Evaluating morphosemantic demotivation through experimental and distributional methods

Alizée Lombard
Université de Fribourg

Marine Wauquier
Université de Paris
LLF, CNRS

Cécile Fabre
Université Toulouse Jean Jaurès
CLLE, CNRS

Nabil Hathout
Université Toulouse Jean Jaurès
CLLE, CNRS

Lydia-Mai Ho-Dac
Université Toulouse Jean Jaurès
CLLE, CNRS

Richard Huyghe
Université de Fribourg

1 Introduction

Demotivation as the obliteration of the morphosemantic relation between a base word and a derivative is part of the lexicalization process (Bauer, 1983; Lipka, 1992; Brinton & Traugott, 2005). It is caused by idiosyncrasic factors such as onomasiological needs, lexical competition, diachronic change, etc. Although morphosemantic demotivation is in itself a gradual process, as shown for example by Roché (2004) in the analysis of French nouns suffixed with *-ier*, it has rarely been evaluated or quantified as such.

Focusing on the morphosemantic demotivation of French nouns derived from verbs, our goal in this presentation is to investigate methods that can be used to measure morphosemantic demotivation. We present and compare two empirical methods, based respectively on experimental and distributional approaches, in order to determine to what extent judgements and corpus-based assessments of demotivation are correlated.

2 Selection of the materials

We define three conditions on base-derivative pairs, identified as C1 for no motivation (e.g. *partage* 'sharing'/*partir* 'leave'), C2 for ongoing demotivation (e.g. *créature* 'creature'/*créer* 'create'), and C3 for synchronically motivated pairs (e.g. *rasoir* 'razor'/*raser* 'shave'). Each condition is populated with pairs extracted from various corpora and lexical resources such as FRCOW16A, Anagrimes[1] and *Trésor de la Langue Française*.

The selected pairs satisfy the formal and semantic criteria of (i) a diachronically attested link between the base and its derivative, and (ii) the absence (in C1 condition) or presence (partial in C2 and complete in C3 conditions) of a semantic relationship between the base and its derivative in synchrony. The selection and classification of the materials in C1/C2/C3 was carried out jointly by six linguists in order to obtain balanced samples, not only in terms of item number, but also of frequency range and suffix representation. Due to their rarity, C1 pairs have served as a reference for selecting materials in the other conditions. 26 pairs were finally selected in each condition. These include a variety of suffixes (*-ade*, *-ance*, *-oir*, *-eur*, *-ette*) and semantic types (action, object, property, location, etc.).

[1]Tool exploiting the French Wiktionary, accessible online at the url `https://anagrimes.toolforge.org/`.

3 Experimental approach

3.1 Design

The aim of the experiment is to measure the degree of demotivation based on French native speakers' judgements about the semantic proximity of the 78 verb-noun pairs[2]. It took the form of two surveys (39 stimuli each, about 15 minutes long) completed online by 411 Bachelor students in the humanities. In order to control for sociological factors, we chose to analyze only the data from native speakers who were no older than 25 years old, which resulted in 150 and 159 responses per survey (N = 309, M_{age} = 19.4, $Range_{age}$ = [17, 25]).

Each pair was presented separately to participants who where asked to evaluate the proximity between the meaning of the base and that of the derivative on a scale from 0 (unrelated meanings) to 6 (highly related meanings). Participants could also indicate if they did not know one or both words presented in each stimulus. The corresponding data were excluded from the analysis (652 trials, i.e. 0.058% of the data). According to our hypotheses, demotivated pairs (C1) should elicit lower scores of proximity than motivated ones (C3). Semi-demotivated pairs (C2) should yield intermediate scores between those of C1 and C3.

3.2 Results

As shown in Figure 1, the experimental results clearly support our hypotheses. Demotivated pairs (C1) obtain the lowest scores (44% of score 0), whereas motivated pairs (C3) have the highest ones (56% of score 6). Scores assigned to semi-demotivated pairs (C2) are rather evenly distributed, which reveals the heterogeneity of this group.

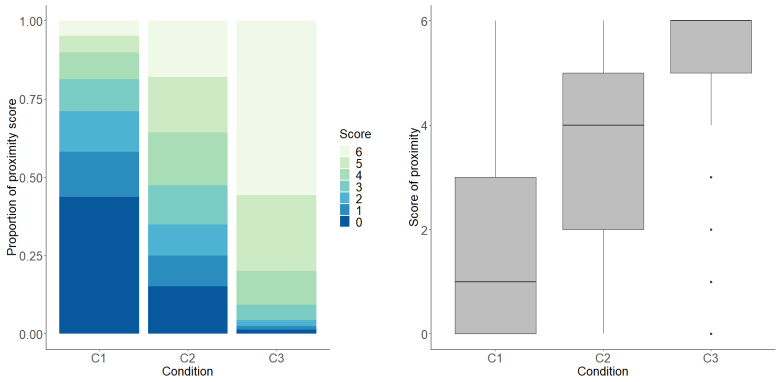

Figure 1: Distribution of experimental scores per condition

These results are analysed through a mixed ordinal logistic regression. The chosen model determines the proximity score as a function of the experimental condition (C1, C2 or C3). It also includes random intercepts per participant and per verb-noun pair. The significance of the predictor effect is determined by a log-likelihood ratio test between models with or without the predictor. The condition appears to have a highly significant effect on the proximity score ($p < 2.2e\text{-}16$). Speakers' intuitions are thus consistent with the experts' judgements.

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

4 Distributional approach

4.1 Design

We use distributional semantics to automatically quantify the semantic similarity between derived nouns and base verbs in the C1, C2 and C3 pairs. Distributional Semantics Models (DSMs) provide a vectorial representation of meaning based on the cooccurrences of a given word in a corpus. In the resulting vector space, the distance between vectors approximates the degree of similarity between the corresponding words.

We compute three distributional measures to estimate the demotivation between a base and its derivative: the proximity score P, the rank of the base in the neighborhood of the derivative (rankB), and the rank of the derivative in the neighborhood of the base (rankD). We expect C3 pairs to display a higher proximity score (i.e. closer to 1), and a higher rank (i.e. closer to 1st rank) than C1 pairs. C2 pairs should display in-between values both in terms of proximity score and ranking.

These measures are computed from a vector space concatenating 5 DSMs trained with Word2Vec (Mikolov et al., 2013) on the Wikipedia lemmatized corpus, consisting of 900 million words. The DSMs training parameters are: CBOW, Negative Sampling, frequency threshold of 5, window size of 5.

4.2 Results

Results are in line with our hypotheses. As shown in Figure 2, the proximity scores increase as the semantic motivation of the pairs increases. The median proximity score is higher for C3 pairs than for C1 pairs, and the median score for C2 falls in between. While C2 median score is close to that of C1, C2 pairs display a much higher dispersion, showing the heterogeneity of C2 as a class, compared to C1 and C3. Similar observations can be made with respect to rankB and rankD depending on the degree of motivation. Figure 2 shows that C3 derivatives appear higher in the base neighborhood than C1 and C2 derivatives. C2 pairs also display a higher dispersion than C1 and C3 pairs.

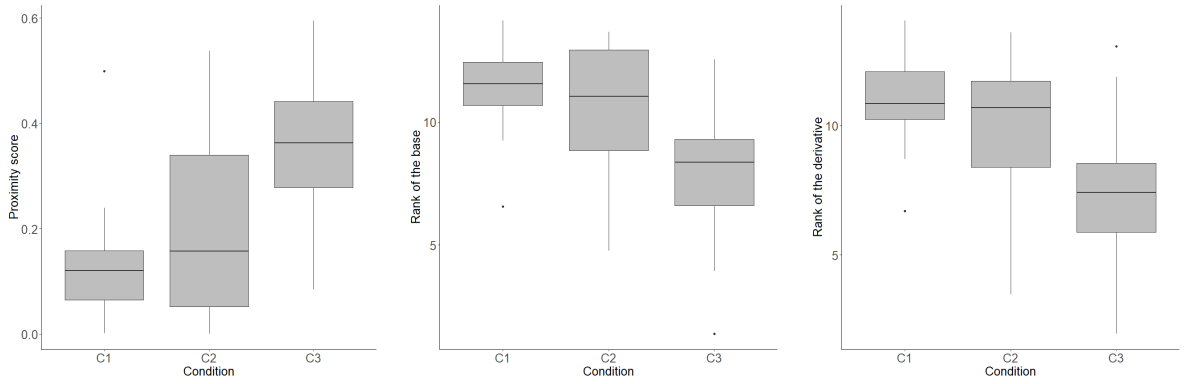

Figure 2: Proximity score (left), ranking (log) of the base in the derivative neighborhood (middle) and ranking (log) of the derivative in the base neighborhood (right) per condition

5 Discussion

The study highlights the psycholinguistic and distributional reality of morphosemantic demotivation, and confirms its gradual properties. Both approaches converge with respect to higher

proximity of motivated pairs (C3) than of demotivated pairs (C1). They also converge in analysing C2 as an intermediate case between C1 and C3, showing in addition a wider variation of C2 items. These observations are consistent with the presumed scalarity of demotivation. We analysed the correlation between experimental and distributional results through a mixed ordinal logistic regression. The model determines the experimental proximity score as a function of the distributional proximity and includes random intercepts per participant and per pair. The effect of the predictor is significant (p = 6.76e-06), showing that the distributional approach is clearly in line with speakers' judgements.

While both measures converge, they also differ to some extent. First, the experimental method seems to provide more accurate results than the distributional one. As can be seen in Figure 3, the experimental score (on the y-axis) allows for a better differentiation of the three conditions than the distributional proximity (on the x-axis). Second, some discrepancies can be observed with respect to the assessment of the demotivation of some specific pairs. For instance, there is a strong disagreement between both scores for pairs such as *peignoir* 'bathrobe'/*peigner* 'comb' (C1) (with a high distributional score, but a low experimental one) and *mouvance* 'movement'/*mouvoir* 'move' (C1) (with a low distributional score, but a high experimental score). These results suggest that in some cases, distributional similarity might be too coarse to grasp the level of demotivation of specific items.

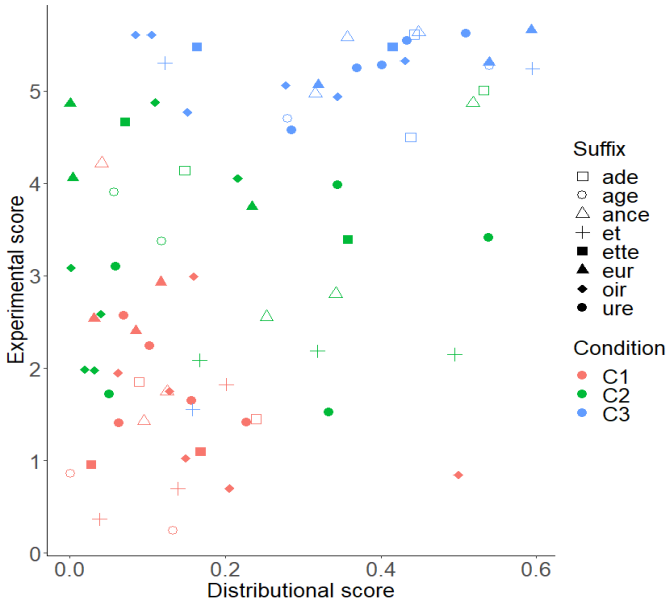

Figure 3: Experimental (0 to 6) and distributional (0 to 1) scores per pair

Although the experimental approach appears to be more accurate, it is more costly and more difficult to extend to a larger scale, especially for quantifying demotivation of previously non-evaluated pairs. By contrast, basing the analysis on the distributional approach would have the advantage of automaticity. It would however need some refinement to neutralize the discrepancies observed between the two methods.

References

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

Brinton, Laurel J & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge University Press.

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

Mikolov, Tomas, Kai Chan, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale.

Roché, Michel. 2004. Mot construit ? mot non construit ? quelques réflexions à partir des dérivés en *-ier*(*e*). *Verbum* 26(2). 459–480.