Lombard Alizée, Wauquier Marine, Fabre Cécile, Hathout Nabil, Ho-Dac Mai & Huyghe Richard. 2022. Evaluating morphosemantic demotivation through experimental and distributional methods. *Lingvisticae Investigationes* 45(1), 83-115.

Evaluating morphosemantic demotivation through experimental and distributional methods

Alizée Lombard*, Marine Wauquier**, Cécile Fabre†, Nabil Hathout†,
Lydia-Mai Ho-Dac†, Richard Huyghe*

*Université de Fribourg - **Université Sorbonne Nouvelle - †Université Toulouse Jean Jaurès

Abstract

The lexicalization of morphologically complex words, i.e. their inclusion in the lexicon, can involve a loss of semantic compositionality. Such a phenomenon, called demotivation, has been overlooked in both morphological and lexical studies, notably regarding its gradual nature. This paper compares two measures of demotivation based on experimental and distributional semantics approaches. It builds on the evaluation of 78 pairs of French verbs and derived nouns selected to represent three levels of demotivation. The comparison of the two approaches using speakers' judgements and word vector similarity indicates convergence on the identification of demotivation degrees within a continuum, while also highlighting specific aspects of each method. The study provides direction to further research on morphosemantic demotivation, bridging together semantic, morphological and methodological considerations.

Keywords: demotivation, semantics, experimental approach, distributional semantics, lexicon

Introduction

The lexicalization of complex words, i.e., their inclusion in the lexicon, can induce a loss of phonological, morphological or semantic compositionality. As noted by many authors (Lipka, 1977, 1992, Bauer, 1983, Corbin, 1987, Blank, 2001, Brinton and Traugott, 2005, Hohenhaus, 2005, Hilpert, 2019; a.o.), conventionalized complex words often have idiosyncratic properties and the outcome of the lexicalization process is difficult to predict. As far as semantic properties are concerned, extralinguistic factors such as onomasiological needs and encyclopedic knowledge, as well as linguistic factors such as synonymy and lexical blocking, can influence the way complex words are fixed and evolve in the lexicon. As a corollary, the semantic analyzability and transparency of lexicalized complex words are highly variable. Some of them, although not completely predictable, maintain a clear semantic relation with their base words, whereas others are more opaque and semantically demotivated. Demotivation itself, as the obliteration of the morphosemantic relation between a base word and a derivative, has received little attention in morphological studies. It can be seen as a gradual phenomenon in which diachronic evolution and semantic change play an important role, but it is not clear how to evaluate such gradualness. The existence of a scale of demotivation has been suggested by authors like Roché (2004), but without any criteria to precisely identify degrees of demotivation between morphologically related words.

In this study, we investigate morphosemantic demotivation to both achieve a better understanding of its gradual nature and explore methods that can be used to evaluate it. To highlight the linguistic and cognitive

aspects of demotivation, we carry out experimental and distributional investigations, based on human judgements and corpus data. Demotivation is a transitional process that results in a loss of semantic transparency. It implies the existence of different states and degrees of semantic relatedness between morphologically related words. Although demotivation as a diachronic process is rarely investigated in a systematic way, studies in psycholinguistics and distributional semantics have previously explored semantic transparency, either when examining its effect on mental representations and cognitive processing (Marslen-Wilson et al. 1994, Longtin et al. 2003, Libben 2010, Gagné et al. 2017, a.o.), or when analyzing the distributional similarity between bases and derivatives (Marelli and Baroni 2015, Padó et al. 2016, Bonami and Paperno 2018, Wauquier 2020, Varvara et al. 2021, a.o.). Building on these studies, we investigate demotivation through its effect on semantic transparency, and through the differences of transparency that can be observed between historically related words. More precisely, we focus on the morphosemantic demotivation of nouns derived from verbs in French. The semantic relatedness of variously motivated verb-noun pairs is empirically studied by using and comparing experimental and computational methods. We examine whether gradual demotivation is consistently identified by speakers and observed in distributional data, and to what extent human and automatic assessments of demotivation converge.

The article is organized as follows. In Section 1, we present the linguistic material used in the study. Sections 2 and 3 are devoted respectively to the experimental and distributional semantics investigations. In Section 4, we compare and discuss the results obtained through both approaches. We conclude our study in Section 5.

1 Material selection

The experimental material assembled for this study is composed of verb-noun pairs that pertain to three categories corresponding to three degrees of semantic (de)motivation:

- C1: complete demotivation (e.g. *partir* 'leave'/*partage*'sharing')
- C2: partial demotivation (e.g. *mouiller* 'wet'/*mouillette* 'bread soldier')
- C3: complete motivation (e.g. *danser* 'dance' /*danseur* 'dancer')

In this section, we describe the characteristics of these verb-noun pairs and present the method used to collect them.

1.1 Characteristics of the verb-noun pairs

The following criteria are valid for all selected verb-noun pairs:

(1) a. The noun can be analyzed as being formally derived from the verb by means of a regular suffixation.
b. The semantic relationship between the verb and the noun is etymologically attested.

The criterion (1-a) indicates that the form of the noun must have a final sequence identical to a deverbal suffix and the remaining part must be identical to one of the verb stems (e.g. *danseur* 'dancer' where *-eur* is the suffix and *dans-* a stem of *danser* 'dance'). Since the psycholinguistic experiment and the computational modeling are performed on orthographic forms, phonological forms are not considered. For example, we did not include a pair like *allouer* 'allocate'/*alouette* 'lark' where (1-a) is verified phonologically but not orthographically because the 'l' is not doubled in the noun. The criterion (1-b) ensures that the formal proximity is not accidental and in particular that the C1 pairs have undergone a demotivation process. We have thus discarded pairs such as *crever* 'die'/*crevette* 'shrimp', where the connection is merely orthographic — the etymon of *crevette* is not linked to *crever* but to *chevrette* 'young goat'.

The three categories are defined as follows. Pairs selected for category C1 are totally demotivated: the meaning of the nouns cannot be analyzed in synchrony as related to the meaning of the verbs in any way, as in *partir* 'leave'/*partage* 'sharing', where the original semantic connection is now totally lost. At the other end of the scale, the meaning of nouns in C3 pairs is totally transparent with respect to the base verbs and the semantic instruction of suffixes. This is for example the case of the pair *danser* 'dance'/*danseur* 'dancer' in which *danseur* denotes 'a person who dances'. Category C2 is intermediate between C1 and C3. The meanings of C2 nouns and verbs are still linked, but in a less obvious way. For example, in *mouiller* 'wet'/*mouillette* 'bread soldier', the noun does not denote an instrument used to wet something, but a small piece of bread that is dipped in a boiled egg. The bread is somehow "wet", even though in French, the verb *mouiller* would not normally be used to describe this type of event. C2 can be considered as a "negative" category for pairs that neither fall under C1 nor C3. As a possibly heterogeneous category linking the two ends of a continuum, C2 provides an insight on the gradualness of demotivation.

The selection of the experimental material was guided by the following principles:

(2) a. An equal number of pairs must be selected for categories C1, C2 and C3.
b. Various suffixes and semantic categories must be represented, in equal proportions in the three categories.
c. The comparability of the pairs must be ensured in terms of word length and frequency.

The first criterion (2-a) is required to facilitate the exploitation of the results in the experimental approach. The second criterion (2-b) makes it possible to introduce as a parameter in the study the suffix and the semantic category of the noun (object, human, location, etc.), to balance the heterogeneity of the three categories. The third criterion (2-c) is intended to control for two types of biases: the number of syllables in paired words, which could influence the assessment of a semantic link between the verb and the noun, and frequency effects, whose impact on both distributional models [Bullinaria and Levy, 2007, Sahlgren and Lenci, 2016] and speakers' mental lexicon [Rayner and Duffy, 1986, Meunier and Segui, 1999, Baayen et al., 2016] is well documented. Because of the rarity of fully demotivated nouns (C1), the strict application of the last two constraints proved to be very difficult to put into practice. As many C1 pairs as possible have been collected and their number and distribution served as a reference for selecting materials in the other categories. Adjusting to the reality of the French lexicon, we built C2 and C3 data sets so as to obtain balanced samples based on the three above-mentioned criteria.

1.2 Data collection

The selection and classification of the materials as C1, C2, or C3 items were carried out jointly by the authors. A list of 16 suffixes that are known to create deverbal nouns in French was first established. [1] We started by selecting C1 pairs, which turned out to be very rare and, as a consequence, required broad-scale lexicons to identify a set as large as possible. We compiled a list of nouns ending with each suffix by using various lexicons extracted from French dictionaries such as GLàFF [Hathout et al., 2014] or Anagrimes[2] and concordances from large corpora such as FRCOW16A [Schäfer and Bildhauer, 2012]. For each suffix, we searched for nouns for which the relation to the verb that once existed in diachrony is not perceptible in synchrony anymore, so as to select the most strongly demotivated possible pairs. In order to satisfy condition (1-b), the etymology of candidate nouns was systematically checked in various dictionaries such as *Trésor de la langue française* [Dendien and Pierrel, 2003] and Wiktionnaire.[3] This step was followed by an adjudication and 26 C1 pairs were retained, as presented in the Appendix. It can be noted that only 8 out of the 16 initial suffixes are represented in the final list (*-ade*, *-age*, *-ance*, *-et*, *-ette*, *-eur*, *-oir*, and *-ure*). These were selected based on the condition that they were instantiated by at least two verb-noun pairs. Isolated pairs, such as *munir* 'supply'/*munition* 'ammunition' for the *-ion* suffix, were thus discarded.

[1] These 16 suffixes are: *-ade*, *-age*, *-aille*, *-aison*, *-ance*, *-ence*, *-erie*, *-et*, *-ette*, *-eur*, *-ion*, *-is*, *-ise*, *-ment*, *-oir*, *-ure*. Note that both masculine *-oir* (*présentoir* 'display stand') and feminine *-oire* (*passoire* 'colander') were initially considered for the *-oir* suffix.

[2] Website accessible at `https://anagrimes.toolforge.org/`.

[3] French Wiktionary accessible at `https://fr.wiktionary.org/`.

In a second step, we selected C2 and C3 pairs with the 8 remaining suffixes, following the same procedure as for C1 pairs, so as to satisfy criteria (2-b) and (2-c). C2 and C3 pairs were selected taking into account the length of each lemma and its frequency in a very large corpus, i.e., the 900-million words French Wikipedia corpus. We favored pairs that resembled C1 pairs with respect to these two characteristics. For example, for the *-age* suffix, the pair *échantillonner* 'sample'/*échantillonnage* 'sampling' was rejected in favor of pairs with 2 syllables such as *passer* 'pass'/*passage* 'passing', to get closer to the C1 and C2 pairs *partir* 'leave'/*partage* 'sharing' and *taper* 'hit'/*tapage* 'disturbance'. Whenever possible, nouns with identical semantic types were selected to ensure semantic homogeneity among C1, C2 and C3 categories. For example, for the polysemous suffix *-oir*, nouns denoting locations (e.g., *dépotoir* 'dump') and instruments (e.g., *sautoir* 'string') are represented in the three categories. The final dataset is given in the Appendix.

2 Experimental approach

Many studies in psycholinguistics investigate the effect of semantic transparency on the recognition and mental representation of morphologically complex words, whether these are derivatives [Marslen-Wilson et al., 1994, Longtin et al., 2003, Kielar and Joanisse, 2011, Smolka et al., 2019, Creemers et al., 2020, a.o.] or compounds [Dohmes et al., 2004, Smolka and Libben, 2017, Park et al., 2020, a.o.]. These studies examine semantic transparency as an explanatory factor for various aspects of word processing, but they do not investigate changes in transparency, nor do they consider the diachronic aspects of transparency loss. Furthermore, most of them operationalize semantic transparency as a binary variable by distinguishing transparent (i.e., compositional) and opaque (i.e., non-compositional) pairs. Other studies contrast transparent pairs with opaque ones consisting of pseudo-derivationally related words (e.g., *corn-corner*) [Rastle et al., 2004, Morris et al., 2013]. In the latter case, paired words do not have any morphological relation, which is an important difference with demotivated pairs of lexemes.

In this study, we explore semantic transparency as the effect of morphosemantic demotivation, and advocate the idea that the semantic relationship between verbs and nouns has to be evaluated in a scalar perspective, in line with the quantitative approach of Gagné et al. [2017]. We conduct an experiment to measure the degree of demotivation between historically related verbs and nouns, based on French native speakers' judgements.[4] According to our hypotheses, demotivated pairs (C1) should elicit judgements of lower semantic proximity than motivated ones (C3), and judgements for semi-demotivated pairs (C2) should fall in-between.

2.1 Method

Before discussing the results, we present our experimental methodology, with regard to participant selection and experimental procedure.

2.1.1 Participants

Four hundred and eleven volunteer Bachelor students[5] from the University of Toulouse - Jean Jaurès (France) with a major in humanities participated in the experiment during two courses in linguistics. Their curriculum included various disciplines in the humanities, such as psychology, sociology, linguistics, and history.

Data were filtered in order to control for different sociological factors. We ensured the homogeneity of participants' age by removing data from people older than 25 years (i.e., 22 participants). Age certainly has an impact on the variation of native intuitions, since there are differences in speakers' lexicons according

[4]The experiment, including hypotheses, procedure, materials, and analysis plan, was preregistered on the OSF platform: `https://osf.io/fbtr6/?view_only=f3e66f3aa5dc4a029d6b059cd2f7039b`.

[5]Most participants were 1st-year Bachelor students, and those who were 2nd-year Bachelor had linguistics as a 'minor' course. This criterion guarantees that the metalinguistic knowledge of the participants is as close to that of naive speakers as possible, and justifies the comparison between participants' and experts' judgements. The impact of linguistic knowledge on demotivation perception could be further investigated in future studies.

to their age. Data from non-native speakers were also excluded (i.e., 80 participants), because of possible unstable intuition and indecisive semantic evaluation. As a result, data collected from 309 French native speakers aged between 17 and 25 years old ($\text{Mean}_{age} = 19.4$) were used for the study.

2.1.2 Procedure

The experiment took the form of two online surveys, each comprising one half of the experimental material in order to reduce possible fatigue effects. Participants completed surveys on their personal computer in approximately 15 minutes without knowing the purpose of the study. After completion, they were given a presentation of the project in the form of a short video.

Each survey included half of the verb-noun pairs selected in the experimental material, i.e., 39 stimuli each. Two training stimuli were displayed before the experimental ones, including a semantically transparent verb-noun pair (e.g., *chanter* 'sing'/*chanteur* 'singer') and a pair of morphologically unrelated words (e.g., *crever* 'die'/*crevette* 'shrimp'). A total of 82 stimuli, including 78 experimental ones, were thus used in the study.

Each verb-noun pair was presented separately on the computer screen with instructions asking participants to evaluate the semantic proximity between the two words on a scale from 0 (unrelated meanings) to 6 (maximal proximity).[6] Participants could also indicate that they did not know one or both words presented in each stimulus, in which case responses were excluded from the analysis (652 trials out of 11,244 were thus discarded, i.e., 5.8% of the data).

We trimmed the data by eliminating all data points with response times longer or shorter than 2 standard deviations from by-participant and by-category means (603 trials out of 10,592, i.e., 5.7% of the data), to ensure that we only consider trials in which participants were fully focused on the experiment. Although this procedure resulted in unbalanced distribution of responses across verb-noun pairs ($\text{Mean}_{obs} = 128$, $\text{Range}_{obs} = [62, 149]$), it does not affect the analysis since observations per level of motivation remain evenly distributed, with 3,210 responses for C1, 3,401 for C2, and 3,378 for C3. The number of responses per suffix and category can be seen in Table 3. Note that among the 309 participants, 29 answered the surveys only partially (from 9% to 91% completion with a mean of 25%). Responses collected in uncompleted surveys were nevertheless used in the analysis (143 trials out of 9,989, i.e., 1.4% of the data).

2.2 Results

This section presents the similarity scores observed per category, a description of the measured data, and inferential statistical analyses. We also discuss the variability of the results per category and per suffix.

As indicated in Figure 1, demotivated pairs (C1) obtain the highest proportion of low proximity scores (0 and 1), with a median of 1 ($\text{N}_{obs} = 3{,}210$, $\text{Range}_{C1prox} = [0, 3]$, $\text{Mean}_{C1prox} = 1.60$), whereas motivated pairs (C3) have the highest proportion of high proximity scores (5 and 6), with a median of 6 ($\text{N}_{obs} = 3{,}378$, $\text{Range}_{C3prox} = [5, 6]$, $\text{Mean}_{C3prox} = 5.18$). Semi-demotivated pairs (C2) have a median of 4 ($\text{N}_{obs} = 3{,}401$, $\text{Range}_{C2prox} = [2, 5]$, $\text{Mean}_{C2prox} = 3.31$), which supports their intermediate status between motivated and demotivated pairs. Note that C2 scores are more evenly distributed than C3 and to a lesser extent C1 scores, which reveals the relative heterogeneity of the former.

Statistical analyses are performed to further investigate the effect of categories C1, C2, and C3 on judgements of semantic proximity. Since the measured variable is ordinal with 7 levels and does not follow a Gaussian distribution, the results are analyzed through mixed-effects ordinal logistic regressions, using the *ordinal* package [Christensen, 2019] in R [R Core Team, 2015]. Three predictors of proximity scores are tested in different models: the category of the verb-noun pair (C1, C2 or C3), the verb log-frequency, and the noun log-frequency (computed from the Wikipedia frequencies mentioned in Section 1). The models also include random intercepts per participant and per verb-noun pair, which are special components in mixed

[6] Exact instructions in French were: *D'après vous, à quel point les sens du nom* ***chanteur*** *et du verbe* ***chanter*** *sont-ils proches ?* 'According to you, how close are the meanings of the noun *chanteur* 'singer' and the verb *chanter* 'sing'?'

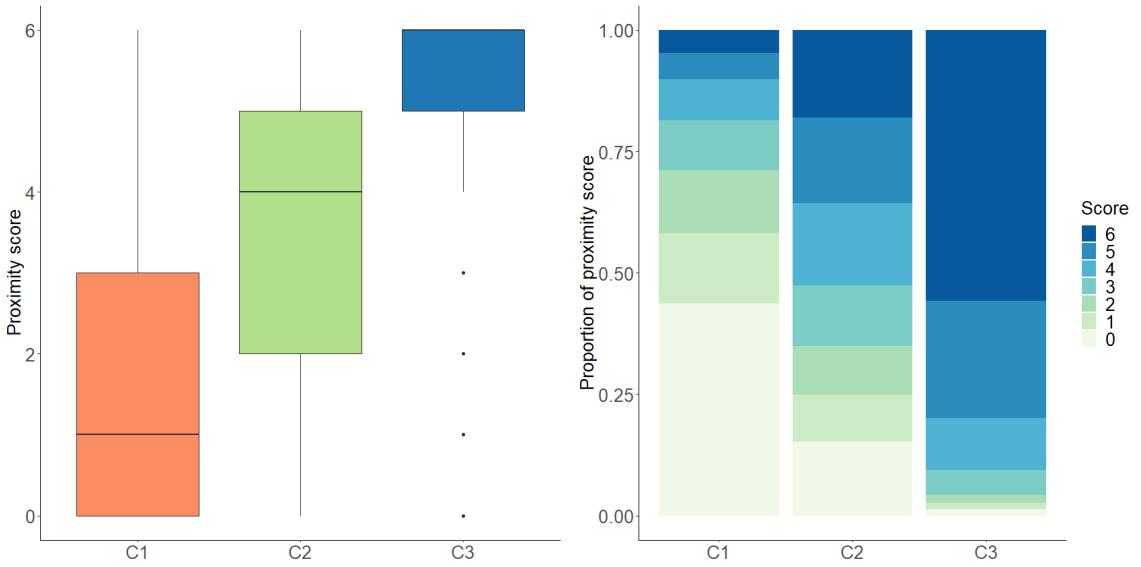

Figure 1: Experimental proximity scores between verbs and nouns per category

regression models that account for individual variation (such as participants assigning overall high or low scores to verb-noun pairs).

Neither verb frequency ($p = .656$) nor noun frequency ($p = .655$) have a significant effect on participants' judgements. A greater or lesser knowledge of base verbs and derived nouns does not directly influence semantic judgements. The model that best fits the data treats proximity scores as a function of the category only (C1, C2 or C3) as shown in Table 1, with a highly significant effect ($p <$ 2.2e-16). According to this model, C1 pairs have a higher probability of obtaining low scores (38% chance of 0 and 20% chance of 1), C2 pairs medium scores (19% chance of 3 and 5, 25% chance of 4), and C3 pairs high scores (28% chance of 5 and 55% chance of 6).

We conducted a post-hoc analysis of the category shifts effects. More precisely, we tested the significance of the pairwise differences between all categories within the regression model, using the function `lsmeans()` of the *lsmeans* package [Lenth, 2016]. The analysis shows that all differences are significant in the model, i.e., between demotivated and semi-demotivated (C1-C2, $p < .0001$), semi-demotivated and motivated (C2-C3, $p < .0001$), and demotivated and motivated pairs (C1-C3, $p < .0001$). Therefore, both the measured data and the inferential analyses confirm that speakers' intuitions about verb-noun semantic proximity are consistent with the experts' classification.

Cat.	**Estimate**	**SE**	***z***	***p***
C1	-2.0662	0.3375	-6.121	9.29e-10
C3	2.4149	0.3376	7.154	8.45e-13

Table 1: Results of the mixed-effects ordinal regression model for experimental scores (with C2 as intercept)

The variation observed within each category calls for further investigation. Figure 2 shows the variability observed between the average scores assigned to the different verb-noun pairs in each category. Motivated pairs (C3) are rather homogeneous, whereas demotivated pairs (C1) and in particular semi-demotivated pairs (C2) are more variable (see Table 2).[7] It can be noted that confidence intervals are lower for motivated pairs (C3) than for demotivated (C1) and semi-demotivated pairs (C2), which confirms a greater variability

[7]The semantically motivated pair *river* 'bind'/*rivet* 'rivet' has a surprisingly low experimental score (1.53), contrasting with all other C3 pairs. This could be explained by a lack of lexical knowledge, given that 46% of the participants declared that they did not know one or both words in this specific pair. The remaining 54% might have considered the frequent metaphorical meaning of *river* as 'focus, stare' rather than the literal meaning, hence the judgement of semantic distance.

in demotivated and semi-demotivated scores. Motivated pairs are clearly identified as such by speakers, whereas less concentrated scores for semi-demotivated and non-motivated pairs reveal the continuous nature of demotivation (see Section 4.2 for a more extended discussion).

Cat.	N_{pair}	Mean	SD
C1	26	1.67	1.61
C2	26	3.28	1.75
C3	26	5.09	1.06

Table 2: Mean and standard deviation of average proximity scores per verb-noun pair

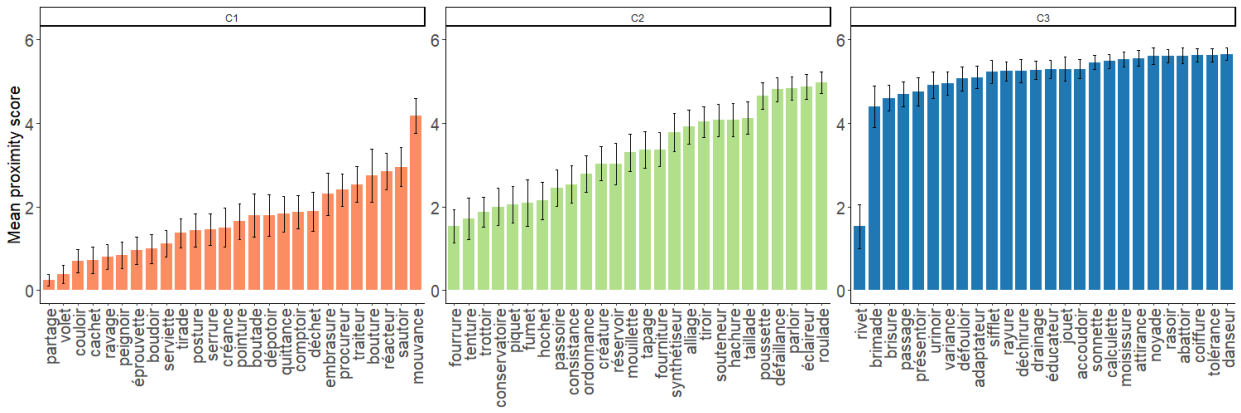

Figure 2: Average proximity score per verb-noun pair with .95 confidence intervals

The variation of experimental scores per suffix can be scrutinized as well. Important differences can be observed among suffixes, especially when comparing the distribution of semi-demotivated pairs (C2) with that of other categories (see Figure 3). In the case of *-age*, *-ade* and *-ette* suffixes, the distribution of proximity scores assigned to semi-demotivated pairs (C2) is more similar to that of motivated (C3) than to that of demotivated (C1) pairs, whereas in the case of *-ure*, *-oir* and *-et*, it is closer to that of demotivated (C1) than to that of motivated (C3) pairs. In the case of *-eur* and *-ance*, C2 scores seem more balanced between the scores of the two other categories. These tendencies are confirmed by the differences between the mean scores obtained for each suffix in each category (see Table 3). Differences between the 8 suffixes can also be observed in global measures of semantic proximity. The suffix *-eur* has the greatest mean proximity score and the lowest SD, and *-ade* and *-ance* are also associated with high means, whereas the mean proximity score of *-et* is very low. These observations suggest that there is an influence of the suffix on speakers' semantic judgements. However, the number of nouns per suffix considered in the experiment is too small to provide strong evidence for such an influence. In particular, the fact that *mouvoir* 'move'/*mouvance* 'trend' and *river* 'bind'/*rivet* 'rivet' are outliers in C1 and C3 respectively could have an effect on the distribution of *-ance* and *-et* items with respect to proximity scores. Further experiments with more linguistic material are therefore needed to confirm our observations here.

3 Automatic measurement of demotivation

The quantification of semantic similarity has benefited from the recent renewal of distributional semantics. Grounded on the distributional hypothesis [Harris, 1954, Firth, 1957], it relies on the idea that the meaning of words is a function of their contexts (i.e., of their linguistic distribution). Distributional Semantics Models (DSMs) provide a vector representation of meaning based on the co-occurrence of words in a corpus. In the resulting vector space, the spatial proximity between vectors approximates the degree of semantic similarity

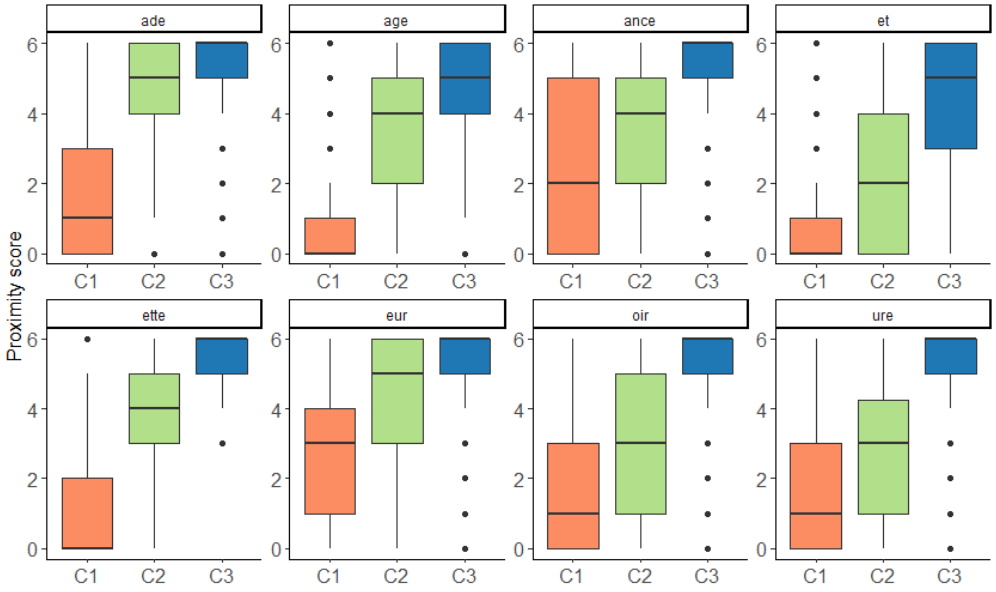

Figure 3: Statistical distribution of experimental scores per suffix

Suffix	**Cat.**	**N**$_{pair}$	**N**$_{obs}$	**Mean**	**SD**	**Mean**	**SD**
	C1	2	207	1.52	1.58		
-*ade*	C2	2	258	4.61	1.43	3.82	2.12
	C3	2	203	5.17	1.28		
	C1	2	263	0.54	1.08		
-*age*	C2	2	268	3.64	1.89	3.06	2.35
	C3	2	265	4.98	1.18		
	C1	3	339	2.48	2.13		
-*ance*	C2	3	408	3.39	2.03	3.84	2.14
	C3	3	416	5.38	0.98		
	C1	3	365	0.87	1.47		
-*et*	C2	3	356	2.09	1.96	2.38	2.33
	C3	3	316	4.43	2.03		
	C1	2	276	1.03	1.43		
-*ette*	C2	2	255	4.00	1.79	3.48	2.32
	C3	2	274	5.46	0.77		
	C1	3	409	2.59	1.87		
-*eur*	C2	3	383	4.26	1.70	4.07	1.94
	C3	3	426	5.33	1.05		
	C1	6	780	1.49	1.81		
-*oir*	C2	6	829	3.02	2.11	3.26	2.31
	C3	6	805	5.21	1.17		
	C1	5	571	1.81	1.87		
-*ure*	C2	5	644	2.81	2.04	3.38	2.24
	C3	5	673	5.25	1.07		

Table 3: Experimental proximity scores per suffix

between the corresponding words by means of a score ranging from 0 (no similarity) to 1 (maximum similarity) [Lenci, 2018, Boleda, 2020], usually based on the cosine or euclidean distance between the vectors.

This approximated degree of similarity can be used as a clue for compositionality and semantic motivation. Although distributional similarity does not strictly equate to compositionality, it can be hypothesized that for a given nominalizing suffix with a given semantic function, the variation in distributional similarity between base verbs and derived nouns will be correlated with variation in compositionality. Indeed, DSMs have often been used to estimate semantic compositionality [Reddy et al., 2011, Marelli and Baroni, 2015, Gagné et al., 2017]. More generally, distributional measures can be employed to capture various types of morphosemantic relations. For instance, Varvara et al. [2021] propose a measure of distributional inclusion to estimate the strength of the relationship between nominalizations and their base verbs, whereas Bonami and Paperno [2018] use difference between vectors to evaluate the consistency of derivational and inflectional relations in terms of shifts in vector space.

In line with these studies, we use distributional semantics to automatically quantify the semantic similarity between verbs and nouns in the C1, C2 and C3 pairs as an approximation of their semantic transparency. We base our assessment on the distance between the vectors of the verb and the noun in a given pair.

3.1 Design

In this study, similarity between a verb and a related noun is computed through two measures of distributional proximity. The first one is the cosine measure between the vectors associated with the verb and the noun, which ranges from 0 for no proximity to 1 for strict identity. The other measure is the rank of a noun among distributional neighbors of the corresponding verb, based on cosine similarity. The closer the rank is to 1, the more similar it is to a given word. Ranking differs from cosine measure in that it is strictly relative. Two vectors might be the closest in the vector space while having a low cosine proximity score, simply because other neighbors all have even lower proximity scores. Thus ranking provides a different perspective from cosine measure, which gives us two easily accessible distributional measures. Henceforth, these measures will be referred to as P for cosine proximity score, rankB for the rank of the verb in the neighborhood of the derived noun, and rankD for the rank of the derived noun in the neighborhood of the verb.

Because fully motivated pairs (C3) are characterized by a higher semantic transparency than (semi-)demotivated C2 and C1 pairs, we expect the former to display a high proximity score P. Likewise, we expect C3 nouns to have a low rank in the verb neighborhood (rankD), and C3 verbs to have a low rank in the noun neighborhood (rankB). On the contrary, we expect demotivated pairs (C1) to have low P scores and C1 nouns (C1 verbs, resp.) to have high rankD values (rankB, resp.). As for semi-demotivated pairs (C2), we expect them to display in-between values, both in terms of proximity scores and rankings.

All these measures are computed from a vector space concatenating 5 DSMs trained with Word2Vec [Mikolov et al., 2013] on the French Wikipedia corpus lemmatized with the Talismane parser [Urieli, 2013]. The DSMs training parameters are: CBOW, Negative Sampling, frequency threshold of 5, window size of 5, vector size of 100 dimensions.

3.2 Results

Figure 4 presents the distribution of the C1, C2 and C3 pairs according to proximity score P. As expected, motivated pairs (C3) display the highest proximity scores, with a median of 0.363 (Range_{C3prox} = [0.085, 0.595], Mean_{C3prox} = 0.351); demotivated pairs (C1) display the lowest proximity scores, with a median of 0.121 (Range_{C1prox} = [0.001, 0.500], Mean_{C1prox} = 0.129); and semi-demotivated pairs (C2) fall in between, with a median of 0.157 (Range_{C2prox} = [<0.001, 0.538], Mean_{C2prox} = 0.206). These results are confirmed by Table 4 which presents the mean and standard deviation of proximity score for the three categories.

Two complementary observations can be made from Figure 4. First, it appears that semi-demotivated pairs (C2) are closer to demotivated pairs (C1) than to motivated pairs (C3) if we consider median proximity scores. Second, semi-demotivated pairs (C2) have a higher dispersion than demotivated pairs (C1) and, to

a lesser extent, than motivated pairs (C3). Demotivated pairs (C1) form the most homogeneous group, and semi-demotivated pairs (C2) the most heterogeneous one. This suggests that demotivated pairs (C1) all tend to be characterized by a low semantic proximity, while semi-demotivated pairs (C2) have a more variable proximity (although low on average), as confirmed by standard deviation in Table 4. It can also be noted that similarity scores overlap between C1 and C2 categories, and to a lesser extent between C2 and C3 categories. This characteristic is much more pronounced than with experimental measures, which suggests that the distinction between the three categories is less clear-cut in the distributional analysis.

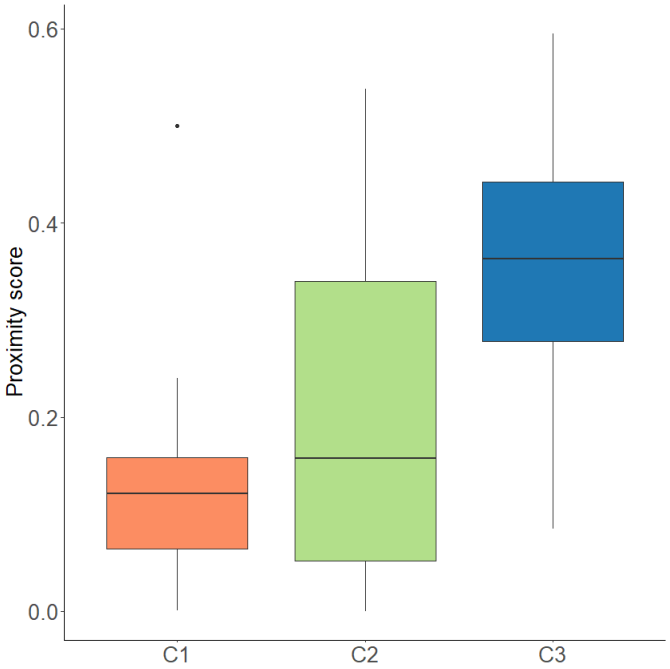

Figure 4: Proximity score P between the verb and its derived noun per category

Category	$\mathbf{N}_{pair}$	**Mean**	**SD**
C1	26	0.129	0.099
C2	26	0.206	0.181
C3	26	0.351	0.151

Table 4: Mean and standard deviation of distributional proximity scores

Similar observations can be made with respect to rankB and rankD. Figure 5 shows that C3 nouns appear at lower ranks in the verb neighborhood (rankD), with a median of 1,682 ($\text{Range}_{C3rankD}$ = [7, 474,913], $\text{Mean}_{C3rankD}$ = 32,478), than C1 nouns, with a median of 51,985 ($\text{Range}_{C1rankD}$ = [805, 1,273,425], $\text{Mean}_{C1rankD}$ = 186,009), and C2 nouns, with a median of 43,903 ($\text{Range}_{C2rankD}$ = [32, 813,588], $\text{Mean}_{C2rankD}$ = 114,817). The reverse is true for the verbs (rankB). C2 pairs also display intermediate values, and a higher dispersion than C1 and C3 pairs. All these results are confirmed by information in Table 5 and are in line with our expectations.

In summary, distributional semantic measures confirm that semi-demotivated pairs (C2) occupy an intermediate position between the two other categories. However, this position is not identical to the one indicated by experimental measures. We discuss this discrepancy in more detail in Section 4.2.

The significance of the observed differences between C1, C2 and C3 categories is tested with a Kruskal-Wallis test using the R `kruskall.test()` function [Hollander and Wolfe, 1973]. Overall, the differences prove to be significant between the three categories for all three measures, with a p-value of 2.2e-05 for P, 2.6e-05

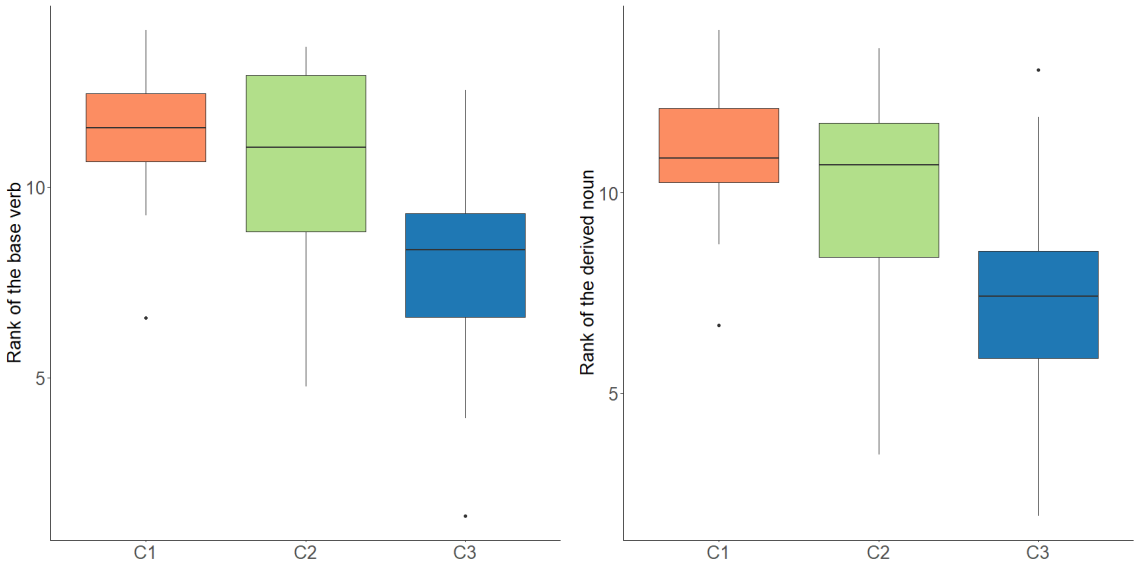

Figure 5: rankB (left) and rankD (right) per category. Ranking is log normalized

	Category	Mean	SD
rankD	C1	186,009	316,016
	C2	114,817	202,261
	C3	32,478	96,891
rankB	C1	277,342	393,835
	C2	226,381	279,137
	C3	32,439	65,447

Table 5: Mean and standard deviation of ranking in the distribution neighborhood

for rankD, and 3.7e-05 for rankB. More specifically, a pairwise comparison of the three categories using the R `pairwise.wilcox.test()` function[8] [Benjamini and Hochberg, 1995] shows that the difference is significant between demotivated (C1) and motivated (C3) pairs (p-value of 1.3e-06, 3.5e-06, and 7.1e-06 for P, rankD, and rankB, resp.), and to a lesser extent between motivated (C3) and semi-demotivated (C2) pairs (p-value of .0037, .0039, and .0025 for P, rankD, and rankB, resp.). It is not significant however between demotivated (C1) and semi-demotivated (C2) pairs, with a p-value of .3039, .1742, and .3039 for P, rankD, and rankB. In other words, all three distributional measures are able to discriminate highly motivated pairs from highly demotivated ones, but fail to properly discriminate highly demotivated pairs (C1) from semi-demotivated ones (C2).

Because our measures only provide one value for each considered pair, the distribution of the proximity and ranking scores cannot be studied pairwise. We can however observe the distribution of the measures depending on the suffix. The mean and standard deviation for proximity scores per suffix and according to the category are given in Table 6. Overall, the mean proximity score per suffix tends to increase with semantic motivation (from C1 to C3), which can be observed for example with *-ade* and *-ure*. However, two suffixes, *-et* and *-oir*, do not follow this trend. For *-et*, C2 nouns are on average closer to their base verb than both demotivated (C1) and motivated (C3) nouns, and they display a lower dispersion. The presence of the outlier *rivet* in C3 pairs, as mentioned in Section 2.2, could motivate this idiosyncratic behavior. In the case of *-oir*, C2 scores are lower than C1 and C3 scores, which may be more difficult to explain. It remains true that these observations cannot be generalized, given the low representativeness of our data with respect to suffix variation.

[8]The p-value adjust method is set to *BH*.

Suffix	Cat.	N_{pair}	Mean	SD
	C1	2	0.164	0.107
-ade	C2	2	0.340	0.272
	C3	2	0.441	0.003
	C1	2	0.067	0.093
-age	C2	2	0.087	0.043
	C3	2	0.410	0.184
	C1	3	0.087	0.043
-ance	C2	3	0.371	0.135
	C3	3	0.374	0.068
	C1	3	0.126	0.082
-et	C2	3	0.327	0.164
	C3	3	0.292	0.263
	C1	2	0.098	0.099
-ette	C2	2	0.214	0.203
	C3	2	0.290	0.178
	C1	3	0.078	0.043
-eur	C2	3	0.080	0.134
	C3	3	0.484	0.145
	C1	6	0.200	0.154
-oir	C2	6	0.070	0.081
	C3	6	0.232	0.140
	C1	5	0.124	0.069
-ure	C2	5	0.265	0.209
	C3	5	0.400	0.083

Table 6: Distribution proximity scores per suffix and category

On a concluding note, the possible correlation between the three distributional measures P, rankB and rankD can be questioned. These measures indeed converge in suggesting that the more the derived noun is motivated with respect to its base verb, the closer they are on a distributional level. Beyond this expected convergence, Figures 4 and 5 highlight the similar distribution of all three scores with respect to the extent of the distinction between C1, C2 and C3 categories, and so does their significance. It suggests that these measures might be redundant when used jointly to account for demotivation phenomena. Their possible correlation is assessed using Spearman's rank correlation coefficient, calculated for the three measures considered in pairs (given that rankD and rankB constitute ordinal variables, and P score a continuous variable). The low p-values ($p <$ 2.2e-16 for all pairings) and the high correlation coefficients observed (0.96 for rankD and rankB, -0.93 for P and rankB, -0.92 for P and rankD) confirm the strong correlation between the three measures, either positive between rankB and rankD or negative between P and rankB or rankD.

Given the strong correlation between the three measures, it can be asked which one represents demotivation the best and should be selected for further statistical modeling, in order to avoid collinearity effets. Based on the classification between C1, C2 and C3 verb-noun pairs, we determine the most predictive measure for demotivation by means of a stepwise multinomial regression using the `multinom` and `step` functions[9] from the R *nnet* package [Ripley, 1996, Venables and Ripley, 2007]. Such regression model allows for the exclusion of features that do not significantly contribute to the model, notably in case of collinearity, so as to keep the most contributing features.

The resulting fitted model only keeps the proximity score P as a predictor, excluding both rankB and

[9]The direction is set to *backward*.

rankD. This confirms the collinearity of the three measures, and indicates that the proximity score P has a higher predictive potential than rankB and rankD. Based on these results, we will only take into consideration the proximity score P when comparing experimental and distributional measures of demotivation, so as to avoid redundancy in the analysis.

4 Discussion

In Sections 2 and 3, we have investigated two possible ways of estimating morphosemantic demotivation, through experimental and distributional approaches, respectively. We now compare these two approaches, analyze their convergences and divergences, and further discuss how they could be combined to provide a reliable assessment of (de)motivation.

4.1 Convergence of experimental and distributional scores

We first compare experimental and distributional scores of demotivation to evaluate their degree of convergence. Figure 6 presents for each verb-noun pair the relationship between the distributional score (from 0 to 1, on the x-axis) and the average experimental score (from 0 to 6, on the y-axis). Categories C1, C2 and C3 are presented in orange, green and blue, respectively. For additional insights, the different suffixes involved are signalled by geometrical forms.

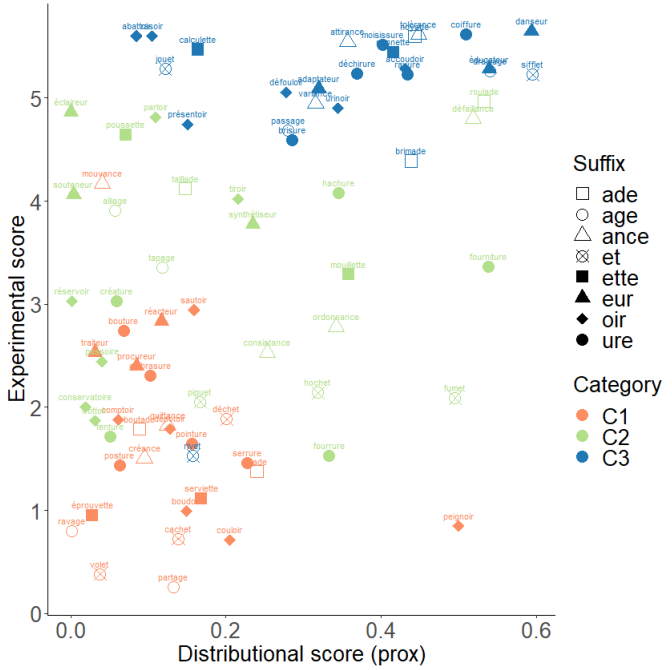

Figure 6: Average experimental (0 to 6) and distributional (0 to 1) scores per pair

Figure 6 shows that there is a moderately strong, positive, nonlinear association between the two scores, with a few potential outliers. Dense point clouds can be seen in the lower left corner and the upper right corner of the plot, corresponding to low vs. high experimental and distributional scores, respectively. It appears that the two methods converge in the identification of low and high values of morphosemantic motivation. Divergent cases are mostly located at the upper left corner of the plot, with high experimental scores and low distributional ones, but a few isolated items can also be found at the lower right of the plot, where distributional scores are high and experimental scores are low.

Despite the divergent points, the two measures are consistently related. A mixed ordinal regression model shows that distributional scores can significantly predict experimental scores (p = 1.08e-05, see Table 7). The model includes the experimental score as a dependent variable, the distributional score as an independent one, as well as random intercepts per participant and per verb-noun pair. According to the model, pairs with a high distributional proximity score have a very high probability of obtaining an experimental score of 5 or 6, whereas pairs with a medium score are more likely to obtain an experimental score of 3, 4 or 5, and pairs with a low score to be somewhat loosely distributed between 0 to 4. In other words, distributional proximity, as a comparative measure of the semantic similarity between two words, seems to be a fairly good indicator of semantic (de)motivation as evaluated by speakers.

Effect	**Estimate**	**Std. Errors**	***z***	***p***
P	4.658	0.988	4.714	1.08e-05

Table 7: Results of the mixed-effects ordinal regression model that determines experimental proximity score as a function of distributional score P

Both methods also confirm the gradualness of demotivation. While this result was to some extent expected with regard to distributional measures [Lenci, 2018], our data empirically support the psycholinguistic reality of such gradualness. More precisely, both methods distinguish degrees of motivation and converge in assigning higher proximity scores to motivated pairs (C3) than to demotivated pairs (C1). C2 appears as an intermediate case in both approaches, in terms of mean values and dispersion, as shown by the higher standard deviation and variance of C2 scores compared to those of C1 and C3. It remains true that no strict distinction between the three categories can be inferred neither from experimental nor from distributional measures. Instead, the three categories differ and overlap more or less depending on the approach adopted. This result strengthens the idea of a continuous scale of semantic demotivation that applies to complex words.

4.2 Discrepancies between the two methods

Although the experimental and distributional methods globally converge, they also differ to some extent with respect to (i) the overall distribution of proximity scores, and (ii) the individual evaluation of some pairs. These discrepancies can be seen in Figure 6. As far as (i) is concerned, the experimental method seems to allow for a finer differentiation of the three categories. C1 (in orange) and C3 (in blue) are clearly distinguished on the y-axis, with C2 (in green) occupying the intermediate space and somehow overlapping with C1. The results obtained with the distributional method are less clear regarding the distinction between the three categories. In particular, C3 covers the whole x-axis (from 0.1 to 0.6) largely overlapping with C2 and C1, while the experimental scores of the C3 pairs are on average higher, considering the difference in scales, and more concentrated towards the upper part of the figure. Nevertheless, the distributional method seems to better identify C1 pairs as demotivated compared to the experimental method: the experimental scores of C1 items are more spread out (from 0 to almost 3) than their distributional scores, which do not exceed 0.2 (apart from one noticeable outlier standing at 0.5).

As for individual discrepancies (ii), two kinds of disagreement between experimental and distributional scores can be identified. On the one hand, there are pairs with a low experimental score and a high distributional one. This is for example the case of the pair *peigner* 'comb'/*peignoir* 'bathrobe' (C1), with an experimental score of 0.85 (out of 6) and a distributional score of 0.50 (out of 1). On the other hand, there are pairs with a high experimental score and a low distributional one, such as *traiter* 'treat'/*traiteur* 'caterer' (C1) or *abattre* 'slaugther'/*abattoir* 'slaughterhouse' (C3), which obtain experimental scores of 2.50 and 5.60 and distributional scores of 0.03 and 0.08, respectively. The latter case is dominant among disagreements, as can be seen in the top left corner of Figure 6.

These discrepancies may be explained by the specific characteristics of each method. The relatively high experimental scores obtained by some demotivated pairs (C1) could be due to the prevalence of regular relationships between form and meaning in the lexicon. Psycholinguistic studies have previously shown the influence of phonological and orthographic structure on the semantic processing of words [see for example

Daneman and Reingold, 2000, Pollatsek et al., 2000, Barca et al., 2016]. It can be extrapolated that formal similarity influences speakers in searching for and rebuilding a semantic link between semantically unrelated but formally similar words. Formal resemblance between verbs and nouns, together with the identification of a suffix-like element (e.g., *-eur* in *traiteur* 'caterer'), could influence speakers' metalinguistic judgements when asked to evaluate the semantic proximity between word pairs.

As for distributional measures, they depend on the similarity of the contexts in which words are used. Proximity scores between verbs and nouns may be influenced by various factors that interfere with semantic transparency. Mere referential proximity between two words can generate high distributional scores, as in the case of *peigner* 'comb'/*peignoir* 'bathrobe' (P = 0.50). While the noun does not any longer denote the instrument used to perform the action described by the verb, both words still belong to the same referential domain — bathroom environment or body care — which can explain their distributional similarity. Human evaluation is arguably more sensitive to the lack of compositionality of *peignoir* 'bathrobe', explaining the rather low experimental score observed.

Lexical ambiguity is also an important factor that could explain specific aspects of distributional measures, especially in the case of motivated pairs (C3). Because the DSMs we used only provide one vector per form, they aggregate the distributional information for all the uses of a word, regardless of its possible ambiguity. This parameter has not been accounted for in our data selection, since we did not beforehand control (a) the number of meanings of each verb and noun nor the semantic overlap between ambiguous base verbs and derived nouns, (b) the frequency of each word sense in the reference corpus. For instance with respect to (a), *rasoir* 'razor' is only related to one sense of *raser* 'shave' (as opposed to *raser* 'bore', *raser* 'raze' and *raser* 'skim'), and the distributional score of this pair is low (0.11), whereas *déchirure* 'tear'/'heartbreak' is related to both senses of *déchirer* 'tear'/'hurt' and has a higher P score (0.37). Sense frequency (b) can be assessed through a sample annotation. The contextual analysis of 100 randomly selected utterances of the verb *raser* shows that the meaning involved in the derivation of *rasoir* 'razor' is minor (27%). Therefore, the vector for *raser* largely includes distributional information that is not related to the targeted meaning, which certainly contributes to the low distributional proximity observed between the verb and the noun. On the contrary, *sonnette* 'doorbell' is related to one relatively frequent meaning of *sonner* 'ring' (present in 40 out of 100 randomly selected utterances of the verb *sonner*) and has a high distributional score (0.42). The impact of lexical ambiguity on distributional models is well known and certainly plays a role in our observations, together with other factors related to the treatment of corpus data, such as lemmatization errors. More generally, corpus bias should be taken in account when using distributional data. Any reference corpus selected for distributional analysis has characteristics that can deviate from speakers' semantic representations of words. This divergence may contribute to explain the differences observed across motivated verb-noun pairs in general, and some of the differences observed for individual pairs.

Finally, it should be reminded that distributional similarity can be related to semantic similarity, but not directly to semantic transparency, and is only used as a proxy to evaluate the latter. The distributional similarity of equally transparent verb-noun pairs may vary depending on the semantic type of the nouns. In particular, deverbal nouns denoting eventualities are semantically more similar to verbs than deverbal nouns denoting entities. As a consequence, eventuality-denoting nouns are expected to be more similar distributionally to the base verbs than entity-denoting nouns. Although such a difference is not easy to evaluate independently of the ambiguity factor, it can be noted that in our data set motivated unambiguous nouns that denote events have a higher average P score and a lower standard deviation (0.424, 0.079 resp.) than motivated unambiguous nouns that denote entities (0.313, 0.179 resp.).[10] Differences in semantic types could thus partly explain the heterogeneity of motivated pairs (C3) with respect to distributional proximity measures.

Convergences and divergences between experimental and distributional methods can be interpreted as an indication of their reliability in the evaluation of morphosemantic demotivation. On the one hand, the higher the distributional similarity, the higher the reliability of the distributional score. As shown by Figure

[10]Although evenly distributed across the three categories, semantic types are not equally represented within each category. In the case of C3 verb-noun pairs, the average P scores are computed based on 15 unambiguous entity-denoting nouns and 6 unambiguous event-denoting nouns. The 5 remaining C3 nouns were not taken into account because they are ambiguous with respect to semantic type, as in the case of *passage* 'pathway'/'passing'.

6, items with the highest distributional scores (ranging from 0.4 to 0.6) tend to have high experimental scores (over 4, with a few exceptions), which can be seen as an indication of strong reliability, whereas the opposite is not observed — items with the lowest distributional scores do not tend to have low experimental scores. Because of the factors mentioned above, some low distributional scores cannot reflect the real degree of (de)motivation of a given verb-noun pair, while there does not seem to be such an effect with high scores. On the other hand, the lower the experimental proximity, the higher the reliability of the experimental score. As shown by Figure 6, items with the lowest experimental score (ranging from 0 to 2) tend to have low distributional scores (from 0 to 0.2), which can be seen as an indication of strong reliability, whereas the opposite is not observed — items with the highest experimental scores do not tend to have high distributional scores. This is due to the existence of items with high experimental and low distributional scores, but also possibly to the bias of speakers tending to identify a semantic relationship between words of similar form, which diminishes the reliability of medium experimental scores. As a consequence, it appears that strong distributional proximity is a solid hint for semantic motivation, whereas low experimental proximity is a solid hint for semantic demotivation. When testing the (de)motivation of a pair of formally related words without knowing anything about the tightness of their semantic relationship, a high distributional score informs us with a high level of certainty about a motivated and analyzable relationship, whereas a low experimental score informs us with a high level of certainty about a demotivated and unanalyzable relationship.

4.3 Going further into the diachronic dimension of demotivation

Demotivation is characterized by diachronic semantic change, affecting the transparency between bases and derivatives. To further explore the diachronic aspects of demotivation, and how diachronic evolution relates to the loss of semantic transparency, we can examine the relation between the lifespan of derivatives and the variation found in our experimental and distributional data. It is known that the more frequent a word is, the more semantically opaque it is [Baayen, 1993]. While frequency is usually asserted in terms of number of occurrences in synchrony, it can also be approached with respect to the span of use. The older a word is, the more likely it is to have undergone semantic shifts through time. Accordingly we hypothesize that the older a derivative is, the more likely it is to be demotivated.

To test this hypothesis, we investigate whether there is a correlation between the date of emergence and the opacity of a given word. More specifically, we assess to what extent C1, C2 and C3 pairs differ with respect to the date at which at least 10 attestations of the derivative can be found in Google Ngrams (following Bonami and Thuilier 2019). A threshold of 10 occurrences ensures that the derivative is not a hapax nor a transcription error and has to some extent entered usage. Note that we focus on derivatives only as we assume that the derivative meaning was built on the verb meaning when it was created, regardless of the possible semantic evolution of the verb. The distribution of attestation dates with respect to the classification of derivatives as demotivated (C1), semi-demotivated (C2) and motivated (C3) is presented in Figure 7.

Figure 7 shows that the dates of the 10 first attestations range from 1580 to 1970 and distribute differently across the three categories. While median dates appear quite similar across the three categories, the distribution extends more towards recent dates for C3 than for C1 and to some extent C2 derivatives. This suggests that C1 derivatives tend to be older than C2 and C3 derivatives. Kruskal-Wallis chi-square tests indicate that the differences observed in attestation dates are significant between the three groups ($\chi^2 = 6.1916$, $p = .04524$, $N_{obs} = 78$). However, a pairwise-comparison between groups using Wilcoxon rank sum test[11] shows that the difference is significant between demotivated (C1) and motivated (C3) pairs only ($p = .048$), as opposed to demotivated (C1) and semi-demotivated (C2) pairs ($p = .147$), and semi-demotivated (C2) and motivated (C3) pairs ($p = .430$).

These results suggest that demotivation can be seen as a function of attestation date: the older a derivative is, the less motivated its relation to the verb is. Accordingly, we expect a similar correlation to be observed with the experimental and distributional scores provided in our study. Yet here again some differences can be observed between the two methods. On the one hand, there is a significant correlation between the

[11]The test was applied with continuity correction, and with the *BH* p-value adjustment method.

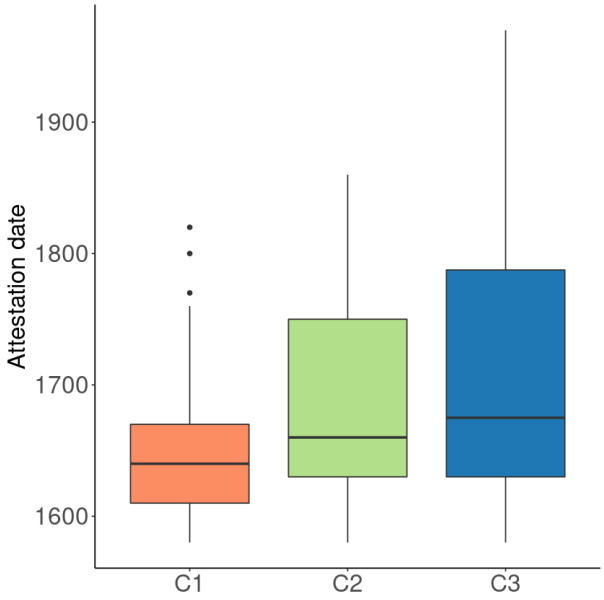

Figure 7: Date of the 10 first attestations per category

experimental score and the attestation date. We first tested a linear regression model to predict the average experimental score of derivatives based on the attestation date, which proved to be significant ($p = .003$). A more detailed examination of the effect of the attestation date on the experimental score is performed by means of a mixed-effects ordinal logistic regression model, with attestation date as the dependent variable[12], and experimental score as the response. The model also includes random intercepts per participant and per verb-noun pair. Its results are presented in Table 8, showing that attestation date has a significant effect on participants' judgements ($p = .00232$). More specifically, the more recent a derivative is, the highest is the experimental score.

Pred.	**Estimate**	**SE**	***z***	***p***
Attestation date	16.064	5.274	3.046	.00232

Table 8: Results of the mixed-effects ordinal regression model for the experimental score based on the attestation date

On the other hand, as far as distributional data are concerned, the training of a linear regression model with attestation date as the dependent variable and distributional scores as the response proves to be inconclusive, as both the model and the predictor are non-significant for all three distributional measures (P score: $p = .5$; rankB: $p = .302$; rankD: $p = .0886$). It appears that attestation date does not allow for a satisfactory prediction of distributional measures. This result confirms the divergence between the distributional and experimental data. Various reasons may explain such a divergence and should be investigated in further studies. A major difference is that the distributional measure provides far less data points (1 per item, 78 in total) than the experimental measure (125 per item on average, 9,989 in total). Combined with the strong overlap observed in the distribution of attestation dates between motivated and demotivated pairs, this lower level of refinement can make the prediction of the distributional score from the attestation date harder than that of the experimental score. This is particularly true with the lower-range attestation date, since we previously reported the fact that lower distributional scores are less reliable than lower experimental scores. This divergence contributes to highlighting the specificity of each method, and attests to the interest of confronting them and possibly combining them to provide a complex measure of semantic transparency.

[12]Attestation date was normalized on a scale from 0 (most ancient) to 1 (most recent) to improve model convergence.

5 Conclusion

In this paper, we have investigated and compared experimental and distributional methods to evaluate the semantic demotivation of complex words. We examined a sample of 78 French deverbal nouns identified as either motivated, semi-demotivated or demotivated with respect to their morphological base. The verb-noun pairs were attributed two scores of semantic proximity: an experimental score based on speakers' judgements and a distributional score based on a vectorial representation of meaning computed from a corpus. Our results emphasize the gradual nature of demotivation both in a linguistic and a psycholinguistic perspective. In the two approaches we adopted, motivated and demotivated pairs constitute the opposite ends of a continuum of semantic relatedness, whereas semi-demotivated pairs are widely spread across the similarity spectrum. The two methods can be seen as complementary. They involve different types of linguistic information, since the experimental method reflects speakers' intuitions, whereas the distributional method reflects corpus uses. It appears that the former allows for a fine-grained assessment of demotivation, but it is costly and difficult to implement on a large scale. The latter has the advantage of automaticity and provides a usage-based assessment, but it suffers from the lack of control over linguistic properties such as lexical ambiguity.

This study is a first attempt to examine demotivation in synchrony using a combination of experimental and computational methods. It naturally calls for a follow-up study focusing on a more diachronic approach to demotivation. The present work provides bases to build on for further investigation. For instance, we plan to extend the linguistic material to be tested. More complex words and more morphological diversity should be considered (e.g., deadjectival nouns, as in *mou* 'soft'/*mollet* 'calf', or denominal nouns, as in *pomme* 'apple'/*pommette* 'cheekbone') to allow for a wider account of complex word demotivation. The material should also be characterized more precisely with respect to semantic properties, including referential type and lexical ambiguity, so as to better understand their potential effect on gradual demotivation. In addition, other approaches to semantic transparency should be explored, such as priming effects with regard to psycholinguistic aspects [Longtin et al., 2003], distributional inclusion [Varvara et al., 2021] and offset vectors [Bonami and Paperno, 2018, Bonami and Tribout, 2021] with regard to computational aspects. Another issue that can be addressed is the identification of the causes of morphosemantic demotivation. The diachronic evolution of affix properties could play an important role in the demotivation process, as suggested by some suffix tendencies observed in Sections 2 and 3. Although the data presented in this study are too scarce to allow for any generalization, properties such as affix productivity should be investigated as possible factors influencing the demotivation of complex words.

Acknowledgements

This research was supported by the France-Switzerland Partenariat Hubert Curien Germaine de Staël program. We are grateful to the two anonymous reviewers for their helpful comments on an earlier version of this paper. All remaining errors are our own.

References

H. Baayen. On frequency, transparency and productivity. In *Yearbook of morphology 1992*, pages 181–208. Springer, 1993.

R. H. Baayen, P. Milin, and M. Ramscar. Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220, 2016.

L. Barca, F. Benedetti, and G. Pezzulo. The effects of phonological similarity on the semantic categorisation of pictorial and lexical stimuli: evidence from continuous behavioural measures. *Journal of Cognitive Psychology*, 28(2):159–170, 2016. doi: 10.1080/20445911.2015.1101117.

L. Bauer. *English word-formation*. Cambridge University Press, 1983.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

A. Blank. *Pathways of lexicalization*, pages 1596–1608. De Gruyter Mouton, 2001. doi: doi:10.1515/9783110194265-049. URL `https://doi.org/10.1515/9783110194265-049`.

G. Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020. doi: 10.1146/annurev-linguistics-011619-030303.

O. Bonami and D. Paperno. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17(2):173–195, 2018.

O. Bonami and J. Thuilier. A statistical approach to rivalry in lexeme formation: French-iser and-ifier. *Word structure*, 12(1):4–41, 2019.

O. Bonami and D. Tribout. Échantinom: a hand-annotated morphological lexicon of French nouns. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 33–42, ATILF & CLLE, Université de Lorraine, Nancy, France, 2021.

L. J. Brinton and E. C. Traugott. *Lexicalization and language change*. Cambridge University Press, 2005.

J. A. Bullinaria and J. P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526, 2007.

R. H. B. Christensen. ordinal—regression models for ordinal data, 2019. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

D. Corbin. *Morphologie dérivationnelle et structuration du lexique*. Mouton De Gruyter, 1987.

A. Creemers, A. Goodwin Davies, R. J. Wilder, M. Tamminga, and D. Embick. Opacity, transparency, and morphological priming: A study of prefixed verbs in dutch. *Journal of Memory and Language*, 110, 2020. doi: 10.1016/j.jml.2019.104055.

M. Daneman and E. M. Reingold. Chapter 17. do readers use phonological codes to activate word meanings? evidence from eye movements. In A. Kennedy, R. Radach, D. Heller, and J. Pynte, editors, *Reading as a perceptual process*, pages 447–474. Elsevier, 2000.

J. Dendien and J.-M. Pierrel. Le trésor de la langue française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement automatique des langues*, 44(2):11–37, 2003.

P. Dohmes, P. Zwitserlood, and J. Bölte. The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language*, 90:203–212, 2004. doi: 10.1016/S0093-934X(03)00433-4.

J. R. Firth. A synopsis of linguistic theory, 1930-1955. In J. R. Firth, editor, *Studies in Linguistic Analysis*, pages 1–32. Basil Blackwell, Oxford, 1957.

C. L. Gagné, T. L. Spalding, and K. A. Nisbet. Processing English compounds: Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics*, 13(2):2–22, 2017.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

N. Hathout, F. Sajous, and B. Calderone. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. ISBN 978-2-9517408-8-4.

M. Hilpert. *Lexicalization in morphology*, pages 1–18. Oxford University Press, 2019.

P. Hohenhaus. Lexicalization and institutionalization. In *Handbook of word-formation*, pages 353–373. Springer, 2005.

M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. John Wiley and Sons, Inc, New York, 1973.

A. Kielar and M. F. Joanisse. The role of semantic and phonological factors in word recognition: An ERP cross-modal priming study of derivational morphology. *Neuropsychologia*, 49:161–177, 2011. doi: 10.1016/j.neuropsychologia.2010.11.027.

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

R. V. Lenth. Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33, 2016. doi: 10.18637/jss.v069.i01.

G. Libben. Compound words, semantic transparency, and morphological transcendence. In S. Olsen, editor, *New impulses in word-formation*, pages 212–232. Buske, Hamburg, 2010.

L. Lipka. Lexikalisierung, idiomatisierung und hypostasierung als probleme einer synchronischen wortbildungslehre. In H. E. Brekle and D. Kastovsky, editors, *Perspektiven der Wortbildungsforschung*, pages 155–164. Bouvier Verlag Herbert Grundmann, Bonn, 1977.

L. Lipka. Lexicalization and institutionalization in English and German. *Linguistica Pragensia/Akademie Ved CR, Ústav pro Jazyk Ceskỳ*, pages 1–13, 1992.

C.-M. Longtin, J. Segui, and H. P. A. Morphological priming without morphological relationship. *Language and Cognitive Processes*, 18(3):313–334, 2003. doi: 10.1080/01690960244000036.

M. Marelli and M. Baroni. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485–515, 2015. doi: 10.1037/a0039267.

W. Marslen-Wilson, L. Komisarjevsky Tyler, R. Waksler, and L. Oldername. Morphology and meaning in the english mental lexicon. *Psychological Review*, 101(1):1–33, 1994.

F. Meunier and J. Segui. Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language*, 41(3):327–344, 1999.

T. Mikolov, K. Chan, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale, 2013.

J. Morris, J. Grainger, and P. J. Holcomb. Tracking the consequences of morpho-orthographic decomposition using ERPs. *Brain Research*, 1529:92–104, 2013. doi: 10.1016/j.brainres.2013.07.016.

S. Padó, A. Herbelot, M. Kisselew, and J. Šnajder. Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, 2016.

J. Park, F. Sana, C. L. Gagné, and T. L. Spalding. Is inhibition involved in the processing of opaque compound words? A study of individual differences. *The Mental Lexicon*, 15(2):258–294, 2020. doi: 10.1075/ml.19011.par.

A. Pollatsek, K. Rayner, and H.-W. Lee. Chapter 15. phonological coding in word perception and reading. In A. Kennedy, R. Radach, D. Heller, and J. Pynte, editors, *Reading as a perceptual process*, pages 339–426. Elsevier, 2000.

R Core Team. R: A language and environment for statistical computing, 2015. URL `http://www.R-project.org/`.

K. Rastle, M. H. Davis, and B. New. The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11(6):1090–1098, 2004.

K. Rayner and S. A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201, 1986.

S. Reddy, D. McCarthy, and S. Manandhar. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, 2011.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

M. Roché. Mot construit ? Mot non construit ? Quelques réflexions à partir des dérivés en -*ier*(*e*). *Verbum*, 26(2):459–480, 7 2004.

M. Sahlgren and A. Lenci. The effects of data size and frequency range on distributional semantic models. *arXiv preprint arXiv:1609.08293*, 2016.

R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey, 2012.

E. Smolka and G. Libben. 'Can you wash off the hogwash?' – semantic transparency of first and second constituents in the processing of german compounds. *Language, Cognition and Neuroscience*, 32(4):514–531, 2017. doi: 10.1080/23273798.2016.1256492.

E. Smolka, G. Libben, and W. U. Dressler. When morphological structure overrides meaning: evidence from german prefix and particle verbs. *Language, Cognition and Neuroscience*, 34(5):599–614, 2019. doi: 10.1080/23273798.2018.1552006.

A. Urieli. *Robust French Syntax Analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, University of Toulouse Jean Jaurès, Toulouse, 2013.

R. Varvara, G. Lapesa, and S. Padó. Grounding semantic transparency in context. *Morphology*, pages 213–234, 2021. doi: 10.1007/s11525-021-09382-w.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 2007.

M. Wauquier. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. PhD thesis, Université Toulouse 2 – Jean Jaurès, 2020.

Appendix: Complete dataset

		C1		C2		C3	
#	Suff.	N	V	N	V	N	V
1	*-ade*	*boutade*	*bouter*	*taillade*	*tailler*	*brimade*	*brimer*
		'joke'	'push out'	'gash'	'prune'	'bullying'	'bully'
2	*-ade*	*tirade*	*tirer*	*roulade*	*rouler*	*noyade*	*noyer*
		'tirade'	'pull'	'roll'	'roll'	'drowning'	'drown'
3	*-age*	*ravage*	*ravir*	*alliage*	*allier*	*drainage*	*drainer*
		'ravage'	'abduct'	'alloy'	'ally'	'drainage'	'drain'
4	*-age*	*partage*	*partir*	*tapage*	*taper*	*passage*	*passer*
		'sharing'	'leave'	'disturbance'	'hit'	'passing'	'pass'
5	*-ance*	*créance*	*croire*	*ordonnance*	*ordonner*	*attirance*	*attirer*
		'debt'	'believe'	'prescription'	'command'	'attraction'	'attract'
6	*-ance*	*quittance*	*quitter*	*consistance*	*consister*	*variance*	*varier*
		'receipt'	'leave'	'consistance'	'consist'	'variance'	'vary'
7	*-ance*	*mouvance*	*mouvoir*	*défaillance*	*défaillir*	*tolérance*	*tolérer*
		'movement'	'move'	'failure'	'faint'	'tolerance'	'tolerate'
8	*-et*	*cachet*	*cacher*	*fumet*	*fumer*	*rivet*	*river*
		'seal'	'hide'	'aroma'	'smoke'	'rivet'	'bind'
9	*-et*	*déchet*	*déchoir*	*hochet*	*hocher*	*sifflet*	*siffler*
		'garbage'	'deprive'	'rattle'	'nod'	'whistle'	'whistle'
10	*-et*	*volet*	*voler*	*piquet*	*piquer*	*jouet*	*jouer*
		'shutter'	'fly'	'stake'	'sting'	'toy'	'play'
11	*-ette*	*éprouvette*	*éprouver*	*mouillette*	*mouiller*	*sonnette*	*sonner*
		'test tube'	'feel'	'bread soldier'	'wet'	'doorbell'	'ring'
12	*-ette*	*serviette*	*servir*	*poussette*	*pousser*	*calculette*	*calculer*
		'towel'	'serve'	'stroller'	'push'	'calculator'	'calculate'
13	*-eur*	*procureur*	*procurer*	*synthétiseur*	*synthétiser*	*danseur*	*danser*
		'prosecutor'	'provide'	'synthesizer'	'synthesize'	'dancer'	'dance'
14	*-eur*	*réacteur*	*réagir*	*éclaireur*	*éclairer*	*éducateur*	*éduquer*
		'reactor'	'react'	'scout'	'light'	'educator'	'educate'
15	*-eur*	*traiteur*	*traiter*	*souteneur*	*soutenir*	*adaptateur*	*adapter*
		'caterer'	'treat'	'pimp'	'support'	'adaptor'	'adapt'
16	*-oir*	*boudoir*	*bouder*	*réservoir*	*réserver*	*urinoir*	*uriner*
		'boudoir'	'sulk'	'tank'	'keep'	'urinal'	'urinate'
17	*-oir*	*comptoir*	*compter*	*passoire*	*passer*	*présentoir*	*présenter*
		'counter'	'count'	'sieve'	'pass'	'display stand'	'display'
18	*-oir*	*couloir*	*couler*	*tiroir*	*tirer*	*abattoir*	*abattre*
		'hallway'	'flow'	'drawer'	'draw'	'slaughterhouse'	'slaughter'
19	*-oir*	*peignoir*	*peigner*	*conservatoire*	*conserver*	*défouloir*	*défouler*
		'robe'	'comb'	'conservatory'	'keep'	'release'	'unwind'
20	*-oir*	*sautoir*	*sauter*	*parloir*	*parler*	*rasoir*	*raser*
		'string'	'jump'	'visiting room'	'speak'	'razor'	'shave'
21	*-oir*	*dépotoir*	*dépoter*	*trottoir*	*trotter*	*accoudoir*	*s'accouder*
		'dump'	'unpot'	'sidewalk'	'trot'	'armrest'	'lean'
22	*-ure*	*bouture*	*bouter*	*fourrure*	*fourrer*	*moisissure*	*moisir*
		'cutting'	'push out'	'fur'	'stuff'	'mold'	'make moldy'
23	*-ure*	*embrasure*	*embraser*	*tenture*	*tendre*	*coiffure*	*coiffer*
		'door frame'	'set alight'	'hanging'	'stretch'	'hairstyle'	'do hair'
24	*-ure*	*pointure*	*pointer*	*créature*	*créer*	*brisure*	*briser*
		'size'	'point'	'creature'	'create'	'fragment'	'shatter'
25	*-ure*	*posture*	*poster*	*hachure*	*hacher*	*rayure*	*rayer*
		'posture'	'poster'	'hatching'	'chop'	'stripe'	'draw lines'
26	*-ure*	*serrure*	*serrer*	*fourniture*	*fournir*	*déchirure*	*déchirer*
		'lock'	'grip'	'supplies'	'supply'	'tear'	'tear'