Distributional Effects of Gender Contrasts Across Categories

Timothee Mickus
Université Paris Diderot,
Laboratoire de
linguistique formelle

Olivier Bonami*
Université Paris Diderot,
Laboratoire de
linguistique formelle

Denis Paperno*
Loria (UMR7503)
CNRS
Université de Lorraine

Abstract

This paper proposes a methodology for comparing grammatical contrasts across categories with the tools of distributional semantics. After outlining why such a comparison is relevant to current theoretical work on gender and other morphosyntactic features, we present intrinsic and extrinsic predictability as instruments for analyzing semantic contrasts between pairs of words. We then apply our method to a dataset of gender pairs of French nouns and adjectives. We find that, while the distributional effect of gender is overall less predictable for nouns than for adjectives, it is heavily influenced by semantic properties of the adjectives.

1 Introduction

Grammatical gender (henceforth **g-gender**) is the phenomenon by which some languages group nouns in classes that exhibit different behavior in agreement, as in French *une$_F$ petite$_F$ table* 'a small table' vs. *un$_M$ petit$_M$ bureau* 'a small desk' (Hockett, 1958). In languages that have such a system, g-gender entertains a complex relationship with the social gender of referents (henceforth **s-gender**). On the one hand, the assignment of g-gender to nouns is often arbitrary. This is massively the case in languages like French, which have only two genders, and need to assign all inanimate nouns to either masculine or feminine. On the other hand, as Corbett (1991) highlights, all g-gender assignment systems have a semantic core, which usually entails lexicalizing different nouns for male and female referents, and assigning them to a matching g-gender (*une$_F$ petite$_F$ fille* 'a small girl' vs. *un$_M$ petit$_M$ garçon* 'a small boy') to masculine g-gender (Corbett, 2013). While this not a categorical rule (some nouns refer to either men or women while having a single gender, e.g. *personne* 'person' is always feminine), it is a very strong tendency.

MAS	FEM	translation
candidat	*candidate*	'candidate'
marchand	*marchande*	'merchant'
infirmier	*infirmière*	'nurse'

Table 1: Sample pairs of human nouns

MAS	FEM	translation
délicat	*délicate*	'delicate'
grand	*grande*	'tall'
plénier	*plénière*	'plenary'

Table 2: Sample pairs of adjectives

In this paper we focus on pairs of morphologically-related nouns such as *candidat, candidate* where g-gender signals s-gender[1]; Table 1 exhibits a few relevant examples. The nature of the relationship between such nouns is an understudied but pressing issue for morphological theory. One position holds that *candidat* and *candidate* are two separate lexical items or *lexemes* (Matthews, 1974), related by derivational morphology (Zwanenburg, 1988). Under such a view, the relationship between the two nouns is similar to that between *danser* 'to dance' and *danseur* 'dancer'. The opposite view holds that *candidat* and *candidate* are forms of the same lexeme, related by inflectional morphology (Bonami and Boyé, in press). Under such a view, the relation between the a masculine and a feminine

*Olivier Bonami and Denis Paperno share senior authorship and are listed in alphabetic order.

[1]More precisely, Burnett and Bonami (in press(a); in press(b)) argue that g-gender carries social meaning rather than denotational meaning: using a feminine signals the speaker's perception of gender-stereotypical properties of the referent, leading to a situation where g-gender and s-gender will mostly match but differ in principled ways in some situations.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 174-184.
New York City, New York, January 3-6, 2019

noun (*candidat* vs. *candidate*) is similar to that between a singular and a plural noun (*candidat* vs. *candidats*) or a masculine and feminine forms of an adjective (*petit* and *petite*).

While these two views make different theoretical predictions on the nature of g-gender systems, they are remarkably difficult to tease apart empirically, given the elusiveness of the empirical divide between inflection and derivation (see e.g. Dressler 1989; Corbett 2010; Spencer 2013). In this paper we build on the well-known observation that inflection is semantically more regular than derivation (Robins, 1959; Matthews, 1974; Wurzel, 1989; Stump, 1998). While the meaning of the English 3SG verb form *dines* can readily be predicted from the meaning of its base form *dine*, the range of meanings of *diner* (including a particular kind of restaurant) is unpredictable. Bonami and Paperno (forthcoming) operationalize this idea by quantifying the diversity of semantic contrasts between pairs of morphologically-related words, and found, consistently with the theoretical literature, that pairs of words in derivational relations contrast in more diverse ways than pairs of word forms related by inflection.

It is not immediately obvious how diverse the semantic contrasts between pairs of gender-contrasting human nouns are. On the one hand, these are systematic enough that dictionaries do not list separate entries for masculine and feminine nouns. On the other hand, the existence of gender biases does lead to some interesting unpredictable differences. For instance, until very recently, masculine nouns referring to a stereotypically male occupation (e.g. *ambassadeur* 'ambassador') were often paired with a feminine noun (e.g. *ambassadrice*) referring to the wife of a man with that occupation, rather than to a woman with that occupation. While social change towards gender equality led to a change in usage in this particular case, the pervasiveness of gender biases leads one to expect differences in meaning or in usage between masculine vs. feminine nouns which have comparable meanings otherwise; cf. Bolukbasi et al. (2016), who highlighted the omnipresence of gender stereotypes in the distribution of English nouns.

This paper compares of the semantic import of g-gender contrasts in human nouns, as illustrated in Table 1, as opposed to g-gender contrasts in adjectives (Table 2). We ask two kinds of questions:

1. Are the semantic contrasts between human nouns **more diverse** than the contrasts between adjectives?

2. Are the semantic contrasts between human nouns **similar** to the semantic contrasts between adjectives?

Different views on the nature of g-gender lead to contradictory predictions as to the answers to questions 1 and 2. Under a naive view of g-gender assignment as completely arbitrary, we would expect to find semantic contrasts among neither nouns nor adjectives, leading to a negative answer to the first question and a positive answer to the second.[2] On the other hand, if g-gender on human nouns does signal s-gender of the referent, then the different takes on the relationship between paired nouns will lead to distinct expectation. If the relation is inflectional, we expect little or no difference between nouns and adjectives, and hence a negative answer to question 1. If it is derivational, following Bonami and Paperno (forthcoming), we expect more irregularity among nouns, and hence a positive answer to question 1. As to question 2, if g-gender signals s-gender, we expect similar contrasts for human nouns and adjectives that modify a nominal expression with human reference. However, we have no such expectation of similarity for those adjective instances that modify an inanimate nominal expression; hence the answer to question 2 should be different for different subsets of adjective usages.

The structure of the paper is as follows. Section 2 presents our new methodology to study morphological contrasts. In Section 3, we test the validity of our methodology by applying it to the study of grammatical gender contrasts in French human nouns (**HN**s) and adjectives and report what differences are observed between these categories. In Section 4, we probe the variability of contrast among adjectives, and in Section 5 we specifically investigate the differences between usages of adjectives to qualify human nouns (**HQA**s) vs. usages of adjectives to qualify non-human nouns (**NHQA**s). Our findings are summarized in Section 6.

[2] In a distributional operationalization, we expect small, erratic differences that cancel each other on average.

2 Methodology

2.1 Framework

In distributional semantics (Lenci, 2018), the meaning of a word is represented in the mathematical form of a vector computed on the basis of the word's contexts of occurrence in a large text corpus. Distributional vectors have a number of applications, ranging from predicting semantic relatedness judgments (Agirre et al., 2009) to initializing neural machine translation systems (Artetxe et al., 2017; Lample et al., 2017). Among the wide range of theoretical and practical applications, distributional semantic modeling has been used in two domains of direct relevance to the present study. First, on the basis of German data, Dye et al. study the relation of distributional similarity and g-gender assigment. Second, distributional methods have been used to characterize natural language morphology, including the issue of semantic transparency in derivation (Marelli and Baroni, 2015), analysis of morphological variation (Varvara, 2017), as well as the nature of the inflectional and derivational relations (Bonami and Paperno, forthcoming).

Our method is closely related to the latter work, and is based on two assumptions. First, following Mikolov et al.'s (2013b) model for solving proportional semantic analogy, we assume that a semantic contrast between two words is represented by the shift between the corresponding word vectors, so that words in identical relations are expected to have similar shifts[3]. The second assumption is that the semantics of a morphological relation can be approached by averaging vector differences for multiple word pairs in the relation. This has the double benefit of cancelling out some of the noise inherent in distributional vectors and evening out variation in the contrast between pairs of words entering the same relation.

Other authors have stressed that some relations were not accurately represented using vector shifts. For instance, Gladkova et al. (2016) have highlighted that "derivational and lexicographic relations remain a major challenge"; Levy and Goldberg (2014) and Linzen (2016) both stress that simple additive models do not suffice to model relations. However, we do expect that even if a linear shift is an imperfect approximation of the relation in a distributional vector space, the regularity of such an approximation still corresponds to the semantic regularity of a specific relation. Therefore we do not make any strong claim regarding the correct representation of a relation in a DSM, but we do assume that the more regular a relation is, the more akin to a linear function – *ie.* a vector offset – its representation will be. More generally, the aim of this work is not to discuss how to accurately capture lexical relations between words, but rather to assess the relative regularity of different relations.

Therefore, although we use an evaluation setup similar to Mikolov et al.'s, our goal here is not to solve the propositional analogy task, but to analyze and assess the predictability of different relations. Hence, we emphasize that we mention intrinsic and extrinsic 'predictions' of word vector values only for the sake of convenience, which is also the reason why we do not employ various improvements on the vector shift method proposed in the literature such as the multiplicative method of Levy and Goldberg (2014), and stick to the simplest, most transparent option suitable for our purposes. Likewise, as we are not attempting to solve the analogy task but merely measuring the relative regularity of different relations, we do not report predictive strength.

2.2 Experimental Procedure

The framework and assumptions we adopt naturally dictate how one can predict the vector of a word from information about a related word. The prediction is based on the computation of the mean shift vector for a morphological relation, as illustrated in Figures 1 and 2. First, we compute the difference between the vectors representing each pair of related words, e.g. $\vec{candidate} - \vec{candidat}$; these **shift vectors** are shown in red for adjectives and in blue for nouns in Figure 1. A shift vector can be seen as a functional representation of the semantic contrast that holds between two words. Second, for each morphological relation, we compute the average of all shift vectors. This gives rise to the mean shift vectors $\vec{m}_A$ for adjective pairs and $\vec{m}_N$ for noun pairs in Figure 2. The mean shift vectors thus represent the average semantic contrast between pairs of words in the relation.

The next step is to use these mean shift vectors

[3] The utility of vector shifts as inputs to word relation classification –Jameel et al. (2018) showed that difference vectors achieve a performance just slightly lower than specialized representations learned for this task– further confirms that, to some extent, they can be used to represent lexical relations.

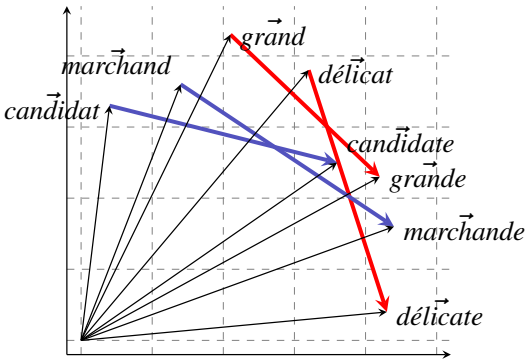

Figure 1: G-gender alternations

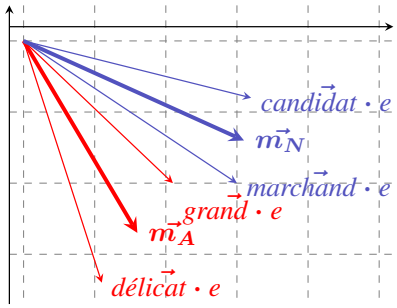

Figure 2: Mean shifts for the two processes

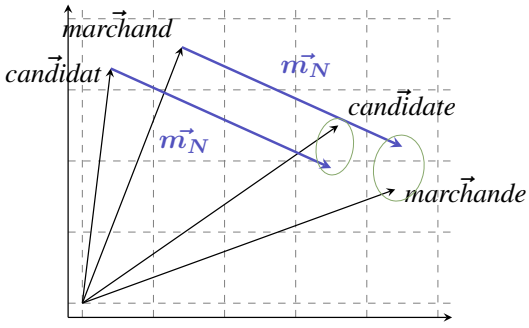

Figure 3: Intrinsic predictions for HNs

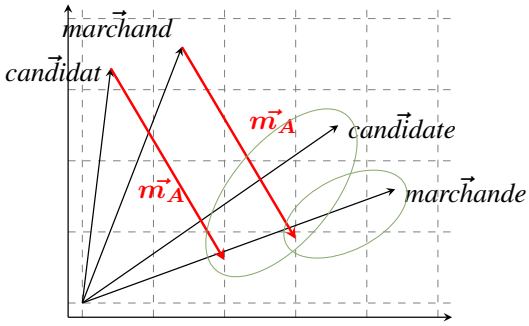

Figure 4: Extrinsic predictions for HNs

as prediction functions. The basic scenario, which we label **intrinsic prediction** (illustrated in Figure 3), can be used to address question 1 above. Given a word w_1 (e.g. the masculine noun *marchand*) participating in some morphological relation R, we add to the vector representation $\vec{w_1}$ of w_1 the mean shift vector for R (here $\vec{m_N}$). This gives us the predicted vector for the morphological alternant w_2 (in this example, the feminine noun *marchande*). We may now assess how far the predicted vector $\vec{w_1} + \vec{m_N}$ falls from the actual observed vector $\vec{w_2}$ of the morphological alternant. Various measures can be used to quantify predictability in this case. In this paper we use two: the Euclidean distance between the predicted and observed vectors for the alternant, and the log rank of the actual vector in terms of distance from the predicted vector within the vector space.

The more diverse the shifts within a relation, the less accurate this intrinsic prediction will be on average. To address question 1, we will therefore compare the quality of intrinsic prediction for pairs of nouns and pairs of adjectives: the answer to it will be positive if prediction is less accurate for nouns than for adjectives.

A different procedure, which we call **extrinsic prediction**, allows us to address question 2 above, and is illustrated in Figure 4. Informally, we test to what extent two relations have the same distributional footprint. Given a word w_1 (e.g. the masculine noun *marchand*) participating in morphological relation R, we add to the vector representation of that word the mean shift vector for the *other* relation R' (here $\vec{m_A}$). If relations R and R' are semantically equivalent (i.e., g-gender alternation has the same effects for nouns and adjectives), then extrinsic prediction should be just as accurate as intrinsic prediction. If, on the other hand, the two relations are not equivalent, we expect extrinsic prediction to be less accurate.

3 Experiment 1: Overall comparison of HNs and adjectives

We start by addressing questions 1 and 2 in broad terms, and will proceed with a more fine-grained analysis of adjectives in subsequent sections.

3.1 Stability of contrast

Research Question The first question addressed in our paper concerns the relative degrees of semantic regularity of gender alternation in human nouns vs. adjectives. We compare the quality of

Measure	t-statistic	p-value
Distance	2.8824	0.0047
Log rank	1.1095	0.2694

Table 3: T-test results for intrinsic predictions of HNs and adjectives in the $\mathcal{M}_1$ model

intrinsic predictions for the two processes.

Materials The corpus used in our experiments is the concatenation of FRWAC (Baroni et al., 2009), FRCOW (Schäfer, 2015) and a dump of French Wikipedia, a total of 14 bln tokens, annotated with Coavoux's (2017) parser. FRWAC was cleaned to remove sentences containing characters not belonging to standard French and duplicate sentences. This corpus was used to compute a word2vec (Mikolov et al., 2013a) model (parameters: CBOW, negative sampling, window of 5), referred to as $\mathcal{M}_1$, where feminine and masculine homographs as well as noun and adjective homographs were disambiguated. HNs were selected from the GLAWI database (Hathout et al., 2014), enriched with information from the Lexeur lexicon (Fabre et al., 2004). To obtain a sample of homogeneous frequency, we only selected word forms occurring between 100 and 1000 times in our corpus. These constraints resulted in 120 HN pairs and 4874 adjective pairs.

Statistical Results We compute the discrepancy between predicted vector and observed vector using both log-normalized rank and euclidean distance[4]. The predictions were compared using a Welch t-test, cf. Table 3. We find that, in terms of distance, adjective predictions are closer to their respective targets than HN predictions.

Discussion The observed difference between adjectives and HNs provides evidence that the semantic relationships between masculine and feminine nouns are less predictable than those between masculine and feminine adjectives. This is coherent with the hypothesis that masculine and feminine nouns are related by derivation, and hence entertain a less regular relation than inflectionally related masculine and feminine adjectives.

[4]We preferred testing both rank and distance measurements over the more widespread cosine similarity measure so as to take into account neighborhood structure (Linzen, 2016).

Measure	t-statistic	p-value
Distance	-5.772	$< 10^{-7}$
Log rank	-6.245	$< 10^{-8}$

Table 4: T-test results for HNs, intrinsic vs. extrinsic prediction in the $\mathcal{M}_1$ model

Measure	t-statistic	p-value
Distance	-39.328	$< 10^{-15}$
Log rank	-33.169	$< 10^{-15}$

Table 5: T-test results for adjectives, intrinsic vs. extrinsic prediction in the $\mathcal{M}_1$ model

3.2 Similarity of Semantic Effects

Research Question Turning to question 2, we test whether gender alternations in nouns and adjectives have the same semantic effect. We do so by comparing the intrinsic and the extrinsic predictions for adjectives, and the intrinsic and the extrinsic predictions for HNs.

Statistical Results Using the same materials as previously, we test whether the two processes yield similar outputs for the same input. Table 4 presents the Welch t-test results for HNs, and Table 5 reports those for adjectives. All tests highlight a significant statistical difference between the intrinsic and the extrinsic predictions: intrinsic predictions always yield lower measurements.

Discussion Comparing predictability measures allows us to test the similarity of the semantic effects of g-gender alternation in HNs and in adjectives. The morphological and syntactic similarity of these processes does not logically imply a semantic identity; the measurements described above provide evidence for the opposite.

Perhaps this decline in predictability indicates a difference between the meaning of g-gender for nouns and adjectives. However, it might also be due to an imbalance in the data: we compared human nouns, where g-gender plausibly signals s-gender, with all adjectives, despite the fact that many adjective tokens describe inanimate entities and hence cannot receive any interpretation in terms of s-gender. To assess this, we need to examine how the human reference of entities described by adjectives influences their distributional properties.

4 Experiment 2: Differences among adjectives

Research Question Adjectives describing primarily humans might be more similar to HNs than other adjectives: they should mostly share the same context and convey similar s-gender information. If so, the more an adjective is used to qualify HNs rather than other nouns, the more similar its gender shift will be to the mean shift of HNs. Adjective shifts can then be expected to express a continuous trend from adjectives primarily used to describe humans (e.g. *talentueux, talentueuse*, 'talented') to those not necessarily describing humans (e.g. *grand, grande*, 'tall'), and to those (almost) never describing human referents (e.g. *plénier, plénière*, 'plenary').

Materials We compute the mean HN gender shift and compare it to the shift vector for each pair of adjectives. The same set of HNs extracted from GLAWI and Lexeur was used as in the previous experiment. We only considered HNs occurring at least 50 times to compute this mean shift. Adjectives were extracted from the GLAWI database; for each adjective we also computed its number of occurrences (in either gender) as the modifier of a HN on the basis of the dependency annotation by Coavoux's (2017) parser. When divided by the total number of occurrences of the adjective, this defines a ratio of usage as qualifying a HN. As with HNs, we only considered adjective forms occurring more than 50 times in our corpus for this experiment, resulting in a total of 15624 adjective pairs.

Statistical Results The hypothesis was tested using a mixed-effects model. Linear effects include log-frequency and shift size (factors of statistical noise), as well as the human qualification ratio. The lexical identity of the adjective was used as a random effect. To obtain a normal distribution, the dependent variable was transformed to $-\log(\log(\frac{1}{\cos(\vec{A^i_f} - \vec{A^i_m}, \vec{m_N})}))$ with $\vec{A^i_f} - \vec{A^i_m}$ the shift for a given adjective and $\vec{m_N}$ the mean noun shift, then rescaled between 0 and 1 for interpretabilty purposes. Note that the dependent variable is monotonic with respect to cosine similarity.

The model was run using the R-Studio LME4 library (Bates et al., 2015), and converged to the results described in 6. All predictors were deemed significant. An analysis of residuals showed that the model is sound and accurate. The quantitatively most important effect is associated with the ratio, which contributes to higher cosine values.

Predictor	Estimate	t-statistic	p-value
Intercept	0.2288	34.35	$< 10^{-15}$
Log freq.	0.0578	42.65	$< 10^{-15}$
Ratio	0.1657	10.20	$< 10^{-15}$
10.20	$< 10^{-15}$		
Shift size	−0.0140	−40.38	$< 10^{-15}$

Table 6: Fixed effects for model of homogeneity of gender contrast in adjectives

Discussion The model highlights the importance of the type of nouns that the adjective qualifies to the semantic effect of the g-gender alternation. It stresses that regularity of gender alternation does not hold in an absolute fashion, and that the semantic contrasts between two related words is modulated by their common lexical semantics.

The precise nature of the sub-regularity indicates that adjective g-gender alternation in some cases resembles that of nouns. This provides an objective basis to disambiguate adjectives according to their usage. We can now use this information to tease apart adjectives which are semantically comparable to human nouns from those that are not.

5 Experiment 3: Comparing HNs to two classes of adjectives

The next experiment aims at studying g-gender alternation within three groups: human noun qualifying adjectives (HQAs), non-human noun qualifying adjectives (NHQAs), and HNs.

5.1 Stability of contrast

Research Question We first compare intrinsic predictions pairwise to assess the relative regularity of our three classes. Since NHQAs modify inanimate or abstract nouns, which do not possess an s-gender, we expect a different degree of regularity within NHQAs than within HQAs.

Materials We extract from the GLAWI database nouns which can only refer to humans, as well as nouns which never refer to humans. We define NHQAs as the adjectives which only qualify nouns that never refer to humans, and HQAs as the adjectives which only qualify nouns that always refer to humans.

Predictors		diff.	Adj. p-val.
HNs	vs. HQAs	-0.00783	0.79589
HNs	vs. NHQAs	0.04846	0.00741
NHQAs	vs. HQAs	-0.05629	0.00005

Table 7: Tukey HSD test results for distance measurements of intrinsic predictions in model $\mathcal{M}_2$

Predictors		Diff.	Adj. p-val.
HNs	vs. HQAs	-0.62057	0.04685
HNs	vs. NHQAs	0.85721	0.03510
NHQAs	vs. HQAs	-1.47778	$< 10^{-15}$

Table 8: Tukey HSD test results for log rank measurements of intrinsic predictions in model $\mathcal{M}_2$

A new model, dubbed $\mathcal{M}_2$, is computed so as to provide a distinct representation for HNs, HQAs and NHQAs. We once again use a word2vec model (CBOW, 5 negative samples, window of 5). In this $\mathcal{M}_2$ model, we disambiguate feminine vs. masculine, adjectives vs. nouns, and HQAs vs. NHQAs vs. ambiguous usages of adjectives.

Consistently with our first experiment, we only consider items occurring between 100 and 1000 times. However, this constraint resulted in sets of very different sizes: 118 noun pairs, 481 HQA pairs and 5074 NHQA pairs. Our NHQA sample is an order of magnitude bigger than the other classes and, more importantly, constitutes less of a natural class semantically. Such a disbalance might impact the results, therefore it was necessary to select the most cohesive group of NHQAs. This was done by retrieving the bottom-most cluster containing enough samples from UPGMA hierarchical clustering[5] and resulted in a set of 101 NHQA pairs.

Statistical Results We compare three morphological processes simultaneously, conducting an analysis of variance (ANOVA) to see if a given measure could discriminate the different processes; if it does, we apply Tukey's Honest Significant Difference (HSD) test to provide estimated factor and adjusted probabilities for each pair of processes. In all the case studies shown here, ANOVAs give strong evidence for differences with the processes ($p < 10^{-4}$, both for Euclidian distance and log rank). We report only HSD test results in the interest of space.

The adjusted p-values for distance of the Tukey HSD test in Table 7 underscore no significant difference between nouns and HQAs, but NHQAs are shown to yield lower measurements than HQAs and HNs; both p-values are under 0.05. This suggests that in terms of distance NHQAs are more regular than nouns and HQAs, and that nouns and HQAs cannot be really distinguished from one another in terms of regularity.

Rank measurements, cf. Table 8, accordingly highlight the same difference between NHQAs on the one hand and nouns and HQAs on the other. Moreover, adjusted p-values indicate that there is a significant difference in measurements not only when comparing nouns and HQAs to NHQAs, but also when comparing nouns to HQAs. This suggests that NHQAs embody the most regular process, followed by HNs, and that HQAs correspond to the least regular process.

Discussion The hypotheses that this experiment tested were that HQA g-gender alternation was more similar to HN g-gender alternation than NHQA due to similarity in lexical meaning, and that NHQA pairs exhibited more regular shifts than HQA pairs.

A significant difference between the two processes was found: NHQAs yield lower rank and distance measurements than HQAs. It is however noteworthy that the difference between the measurements for the two processes is very small, so the distinction it introduces is subtle.

Another point of interest is that HN pairs exhibit more regularity than HQAs, but less than NHQAs. This suggests that, when compared to semantically comparable adjectives, human nouns fall within the scope of semantic regularity expected for inflectional alternations. This more careful experiment hence disproves the tentative conclusions reached after experiment 1: if anything, distributional evidence points to an inflectional status for g-gender alternations in human nouns.

5.2 Similarity of Semantic Effects

Research Question We now turn to the comparison of extrinsic and intrinsic predictions. The expectation is that g-gender alternation of HNs is more similar to that of HQAs than to the alterna-

[5] We tested using a Tukey HSD test whether class influenced pairwise distances measures within groups of distinct gender and class. NHQAs initially introduced a difference of ten times what we observed for other classes. With clustering, we observed a variation of means of 0.016 ± 0.008. Applying the same constraint to HQAs as well made HQAs overly cohesive but did not substantially affect the results.

Predictors	Diff.	Adj. p-val.
nouns vs. HQAs	-0.10139	$< 10^{-6}$
nouns vs. NHQAs	-0.11102	$< 10^{-15}$
NHQAs vs. HQAs	0.00963	0.86564

(a) Tukey HSD test results for distance measurements

Predictors	Diff.	Adj. p-val.
nouns vs. HQAs	-2.00157	$< 10^{-6}$
nouns vs. NHQAs	-3.19196	$< 10^{-15}$
NHQAs vs. HQAs	1.19039	0.00427

(b) Tukey HSD test results for log rank measurements

Table 9: Tukey HSD tests results for HNs, intrinsic vs. extrinsic prediction in model $\mathcal{M}_2$

Predictors	Diff.	Adj. p-val.
HQAs vs. nouns	-0.07383	$< 10^{-15}$
HQAs vs. NHQAs	-0.08725	$< 10^{-15}$
nouns vs. NHQAs	-0.01342	0.22433

(a) Tukey HSD test results for distance measurements

Predictors	Diff.	Adj. p-val.
HQAs vs. nouns	-0.99466	$< 10^{-15}$
HQAs vs. NHQAs	-2.75437	$< 10^{-15}$
nouns vs. NHQAs	-1.75971	$< 10^{-15}$

(b) Tukey HSD test results for log rank measurements

Table 10: Tukey HSD tests results for HQAs, intrinsic vs. extrinsic prediction in model $\mathcal{M}_2$

tion of NHQAs, due to their greater semantic relatedness; moreover, if g-gender alternation in nouns is indeed derivational, we expect nouns to diverge significantly from adjectives.

Statistical Results We start with examining predictions for HNs. An ANOVA shows that both distance and log-scaled rank measurements highlight a variation among the different prediction setups. We thus perform Tukey HSD tests, summarized in Table 9, to study more precisely what these differences in distance and log rank entail. From distance measurements (cf. 9a), we see that using information from either HQA g-gender alternation or NHQA g-gender alternation clearly deteriorates the measurements for HNs. Moreover, no significant statistical effect is attested when comparing the two extrinsic predictions.

When studying log rank measurements (cf. 9b), the same deterioration can be observed. However, log rank measurements also reveal a significant difference when comparing the two extrinsic predictions: NHQAs yield even higher measurements than HQAs, suggesting that the semantics of HQA alternation are more similar to that of HNs than the gender alternation of NHQAs.

The second group of extrinsic predictions concerns HQAs. Tukey HSD tests in Table 10 display the same results as with nouns, both in terms of distance (cf. 10a) and in terms of log-rank (cf. 10b) which clearly indicate that noun-based extrinsic predictions are a better fit than NHQA-based extrinsic predictions. This might entail that there is a gradation of semantic effects' similarity: HNs would be more similar to HQAs than NHQAs.

The last group of extrinsic predictions are those

Predictors	Diff.	Adj. p-val.
NHQAs vs. nouns	-0.34385	$< 10^{-15}$
NHQAs vs. HQAs	-0.41443	$< 10^{-15}$
nouns vs. HQAs	-0.07061	$< 10^{-06}$

(a) Tukey HSD test results for distance measurements

Predictors	Diff.	Adj. p-val.
NHQAs vs. nouns	-5.49577	$< 10^{-15}$
NHQAs vs. HQAs	-7.24153	$< 10^{-15}$
nouns vs. HQAs	-1.74575	$< 10^{-6}$

(b) Tukey HSD test results for log rank measurements

Table 11: Tukey HSD tests results for NHQAs, intrinsic vs. extrinsic prediction in model $\mathcal{M}_2$

for NHQAs. As previously, after an ANOVA, a Tukey HSD test is conducted for each pair of measurements; the results are summarized in Table 11. When studying either distance (cf. 11a) or log-rank (cf. 11b) variation, we observe both that intrinsic prediction performs better than extrinsic predictions, and that the extrinsic prediction based on HNs yields better measurements than the one based on HQAs. This implies that the semantic effects of NHQAs are more similar to those of HNs than to those of HQAs.

Discussion Intrinsic prediction is systematically better than any of the extrinsic predictions, highlighting that all three groups embody different semantic processes. We, however, observe a gradient: gender alternations for HQAs and HNs are more similar to each other than to NHQAs, and HNs are somewhere in between the two classes of adjectives. HNs and HQAs form a more cohesive group from which NHQAs differ systematically.

Although the cohesiveness of HNs and HQAs might be explained by the mechanics of distributional semantics, this in and of itself does not suffice to explain the gradient effect we observe. One could tentatively conclude from these facts that s-gender plays a greater role for HQAs compared to HNs. The concreteness of HNs may entail that all speakers agree on their semantics: a woman manning a checkout desk shall necessarily be *une caissière*; on the other hand, whether to use a specific adjective, such as *délicat*, to qualify a human referent, depends on the speaker's judgment which can be sensitive to s-gender. In other words, the person at the cash register is objectively a cashier regardless of their s-gender, but the standards of being *delicate* can be different for men and women, which in turn might explain the relatively idiosyncratic character of g-gender alternation in HQAs.

6 General Discussion

In this paper, we have detailed a data-driven methodology which enables a comparison of the distributional effects of a grammatical feature across categories. This methodology has allowed us to make several observations on g-gender alternation in HNs and adjectives.

In the first set of experiments, we compared the grammatical gender feature of French HNs to its counterpart in adjectives. We observed a greater semantic regularity in adjectives, which we tentatively attributed to the status of the gender distinction in nouns vs. adjectives: pairs of nouns are related by derivation, but pairs of adjectives are related by inflection. In addition, the comparison of intrinsic and extrinsic predictions highlighted a clear semantic difference between g-gender contrasts in nouns and adjectives.

Experiment 2 showed that another factor comes into play: the shift for an average adjective pair is more likely to resemble that of nouns, when the adjective pair itself is used to qualify HNs. This lead us to compare HNs with HQAs and NHQAs in the following section.

In the last set of experiments, g-gender variation in HNs was shown to be more semantically regular than in HQAs. Hence the provisional conclusion of the first set of experiments was disproved: the apparent semantic irregularity of HNs was due to comparing them with a semantically discommensurate class of adjectives. On the other hand, the experiment still highlighted that all three groups constituted distinct processes.

Although all three types of gender shifts significantly differed from one another, we observed that HQAs and HNs formed a more cohesive group. We can derive two conclusions from this. First, g-gender alternation within a category (adjectives) can vary more than across categories (adjectives vs. nouns); second, correlation with s-gender in both nouns and adjectives lead to a greater commonality in gender alternations. This result corroborates, on the basis of distributional data, the mounting sociolinguistic (McConnell-Ginet, 2013, a.o.) and psycholinguistic (Gygax et al., 2012, a.o.) evidence that, when referring to humans, g-gender always has some interpretive effects. In addition, we have shown that g-gender alternation of HNs is more regular than that of some adjectives. Given that inflection is assumed to be more semantically regular than derivation (Robins, 1959; Dressler, 1989), this suggests that gender alternations in nouns should be seen as inflectional, as argued on independent grounds by (Bonami and Boyé, in press).

This work has addressed theoretical issues regarding one grammatical feature from a data-driven perspective. Future research will determine to what extent our results are specific to gender or generalize to other grammatical features such as number. Complementarily, we plan to look in more detail at the specific contribution of gender in languages where the relationship between g-gender and s-gender is different than in French. Finally, we plan to test the potential impact of social bias on the usage of gender forms of HQAs, which was suggested in the preceding section, both from the distributional and from the psycholinguistic point of view.

Acknowledgments

We thank Heather Burnett, Alessandro Lenci, Enrico Santus, and three anonymous reviewers for useful comments on previous versions of this paper. The work was supported by two public grants overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program: Labex *Empirical Foundations of Linguistics* (reference: ANR-10-LABX-0083) and Idex *Lorraine Université d'Excellence* (reference: ANR-15-IDEX-0004). Support also came from the CNRS PEPS grant *ReSeRVe*.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, Kyunghyun Cho, and CIFAR Azrieli Global Scholar. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web:a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Olivier Bonami and Gilles Boyé. in press. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond, and Andrew Hippisley, editors, *Perspectives on Morphology*. Edinburgh University Press.

Olivier Bonami and Denis Paperno. forthcoming. A characterisation of the inflection-derivation opposition in a distributional vector space. *Lingua e Langaggio*. Forthcoming.

Heather Burnett and Olivier Bonami. in press(a). A conceptual spaces model of socially motivated language change. In *Proceedings of the 2nd Meeting of the Society for Computation in Linguistics*.

Heather Burnett and Olivier Bonami. in press(b). Linguistic prescription, ideological structure and the actuation of linguistic changes: Grammatical gender in french parliamentary debates. *Language in Society*, 48.

Maximin Coavoux. 2017. *Discontinuous Constituency Parsing of Morphologically Rich Languages*. Ph.D. thesis, Univ Paris Diderot, Sorbonne Paris Cité.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press., Cambridge.

Greville G. Corbett. 2010. Canonical derivational morphology. *Word Structure*, 3:141–155.

Greville G. Corbett. 2013. Sex-based and non-sex-based gender systems. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Wolfgang U. Dressler. 1989. Prototypical differences between inflection and derivation. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 42:3–10.

Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. *A functional theory of gender paradigms*. Brill, Leiden.

Cécile Fabre, Franck Floricic, and Nabil 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*. ERSS, Université de Toulouse II-Le Mirail.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *SRW@HLT-NAACL*.

Pascal Gygax, Ute Gabriel, Arik Lévy, Eva Pool, Marjorie Grivel, and Elena Pedrazzini. 2012. The masculine form and its competing interpretations in french: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology*, 24(4):395–408.

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary. In *Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, pages 65–74, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Charles F. Hockett. 1958. *A course in modern linguistics*. New York: Macmillan.

Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. 2018. Unsupervised learning of distributional relation vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180. Association for Computational Linguistics.

T. Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *ArXiv e-prints*.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122 3:485–515.

P. H. Matthews. 1974. *Morphology*. Cambridge University Press, Cambridge.

Sally McConnell-Ginet. 2013. Gender and its relation to sex: The myth of natural gender. In Greville Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton Berlin.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

R. H. Robins. 1959. In defense of WP. *Transactions of the Philological Society*, 58:116–144.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Andrew Spencer. 2013. *Lexical relatedness: a paradigm-based model*. Oxford University Press, Oxford.

Gregory Stump. 1998. *The handbook of morphology*, chapter Inflection. Oxford: Blackwell.

Rossella Varvara. 2017. *Verbs as nouns: empirical investigations on event-denoting nominalizations*. Ph.D. thesis, Università degli Studi di Trento.

Wolfgang Ulrich Wurzel. 1989. *Inflectional Morphology and Naturalness*. Kluwer, Dordrecht.

Wiecher Zwanenburg. 1988. *Aspects de linguistique franaise.*, chapter Flexion et dérivation: le féminin en franais.