaires: Les nominalisations du français en -age et en -ment. In: F. Neveu, B. Harmegnies, L. Hriba & S. Prévost (éds), *Congrès Mondial de Linguistique Française* (CMLF), July 2018, Mons, Belgium: EDP Sciences.

DARMESTETER A. (1877) *De la création actuelle des mots nouveaux dans la langue française et des lois qui la régissent*. Paris: F. Vieweg.

DON J. (2004). Categories in the Lexicon. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 42(5): 931-956.

DUBOIS J. (1962). *Etude sur la dérivation suffixale en français moderne et contemporain*. Paris: Larousse.

FERRET K., SOARE E. & VILLOING F. (2010). Rivalry between French -age and -ée: the role of grammatical aspect in nominalization. In M. Aloni, H. Bastiaanse, T. De Jager & K. Schultz (eds), *Logic, language and meaning*, 17th Amsterdam Colloquium, The Netherlands, December 2009, Revised Selected Papers, Lecture Notes in Computer Science (Vol. 6042), Berlin: Springer, 284-295.

FERRET K. & VILLOING F. (2012). L'aspect grammatical dans les nominalisations en français: les déverbaux en -age et -ée. *Lexique* (20), 73-127.

FIRTH J. R. (1957). A synopsis of linguistic theory, 1930-1955. Reprinted in: Palmer, F. R. (ed.) (1968). *Selected Papers of J. R. Firth 1952-59*, London: Longmans, 168-205.

FRADIN B. (2014). La variante et le double. In: F. Villoing, S. David & S. Leroy (éds), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux* Nanterre: Presses Universitaires de Paris Ouest, 109-147.

FRADIN B. (2019). Competition in derivation: What can we learn from French doublets in-age and -ment ? In: F. Rainer, F. Gardani, W. U. Dressler & H. C. Luschützky (eds), *Competition in Inflection and Word-Formation*. Studies in Morphology, vol. 5., Cham: Springer, 67-93.

HARRIS Z. S. (1954). Distributional structure. *Word*, vol. 10, no 2-3, 146-162.

KERLEROUX F. (1999). Identification d'un procédé morphologique: la conversion. *Faits de langue* n°14, 89-100.

KISSELEW M., RIMELL L., PALMER A., & PADO S. (2016). Predicting the direction of derivation in English conversion. In: M. Elsner, S. Kuebler (eds), *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. 2016, 93-98.

LANDAUER T. K., FOLTZ P. W., & LAHAM D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 1998, vol. 25, no 2-3, 259-284.

LIEBER, R. (2005). English Word-Formation Processes. In : P. Štekauer & R. Lieber, (eds), Handbook of Word-Formation, Studies in Natural Language and Linguistic Theory, vol. 64. Dordrecht: Springer Netherlands, 375-427.

LINDSAY M. & ARONOFF M. (2013). Natural selection in self-organizing morphological systems. In: N. Hathout, F. Montermini, J. Tseng (eds), *Morphology in Toulouse: Selected Proceedings of Décembrettes*, 2013, vol. 7, 133-153.

MARANTZ A. (1997). No Escape From Syntax: Don't Try Morphological Analysis in the Privacy of Your Own Lexicon. In: A. Dimitriadis *et al.* (eds), *Proceedings*

of the 1998 Penn Linguistics Colloquium (University of Pennsylvania working papers in Linguistics: vol. 4.2), 201-225.

MARTIN F. (2010). The semantics of eventive suffixes in French. In: M. Rathert, A. Alexiadou (eds), *The Semantics of* Nominalizations *across Languages and Frameworks*. Berlin: Mouton de Gruyter, 109-141.

MARZO D. (2013). Italian verb to noun conversion: the case of nouns in *-a* deriving from verbs of the 2nd and 3rd conjugation. *Revista de Estudos Linguísticos da Univerdade do Porto*, vol. 8, 869-87.

MIKOLOV T., CHEN K., CORRADO G., & DEAN, J. (2013). Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at ICLR*. arXiv:1301.3781v1.

MISSUD A., AMSILI P. & VILLOING F. (2020). VerNom: une base de paires morphologiques acquise sur très gros corpus. In: C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla & S. Schneider (eds), *Actes de la 27ème conférence sur le Traitement Automatique des Langues Naturelles* (TALN 2020), 8-19 juin 2020, Nancy: ATALA, 305-313.

NYROP C. (1936). *Grammaire historique de la langue française* III, "Formation des mots". Geneva: Slatkine Reprints.

PLAG I. (2003). *Word-formation in English* (CAMBRIDGE TEXTBOOKS IN LINGUISTICS). Cambridge: Cambridge University Press.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In: C. S. Mellish (ed), *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, Montreal, August 1995, 448-453.

SCHÄFER R. & BILDHAUER F. (2012). Building large corpora from the web using a new efficient tool chain. *LREC*. 2012, 486-493.

SCHÄFER R. (2015). Processing and querying large web corpora with the COW14 architecture. In: P. Banski, H. Biber, E. Breiteneder, M. Kupietz, H. Längen & A. Witt (eds), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster: UCREL IDS, 28-34.

TRIBOUT D. (2010). *Noun to verb and verb to noun conversions in French*. PhD Dissertation, Université Paris Diderot (Paris 7).

TRIBOUT D. (2012). Verbal stem space and verb to noun conversion in French. *Word Structure*, 5 (1), 109-128.

TRIBOUT D. (2015). Problèmes de compositionnalité en morphologie dérivationnelle: le cas de la conversion. Verbum XXXVII(2), 235-255.

UTH M. (2010). The rivalry of the French nominalization suffixes -age and -ment from a diachronic perspective. In: M. Rathert, A., Alexiadou (eds), *The Semantics of Nominalizations across Languages and Frameworks*, Berlin: Mouton de Gruyter, 215-244.

VALERA S. (2014). Conversion. In: R. Lieber, P. Štekauer (eds), The Oxford Handbook of Derivational Morphology. Oxford: Oxford University Press, 154-168.

WAUQUIER, M., HATHOUT, N. & FABRE, C. (2019). Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns. In: J. Audring, N. Koutsoukos, C. Manoulidou (eds), *Rules, patterns, schemas and analogy: Online Proceedings of the 12th Mediterranean Morphology Meetings, University of Ljubljana, June 27-30, 2019*, vol. 12. University of Patras, 111-121.

WAUQUIER M., FABRE C., HATHOUT N. (forthcoming). Différenciation des noms d'action dérivés: le facteur de technicité étudié en corpus In: Frérot, C., Pecman, M. (éds), *Les corpus numériques à la modélisation linguistique en langues de spécialité*, Grenoble: UGA éditions, 1-13.