Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns

Marine Wauquier, Nabil Hathout, Cécile Fabre

Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns

Marine Wauquier Nabil Hathout Cécile Fabre
CLLE, CNRS, University Toulouse - Jean Jaurès
marine.wauquier@univ-tlse2.fr nabil.hathout@univ-tlse2.fr cecile.fabre@univ-tlse2.fr

1. Introduction

The Distributional Hypothesis, proposed by Harris (1954), Firth (1957), and Miller and Charles (1991) among others, states that the semantic proximity between words is reflected in the proximity of their distribution. This principle has been captured in distributional semantics models (DSMs) where words are represented as context vectors (Sahlgren 2008; Lenci 2018; Boleda 2020). In these models, the Euclidian distance between two vectors gives a measure for the semantic proximity of the represented words. On this basis, the geometrical representation of the meaning allows various mathematical operations on vectors that can be used to approximate semantic operations such as compositionality, disambiguation, analogy, etc. (Baroni et al. 2014).

This distributional approach of meaning is widely used in Natural Language Processing (Fabre & Lenci 2015), and plays an increasingly important role in linguistics (Lenci 2018, Boleda 2020). It has proven to be successful in the analysis of various linguistic phenomena such as meaning shifts (Kulkarni et al. 2015; Hamilton et al. 2016), the semantic irregularity of derivation (Bonami & Paperno 2018) or the comparison of gender contrasts for nouns and adjectives (Mickus et al. 2019).

With similar objectives, we present a large-scale study of the distributional properties of derived agent nouns in French. We consider distributional semantics as a tool to re-examine some classical morphological questions with an extensive approach of morphology (Hathout et al. 2008). To this end, we perform the analysis of a large number of contexts in order to define distributional profiles of lexemes and families of lexemes. The objective of this study is thus twofold: we aim at validating and renewing some linguistic hypotheses, while defining the operational conditions of DSMs use for linguistic analysis.

More precisely, could a large scale analysis of the distributional properties of derived lexemes give access to the semantic specialization of suffixes within derivational families? We look at the comparison, in French, between masculine agent nouns in *-eur* derived from verbs (e.g. *chanteur* ‘singer’, *administrateur* ‘administrator’) and their feminine equivalents in *-euse* and *-rice* (e.g. *chanteuse* ‘female singer’, *administratrice* ‘female administrator’). We investigate the ability of DSMs to capture the agentive meaning of these suffixes and to identify the specific profile of each of them.

In the following (section 2), we first review the criteria traditionally used for the semantic distinction between these suffixes. We then present our experimental setup (section 3) and the first results of the study, based on a distributional abstract representation at the level of the lexemes families (section 4).

MMM12 Online proceedings 2020
J. Audring, N. Koutsoukos & C. Manouilidou (Eds)

2. Agentive suffixation in *-eur*, *-euse* and *-rice*

Work related to suffixal competition aims to explain how one form prevails over the other in pairs like *-ee* (*attendee*) and *-er* (*attender*) in English (Heyvaert 2011) or *-iste* (*chimiste* 'chimist') and *-ien* (*physicien* 'physicist') in French (Lignon 2007). In our study, we focus on the French agentive suffix *-eur* and its rival feminine equivalents *-euse* and *-rice* such as (*sculpteur* 'sculptor', *sculpteuse* and *sculptrice* 'female sculptor'). To the best of our knowledge, very few studies compare feminine and masculine equivalents on one hand, and the feminine suffixes with each other on the other hand, except in psycholinguistic or sociolinguistic studies (Burr 2003; Lenoble-Pinson 2008).

These three suffixes coin agent nouns (*acheteur* 'buyer') or instrument nouns (*réfrigérateur* 'refrigerator') derived from verbs (*acheter* 'to buy', *réfrigérer* 'to refrigerate'), and in some cases from nouns (*camion* 'truck' → *camionneur* 'truck driver'). An agent noun designates the animate entity who intentionally undertakes the action described by the base verb, while an instrument noun designates the prototypical artefact used to undertake the action (Huyghe & Tribout 2015).

The semantic distinction between the masculine suffix and the feminine suffixes has evolved over the time. Initially, the *-eur* suffix formed agent nouns (*moissonneur* 'harvester') while *-euse* and *-rice* formed instrument or tool nouns (*moissonneuse* 'combine harvester') derived from the same base (*moissonner* 'to harvest') (Dubois 1962). This distinction is said to have gradually disappeared as the use of machines and the automation of work increased (Dubois 1962), but no diachronic study has confirmed this hypothesis yet.

The masculine and feminine suffixes *-eur*, *-euse* and *-rice* differ with respect to the referential gender (the gender of the denoted person) of the resulting agent noun. In the same way as the suffix *-trice* in Italian (*attrice* 'actress'), *-in* in German (*Autorin* 'female author') or *-ess* in English (*huntress*), the *-euse* and *-rice* suffixes indicate the feminine gender of the agent noun. Various studies have highlighted additional semantic values of the feminine linked to cultural expectations, as illustrated by the distinction between *mister* and *mistress* in English (Hellinger 2001; Marcato & Thüne 2002;), or *entraîneur* 'coach' and *entraîneuse* 'coach/barmaid' in French. However, these studies are sparse and focus mainly on formal rather than semantic aspects (Schafroth 2001).

When *-euse* and *-rice* coexist, the two suffixes are also not strictly identical: they bear sociolinguistic connotations (Dawes 2003). The suffix *-rice* is considered nobler and more gratifying than the suffix *-euse*, seen as depreciative (Houdebine-Gravaud 1998; Dawes 2003; Lenoble-Pinson 2008), as illustrated by nouns like *directrice* 'female manager' and *sénatrice* 'female senator' on one hand and nouns like *coiffeuse* 'female hairdresser' and *vendeuse* 'female seller' on the other hand. This tendency also exists in other Romance or Germanic languages: the French suffix *-esse* and its equivalents in Italian *-essa*, Romanian *-esa* and German *-ess* all have strong connotations (Meurice 2001; Marcato & Thüne 2002; Bußmann & Hellinger 2003; Dawes 2003). These connotations are either sexual or depreciative. Note that suffixes with no specific connotation also exist, like the Italian *-trice* and German *-in*.

These previous studies are mainly based on the analysis of a limited set of examples. With the distributional semantics approach we have the opportunity to extend them on a larger scale. Few works have used this method. Among them, we can cite Zeller et al. (2014) who show that the difference of referential gender is correlated to a variation in distributional proximity between the masculine and feminine agent nouns. Varvara et al. (2016) use distributional information to distinguish between alternative nominalization processes in German. Mickus et al. (2019) compare the gender alternation in derived nouns and adjectives and show that the semantic contrast is more regular for inflection than derivation.

In the present study, we examine the hypothesis that the nouns suffixed in *-euse* and *-rice* are the feminine equivalents of the *-eur* derived agent nouns on a semantic level, and that a distinction between the two feminine suffixes can be established on a distributional basis. Our main contributions are: (i) the use of operational semantic representations that can be easily compared and (ii) the treatment of large set of derivational relations in order to get better grounded results.

3. Experimental setup

In this study, we combine various sources of information: results from distributional semantic tools are confronted to linguistic knowledge validated by experts. The word2vec tool provides the semantic representations; the linguistic resource Lexeur provides the morphological descriptions.

3.1 Lexeur

We base our study on a derivational linguistic resource, Lexeur, containing 5974 agent nouns suffixed in *-eur*. This resource is an inventory of agent nouns in *-eur* within their derivational families (Fabre et al. 2004). Nouns were taken from the French dictionary *Trésor de la Langue Française*, and completed by words harvested from the web. In Lexeur, each noun in *-eur* is manually associated to a part of its derivational family composed by its base (whether verbal or nominal) and a list of nominalizations of its base. The resource was then completed by the feminine equivalents in *-euse* and *-rice* of the agent nouns in *-eur*. This addition has been performed within the Démonette project (Hathout & Namer 2014). Each lexeme of the resource is associated with a morphosyntactic description. Five entries of Lexeur are shown in table 1.

Table 1: 4 entries from Lexeur

fraudeur 'fraudster'	*fraudeuse* 'female fraudster'	*frauder* 'to defraud'	V	*fraude* 'fraud'
agrimenseur 'land-surveyor'	*agrimenseuse* 'female land-surveyor'	Ø	Ø	*agrimensation* 'land-surveying'
sculpteur 'sculptor'	*sculpteuse*; *sculptrice* 'female sculptor'	*sculpter* 'to sculpt'	V	*sculpture* 'sculpture'; *sculptage* 'sculpting'
auto-stoppeur 'hitchhicker'	*auto-stoppeuse* 'female hitchkicker'	*auto-stop* 'hitchkicking'	N	Ø

Table 1 shows the diversity of the derivational families included in Lexeur: some are complete such as *sculpteur*, other are incomplete, such as *agrimenseur* (with no identified base) or *auto-stoppeur* (which has an agent derivative but no other nominal). 78% of the agent nouns in the resource are deverbal, 14% are derived from a noun, and 8% are not associated with a base of any kind (such as *agrimenseur*). All the *-eur* agent nouns have at least one feminine equivalent, but the suffixes *-euse* and *-rice* are not represented in the same proportions: there are 3 times as many feminine agent nouns in *-euse* as in *-rice* (4542 vs. 1514). Only 1% of the agent nouns in *-eur* have both a *-euse* and a *-rice* feminine equivalent, such as *sculpteur* in table 1. Nouns for which several bases can be identified, such as *chasseur* 'hunter' (which can be derived from *chasser* 'to hunt' and *chasse* 'hunt') or *inflammateur*

'igniter' (from *inflammer* 'to ignite' and *enflammer* 'to ignite'), have distinct entries, one for each base.

We refer broadly to the *-eur* nouns as agent nouns, but Lexeur actually contains both agents (*chanteur* 'singer') and instruments (*détonateur* 'detonator'). Moreover, the lexemes present in the resource are polysemous to various degrees (*navigateur* both meaning 'sailor' and 'browser').

3.2 Distributional Semantic Models

Distributional Semantic Models (DSMs) are mathematical representations of word meanings based on their distribution in a given corpus, where each word is represented by a vector of real numbers. The first models are said to be count models as each dimension of a vector represents the degree of association of the word with each context in the corpus. Count-based models generally undergo a reduction of dimensions to make the matrix denser.

More recently, predictive models have been proposed. These models are computed by neural network-based tools such as word2vec (Mikolov et al. 2013) and fastText (Bojanowski et al. 2016). These models are trained to predict the words susceptible to appear in a given context through unsupervised machine learning. Predictive models have been popularized because of their performances, their efficiency in terms of processing cost, and their usability. These advantages come along with a greater degree of opacity. Contrary to count models, where each dimension is identifiable and corresponds to a particular context, the compression of the distributional information in few hundreds of dimensions makes it non interpretable.

The word2vec tool provides dense vector representations (or word embeddings) for each word in a given corpus based on its distribution. These representations can be used to determine the distributional neighbors of a word, to compute word similarity, to solve analogies. The score of similarity is based on the cosine distance of their vectors and ranges from 0 (no proximity) to 1 (strict equality).

The use of such distributional tools requires large corpora. We choose to work with the French Wikipedia corpus, built from the version of 2018, which contains about 900 million words. This choice is supported by our desire to have a vocabulary as large and diverse as possible, coming from various domains, in line with the vocabulary found in Lexeur.

In this study, the distributional model is computed from lemmatized words, with no consideration for the syntactic relations. We built five matrices[1], using the same default parameters for all of them: CBOW architecture, the training algorithm Negative Sampling, a frequency threshold of 5, an under sampling threshold of frequent words of 10-3, a window size of maximum 5, and 100 dimensions.

4. Prototypical lexical meaning

One of our goals is to build a representation of the semantic instruction prototypically associated to a given suffixation. Since this semantic abstraction is not directly instantiated in the corpus, we cannot compute its vector representation as we do for any other word. We only have access to the vectors of the lexemes constructed by this suffixation.

[1] It has been recently shown that predict models suffer from some instability due to stochastic methods implied in the unsupervised training (Pierrejean & Tanguy 2018). This instability results in variation in the space and its organization, meaning that the proximity between vectors may vary slightly from one model to another. To assess the validity of the results, it is advised to average the results over several DSMs. We arbitrarily choose to average the results over 5 models.

4.1 Prototypical representation of derivatives

We use a notion of prototypical derivative[2] which we consider as the representative of a given class defined by the suffixation and which encapsulates the semantics of this class. We define the meaning of this representative as the mean of the meanings of the words coined by this suffixation (we are following to this end Kintsch 2001 work on predication). The vector (or centroid) $\vec{D}$ of the prototypical derivative of a given suffixation *suff* is computed on the basis of the mean of the vectors $\overrightarrow{\text{Nsuff}_i}$ of the words constructed with *suff* as indicated in (1).

(1) $$\vec{D} = \frac{\sum_{i=1}^{n} \overrightarrow{Nsuff_i}}{n}$$

Once this centroid has been calculated, we assess the semantics it encapsulates through the observation of its 100 nearest neighbors. We consider these nearest neighbors as representatives of the morphosemantic class represented by the centroid.

We operationalize this as follows. We compute a centroid by suffix for each of the 5 models. In concrete terms, we average the 1675 masculine deverbal agent nouns in *-eur* from Lexeur whose frequency is greater or equal to 5, 302 feminine deverbal agent nouns in *-euse*, and 73 feminine deverbal agent nouns in *-rice*. As we intend to make a qualitative analysis of the 100 nearest neighbors of the centroid, we average the closest neighbors over the 5 models by keeping the 100 neighbors whose average proximity to the centroid over the 5 models is the highest. Because the corpus is lemmatized but not tagged, there is no restriction on the PoS of the neighbors.

4.2 The agentive meaning

Our aim is to assess whether DSMs can grasp the agentive meaning of *-eur* suffixation. We analyzed the 100 neighbors of the *-eur* centroid. The first 20 are reported[3] in table 2.

Table 2: 20 closest neighbors to the *-eur* centroid over 5 models

Neighbor		**Cosine**	**Neighbor**		**Cosine**
aspirateur	'vacuum cleaner'	0.685	*manipulateur*	'manipulator'	0.630
plombier	'plumber'	0.676	*bricoleur*	'handyman'	0.622
client	'client'	0.656	*rabatteur*	'beater, tout' 'reel'	0.621
machiniste	'machinist'	0.655	*soudeur*	'welder'	0.620
pousseur	'pusher'	0.655	*nettoyeur*	'cleaner'	0.619
mécano	'mechanic'	0.646	*mouchard*	'informer' 'cookie'	0.617
garagiste	'mechanic'	0.641	*opérateur*	'operator'	0.617
conducteur	'driver' 'conductor'	0.636	*électro-aimant*	'electromagnet'	0.616
déménageur	'mover'	0.634	*vibrateur*	'vibrator'	0.615

[2] We use the notion of prototype with respect to the idea of a gradual categorization (Kleiber 1990). We are trying to describe the derivative which would instantiate as much characteristic features of a given derivational category as possible.

[3] We do not report all the neighbors for practical reasons. The details of all the results of this study are however available at https://github.com/mwauquier/phdthesis/tree/master/comparison_agentive_suffixes .

lance-pierre	'slingshot'	0.630	*interrupteur*	'interrupter' 'switch'	0.613

First, we notice that not all the 100 nearest neighbors are nouns used for the computation of the centroid: the overlap rate is 45%. This overlap corresponds to the neighbors are constructed by the *-eur* suffixation. As for the rest, we count 16 compounds (mostly instruments like *marteau-piqueur* 'jackhammer', except *contremaître* 'foreman'), 5 borrowings (mostly instruments such as *pacemaker*), 9 simplex nouns (*client* 'customer'), and nouns coined by other agentive suffixes (*mouchard* 'informer', *technicien* 'technician', *armurier* 'gunsmith', *garagiste* 'mechanic' among others).

In order to assess the extent to which the centroid encapsulates the agentive meaning, we carried out a coarse-grained annotation of the semantic type of the neighbors[4], with the 4 following categories:

(i) "agent" for nouns denoting a human[5] described in relation to an action (*plombier* 'plumber');
(ii) "instrument" for nouns denoting material or immaterial objects described in relation to an action (*aspirateur* 'vacuum cleaner');
(iii) "polysemous" for nouns which allow for both the agent (human) and instrument (object) readings (*conducteur* 'driver'/'conductor');
(iv) "other" for all the remaining neighbors.

Among the 100 neighbors, we find 43 instruments, 32 agents, 21 polysemous nouns, and 4 neighbors that fall under the category "other". These nouns are *chien* 'dog', *tranquillisant* 'sedative', *gadget* gadget', and *accessoire* 'accessory'. While they do not strictly fall within the definition of agent or instrument nouns, they certainly share some of their semantic properties.

This first analysis suggests that the *-eur* centroid encapsulates both the agentive and instrumental meanings associated to this suffixation (96 out of 100 neighbors).

4.3 Feminine agent nouns

The results we just presented show that the agentive meaning of the *-eur* suffixation is visible to a certain extent in the average distribution of the agent nouns in *-eur*. We hypothesize that this agentive component is also salient in the average distribution of the feminine agent nouns in *-euse* and *-rice*. However we saw in section 2 that these two suffixes exhibit different behaviors and connotations on a semantic level with respect to the connotation associated with the agents denoted by the nouns. We can then hypothesize that the agentive dimension will not appear equally in the distribution of nouns in *-euse* and *-rice* and that their centroids' neighbors will delineate two different distributional profiles.

As for the agent nouns in *-euse*, we first analyze the 100 nearest neighbors of the centroid of the 302 feminine agent nouns in *-euse*. Table 3 presents the first 20 first neighbors of this centroid.

[4] We annotate the semantic type of the neighbor based on its nominal use if the word has multiple PoS, like *conducteur* 'driver'/ 'conductor', which also can be an adjective. We annotate the word as "other" if there is no nominal use.

[5] We assume here that agent nouns have human referents, as we wish to distinguish agents from instruments and inanimate effectuators.

Table 3: 20 closest neighbors to the *-euse* centroid over 5 models

Neighbor		**Cosine**	**Neighbor**		**Cosine**
cuisinière	'female cook' 'cooker'	0.702	*serveuse*	'waitress'	0.641
coiffeuse	'female hairdresser' 'dressing table'	0.686	*râpe*	'grater'	0.640
manucure	'manicurist' 'manicure'	0.683	*batteuse*	'female drummer' 'threshing machine'	0.638
ballerine	'ballerina' 'ballet shoe'	0.662	*chatte*	'female cat'	0.633
tireuse	'female marksman' 'tap'	0.661	*sauteuse*	'female jumper' 'frying pan'	0.629
stripteaseuse	'female striptease artist'	0.658	*cafetière*	'female cafe owner' 'coffee maker'	0.628
gitane	'gypsy woman'	0.658	*chauffeuse*	'female driver' 'low chair'	0.626
rôtissoire	'rotisserie'	0.648	*blonde*	'blonde'	0.626
lavallière	'ascot tie'	0.648	*cover-girl*	'cover girl'	0.622
barmaid	'female bartender'	0.644	*jolie*	'pretty'	0.616

Just as we saw for the *-eur* centroid's neighbors, the neighbors of the *-euse* centroid display various morphological types. We find only 27 nouns suffixed in *-euse* within the 100 neighbors of the centroid, the base being either a verb (*serveuse* 'waitress' → *servir* 'to serve') or a noun (*stripteaseuse* 'female striptease artist'→ *striptease* 'striptease'). The targeted suffixation is thus less salient than for *-eur*. The other suffixations represented in the neighborhood include *-ière* (*cafetière* 'café owner' / 'coffee maker'), *-oire* (*rôtissoire* 'rotisserie'), *-iste* (*modiste* 'milliner', *standardiste* 'receptionist'), *-arde* (*fêtarde* 'female partygoer'), and converts in relation with verbs (*râpe* 'grater' ↔ *râper* 'to grate') and with adjectives (*jolie* 'pretty'). The *-euse* centroid neighborhood also includes simplex nouns such as *gitane* 'gypsy woman' or *poupée* 'doll', and borrowed nouns (*cover-girl* 'cover girl', *chapka* 'ushanka'). The relative homogeneity observed around the *-eur* centroid neighborhood is weaker for *-euse*. This suggests that the semantics delineated by the *-euse* centroid neighbors is not specifically associated to the suffix *-euse*.

On a semantic level, the neighbors for the *-euse* centroid differ slightly from the *-eur* centroid. We still find agent, instrument and polysemous nouns, but not in the same proportions. They include 20 agent nouns (*serveuse* 'waitress', *pêcheuse* 'female fisherman'), 18 instrument nouns (*essoreuse* 'spin dryer', *mortaiseuse* 'mortiser'), 23 polysemous nouns with both agent and instrument readings (*cafetière* 'coffee maker' / 'female cafe owner', *perceuse* 'female piercing artist' / 'drill') and 39 neighbors that fall in another semantic category, such as *chatte* 'female cat'. The semantic distribution for *-euse* neighbors is significantly different from that of *-eur* neighbors (Pearson chi-squared p-value <0.01). This seems to indicate that the semantics of *-euse* suffixation does not just have the agentive and instrument components. We can hypothesize that this additional component corresponds to the semantic specificity of *-euse* suffixation.

Indeed the 39 neighbors labeled 'other' display some regularities, with numerous animals (*chatte* 'female cat', *hérissonne* 'female hedgehog', *ponette* 'female pony', *tigresse* 'tigress', *crevette* 'shrimp', *cochonne* 'female pig'), pieces of clothing (*jupe* 'skirt', *salopette* 'overalls', *doudoune* 'winter jacket', *gourmette* 'chain bracelet'), or food (*tartelette* 'tartlet') and nouns denoting humans which can hardly be described as agents (*midinette* 'starry-eyed girl', *mémère* 'old lady').

Note that many of those nouns have a negative reading accounting for mostly sexual or connotated features or behaviors: *chatte* or *doudoune* also are in French slang terms to describe female anatomical attributes (respectively 'pussy' and 'big breast'), *tigresse* and *cochonne* 'sex maniac' are used to denote women based on their sexual appetite, and *bimbo* 'bimbo', *brune* 'brunette' and *jolie* 'sweetheart' (when used as a noun) refer to women on the basis of their physical appearance. As for the object and instruments nouns, they are strongly associated with fashion and with the culinary field, which are stereotypically considered to be feminine. Even if chefs and fashion designers are mostly men, objects like *jupe* 'skirt', *coiffeuse* 'dressing table', *manucure* 'manicure', *bigoudi* 'hair curler' or *guêpière* 'bodice' are mainly associated with women.

Overall, we notice that the neighbors denoting human or agent (animate more generally) mostly, if not exclusively, denote female referents. When there is no hint of the referential gender based on the suffixation, as in the case of *modiste* 'milliner', *fleuriste* 'florist', *standardiste* 'receptionist', *manucure* 'manicurist', or *dactylo* 'typist', there still are sociocultural expectations regarding the gender of the human denoted, mostly in favor of the female one.

There are however some neighbors, mostly instruments, that do not seem to be specifically linked to feminine referents or activities such as *tondeuse* 'lawnmower', *dameuse* 'snow groomer', and *batteuse* 'threshing machine', which are technical entities. These nouns emphasize the instrument dimension of the *-euse* centroid and reflect a well-documented use of the suffix as already discussed in section 2.

In summary, the distributional behavior of *-euse* and *-eur* centroids differs with respect to the agentivity (or instrumentality) of the neighbors, the *-euse* suffix exhibiting a greater degree of heterogeneity. This heterogeneity may be explained by the feminine dimension of *-euse* suffixation. However this difference does not exclusively rely on the referential gender of the agent nouns, since sexual or connotated meanings emerge. It tends to show that feminine agents are described with respect to their bodies and behaviors, while masculine agent nouns are used in a more neutral way to describe professions or status.

We now turn to the analysis of the 100 nearest neighbors of the centroid computed for the 73 feminine agent nouns in *-rice*. An overview of the first 20 neighbors is given in table 4.

The neighbors of the *-rice* centroid display a morphological variety similar to the one of *-euse*: only 34% of the neighbors are constructed with the *-rice*, *-euse* or *-eure* suffixes (respectively 17%, 13% and 4%), as illustrated by *médiatrice* 'female mediator', *régisseuse* 'female manager' and *pasteure* 'female pastor'). Among the other derived neighbors, we notably find the suffixes *-ienne* (*plasticienne* 'female plastic artist', *généticienne* 'female geneticist'), *-ière* (*parolière* 'female lyrics writer') and *-iste*. (*modiste* 'milliner') which all are agent nouns. We also find compounds such as *auteure-compositrice* 'female singer-songwriter' or neoclassical compounds such as *synostose* 'synostosis'. Like for *-euse*, the neighborhood also includes simplex nouns (*agente* 'female officer', *cheffe* 'female manager') and borrowed nouns (*scripte* 'script girl').

Table 4: 20 closest neighbors to the *-rice* centroid over 5 models

Neighbor		**Cosine**	**Neighbor**		**Cosine**
cofondatrice	'female cofounder'	0.681	*plasticienne*	'female plastic artist'	0.567
co-fondatrice	'female co-founder'	0.637	*youtubeuse*	'female YouTuber'	0.554
fondatrice	'female founder'	0.621	*standardiste*	'receptionist'	0.547
traductrice	'female translator'	0.597	*Réso-Liain*	-	0.544
sculptrice	'female sculptor'	0.596	*poétesse*	'poetess'	0.544
directrice	'female director'	0.593	*danseuse*	'female dancer'	0.543

ingénieure	'female engineer'	0.585	*astrophysicienne*	'female astrophysicist'	0.541
esthéticien	'esthetician'	0.579	*entrepreneure*	'businesswoman'	0.540
professeure	'female professor'	0.570	*blogueuse*	'female blogger'	0.538
cheffe	'female manager'	0.567	*angiopathie*	'angiopathy'	0.535

On the semantic level, the neighbors can be characterized with respect to the 4 semantic types defined in 4.2: there are 47 agent nouns, 3 instrument nouns, 7 polysemous nouns and 43 other nouns. This distribution differs significantly from that of *-euse* (Pearson chi-squared p-value <0.01) and of *-eur* (Pearson chi-squared p-value <0.01) at this level of annotation: there are far more agents in the neighborhood of the *-rice* centroid than for *-euse* and *-eur*, less instruments and less polysemous nouns. The agentive feature seems therefore more salient for *-rice* than for *-euse*, and the instrumental dimension almost nonexistent.

If we take a look at the nouns labeled ', we see that they are as numerous as for *-euse*, the same tendency can be observed, with the difference that there are no objects and there are fewer human nouns (*petite-nièce* 'great-niece', *canadienne* 'female Canadian'). Interestingly, the human nouns do not refer to sexual behavior or physical appeareance as we observed for *-euse*, except for the noun *transsexuelle* 'transgender woman'. The strong connotation found for the *-euse* centroid is not noticeable.

In contrast, we have a lot of nouns from the scientific and medical domains such as diseases (*tétraparésie* 'tetraparesis', *gliose* 'gliosis', *acanthose* 'acanthosis'), entities related to anatomy (*synostose* 'synostosis', *intrafusales* 'intrafusal', *paracrine* 'paracrine'), molecules and substances, such as *flavoprotéine* 'flavoprotein', *neurohormone* 'neurohormone', *thyronamine* 'thyronamine' or *radiosource* 'radiosource'. The degree of agentivity or non agentivity of these nouns is debated in the literature: while they have no animacy or intentionality, they might be considered as effectuators insofar as they deploy energy to autonomously carry out an action (Van Valin & Lapolla 1997). On this basis, if effectuators are considered to some extent as agents, the presence of such neighbors reinforce the agentive dimension exhibited by the *-rice* centroid.

As for *-euse*, the animate neighbors mostly refer to feminine referents. Interestingly, we find a few neighbors which are grammatically masculine, such as *costumier* 'costumer' or *esthéticien* 'esthetician'. These two nouns appear in two different situations in the corpus. In the first case, the masculine agent noun can actually be a feminine agent noun wrongly lemmatized. In the second case, the masculine agent noun is correctly lemmatized, and the referent is a man. As for *esthéticien* in particular, it is interesting to note that the feminine equivalent *esthéticienne* mostly refers to the woman who gives beauty treatment (as in 'beautician'), whereas the masculine form mostly (but not exclusively) refers in the corpus to a specialist in artistic theory (as in 'aesthetician'). The fact that the masculine 'nobler' equivalent is found in the *-rice* centroid neighborhood but not in the one of *-euse* accentuates the difference of connotation observed between *-euse* and *-rice*, namely that the agents denoted by the animated nouns are more socioculturally valued for the *-rice* centroid than for the *-euse* centroid.

We note that the *-euse* and *-rice* centroids share 11 neighbors. They are listed in (2).

(2) *coiffeuse* 'female hairdresser', *manucure* 'manicure', *ballerine* 'ballerina/ballet shoe', *barmaid* 'female bartender', *modiste* 'milliner', *standardiste* 'receptionist', *Youtubeuse* 'female YouTuber', *dactylo* 'typist', *call-girl* 'call girl', *snowboardeuse* 'female snowboarder', *hackeuse* 'female hacker'

They all are human agent nouns or nouns allowing both the agentive and instrumental readings. With the agentive readings, most of them refer unambiguously to a feminine agent, except for *manucure*, *modiste* and *standardiste*, which are gender neutral although there are strong sociocultural expectations regarding the feminine gender of the referent. The semantic content shared by the two centroids seems to include both the agentive and the feminine meanings. This goes in line with previous observations, as we saw that *-euse* and *-rice* neighborhoods contained agent nouns, and that *-euse* is more strongly influenced by the instrumental dimension.

5. Conclusion

We have seen in this article that DSMs are an effective tool for the study of morphosemantic classes on large corpora. We represented the meaning of the French *-eur* suffixation based on the average distribution of the *-eur* agent nouns. With the same methodology, we compared the semantic features of the *-euse* and *-rice* suffixations, and confirmed at a larger scale differences previously discussed in empirical study, through the examination of few hundreds of distributional neighbors. Our study of the average distribution of the agent nouns in *-euse* and *-rice* shows that *-euse* suffixation has a stronger instrumental meaning and conveys sexual and biased constructions, while *-rice* suffixation is mostly associated to agentivity and displays a higher degree of valorization, as shown by its frequent use in sciences and medicine.

We intend to develop further the methodological aspects of this study by reconsidering the experimental settings of the analysis. First, we want to reduplicate this study with another corpus to evaluate to what extent the results depend on the corpus. Additionally, we chose to study all agent nouns with frequency greater or equal to 5, in order to have a large set of derivatives, but this low frequency threshold may have undesired consequences on the distributional representation. This issue involves reconsidering the size of the samples used: we computed centroids based respectively on 1675, 302 and 73 nouns, and the increase of the frequency threshold would accentuate the disproportion already displayed by our data. We still have to assess the impact of these parameters on our results.

Acknowledgments

The DSMs were built using the OSIRIM computing platform that is administered by the IRIT computer science lab and supported by the National Center for Scientific Research (CNRS), the Région Midi-Pyrénées, the French Government, and the European Regional Development Fund (ERDF).

References

Baroni, M.; Bernardi, R. & R. Zamparelli. 2014. Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, 9: 241-346.

Bojanowski, P.; Grave, E.; Joulin, A. & T. Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Boleda, G. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*: Accepted. DOI: 10.1146/annurev-linguistics-011619-030303

Bonami, O. & D. Paperno. 2018. Inflection vs. Derivation in a Distributional Vector Space. *Lingue e linguaggio*, 17 (2): 173-196.

Burr, E. 2003. Gender and language politics in France. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 3: 119-139.

Bußmann, H. & Hellinger, M. (2003). Engendering Female Visibility in German. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 3: 141-174.

Dawes, E. 2003. La féminisation des titres et fonctions dans la Francophonie : de la morphologie à l'idéologie. *Ethnologies, Association Canadienne d'Ethnologie et de Folklore*, 25: 195-213.

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.

Fabre, C. ; Floricic, F. & N. Hathout. 2004. *Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur*. Communication aux journées d'étude sur La place des méthodes quantitatives dans le travail du linguiste. Université de Toulouse II-Le Mirail.

Fabre, C. & A. Lenci. 2015. Distributional Semantics Today - Introduction to the special issue. *Traitement Automatique des Langues*, 56(2): 7-20.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. In J. R. Firth (Eds), *Studies in Linguistic Analysis*. Oxford: Blackwell, 1-32.

Hamilton, W. L.; Leskovec, J. & D. Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, 1486-1501.

Harris, Z. S. 1954. Distributional Structure. *Word*, 10: 146-162.

Hathout, N.; Montermini, F. & L. Tanguy. 2008. Extensive data for morphology: using the World Wide Web. *Journal of French Language Studies*, *18*(1): 67-85.

Hathout, N. & F. Namer. 2014. Démonette, a French Derivational Morpho-semantic Network. *Linguistic Issues in Language Technology*, 11(5): 125-168.

Hellinger, M. 2001. English-Gender in a Global Language. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 1: 105-113.

Heyvaert, L. 2011. Attenders or Attendees? Deverbal -ee and-er Variants in English. *Journal of Pragmatics*, 43(1): 62-72.

Houdebine-Gravaud, A. M. 1998. L'imaginaire linguistique: questions au modèle et applications actuelles. *Limbaje şi comunicare*, 9-32.

Huyghe, R. & D. Tribout. 2015. Noms d'agents et noms d'instruments: le cas des déverbaux en -eur. *Langue française*, 99-112.

Kleiber, G. 1990. *La sémantique du prototype: catégories et sens lexical*. Paris : Presses Universitaires de France.

Kintsch, W. 2001. Predication. *Cognitive science*, 25: 173-202.

Kulkarni, V.; Al-Rfou, R.; Perozzi, B. & S. Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th Internation Conference on World Wide Web*, 625-635.

Lenci, A. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4: 151-171.

Lenoble-Pinson, M. 2008. Mettre au féminin les noms de métier : résistances culturelles et sociolinguistiques. *Le français aujourd'hui*, 73-79.

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*. Université Marc Bloch - Strasbourg 2, 287-295.

Marcato, G. & E. M. Thüne. 2002. Gender and Female Visibility in Italian. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 2: 187-217.

Meurice, F. 2001. Deconstructing Gender - The Case of Romanian. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 1: 229-252.

Mickus, T.; Bonami, O. & D. Paperno. 2019. Distributional Effects of Gender Contrasts Across Categories. *Proceedings of the Society for Computation in Linguistics*, 2, 174-184.

Mikolov, T.; Chen, K.; Corrado, G. & J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Miller, G. A. & W. G. Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and cognitive processes*, Taylor & Francis, 6: 1-28.

Pierrejean, B. & L. Tanguy. 2018. Towards Qualitative Word Embeddings Evaluation: MeasuringNeighbors Variation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. New-Orleans, 32-39.

Sahlgren, M. 2008. The Distributional Hypothesis. *Italian Journal of Linguistics*, 20(1): 33-54.

Schafroth, E. 2003. Gender in French - Structural Properties, Incongruences and Asymmetries. In M. Hellinger & H. Bußmann (Eds), *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company, 3: 87-117.

Van Valin, R. & R. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

Varvara, R.; Lapesa, G. & S. Padó. 2016. Quantifying regularity in morphological processes: An ongoing study on nominalization in German. In *Proceedings of ESSLLI DSALT Workshop: Distributional Semantics and Semantic Theory*.

Zeller, B. D.; Padó, S. & J. Snajder. 2014. Towards Semantic Validation of a Derivational Lexicon. In *Proceedings of the 25th International Conference on Computational Linguistics* (COLING). Dublin, 1728-1739.