Grammatical gender contrasts in French: a distributional semantic approach

MICKUS, Timothee

Supervisors: BONAMI, Olivier (LLF)
PAPERNO, Denis (LORIA)

Résumé

Le présent mémoire étudie l'assignation du genre grammatical dans une perspective sémantique distributionnelle. Le genre grammatical est un système de classification des noms basé sur l'accord (F. Hockett, 1958). Il convient de le distinguer du genre social, qui correspond à la valuation sociale de caractéristiques biologiques relevant du sexe (Eckert et McConnel-Ginnet, 2003).

Les noms humains en français correspondent à un point de contact entre ces deux concepts : prototypiquement, les noms humains féminins (par exemple *la présidente*) réfèrent à des femmes, alors que les noms humains masculins peuvent référer soit à des hommes, soit à des femmes (*le président*, *cf. Monsieur le président, Madame le président*). Cette asymétrie est révélatrice de deux tendances prescriptives simultanées du français contemporain : l'une différencie le genre social à l'aide de formes de noms humains d'un genre grammatical spécifique, l'autre utilise une seule forme dont le genre grammatical est fixe.

La systématicité de la première tendance pousse certains auteurs (Bonami et Boyé, à para.) à supposer que cette alternance en genre grammatical est flexionelle, ce qui va à l'encontre de la vision traditionnelle qui veut qu'un lexème soit doté d'un genre grammatical propre (par exemple Zwanenburg (1988)), et donc que l'alternance en genre grammatical soit dérivationnelle. Ce mémoire se propose de contribuer à départager ces deux hypothèses.

Stump (1998) propose cinq critères pour distinguer les processus flexionnels des processus dérivationnels. L'un d'entre eux stipule que les effets sémantiques d'un processus sont davantage réguliers si celui-ci est flexionnel.

Des travaux précédents (Bonami et Paperno, à para.) ont établi une méthodologie basée sur les modèles de sémantique distributionnelle (DSM) pour comparer la régularité sémantique de deux processus à l'aide de paires contrastives $\langle pivot, comp._1, comp._2 \rangle$, telles que *pivot* et $comp._1$ correspondent à un processus et *pivot* et $comp._2$ à l'autre. Pour calculer le DSM un corpus de 14 milliards de mots a été constitué en concaténant un dump du wikipédia français (Coavoux, 2017), frWaC (Baroni et al., 2009) et frCoW (Schäfer, 2015). Un lexique de noms humains a été extrait d'un réseau sémantique tiré du wiktionnaire (Muller, Hathout et Gaume, 2006). La première expérience compare l'alternance en genre grammatical des noms humains à la dérivation des noms d'agents déverbaux. Les résultats indiquent clairement qu'on observe davantage de régularité pour l'alternance en genre grammatical des noms humains, en particulier si l'on tient compte du bruit statistique.

Mais cette première expérience ne permet pas de tracer la frontière de régularité qui départagerait flexion et dérivation. C'est pourquoi il faut comparer les noms humains à un processus flexionnel : nous les comparons à l'alternance en genre grammatical des adjectifs. Nous définissions et utilisons une nouvelle méthodologie qui se passe de paires contrastives : nous comparons les processus deux à deux directement en mesurant leurs performances dans des tâches prédictives soit intrinsèques (lorsque la fonction de prédiction est calculée à partir du processus sur lequel elle sera appliquée) soit extrinsèques (lorsque la fonction de prédiction est calculée à partir de l'autre processus que celui sur lequel elle sera appliquée). Les premières donnent une mesure de la régularité d'un processus, les secondes indiquent la similarité des effets sémantiques des deux processus.

Comparant l'alternance en genre grammatical des noms humains à celle des adjectifs, on observe une plus grande régularité pour les adjectifs, ainsi qu'une claire différence d'effets sémantiques qu'on peut peut-être imputer à la différence entre flexion inhérente et flexion contextuelle (Booij, 1995). Cependant on observe que plus un adjectif est fréquemment utilisé pour qualifier des noms humains, plus son alternance ressemble à celles des noms humains. Comparant les noms humains aux adjectifs ne qualifiant que des noms humains, la différence de régularité précédemment établie s'évanouit. On observe même une plus grande différence entre adjectifs ne qualifiant que des noms humains et adjectifs ne qualifiant que des noms non-humains qu'entre adjectifs ne qualifiant que des noms humains et noms humains, et même (dans une moindre mesure) qu'entre les adjectifs ne qualifiant que des noms non-humains et les noms humains, suggérant que les adjectifs ne qualifiant que des noms humains sont davantage corrélés à des facteurs sociaux que les noms humains.

De ceci on peut conclure que soit, *contra* Zwanenburg (1988) et *pro* Bonami et Boyé (à para.), l'alternance en genre grammatical des noms humains est flexionnelle, soit le critère de régularité de Stump (1998) doit être révisé. Si l'on accepte la première alternative, les expériences conduites indiquent que le statut du genre grammatical comme système de classification ne peut être maintenu que si l'on distingue bien classification de partition.

Abstract

The present thesis studies grammatical gender assignment in a distributional semantics perspective. Grammatical gender is a noun classification system based on agreement (F. Hockett, 1958). It is to be distinguished from social gender, which corresponds to the social valuation of sexual biologic characteristics (Eckert and McConnel-Ginnet, 2003).

Human nouns in French correspond to a point of contact between these two concepts: prototypically, feminine human nouns (eg. *la présidente* 'the$_{F}$ female president') refer to women, whereas masculine human nouns can refer to either men or women (eg. *le président* 'the$_{M}$ president', *cf. Monsieur le président, Madame le président*, 'Mister President, Madam President'). This asymmetry reveals two simultaneous prescriptives tendencies of contemporary French: the one discriminates social gender using forms of human nouns of a specific grammatical gender, the other uses a single form with a fixed grammatical gender.

Due to the systematicity of the first tendency, some authors (Bonami and Boyé, forthcoming) suppose that this grammatical gender alternation is inflectional, which goes against the traditional view that assumes any lexeme is attributed a given grammatical gender (eg. Zwanenburg (1988)) and that therefore grammatical gender alternation is derivational. This thesis aims at testing which hypothesis fits best to the observed facts.

Stump (1998) proposes five criteria to distinguish inflectional and derivational processes. One of them stipulates that the semantic effects of an inflectional process are more regular than that of a derivational process.

Earlier works (Bonami and Paperno, forthcoming) established a methodology based on distributional semantics models (DSM) to compare the semantic regularity of two processes by means of contrastive pairs $\langle pivot, comp._1, comp._2 \rangle$, where *pivot* and $comp._1$ correspond to one process and *pivot* and $comp._2$ to the other. To compute the DSM, a 14 billion words corpus was constituted by concatenating a French wikipedia dump (Coavoux, 2017), frWaC (Baroni et al., 2009) and frCoW (Schäfer, 2015). A lexicon of human nouns was extracted from a semantic network derived from the French Wiktionary (Muller, Hathout, and Gaume, 2006). The first experiment compares grammatical gender alternation of human nouns to deverbal agent nouns derivation. Its results clearly indicate that more regularity is attested for grammatical gender alternation of human nouns, especially when taking statistical noise into account.

However, this first experiment isn't sufficient to set a regularity threshold that would tease apart inflection and derivation. One therefore needs to compare human nouns to an inflectional process: we compare them to grammatical gender alternation of adjectives. We define and employ a new methodology which does not rely on contrastive pairs: we directly compare processes two-by-two by measuring their performances in predictive tasks, either instrinsic (when the predictive function is computed from the process on which it will be applied) or extrinsic (when the predictive function is computed from the other process than the one on which it will be applied). The former measure the regularity of a process, the latter indicate the similarity of the semantic effects of the two processes.

When comparing grammatical gender alternation of human nouns to that of adjectives, one can observe a greater regularity for adjectives, as well as a clear difference in terms of semantic effects — which may perhaps be imputed to the difference between inherent inflection and contextual inflection (Booij, 1995). However, one can also observe that the more an adjective is used to modify human nouns, the more its alternation resembles that of human nouns. When comparing human nouns to adjectives qualifying human nouns only, the previously established difference in terms of regularity vanishes. One can even observe that the difference between adjectives qualifying human nouns only and adjectives qualifying non-human nouns only is greater than that between adjectives qualifying human nouns only and human nouns, and what's more greater than that between adjectives qualifying non-human nouns only and human nouns (although in lesser proportions). This suggests that adjectives qualifying human nouns only are more correlated to social factors than human nouns.

From these experiments, one can conclude either that — *contra* Zwanenburg (1988) and *pro* Bonami and Boyé (forthcoming) — grammatical gender alternation of human nouns is inflectional, or that the semantic regularity criterion of Stump (1998) needs to revised. If the first alternative is to be accepted, the conducted experiments entail that the status of grammatical gender as a classification system can only be maintained if classification is distinguished from partition.

Contents

List of Figures

List of Tables

1 Introduction

This study concerns itself with grammatical gender assignment within a distributional semantics framework. As such, it stands at the crossroad between two different domains: on the one hand, morphological studies based on distributional semantics models — which are strikingly rare —, on the other, the interface between social gender and grammatical gender — which has been researched from a wide array of different perspectives, ranging from psychology to sociolinguistics and to morphosyntactic typology. Therefore, the present study concerns itself with both empirical methodology — as it employs a fairly recent methodology to study a widely debated phenomenon — and theoretical implications — as it brings about a new array of observations to discuss this very phenomenon.

In particular, this study addresses three questions:

- Is distributional semantics a sound model of theoretically abstruse morphological processes?
- Is the grammatical gender alternation of French human noun more akin to inflection or derivation?
- What role does social gender play in grammatical gender assignment in French?

This study therefore tackles complex issues regarding gender simultaneously, and cannot aim at completeness.

The term of "gender" itself is ambiguous: it can be defined either as a social construct, or as a linguistic phenomenon. The linguistic phenomenon of grammatical gender or g-gender is a noun classifying system that is attested in some languages only (Corbett, 1991). The social construct of social gender or s-gender corresponds to stereotypical identities assigned to humans on the basis of their biological sex, and is pervasive to all cultures (Eckert and McConnel-Ginnet, 2003). It is therefore difficult to study gender as a whole — this is why this thesis aims at studying gender within a restricted scope: that of grammatical gender assignment to French human nouns using the framework of distributional semantics.

Human nouns correspond exactly to the point of contact between the two distinct concepts of social and grammatical gender. For instance, they have been the subject of opposite linguistic prescriptions in recent years: the one, conservative, dictates the use of a single form regardless of the social gender of its referent (thus leading to occurrences such as *Madame le Président*, 'Madam President$_{\text{MASC}}$'), the other tends to associate distinct forms to distinctively socially gendered individuals, using grammatical gender to match a form to a social category (thus contrasting *Monsieur le Président*, 'Mister President$_{\text{MASC}}$' with *Madame la Présidente*, 'Madam (female) President$_{\text{FEM}}$'). In fact, that we speak of "feminine" and "masculine" gender is due to the fact that the first is prototypically assigned to feminine socially gendered individuals, whether the second prototypically refers to masculine socially gendered individuals.

Yet, one should not confuse grammatical and social gender, as these concepts are clearly independent one from the other — social gender is attested for languages lacking grammatical gender such as English, and grammatical gender systems may also encompass genders prototypically referring for instance to inanimates, animals, or edible things (Corbett, 1991). For instance, whereas in French the feminine g-gender (*la table*, 'the table') contrasts with the masculine g-gender (*le tableau*, 'the board'), Oneida (Iroquian) possesses four distinct g-genders: one for inanimates, one for females and animals, one for females and indefinite referents and one for masculine (Michelson, 2015). What is more, grammatical gender assignment is always at least to some extent arbitrary: there is no reason why a table should be feminine, whereas a board should be masculine.

Moreover, some authors have discussed the status of the grammatical gender alternation of French human nouns: although standardly analyzed as a case of derivation (Stump, 1998), the regularity and systematicity of the process may prove enough to classify it as inflectional (Bonami and Boyé, forthcoming). The complexity of the phenomenon at hand entails that it should be studied from various perspectives. In this respect, the present study will also focus on presenting a wide array of methodological protocols, so as to provide the debate with factual, objective observations of the relevant data. It is rarely the case that a study addressing morphological issues does so from a data-driven perspective — research in morphology tends to be focused either on descriptive statistics or on theoretical models. Studies discussing the distinction between inflection and derivation in particular are very rarely based on statistical measurements; likewise, the elusive nature of gender often entails that only a handful cases be considered. There is therefore a need for a new methodology, one that could study the theoretical assumptions in the light of objective facts abstracted from data.

In addition, as the study of grammatical gender assignment in human nouns is also obfuscated by the social bias that comes with it, it is all the more preferable when studying grammatical gender to adopt a data-driven framework so as to avoid as much as possible any bias — whether conscious or not. Moreover it is only when adopting a standpoint agnostic with respect to the actual structure of the data when debating the morphological nature of this process — *id est*, whether it is inflectional or derivational — that sound, objective, factual arguments can be brought about. In this respect, this research aims at producing observations covering as many cases as possible, which can only prove helpful to the study of grammatical gender assignment.

Studying words from a statistical standpoint can be done either by considering the statistical distribution as a whole, or by studying the context of apparition of each word. As this thesis deals with grammatical gender, which is traditionally defined using agreement, *id est* through its effects on its syntactic context (F. Hockett, 1958), a more accurate description can be captured from studying the context of g-gendered words. From this stems that distributional semantics models, which provide representations of words based on their contexts (Firth, 1957), are apt to describe the dynamics of gender assignment.

The present study will therefore focus on grammatical gender assignment within the framework of distributional semantics. One of its goals is to provide a new perspective on the central question at hand, as well as an inkling of the strength and weaknesses of a distributional semantics approach to the study of morphology. Section 2 will cover a general presentation of the problem at hand from a linguistic point of view, as well as a presentation of distributional semantics as a linguistic theory and of the mathematical mechanisms on which they are built. Following this presentation, in section 3 we will delve into a description of the resources and the necessary preparatory work so as to obtain a lexicon of French human nouns. The next two sections will compare human nouns g-gender assignment to other morphological phenomena: section 4 will detail a comparison of the morphological process of grammatical gender alternation to a typical derivational process, deverbal agent noun formation, while section 5 will compare it to a typical inflectional process, adjectives gender alternation. Lastly, section 6 concludes this study by summarizing the empirical observations conducted and their theoretical implications.

2 General Presentation

Gender in linguistics has recently gathered quite a lot of interest.

Gender, in a word, is elusive. Acquiring the grammatical gender of words in a given language is one of the most difficult tasks a non-native speaker is confronted with upon learning. The sociolinguistic take on gender, on the other hand, has proven that this notion is riddled with preconceptions that are hard to get away from even for the researchers themselves. The mechanisms behind the assignment of a given gender to a noun take roots in linguistic domains as varied as phonology, semantics and pragmatics. The term itself of gender might refer to vastly differing objects of studies depending on the specific field of the researcher. All in all, gender has yet to receive a coherent cross-domain definition.

French human nouns are at the crossroad of many of these domains. For one thing, their usage has been rapidly evolving in the last few years, consequence to their being invested in in public debates. The prescriptive push of a part of the society towards gender-sensitive human nouns has been widely displayed and relayed, which in turn amplified the importance of said human nouns in French sociolinguistics. For another, their very definition as *human nouns* calls on notions both morphosyntactic and semantic. They therefore seem a very natural candidate to pick when studying more closely the role of gender in French, and help highlighting the main landmarks of the study of gender in linguistics.

2.1 Grammatical Gender vs. Social Gender

In recent years, the concept of gender has proved to be a topic of heated debate, both in academic and social circles. From a linguistic point of view, two conceptions of what "gender" can refer to are opposed: they can be roughly distinguished as **social gender** on the one hand and **grammatical gender** on the other hand. In order to avoid any confusion, they will be hereafter abbreviated **s-gender** and **g-gender** respectively.

Social gender is defined as a social categorization based for the most part on biological differences between the sexes and reproductive capacity. To quote Eckert and McConnel-Ginnet (2003): "Sex is a biological categorization based primarily on reproductive potential, whereas gender is the social elaboration of biological sex".

On the other hand, grammatical gender refers to noun classification. A definition often quoted in literature can be found in F. Hockett (1958): gender is "classes of nouns reflected in the behaviour of associated words".

These two distinct definitions stem from distinct disciplines: the social definition of gender comes from sociology, whereas the grammatical definition of gender was inherited from traditional grammar. Rather than contradictory, these conceptions are complementary. Social gender and grammatical gender do not refer to the same set of observed facts, however some facts are relevant to both. On the one hand, grammatical gender is mostly a morphological and syntactic issue. The "behaviour of associated words" that Hockett mentions in most cases corresponds to agreement. Although some phonological and semantic cues can be taken into account when assigning a given grammatical gender, this assignment *per se* is arbitrary: compare for instance German *Mond* ('moon', masculine) and *Sohne* ('sun', feminine) vs. French *lune* ('moon', feminine) and *soleil* ('sun', masculine). On the other hand, facts relevant to social gender can be found even in languages that do not have a grammatical gender system. The most commonly studied example corresponding to such a case is, obviously, English.

This mutual ignorance would matter less if the observed facts did not intersect, however most grammatical gender systems distinguish classes based on a social valuation of the differences between sexes (Corbett, 1991). French, in particular, is one example of a language with two g-genders based on such a social valuation. One clear sign of its importance is the treatment of human nouns. Neologisms, as much as possible, are assigned a g-gender corresponding to the social perception of the yet unnamed item (Bonami and Boyé, forthcoming). New occupations, in particular, tend to give rise to two forms applying to either man or woman — for instance *joggeur* ('male runner', masc) vs. *joggeuse* ('female runner', fem) — or common gender nouns such as *biologiste* ('biologist'), *id est* nouns with a single form used in both g-genders.[1] In recent years, these human nouns in French have been a focal point of social prescription

[1]The French tradition also refers to common gender nouns as 'epicenes'. On the other hand, Corbett (1991) distinguishes 'epicenes' from 'common nouns', in that "[c]ommon nouns take two different sets of agreement forms, epicene nouns take only one, though they denote beings of either sex". He also calls "motion' nouns' nouns with an overt morphological gender marker, such as Spanish *hijo*, *hija* ('son', 'daughter'). Following Corbett's (1991) terminology, g-gender alternating human nouns in French are either common gender nouns (*biologiste*) or 'motion' nouns (*joggeur, joggeuse*), but not epicenes.

and debates. Not only neologisms, but also older nouns have been reinterpreted to fit that scheme of two forms. For instance the word *ministre* ('minister') is nowadays used both as a feminine or a masculine (Burnett and Bonami, submitted), the feminine form being reserved to cases when said minister is a woman — although the original masculine form can also be used when referring to a female minister, giving rise to an asymmetry in the semantics of the two nouns.

This pattern of twin human nouns in French is closely related, from a morphological point of view, to the g-gender variation of French adjectives. The same affixal strategies can be observed for both pairs of human nouns and adjectival paradigms. For instance *vendeur* ('seller, man', masc) and *vendeuse* ('seller, woman', fem) are formally similar to *enjôleur, enjôleuse* ('charming, cajoling', adj). However, as Bonami and Boyé (forthcoming) highlighted, this observation leads naturally to reflect on the fact that French nouns might not all share the same paradigm shape, especially since common gender seems to be the default strategy for g-gender assignment of French human nouns. In other words, are pairs of human nouns two forms of one lexeme related by inflection, or are they distinct lexemes related by derivation? This in turn questions an assumption generally made when dealing with g-gender: how come, if grammatical gender is to be conceived as a noun-classifying system, that a single word can be assigned to two g-genders depending on social cues?

2.2 Grammatical Gender as a Challenge to Morphology

Figure 1: Opposing views of human nouns g-gender assignment

G-gender in French human noun also exhibits a few noteworthy characteristics that generally conflict or at least question traditional morphological points of view. In particular, the standard standpoint is to treat the alternation of g-gender in human nouns as a derivational process, rather than an inflectional process as was proposed in Bonami and Boyé (forthcoming). This standard view is due to the conception that g-gender is either inherent to the lexeme (as it is for instance for German nouns in general), or specified by agreement (as is the case for German adjectives). Figure 1 represents the contrast between this traditional perspective and that of Bonami and Boyé (forthcoming). On the left, as per the former, two lexemes (highlighted in red) are related by a derivational morphological process — whereas on the right, illustrating the latter, specific word-forms (in blue) of a single lexeme underspecified for g-gender are related by the fact that they are members of the same inflectional paradigm.

An advocate of the traditional view, as summarized on the left part of Figure 1, can be found in Zwanenburg (1988). His position is that the g-gender of a human noun is entirely specified by its lexeme. Discussing an example containing the plural form *féministes* of the human noun *féministe* ('feminist woman'), he writes:

> "[L]e trait féminin est un trait que le nom *la féministe* porte dans le lexique et qu'il introduit tel quel dans la phrase. Mais le trait pluriel est un trait qui est attribué au nom dans la D-structure sur la base du choix nécessaire entre singulier et pluriel. Ce n'est pas un trait proprement inhérent au mot, mais un trait imposé par la dérivation de la phrase en D-structure."[2]

The traditional view that Zwanenburg (1988) holds will be referred to as the **derivational analysis** of the g-gender alternation of human nouns: if the trait related to g-gender is retrieved from the lexicon representation of a word, then it should follow that g-gender alternation of human nouns is a morphological process involving two lexemes — which entails it is derivational.

On the other hand, Bonami and Boyé (forthcoming), whose position is depicted in the right hand part of Figure 1, argue that viewing human nouns as derivationally–related "completely fail[s] to capture the

[2] "The feminine trait is a trait that the noun *la féministe* ('the$_F$ feminist') carries in the lexicon and that it introduces as such in the sentence. However the plural trait is attributed to the noun in the D-structure on the basis of the necessary choice between singular and plural. It is not a trait properly inherent to the word, but a trait imposed by the derivation of the sentence in the D-structure."

shape of the French morphological system", and therefore take pairs of human nouns to be inflectionally related. In the following, we will refer to this position as the **inflectional analysis** of the g-gender alternation of human nouns.

As such, studying whether g-gender alternation is inflectional or derivational should provide a criterion to invalidate either one of these two contrasting analyses. On the other hand, it is important to note that both analyses are not free from theoretic implications.

2.2.1 Inflection vs. Derivation

The question of how to distinguish derivational from inflectional processes was addressed in Stump (1998). In this consensual summary of the relevant previous literature, the author discusses five criteria which, taken together, should discriminate inflectional from derivational processes.

Change in lexical meaning or part of speech The first criterion relates to the fact that transcategorial processes are necessarily derivational, or as Stump put it:

> "Two expressions related by principles of derivation may differ in their lexical meaning, their part-of-speech membership, or both; but two expressions belonging to the same inflectional paradigm will share both their lexical meaning and their part of speech."

Although, as the author himself highlights, this fact is challenged by transcategorial processes traditionally labelled as inflectional such as participles — verbal forms displaying many traits commonly found in adjectives — this generally correlates with the intuition that various forms of a same lexeme should hold the same part of speech. Another point that Stump mentions is that this first criterion also fails to characterize any non-transcategorial processes which, crucially, is the case of human nouns gender alternation.

Syntactic determination The second criterion proposed by Stump is that syntactic constraints only apply to inflectional processes:

> "A lexeme's syntactic context may require that it be realized by a particular word in its paradigm, but never requires that the lexeme itself belong to a particular class of derivatives."

From a view strictly concerned with syntax, one can consider that this criterion does not always apply to inflectional processes — in many instances, there is no practical syntactic cue from which one can discriminate the use of a form rather than the other: to make a specific case, an English perfect, rather than a present tense, in a simple clause. Stump also agrees that "not all inflectional morphology is directly relevant to syntax", citing as example the thematic vowels of the latin conjugations and declensions. When it comes to g-gender alternation, the syntactic context may or may not constraint the form, depending on the syntactic theory: as French distinguishes determiners by g-gender, a strictly linear account of a sentence therefore provides a strong constraint; on the other hand, if the g-gender of the determiner is only an extraneous property inherited by agreement, then the case still needs to be made. This also indicates that studying morphological processes should be done in a syntactically agnostic framework — or at least one consensual in that respect. Therefore, although this criterion is pertinent to the present study, it also imposes some constraint to the theoretical framework from which the process can be studied.

Productivity Thirdly, Stump mentions that an inflectional process almost always hold for all members of its relevant part of speech, whereas derivational processes generally don't:

> "Inflection is generally more productive than derivation."

His point here is that virtually every English noun has a plural, whereas not all verbs have a deverbal agent noun. This however begs the question of defective paradigms and periphrastic forms, which are issues closely related to that of paradigm uniformity (see below, 2.2.2). In the specific case at hand in this study, the productivity of the process is limited by a semantic factor, that the noun refers to a human — however the same can be found for comparatives and superlatives of adjectives which hold no sort of gradability, as for instance *tall*, *taller* vs. *dead*, **deader*. As for how systematically this process can be employed, although as Burnett and Bonami (submitted) showed the presence of two sociolinguistic prescriptive tendencies, the criterion to distinguish them is a proportion of usage rather than grammaticality judgements. This therefore indicates that the process is rooted firmly enough to be systematically employed.

Semantic regularity The fourth criterion is that of semantic regularity:

> "Inflection is semantically more regular than derivation."

To illustrate this criterion, Stump compares the English third person suffix *-s* and the English denominal verb suffix *-ize*, arguing that the former has the same effect from a verb to another, whereas the latter presented contrasted effects depending on the noun it was related to (Stump's selected example were *winterize*, 'prepare for winter', *hospitalize*, 'put into a hospital', and *vaporize*, 'turn into vapor'). This is closely related to the idea that inflection is intrinsic to the lexeme, whereas derivation produces lexemes from other lexemes — thus the shape of the paradigm itself guarantees the regularity of the operation. In that respect, semantic regularity is related to productivity. Although this account has been challenged, especially by Štekauer (2014), this criterion might prove useful to the current study as long as this notion of semantic regularity can be operationalized.

Closure The last criterion mentioned in this article is that of closure, *id est*, the possibility for a given derived form to serve as base for further morphological processes.

> "Inflection closes words to further derivation, while derivation does not."

Although quite attractive, this criterion is, once again, not absolutely applicable. For instance, one of the formal means used to produce g-gender alternation is the paired deverbal agentive suffixes *-eur* and *-euse*. However, deriving an agent noun with either suffix seems to close the word to further processes: applying the denominal verb suffix *-iser* to a feminine agent noun in *-euse* sounds extremely unnatural, whereas the masculine form, which would yield *-oriser* is at best attested marginally. Although this informal outlook is in all respect rather informative, it only brings a negative proof from negative facts — which implies that this criterion is difficult to study.

Generally speaking, Stump's five criteria do not provide a clearcut distinction, but rather some sort of guideline to establish this distinction — the author himself writes: "the logic of inflection does not entail that the five criteria [...] should partition morphological phenomena along the same boundary; the extent to which the criteria do coincide therefore suggests that a number of independent morpholexical principles are sensitive to (if not categorically constrained by) the distinction between inflection and derivation".

Interestingly enough, and as Stump also remarks, another conclusion is possible — that the distinction between inflection and derivation is not motivated by the data. One proponent of this alternative conclusion is Štekauer. In Štekauer (2014), discussing the notion of "derivational paradigm", he writes:

> "In contrast to the traditional false belief of a clear-cut boundary between them [inflection and derivation], cross-linguistic research in recent decades has shown that these differences are not of a disjunctive nature. On the contrary, the boundaries between inflection and derivation are rather fuzzy. The inflection–derivation relations can thus be represented as a line with more typical (prototypical) and less typical (peripheral) cases along the inflection–derivation continuum."

Very little quantitative analysis has been made to contrast these two alternative conclusions. One article pertaining to such analyses is Bonami and Paperno (forthcoming), using a mathematical representation of words as vectors to show the difference between inflectional and derivational processes. The only criteria from Stump (1998) which can be subjected to such analyses are that of syntactic determination and semantic regularity. The present study will therefore focus on these two specific criteria so as to determine objectively whether the g-gender alternation of human nouns is derivational or inflectional in nature.

2.2.2 Paradigm Uniformity

Another element that has been briefly mentionned in the previous section and of relevance to this case is that of paradigm uniformity. That grammatical gender for French human nouns be conceived as an inflectional distinction, like number is to French nouns in general, implies an important case of paradigm shape variation for French nouns. Although, as Bonami and Boyé (forthcoming) remind, the same non-uniformity can be found for intransitive verbs in languages with object agreement — for example Nahuatl distinguishes transitive verbs prefixed either with a pronominal prefix or using *-te* and *-tla* for indefinite

object when needed from intransitive verbs who never exhibit such prefixes — the case is more rarely made when it comes to noun. Zwanenburg (1988) underscores that, if inflectional morphology is relevant to syntax, then "the only thing that is pertinent for syntax is that *épouse* ('spouse$_F$'), like *femme* ('woman'), provides its feminine gender which can induce agreement". In that respect, he believes that g-gender alternation of human nouns in French should be analyzed as derivational. An additional argument that he mentions is that g-gender alternation in human nouns is not as productive as the inflectional g-gender alternation of adjectives :

> "En outre on n'expliquerait pas de cette façon pourquoi la formation des noms animés féminins à l'aide de *-e* (et parfois d'autres suffixes) présente des restrictions de productivité que ne connaissent pas les formes féminines des adjectifs. De telles restrictions sont inattendues au niveau de la syntaxe, comme le montre la flexion du pluriel, applicable en principe à chaque nom."[3]

If grammatical gender for French human nouns is to be treated as an inflectional morphological process, then one needs to reconsider not only the role of g-gender as a noun-classifying system and the shape of the French morphological paradigm. One would then have to decide whether g-gender alternating French human nouns formally unvarying, *id est* common gender nouns (eg. *biologiste*), are to be treated in the same manner as g-gender alternating French human nouns formally varying (eg., *vendeuse*, *vendeur*).

Bonami and Boyé (forthcoming) summarize the possible implications thusly: either common gender nouns are derivationally related — in which case their prevalence in the lexicon would imply a highly productive derivational rule converting any human noun of a given grammatical gender to the other, and one would need to explain why it doesn't apply to words like *vendeur* — or they are inflectionally related — in which case they correspond either to different occurrences of a lexeme underspecified for gender or to syncretic forms of different paradigm cells. Likewise, for non–common gender human nouns, two possibilities can be envisioned: they are derivationally related, or they are inflectionally related and correspond to different paradigm cells.

Zwanenburg (1988) on the other hand argues that considering human nouns as inflectionally related would "constitue a complication of the grammar":

> "ll faudrait, en effet, compliquer la syntaxe d'une façon ad hoc en introduisant au niveau de la D-structure le choix du genre pour une classe particulière de noms, à savoir les noms animés, à la façon dont on choisit le nombre pour l'ensemble des noms. Et le lexique n'en serait pas simplifié. D'abord on perdrait la généralisation que les noms français ont un genre lexical, puisqu'un nom comme *époux* serait sans genre, et obtiendrait son genre masculin ou féminin en syntaxe. En outre le lexique devrait tout de même comporter une règle d'affixation dérivationnelle[…]"[4]

Moreover, he argues that the distinction between human nouns is semantic in nature, whereas g-gender alternative forms of adjectives only vary due to syntax. Finally, Zwanenburg believes that if human nouns possess a g-gender alternating paradigm, such as what Bonami and Boyé (forthcoming) advocate, then human nouns would be unspecified for g-gender, which entails that the classification system of French g-gender as a whole would no longer be pertinent. There are however two possible solutions to this issue: either by stating that French possess three genders — and that therefore human nouns hold a specific g-gender, which is neither feminine nor masculine, and which can trigger either masculine or feminine agreement —, or by saying that classification does not ncessary entail partition — *id est* that a noun can possess multiple genders.

Crucially enough, the question of paradigm uniformity is also central to one of the criteria from Stump (1998): that of productivity. If a large-scale case of non-paradigm uniformity is to be proved, *id est*, if the g-gender alternation is an inflectional process, then at least two major questions directly relevant to Stump

[3] "Moreover, one cannot explain in this manner why the formation of feminine animate nouns by means of *-e* (and cometimes other suffixes) presents restrictions in terms of productivity which the feminine forms of the adjectives do not know. Such restrictions are unexpected at the level of syntax, as shows the inflection of plural, which can in principle be applied to any noun."

[4] "Indeed, one would need to complicate the syntax in an *ad hoc* manner by introducing at the level of the D-structure the choice of gender for a particular class of nouns, viz. animate nouns, in the same manner that number is chosen for all nouns. And this wouldn't simplify the lexicon. Firstly, we would lose the generalisation that French nouns have a lexical gender, since a noun like *époux* ('spouse$_M$') would be without gender, and would gain its masculine or feminine gender from syntax. Moreover the lexicon would still need to contain a derivational affixation rule[…]'

(1998) arise: should one consider human nouns to be of the same part of speech than the other common nouns? And if not, should this criterion be reformulated, undermined, or abandonned?

2.3 The theoretical framework of distributional semantics

One of the technical difficulties inherent to this topic of research is to convincingly model processes, and not only word representations, in a fashion removed from any morphological and syntactic theories. This problem is best addressed with models that provide representations of words within a fairly unconstrained framework: while studying the morphological processes themselves, it is best to assume as little as possible about what they consist of. Following this, a very natural choice of modelization is that of distributional semantics models and compositional semantics models which postulate very little about the inherent morphology and syntax of words. A natural extension to such studies that is also possible thanks to the use of distributional semantics is to evaluate the role of g-gender on a scale more global than that of French human nouns. French human nouns, as was already discussed, are situated at the crossroad between many linguistic domains. Some properties of grammatical gender in French might therefore not be best represented by French human nouns: for example one can expect that the influence of social cues might be less relevant in general than it is to human nouns in particular. Distributional semantics models (hereafter DSM), since they provide a representation for every possible word, may prove to be a very handy tool for furthering research on this topic.

2.3.1 Origins of DSM

The present study therefore addresses the subject of g-gender in French from a specific perspective, that of Distributional Semantics. This framework is based on the assumption that the meaning of a word can be derived from the contexts in which it occurs, or in the words of Firth (1957): "You shall know a word by the company it keeps". The idea is that words of similar meanings will occur in similar contexts: for instance, since dogs are creatures that bark, while elephants are not, the co-occurrence of the words *dog* and *bark* is likely to be more frequent than that of *elephant* and *bark*. Inasmuch, the meaning of *bark* will be more closely related to *dog* than to *elephant*.

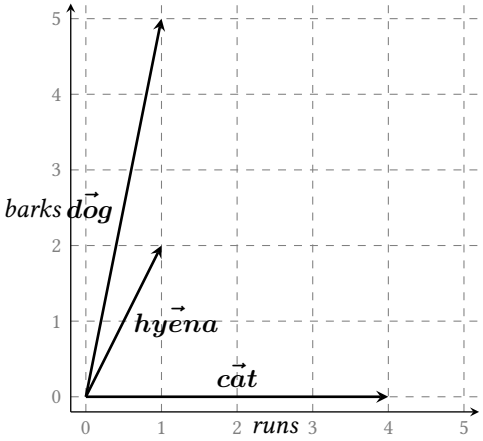
This approach takes root from works ranging from Wittgenstein (1921), who described the meaning of a word as a function of its usage, to Quine (1960) arguing for a holistic account of semantics. These philosophers denounced the concept of a language independent meaning, on the basis of the fact that the possible correct usages of a word are constrained by the existence of other words which may compete for external reference, and that the exact meaning itself can only be understood when one knows the language in its entirety. The linguistic formalization of the approach dates back to the late 1990s — especially of note is Latent Semantics Analysis (Landauer and Dumais, 1997) — building upon earlier works of frequency counting and information retrieval. Varied architectures have been researched in the following decades, including Non-negative Matrix Factorization, Latent Dirichlet Allocation, as well as log bi-linear probabilistic models and word embeddings from neural networks.

A more formal definition of what distributional semantics generally entail is that they associate words from a given language with vectors derived from their usage. Although the definition generally given is more lax, allowing in theory any model of word meaning that builds upon the idea that the meaning of a word can be retrieved from the context in which it occurs, vector spaces are one of the most natural way of satisfying such conditions. The vector associated with a word will encode its contexts of occurrence. One of the most naive ways to construct such vectors is to compute a word co-occurrence matrix, and consider each row as the vector for the corresponding word. However the exact computation of such a space is what distinguishes models from one another; what matters chiefly is the characteristics intrinsic to vector spaces. Vector spaces allow for a wide range of mathematical operations — from angle measurement and normalization to multi-linear functions and tensor products — which can be likened to a wide range of linguistic operations — from relatedness judgment to syntactic composition. In particular, the relatedness of the meanings of two words can be quantified in various ways — most often, it is likened to the distance of two points or the angle of two vectors in the space.

2.3.2 Practical Intuition

Figure 2 shows a toy example of what a DSM can be. In this example taken from Baroni, Bernardi, and Zamparelli (2014), the vectors for the words *dog*, *hyena* and *cat* are computed from two observations of

	runs	*barks*
dog	1	5
hyena	1	2
cat	4	0



from Baroni, Bernardi, and Zamparelli (2014)

Figure 2: Toy DSM example

their contexts: how many times the word cooccurs with *runs*, and how many times it cooccurs with *barks*; the mock observations are reported in the table on the left. These observations can be quantified in a trivial manner, and therefore, vectors associated with each of the three words can be computed and drawn on a plane, as can be seen on the right hand part of the figure. Studying this representation, one can see that the meaning of *hyena* is somewhere between that of *dog* and *cat*.

Another interesting point is that the frequency of the word has an impact on its vector representation — in this toy example, the vector for *dog*, computed from six observations, is longer than the vector for the rarer *hyena*, computed from only three observations. The literature therefore provides different techniques to compare two vectors while taking this difference in length into account. The most natural way would be normalizing vectors — *id est*, rescaling so that they all have the same unitary length —, another is to focus on angle measurements.

It is however crucial to note that this effect on the length of the vector is not the only effect that frequency has — especially since vector spaces are generally computed over a hundred dimensions, and not two like in this toy example, the more frequent the item, the more varied the situations it can be observed in, the fewer null dimensions it has, which mathematically leads a frequent word to produce a vector near from any other vector. See also section 2.4 for a more formal approach of this phenomenon. The key point here is that, when studying DSMs, word frequency needs be controled.

A last point worthy of attention is that the DSM here presented is defined only on two dimensions. As was said before, actual DSMs are often computed from a general cooccurrence matrix, which leads to vector spaces of very high dimensionality populated with "sparse" vectors with null values for most dimensions. As these vector spaces are generally unusable from a computational standpoint, they are often mapped to vector spaces of lower dimensionality using what is called "dimensionality reduction techniques". See also section 2.4 for more details.

2.3.3 Usages of DSM in linguistics

Although Distributional semantics constitute a fairly recent approach in linguistics, DSMs have already been used in a wide array of linguistic studies, ranging from psycholinguistics accounts to sentence-building models.

In psycholinguistic studies, authors generally attempt to liken the semantic relatedness of words to the distance or angle of their associated vectors, for instance through measures of priming effects, as can be seen for instance in Marelli and Baroni (2015). Other uses of DSMs include mental maps (see for instance Louwerse and Zwaan (2009) and Louwerse and Benesh (2012)), as well as the question of "grounding" the word-to-word relation encoded in a DSM as a cognitive perception (*cf.* Lenci (2008), Lazaridou, Bruni, and Baroni (2014) or Kiela and Clark (2015)).

As they are semantic in nature, DSMs are also used in more traditonnal semantic tasks, such as hypernym detection (Roller, Erk, and Boleda (2014), Santus et al. (2014), Santus et al. (2016) for instance); other noteworthy papers concern attempts to conciliate DSMs with more traditional semantic theories (see

Herbelot and Vecchi (2015), Kruszewski, Paperno, and Baroni (2015) or Kruszewski et al. (2016)).

Some sociolinguistic studies also mention, for instance, the importance of *s-gender* bias in word embeddings as encoded in a DSM: one example is Bolukbasi et al. (2016).

An important group of papers address the problem of phrase building, from both a semantic and a syntactic point of view. The general approach is described in Baroni, Bernardi, and Zamparelli (2014), other relevant papers include Mitchell and Lapata (2008), Baroni et al. (2014) and Paperno et al. (2014). They are also widely used as a preprocessing step for other NLP tasks (Mikolov et al., 2017).

More relevant to the present study, some works address morphological issues using DSM. As standard DSM are word-based, linguistic studies that focus on issues below the word level would not intuitively rely on such models, which may explain why morphology-related papers are less common as compared to other fields. However this mismatch is apparent only, and DSM have been used successfully to explicit morphological facts in at least two ways: either by contrasting vector pairs, as can be seen in Varvara (2017), Wauquier (2017) or Bonami and Paperno (forthcoming), or by considering morphological processes as a compositional problem as can be seen in Marelli and Baroni (2015). Interestingly, DSMs addressing either syntactic and morphological composition often employ the same methodological means — namely, they provide a functional representation of dependents.

2.3.4 Applications to morphology

In subsubsection 2.3.2, we introduced the framework of distributional semantics, in which each word is associated with a vector. However, morphological processes can only be grasped through the contrast between two words (or more) : the very fact that words like *vendeur* contrast with word likes *vendeuse* defines the morphological process of g-gender alternation of human nouns. In that respect, there is a need to derive representations for processes from the word-level representations of DSMs. One potential solution is to look at vector offsets: as processes involve paired words, the offset, or the shift from the vector representation of one to that of the other should represent the effects of the process itself.

As DSMs represent words as vectors, the shift can be computed as a difference between vectors. The visual representation shown in Figure 3 might clarify this. The vector $\vec{s}$ corresponding to the shift from

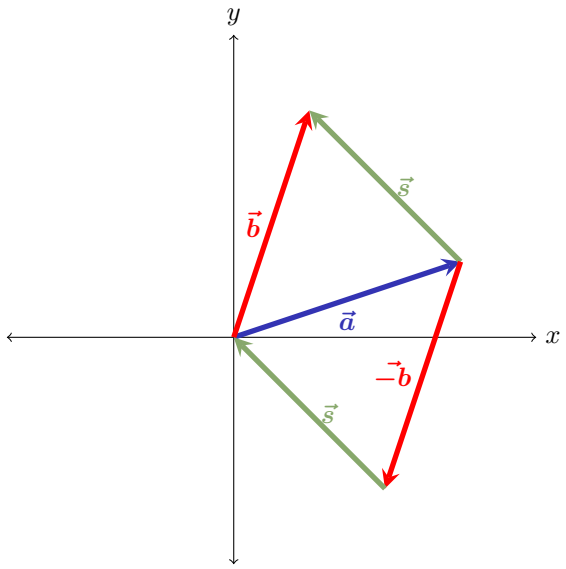


Figure 3: Shift seen as vector difference

the vector $\vec{a}$ to the vector $\vec{b}$ can be computed by subtracting $\vec{b}$ from $\vec{a}$, which is equivalent to adding its inverse $-\vec{b}$.

This property has been put to use to express linguistic regularities. Especially relevant to this study is the paper of Mikolov, Yih, and Zweig (2013), the main intuitions of which are depicted in Figure 4. In their article, Mikolov and colleagues discussed the fact that constant shifts of meaning could be approximated by a vector shift. For instance, as seen on the leftmost part of the figure, the same shift vector can be computed from *wo$\vec{m}$an* to *m$\vec{a}$n*, from *qu$\vec{e}$en* to *k$\vec{i}$ng*, or from *a$\vec{u}$nt* to *$\vec{u}$ncle*. As the shift in the meanings of

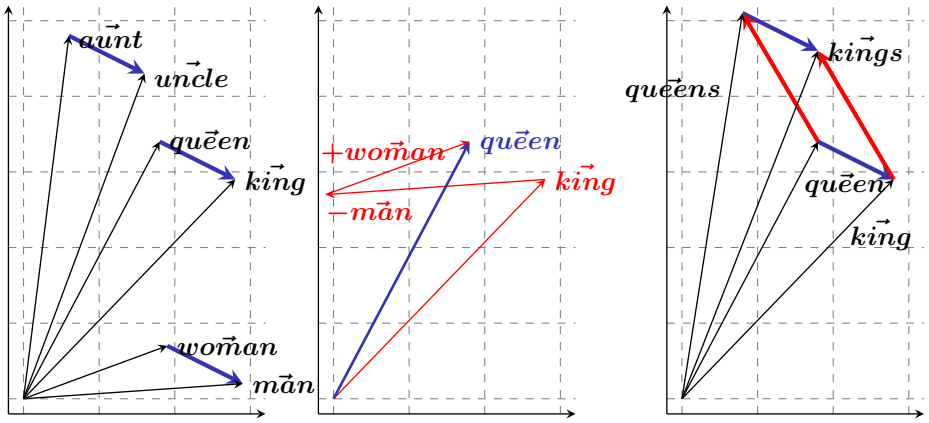


Figure 4: Linguistic regularities and DSM

each of these pairs of words is constant as well — one referring to female and the other to male —, this suggest that formal analogy can be operationalized using vectorial algebra: as can be seen in the middle of Figure 4, that *queen* is to king what *woman* is to *man* can be written as:

$$\vec{queen} \approx \vec{king} - \vec{man} + \vec{woman} \tag{1}$$

As can be seen, the blue vector for *queen* is equal to the sum of the other vectors in red. Crucially, as is illustrated in the rightmost part of the figure, this operationalization of formal analogy also holds for morphological processes, such as the singular-plural alternation. This very characteristic of DSMs is crucial to this study, as it provides a simple, straightforward manner of representing a given morphological process — as the mean shift of all paired vectors that embody it — as well as obvious criteria to distinguish more regular from less regular processes — by studying how adequately this mean shift vector represents each and every shift.

2.4 Mathematical definition of vector spaces and DSM

2.4.1 Mathematical Definition

From a mathematical point of view, a vector space V over a field F is a set possessing both an addition $V \times V \rightarrow V$ (two vectors added give a vector) and a scalar multiplication $F \times V \rightarrow V$ (the multiplication of a vector by a scalar is a vector). The addition possesses an identity element: $\vec{0}$ defined the property that the addition of any vector with the identity element equals that vector itself:

$$\forall \vec{V}\ \vec{V} + \vec{0} = \vec{V} \tag{2}$$

Inverse elements are defined for the addition:

$$\forall \vec{V}\ \exists \vec{V^{-1}} \text{ such as } \vec{V} + \vec{V^{-1}} = \vec{0} \tag{3}$$

That is to say the addition of a vector and its inverse equals the identity element. It is also associative

$$\vec{U} + (\vec{V} + \vec{W}) = (\vec{U} + \vec{V}) + \vec{W} \tag{4}$$

and commutative

$$\vec{V} + \vec{U} = \vec{U} + \vec{V} \tag{5}$$

which implies that the order in which vectors are added has no impact on the result. The vector space also verifies the following axioms:

$$(nm)\vec{V} = n(m\vec{V}) \tag{6}$$

id est, the scalar multiplication is compatible with the field multiplication, and

$$1\vec{V} = \vec{V} \tag{7}$$

which means the identity element of the field multiplication is the identity element for the scalar multiplication. Lastly, the scalar multiplication must be distributive over both field and vector additions

$$n(\vec{V} + \vec{U}) = n\vec{V} + n\vec{U} \tag{8}$$

and

$$(n + m)\vec{V} = n\vec{V} + m\vec{V} \tag{9}$$

In the case of DSMs, authors often insist that the vectors must be associated with specific words or terms, highlighting that two mathematically equivalent vector spaces where vectors are associated to different words should not receive the same interpretation, since measures regarding a *word* will differ in the two spaces.

In a more practical, tangible way, vectors in vector spaces are generally represented as a a ordered list of real numbers of a given length. The members of the list are referred to as components, and their ordinal as dimensions. The length of the list itself defines the dimensionality of the vector. All vectors in a given vector space must have the same length, that is to say dimensionality is characteristic to the vector space. Vector addition is furthermore computed as component-wise addition, and scalar multiplication as the multiplication of each component with the scalar. The identity element $\vec{0}$, often referred to as the origin of the vector space, is therefore implicitly given as the vectors for which all components equal zero, and the field on which the vector is based is that of real numbers algebra. Such a configuration give rise to a few specific characteristics which are not intrinsic to all vector spaces: for instance, scalar multiplication of any vector with the absorbing element of the field multiplication gives the addition identity element: $\forall \vec{V}\ 0\vec{V} = \vec{0}$.

2.4.2 Commonplace Vector Spaces Functions

Most models that are based on vector spaces use cosine as a measure of similarity of direction and define a distance function, generally the classical Euclidean distance, in order to analyze the relative disposition of vectors. In the case of vector spaces, the cosine of the angle θ given by two vectors $\vec{v}$ and $\vec{u}$ can be computed as:

$$cos(\theta) = cos(\vec{v}, \vec{u}) = \frac{\vec{v} \cdot \vec{u}}{|\vec{v}||\vec{u}|} = \frac{\sum_{i=1}^{d} v_i \times u_i}{\sqrt{\sum_{i=1}^{d} v_i^2} \times \sqrt{\sum_{i=1}^{d} u_i^2}} \tag{10}$$

In particular, should two vectors be equal, their cosine would be equal to 1:

$$cos(\vec{v}, \vec{v}) = \frac{\vec{v} \cdot \vec{v}}{|\vec{v}||\vec{v}|} = \frac{\sum_{i=1}^{d} v_i^2}{\sqrt{\sum_{i=1}^{d} v_i^2} \times \sqrt{\sum_{i=1}^{d} v_i^2}} = \frac{\sum_{i=1}^{d} v_i^2}{\sum_{i=1}^{d} v_i^2} = 1 \tag{11}$$

Hence, the cosine can be and often is used as a similarity measure. However a caveat needs to be made here, as this is in fact a specific case of a more general implication: if a vector can be expressed as the scalar product of another $\vec{u} = k\vec{v}$, then their cosine will be equal to 1.

$$\begin{aligned} cos(\vec{v}, k\vec{v}) &= \frac{\vec{v} \cdot k\vec{v}}{|\vec{v}||k\vec{v}|} = \frac{\sum_{i=1}^{d} v_i^2 \times k}{\sqrt{\sum_{i=1}^{d} v_i^2} \times \sqrt{\sum_{i=1}^{d} (v_i \times k)^2}} \\ &= \frac{k \times \sum_{i=1}^{d} v_i^2}{\sqrt{\sum_{i=1}^{d} v_i^2} \times \sqrt{k^2 \sum_{i=1}^{d} v_i^2}} = \frac{k \times \sum_{i=1}^{d} v_i^2}{k \times \sum_{i=1}^{d} v_i^2} = 1 \end{aligned} \tag{12}$$

This is why the cosine is a mathematical measure of the angle between the vectors, since two vectors equal modulo a scalar are codirectional. In the case of DSM, as they are in general normalized — *id est*, all of their vectors being of the same length —, in most practical cases a cosine equal to 1 does imply an equality of the two vectors.

This definition of the cosine makes use of the norm of a vector $|\vec{v}|$, here defined as a Euclidean norm by the function:

$$|\vec{v}| = \sqrt{\sum_{i=1}^{d} |v_i|^2} \tag{13}$$

A norm can be understood as the distance between the origin of the vector space and the point described by the coordinates of the vectors, *id est*, as the length of the vector.

The Euclidean norm is in particular related to the Euclidean distance, which can be conceived as the Euclidean norm of the shift from one vector to the other:

$$dist(\vec{v}, \vec{u}) = \sqrt{\sum_{i=1}^{d} |v_i - u_i|^2} \tag{14}$$

Should $\vec{u}$ be the origin, then the distance between $\vec{v}$ and $\vec{u}$ the origin would correspond to the norm of $\vec{v}$.[5]

When it comes to DSMs, the distance (equation 14) and cosine (equation 10) are specifically impacted by the frequency of the words. In the most naive models, such as what was presented in section 2.3.2, the frequency has a direct effect on the length of the vector, and therefore the distance to the vector of a frequent word is likely to be greater than that to any other. A second effect of frequency — which can also be attested in more elaborate DSMs — is that more frequent items tend to cluster together. Consider $\vec{1}$, the vector whose dimensions are all equal to 1; it is obvious that the more non-zero dimension a given vector $\vec{v}$ has, the higher the cosine of $\vec{1}$ and $\vec{v}$ will be — and before any dimensionality reduction technique is applied, DSMs tend to be populated with very sparse vectors, *id est*, vectors for which almost all dimensions are zero. This also holds true for vector spaces with normalized lengths; authors, for instance Faruqui et al. (2016) or Schnabel et al. (2015), have also shown that the effects of frequency persist with dimensionality reduction techniques and, moreover, that it can also be attested in models not directly based on frequency counts. This implies that word frequency plays an important role and that, as was indicated in a previous section, any experiment involving a vector space should take into account the effects that frequency can have.

2.4.3 Modelizations

The computation methods employed by researchers, as said previously, vary greatly. A brief overview show that the vectors can be obtained from at least three types of structures: term-to-document matrices

[5]The Euclidean distance is only one of the possible distances that can be defined for a vector space. The Manhattan distance is another way of estimating differences between vectors:

$$dist(\vec{v}, \vec{u}) = \sum_{i=1}^{d} |v_i - u_i| \tag{15}$$

The Manhattan and Euclidean distances are specific cases of p-distances, where p equals respectively 1 and 2:

$$dist(\vec{v}, \vec{u}) = \sqrt[p]{\sum_{i=1}^{d} |v_i - u_i|^p} \tag{16}$$

The larger the select value for p, the more emphasis there will be on greater components. In fact, the function tends toward selecting the maximum component of the vector for greater values of p, thus providing a definition for the infinity p-distance:

$$\lim_{p \to \infty} |\vec{v}|_p = |\vec{v}|_\infty = \max(|v_1|, \ldots, |v_d|) \tag{17}$$

Norms derived from p distances are noted as l^p norms:

$$|\vec{v}|_p = \sqrt[p]{\sum_{i=1}^{d} |v_i|^p} \tag{18}$$

Hence the Euclidean norm is also referred to as the l^2 norm, and can also be noted as $|\vec{v}|_2$. In the particular case of this study, only the Euclidean distance and the related Euclidean norm from it will be used.

(for instance LSA: Landauer and Dumais (1997)), term-to-term matrices (GloVe: Pennington, Socher, and Manning (2014)), and neural embeddings (word2vec: Mikolov et al. (2013)). So as to avoid sparse representations of words (which aren't tractable), weighting of components (through tf-idf, point-wise mutual information or normalization) and dimensionality reduction techniques (singular value decomposition, primary component analysis, latent Dirichlet allocation, non-negative matrix factorization…) are usually applied to these matrices. In the case of neural embeddings, the weighting and dimensionality reduction are done implicitly, in the sense that the model learns to apply these operations although they are not necessarily predefinied by its architecture. The following is a brief review of the various models broached in the relevant literature.

Latent Semantics Analysis Better known under the acronym LSA, Latent Semantics Analysis, described in Landauer and Dumais (1997), is one of the earliest DSMs described in the literature. It is based on term-to-document matrix. Each word corresponds to a row of the matrix, and each document to a column. The values for each cell correspond to the point-wise mutual information of the word and the document. It builds upon an earlier similar architecture, on which Landauer had previously worked, that was specifically conceived for Information Retrieval tasks: Latent Semantics Indexing. The conceptualization of the context of a word as the documents in which it appears can be therefore explained when comparing the aim of this previous architecture: Information Retrieval aims at retrieving a specific resource using the given informational cues. Landauer & Dumais applied an algorithm for dimensionality reduction: Singular Value Decomposition (SVD), in which the original matrix of shape $W \times D$ is decomposed as three matrices of shape $W \times W$, $W \times D$ and $D \times D$. The second matrix of shape $W \times D$ is afterwards truncated to a given number K of columns, thus giving an approximated version of the distribution of element from the original vector space of dimensionality D in a space of dimensionality K. The authors conceive the dimensionality reduction applied as an important theoretical step, corresponding to abstracting information on correlations from the original data. The authors have also shown that applying no dimensionality reduction techniques leads in many cases to poorer results.

NMF Non-negative Matrix Factorization (NMF), also known as Topic models (TM), introduced by Griffiths, Steyvers, and Tenenbaum (2007), differs mainly from LSA by giving the matrix decomposition a probabilistic interpretation. Like LSA, NMF starts by computing a term-to-document cooccurrence matrix. However, the decomposition is based on a latent Dirichlet allocation (LDA), which allows to infer from the documents a probability distribution over their "topics". Any word can afterwards be associated to its probability distribution over all possible topics. Whereas dimensionaility reduction in LSA was conducted using a purely mathematical algorithm, SVD, the LDA process of NMF can be more easily understood as a linguistic computation.

GloVe GloVe, which stands for "Global Vectors", corresponds to a DSM architecture that was presented by Pennington, Socher, and Manning (2014). Unlike TM and LSA, GloVe is a neural based DSM trained directly to decompose the word-to-word cooccurrence matrix. More formally, it is a log bi-linear model with a weighted least-squares objective. It is important to note that GloVe combines a "global matrix" approach (as is the case in LSA and TM which are based on global term-to-document matrices) with "local context window" methods: using a word-to-word cooccurrence matrix allows for matrix manipulation algorithms to be applied, and populating the global matrix with local cooccurrence has been shown to improve performances on word-level analogy tasks, such as what was presented in Figure 4.

word2vec word2vec is a word embeddings model — or more accurately, a group of related word embeddings models — presented in Mikolov et al. (2013). Like GloVe, the word vectors are obtained using a neural network: the training consists in matching a word and its context, using either or both of two training algorithms called "hierarchical softmax" and "negative sampling". The former maximizes an approximation of the conditional log-likelihood, while the latter aims at minimizing the log-likelihood of negative instances that were sampled. Two model architectures are also proposed, called "CBOW" and "skip-gram". The acronym CBOW stands for Continuous Bag Of Words. This architecure corresponds to predicting a given word from its context. It does not take into account the order in which the context words appear in. The "skip-gram" architecture, on the other hand, aims at predicting a context from a

given word. In this architecture, the relative order of contextual words therefore matters. In both cases, the vectors retrieved are computed as the weight matrix of the hidden layer.

The selected model for this study was a word2vec model, as it is one of the easily obtainable models often cited. One specificity of word2vec is that the associated vectors are the word embeddings learned from a neural network performing either of two tasks: learning to predict a word from the context in which it occurs, and learning to predict the context from a given word. In both cases, the vectors are retrieved from an inner layer that was trained for a functional mapping of a word to its contexts of occurrences thus capturing the abstract regularities of such a mapping.

3 Building Resources

Two resources were required to study the distribution of g-gender in human nouns: a corpus from which vectors would be derived, and an annotated lexicon to select suitable words for the experiments. The resources here described will be used in all the following experiments.

3.1 Corpus

Firstly as was stated before, DSMs are generally conceived as a mapping of words to suitable vector representations of their contexts. Using more frequently used terms of machine learning, the model is trained on the contexts to produce the vector space. A large corpus is generally needed to compute a DSM, since they are supposed to assign a vector to any word no matter how rare it is. This often leads to a problem of data sparsity: to calculate a vector, its context of occurrence needs to be known, and the more numerous its occurrences are, the more certain its context. Conversely, rare words tend to have poor vector representations. The obvious and naive way out is to simply provide more data, *ide est* a larger corpus on which to train the DSM.

This specific study used a composite corpus, that is to say, the concatenation of different corpora: a French wikipedia dump that was fed to the parser described in Coavoux (2017) (which shall be referred to as FRWIKI), as well as the FRWAC (introduced in Baroni et al. (2009)) and the FRCOW (described in Schäfer (2015)) corpora. The FRCOW and the FRWAC corpora are two comparable web corpora; their main difference is that the FRCOW corpus is more voluminous and recent than the FRWAC corpus.

3.2 Corpus manipulation

Although all these corpora provided g-gender annotation for common nouns, the FRWAC corpus exhibited some obvious errors when it came to g-gender annotations and PoS-tagging. It was therefore deemed necessary, in the context of this particular study, to re-tag the corpora using the parser from Coavoux (2017) in the interest of tagging accuracy (Coavoux (2017) reports 97 % correct g-gender tagging) and uniformity of annotations. The FRWIKI corpus had already been parsed with this algorithm, which meant that only the two other corpora, FRWAC and FRCOW, were re-parsed. The FRWAC corpus also required some cleaning to remove sentences containing non standard French characters and duplicate sentences, much in the spirit of what Faaß and Eckart (2013) did when creating the SDEWAC out of the DEWAC corpus.

A manual validation of a small sub-sample of a hundred words from each corpora lead to similar figures for g-gender annotation as what was reported in Coavoux (2017) for the FrenchTreeBank corpus. The composite corpus totals roughly 14 billion tokens (12 billions from FRCOW, 1.4 from the cleaned FRWAC, 0.8 from the FRWIKI corpus). Various sets of statistics were computed on this composite corpus, which were used throughout the experiments, in particular as cues for establishing pairs of comparable words. These sets include a count of occurrences per types, as well as a count of types per PoS.

3.3 Lexicon

The second resource that was required for this study was a lexicon of test words — specifically French human nouns as well as morphologically related words. Multiple resources were combined for that purpose.

GLAWI & GLÀFF The GLAWI lexicon, described in Hathout, Sajous, and Calderone (2014a), Sajous and Hathout (2015) and Hathout and Sajous (2016), was built using the GLÀFF (see Hathout, Sajous, and Calderone (2014b), Sajous, Hathout, and Calderone (2014) and Sajous, Hathout, and Calderone (2013)) and the Wiktionary as a XML-structured document describing for each token its possible parts of speech, and for each of these parts of speech, its paradigm according to GLÀFF and its wiktionary definitions. At times, the experiments described in this paper required information more easily accessible through the GLÀFF lexicon than through the GLAWI, especially when it came to paradigm structures.

Lexeur The Lexeur lexicon lists deverbal nouns suffixed with *-eur*, their feminine counterpart (either suffixed with *-rice* or *-euse*), their base verb as well related action nouns. (Fabre, Floricic, and Hathout, 2004)

3.4 Extracting human nouns from the lexica

These various lexica as they are do not directly indicate whether the nouns they list refer to human or not. Therefore these lexica were merged, and then the data they contained was subjected to some extrapolations in order to extract nouns referring to humans only, which were needed to conduct our research. In particular, although derivational relations holding between verbs and nouns were available through the use of the Lexeur lexicon, as deverbal nouns ending with the *-eur* suffix and its feminine counterparts *-euse* or *-rice* exhibit a variety of uses that can be for the most part subsumed in two categories: agentive meanings (for instance *voleur*, 'thief', from *voler* 'to steal') and instrumental meanings (*agrafeuse*, 'stapler', from *agrafer*, 'to staple, to clasp'). In other words, the Lexeur lexicon contains morphologically homogenous words, but not semantically homogeneous words, and listed items may or may not be relevant to the study of gender in French human nouns.

abandonnataire	1. Personne qui reçoit des biens abandonnés.
ouvrier	1. Personne qui, moyennant salaire, effectue un travail généralement manuel [...]. 2. Celui qui œuvre à une chose ; auteur ; artisan.
allumeur	1. Celui qui était (autrefois) chargé d'allumer quelque chose. 2. Dispositif du système d'allumage d'un moteur à explosion [...]. 3. Homme qui cherche à exciter le désir.
agrafeuse	1. Outil permettant de fixer ensemble, à l'aide d'une agrafe, des objets de faible épaisseur.

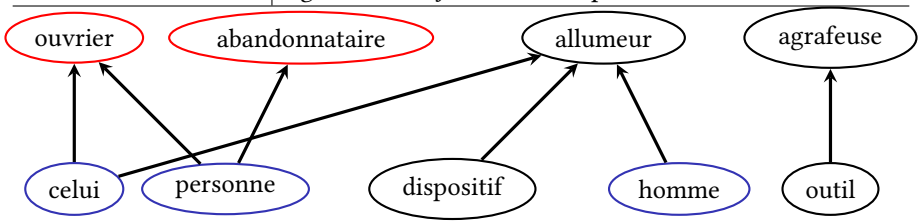


Figure 5: Dynamic human nouns extraction

To circumvent such limitations, one can infer whether a noun refers to a human or not not using the Wiktionary entries as listed in the GLAWI lexicon. Specifically, as is noted in Muller, Hathout, and Gaume (2006), the first content word (*defininiens*) of a definition contributes generally more to the meaning of the defined word (*definiendum*) than the second, and the second more than the third, and so on. From this, it follows that if the first *definiens* for a given *defininendum* is a human noun, then it is more probable that said *definiendum* is a human noun as well. Therefore, if all the definitions of a noun start with a human noun it is highly probable that this word is a human noun. Such a relation is also transitive in a mathematical sense: if all definitions of a *definiendum* start with *definientia* who in turn all have their definitions starting with human nouns, then it follows that this *definiendum* has in all probability human nouns for first *definientia* and that therefore he is an human noun as well.

The GLAWI lexicon was therefore enriched with such information derived from computing the closure for a given small set of seed words (*homme* ('man'), *femme* ('woman'), *celui* ('the one who', masc.), *celle* ('the one who', fem.), *qui* ('who'), *personne* ('person'), *humain* ('human'), and their plurals). The seeds contained very general human nouns as well as pronouns that strictly refer to humans in the context of dictionary definitions. This process is exemplified in Figure 5, where the fact that the words circled in red refer only to humans can be infered from the fact that the words circled in blue should only refer to humans in the context of the first word of a dictionary definition, effectively selecting human words like *ouvrier* ('worker') or *abandonnataire* ('abandonee') and leaving out words as *allumeur* ('person in charge of lighting up', 'ignition system', 'male flirt'), which may or may not refer to humans, or *agrafeuse* ('stapler'), which never refer to humans.

The closure of this transitive relation formed a set of 15222 lexemes, all of which have high probability of referring exclusively to humans. A informal manual evaluation over a hundred randomly sampled items only found one obvious non-human in the sample. The set can be dramatically enlarged, to a total

of 95890 nouns, if the seeds contained as well *habitant* ('inhabitant', masc), *habitante* ('inhabitant', fem), and their plurals, words that add demonyms to the closure.

Human reference can thus be inferred from the Glawi lexicon, which in turn can help classify the deverbal nouns listed in the Lexeur lexicon. When an *-eur* suffixed word from lexeur could be found in the list of nouns from GLAWI, all information from Lexeur was added. In experiments, feminine counterparts were selected if and only if the masculine counterpart was in the computed closure (so that a pair could be constructed) and they either were listed in GLAWI with no definition starting with something external to the computed closure (in which case they were most certainly human nouns), or absent from GLAWI (under the assumption that if they aren't listed in such a wide coverage dictionary, then they must be either very rare or strict feminine counterparts to human nouns). Finally, so as to prove that the previously computed closure can help to disambiguate the instrumental meanings from the agentive meanings, a manual evaluation showed that only one of a hundred items exhibited an instrumental meaning in its usage along with its socially prescribed agentive meaning.

4 Comparing human nouns g-gender alternation to deverbal agent noun derivation

The usual position concerning g-gender variation of French human nouns is that this morphological process is derivational. Since this study aims at providing objective factual observations, comparing human noun g-gender variation to a canonical derivational process is an obvious first step. If no difference in terms of semantic regularity can be found between the two processes, then one would be inclined to discard the inflectional analysis (*cf.* subsection 2.2) expressed in Bonami and Boyé (forthcoming).

As for which derivational process should be compared to the human noun g-gender alternation, one natural canditate to select is that of deverbal agent noun formation as it necessarily involves a human noun — the deverbal agent noun — which is therefore subject to g-gender alternation in French. The subsequent experiments will therefore focus on comparing human nouns g-gender alternation to deverbal agent noun formation, from different perspectives. Due to limitations of available data, the focus will be restricted to those corresponding to *-eur* masculine agent nouns.

4.1 First experiment : measuring the variation

The first experiment aimed at computing, using a DSM, if the variation between two g-gender alternating deverbal agent nouns was less than the variation between a deverbal agent noun and its base verb. The expected result would be that, seeing their morphological similarity to adjective paradigms, the vectors for paired agent nouns vary less than vectors for paired deverbal nouns and their base verb. From a methodological point of view, this first experiment reproduced the measurements described in Bonami and Paperno (forthcoming): since the theoretical framework of this present study is inspired by these authors', partially reproducing their experiments provides a natural starting point.

4.1.1 Calculating the Variation

The first experiment followed the protocol described by Bonami and Paperno (forthcoming). In this article, the athors compare the regularity of two morphological processes by constituting ordered triples of the shape $\langle pivot, comp._1, comp._2 \rangle$, where *pivot* and $comp._1$ instantiate the first process, and *pivot* and $comp._2$ instantiate the second process. The general assumption is that, if the semantic regularity criterion from Stump (1998) is valid (*cf.* subsubsection 2.2.1), then the shifts of paired vectors instantiating an inflectional process should be more regular — *id est* stray less off the mean shift for this process — than the shifts of paired vectors instantiating a derivational process.

In the case of this experiment, the items were therefore of the shape of ordered triples: ⟨*feminine agent noun*, *masculine agent noun*, *base verb*⟩. As shown in Figure 6, the selected triples can also be seen as

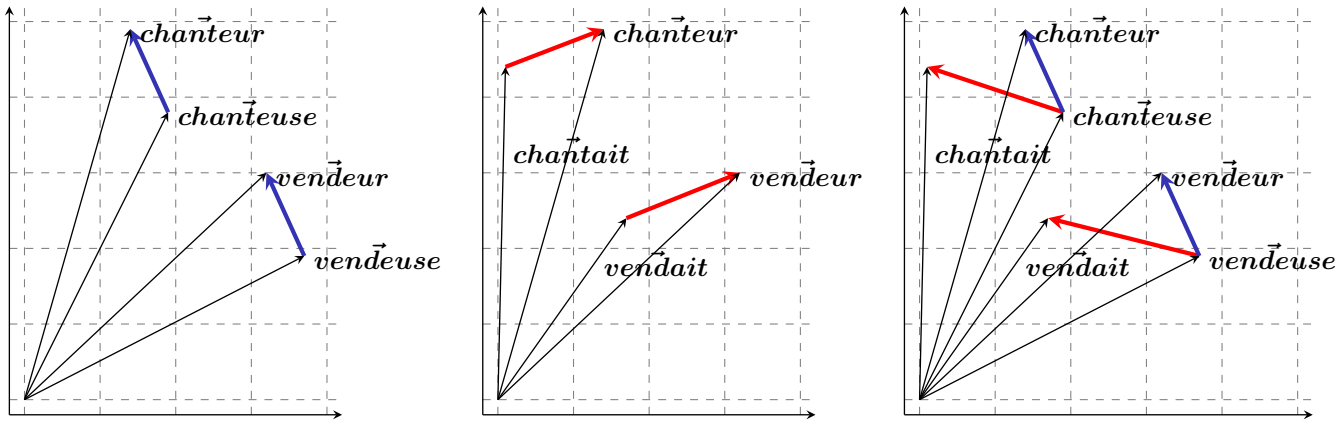


Figure 6: Paired vectors per process and contrastive pairs

contrastive pairs, so as to make sure to negate all possible effects of the stem itself: the leftmost and center figures show the distinct processes studied, embodied by some selected relations, whereas the rightmost figure shows that, when merged, the two sets form contrastive pairs.

The general idea of the experiment is to compute various tests for each pair that will provide each pair with a certain numeric value; therefore each of the two morphological processes, by proxy of the set of paired vectors that embody it, are provided with a numeric distribution that can be compared to the other morphological process. Thus each triple should rather be understood as two pairs: masculine paired with feminine agent nouns, and a feminine deverbal agent noun paired with its verb base. The actual comparisons between processes will therefore involve two sets of measures of variation at each time: one relative to masculine-feminine pairs, and one relative to deverbal-verb pairs. The methodology adopted is therefore to first assess the variation amongst a given process and then to compare to one another the assessed variations of the different process.

As feminine agent nouns appear in both the deverbal-verb pairs and the masculine-feminine human noun pairs, they will be referred to as pivot word. Masculine human nouns and base verbs, on the other hand, will be referred to as compared items or compared words. The pairs associated with a given morphological process — in the case of this experiment, either the deverbal agent noun formation or the human noun g-gender alternation — will be said to embody said process. A morphological process can therefore be represented by the set of paired words that embody it.

The term "variation" requires a more formal mathematical definition. There are two different ways of assessing variation amongst a process. The first is to compare the shift of a pair of vectors to the mean shift vector. The second is to compare directly how systematic the contrast between the two members of each pair is.

As for the first way, the shift vector can be defined as the vector equivalent to the translation needed to move from one point to another, *id est*, it is the actual semantic effect for that pair of the process that the pair embody. If the starting point is the one corresponding to the pivot word $\vec{p}$, and the ending point is the one corresponding to the vector for the compared word $\vec{c}$, then the shift vector $\vec{s}$ can be computed as $\vec{s} = \vec{p} - \vec{c}$. This correctly defines the translation from $\vec{p}$ to $\vec{c}$, as was explained previously by Figure 3.

The mean shift vector from masculine to feminine, therefore, is the vector that averages the shift for all pairs of feminine and masculine agent nouns in the selected triples:

$$\vec{m_1} = \frac{\sum\limits_{\textit{triples}} \vec{p} - \vec{c_1}}{\#\textit{triples}} \tag{19}$$

where *triples* is the set of all selected triples (*cf.* supra), $\vec{p}$ represents the vector for a pivot word (in this case, a feminine agent noun) and $\vec{c_1}$ the first item to compare — viz. the counterpart masculine agent noun. Likewise, the mean shift vectors from feminine to verbs is computed as

$$\vec{m_2} = \frac{\sum\limits_{\textit{triples}} \vec{p} - \vec{c_2}}{\#\textit{triples}} \tag{20}$$

where $\vec{c_2}$ is the vector associated to the second item to compare — here, a given verb form. The mean shift vector $\vec{m_\mu}$ for the set of paired vectors embodying a given process P_μ can also be seen as the average semantic effect of process P_μ. Studying how different from the average shift vector $\vec{m_\mu}$ the actual shift vectors $\vec{p} - \vec{c_\mu}$ for the pairs embodying the process P_μ are therefore gives an assessment of the regularity of its semantic effects : the less variation there is, the more semantically regular the process is.

Once these mean shift vectors are computed, the variation compared to this mean can be studied, using either an angle measure or a distance measure.

Shift vectors of each pair can be mapped to their distance to the mean shift vector for the process. which can be written more formally:

$$C_{1_D} = \{\textit{distance}(\vec{p} - \vec{c_1}, \vec{m_1}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{21}$$

for the masculine nouns, and likewise for the verbs:

$$C_{2_D} = \{\textit{distance}(\vec{p} - \vec{c_2}, \vec{m_2}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{22}$$

where the shift is computed as the subtraction of one compared vector ($\vec{c_1}$ or $\vec{c_2}$ being respectively masculine or base verb) from the pivot vector ($\vec{p}$), and the distance as an Euclidean distance — which, as was noted previously (*cf.* equation 13), is equivalent to the l^2 norm of the subtraction of one vector from the other. These sets can be seen as containing information relevant to the variation of shifts in terms of

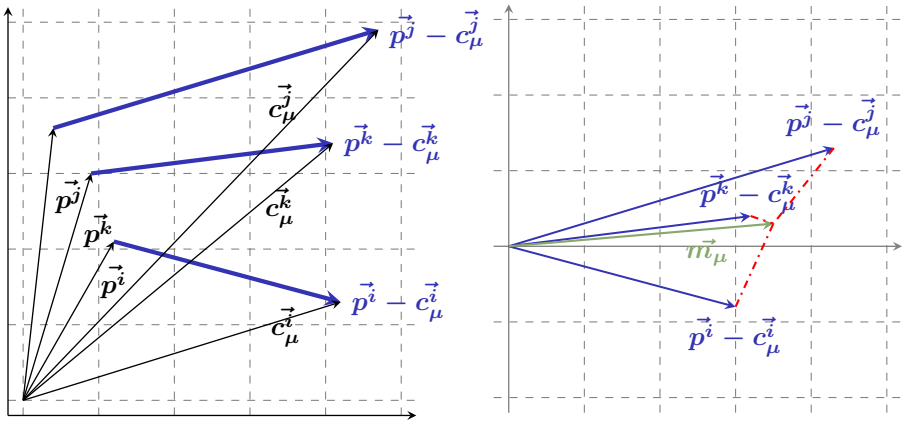


Figure 7: Pair shifts distances to the mean shift

distance from the mean shift, or, in an equivalent manner, in terms of length of the difference between the mean shift and each pair shift.

This measure is described in Figure 7: on the left, a shift $\vec{p} - \vec{c_\mu}$ (in blue) is computed for each pair $\vec{p}, \vec{c_\mu}$ of vectors (in black); on the right, the distance (viz., the length of the dash-dotted line in red) to the average shift vector $\vec{m_\mu}$ (in green) is computed for each shift vector (in blue). Some shifts are more similar to the mean shift, for instance in the figure the $\vec{p^k} - \vec{c^k_\mu}$ shift is closer to the $\vec{m_\mu}$ mean shift than the two others, which leads to lower distance measurements for those specific shifts. Therefore, the more similar the semantic effects of a pair to the average semantic effects of the process it embodies is, the lower the distance of the paired words shift to the mean shift will be. It therefore implies that the more regular a process is, the lower the distance values throughout the set of associated distance measurements will be.

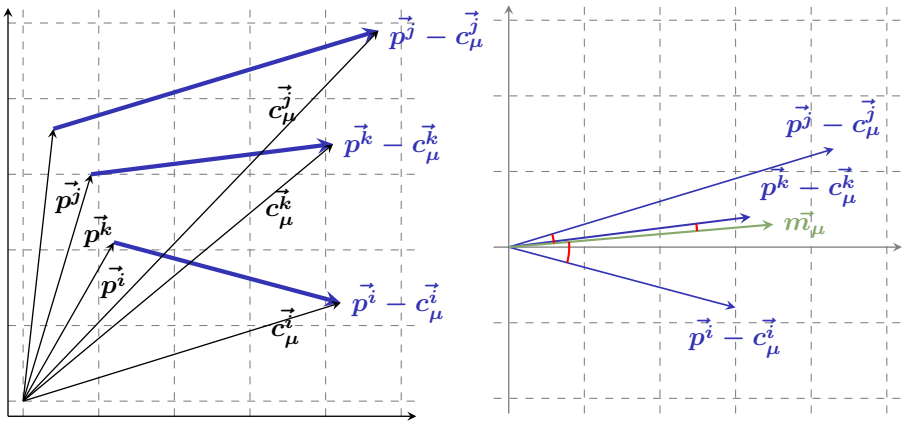


Figure 8: Pair shifts angles to the mean shift

Another quantitative measure of variation is the variation of angle to the mean shift vector. This variation of angle is captured through a measure of the cosine, hence the sets

$$C_{1_{MC}} = \{cos(\vec{p} - \vec{c_1}, \vec{m_1}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{23}$$

defining the variation of angle relatively to the mean shift for the masculine class, and likewise for the verbs

$$C_{2_{MC}} = \{cos(\vec{p} - \vec{c_2}, \vec{m_2}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{24}$$

Figure 8 provides an illustration of this second measure. As in the previous figure, the left part indicates how shift vectors in blue are computed from each pair of vectors in black. The right subfigure shows the same mean shift in green; the significant difference is that the measure this time is the cosine, which corresponds to the angle in red that a given shift in blue forms with the mean shift in green. Once again,

some shifts — viz., $\vec{p^k} - \vec{c_\mu^k}$ — will be closer to the mean shift than others, which will lead to smaller angles relative to the mean shift for these shifts and therefore to higher cosine values. Therefore a process can be considered all the more regular that the values throughout the set of associated shift cosine measurements are high.

These two measures of distance and angle compared to the mean shift should therefore provide the same sort of information, as they describe the same variation of shifts from the complementary points of view of distance and angle measurements: if the set of distances to the mean shift contains only high values, then the set of shift cosine measurements should only contain low values, and vice-versa. Comparing the degree of regularity of the two processes can therefore be done by studying the distribution of their associated measurements: the process that will produce lower distances measurements and higher shift cosine measurements can be thought of as more regular.

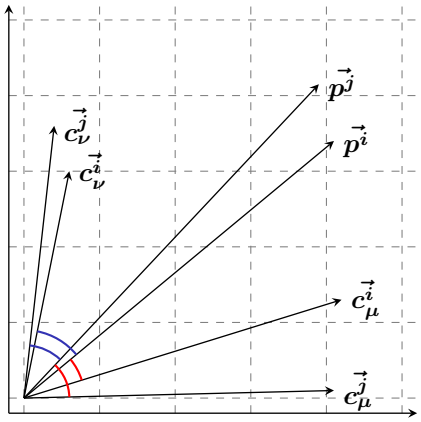


Figure 9: Angles of paired vectors

Another way to assess the variation amongst the set of paired vectors embodying a given process is to directly compare the distribution of angles for each pair of vectors — which gives for feminine nouns the set

$$C_{1_{PC}} = \{cos(\vec{c_1}, \vec{p}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{25}$$

and likewise for verbs :

$$C_{2_{PC}} = \{cos(\vec{c_2}, \vec{p}) | \langle \vec{p}, \vec{c_1}, \vec{c_2} \rangle \in \textit{triples}\} \tag{26}$$

The cosine is computed as previously described in equation 10.

A more visual representation of this measure can be found in Figure 9: for each triple, the angle that each vector $\vec{c_\mu}$ corresponding to a specific morphological process P_μ forms with its associated pivot $\vec{p}$ provides a set of measures (in red) for process P_μ, whereas the angle that each vector $\vec{c_\nu}$ correspond to the other morphological process P_ν forms with its associated pivot $\vec{p}$ provides a second set of measures (in blue) for process P_ν. In this figure, the angles in red, corresponding to process P_μ, vary more than those in blue, associated with P_ν — which implies that the process P_μ is somewhat less regular than the process P_ν.

It is important to note that the difference between the variation of angle (or distance) compared to the mean and the variation of angle compared to the paired feminine human noun vector capture two different things: the first reflects how varied in angle the vectors of the set are, whereas the second reflects how divergent from its paired feminine noun vector a given word vector is or so to say the absolute divergence. As such, this third measure is not as closely related to the other two as these two are with one another. This can also be seen in the formulae through the fact that this last measurement does not involve the shift $\vec{p} - \vec{c_\mu}$ nor the mean $\vec{m_\mu}$ but rather the elements of the triple $\vec{p}$ and $\vec{c_\mu}$ directly.

4.1.2 Selecting Data

The triples were selected from the Lexeur dataset, provided the *-eur* suffixed nouns were listed in the GLAWI closure (*cf.* subsection 3.4). That is to say, masculine and feminine deverbal nouns that were selected were defined using a word referring to a human, which removes any potential instrumental noun from the set.

Following the methodology of Bonami and Paperno (forthcoming), three filters were also applied to constrain the data to be comparable statistically. Firstly, all words must be above an absolute frequency, which was set at 50. This ensures that all words are frequent enough for their associated vectors to represent their contexts well.

Secondly, it was required that the ratio of absolute frequencies be between 1/5 and 5, that is: $\frac{1}{5} < \frac{\#(masc)}{\#(verb)} < 5$. This controls that items in a given triple do not have frequencies too dissimilar to be studied.

Thirdly, the median of the ratios computed for all frequencies must be 1. This last condition along with the previous one ensure that the triples are of a homogeneous enough distribution, centered around 1, to be compared.

These conditions resulted in a set of 112 triples, which is large enough to allow for statistical analysis. So as to obtain enough data, the feminine agent noun was chosen as the pivot (the common word between the two pairs of word): because the filters applied are more lenient for the pivot than for the compared items, more triples could be kept when select as a pivot the least frequent items of the three, *id est* the feminine agent noun. Moreover, varying the verb form of the triple according to the GLÀFF paradigms can change the size of the set greatly. The choice of using the indicative imperfect third singular form is due to the fact that it was the unambiguous form that provided the largest set of candidates under these constraints.

4.1.3 Results

The sets are to be compared pairwise: the sets of distance measures for human noun pairs C_{1_D} with the set of distance measures for deverbal agent noun derivation pairs C_{2_D}, the set of shift cosine measures for noun pairs $C_{1_{MC}}$ with the set of shift cosine measures for deverbal agent noun derivation pairs $C_{2_{MC}}$, and the set of paired vectors cosine measures for human noun pairs $C_{1_{PC}}$ with the set of paired vectors cosine measures for deverbal agent noun derivation pairs $C_{2_{PC}}$. The expectations are both that measures from each paired sets differ significantly from one another and that the distance from the mean shift be lower and cosine measures be higher when it comes to masculine forms in general.

The comparisons are computed as a dependent paired samples Student's t-test. This test evaluates as a null hypothesis that the data is distributed in the same way in both sets, or more precisely that the same average value is to be found for both sets. [6]

Sets compared		t-statistics	p-values
Pair shifts distances to the mean	C_{1_D}, C_{2_D}	-6.49	2.5073×10^{-9}
Pair shifts angles to the mean	$C_{1_{MC}}, C_{2_{MC}}$	-4.14	68.317×10^{-6}
Angles of paired vectors	$C_{1_{PC}}, C_{2_{PC}}$	11.48	13.547×10^{-21}

Table 1: Results for first experiment

The general results for the t-test are given in Table 1. Negative t-statistics indicate lower measures for the human noun pairs than for the deverbal agent noun derivation pairs, whereas positive values indicate higher measure for the human noun pairs than for the deverbal agent noun derivation pairs. The p-values for each test is very low ($p < 10e - 6$), *id est* the null hypothesis is most certainly false, and the distributions of paired sets differ significantly from one another.

[6] The t-test evaluates the ratio of the difference between the observed and expected values, and the standard deviation over the square root of the number of samples. Conceptually, it is akin to evaluating the average gap between expectation and reality, taking into account how wide the gap is for each observation. It is formalized with the formula:

$$t = \frac{\bar{d}}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{27}$$

where $\bar{d}$ is the mean difference of paired samples: $\bar{d} = \frac{\sum_n d_i}{n}$, $\hat{\sigma}$ the standard deviation: $\hat{\sigma} = \sqrt{\frac{\sum_n (d_i - \bar{d})^2}{n-1}}$, and where d_i is computed as the difference between the i$^{\text{th}}$ paired samples. The p-value for a test is defined by the probability of observing the t-statistic t under the null hypothesis:

$$p = 2Pr(T > |t|) \tag{28}$$

The use of a dependent t-test is justified by the fact that data was paired — in the case of this experiment, it was paired using the same word as a pivot: every feminine human noun paired with a given masculine form was also paired with a base verb as its deverbal agent noun.

In all cases, the tests show that deriving a deverbal noun from a verb base differs significantly from pairing a feminine human noun to a masculine human noun. As can be seen from the first row, shift vectors between feminine agent nouns and verbs are significantly more varied than those between feminine and masculine human nouns. The second row indicates a significant difference between shift cosine measures of human noun pairs and shift cosine measures of deverbal agent noun pairs, although the sign of the t-statistics is not as expected: if human noun pairs are more regular, they should have higher shift cosine measures, and therefore the t-statistics should be positive. This question needs and will be studied further. Finally, the third row indicates that paired vector cosines are higher when it comes to human noun vector pairs than when it comes to pair embodying the deverbal agent noun derivation process.

Therefore, be it in shift variance, angle formed with the mean shift, or angle formed with the paired word, measures from verbs and masculine forms sets diverge significantly. Although the results are generally going in the expected direction, they only give an external, monolithic point of view: the exact difference, and how to interpret it, cannot be captured by only checking the p-value and the t-statistics. Therefore, although these results are as expected, a more in-depth analysis is still useful in order to make sense of the measures. A visual representation of the measures for each paired set, using a boxchart diagram, may prove of help to study and compare the distribution of the three measures for human noun pairs and agent derivation pairs.

The first measure is the variance of the shifts — *id est* the distance to the mean shift —, corresponding to sets C_{1_D} and C_{2_D}. Each element in a set corresponds to the norm of a given shift for a pair of words to the mean shift for all pairs of words. The figures give an idea of how grouped the vectors are in terms of distance. Figure 10 displays obviously enough that masculine forms are in general nearer from the mean

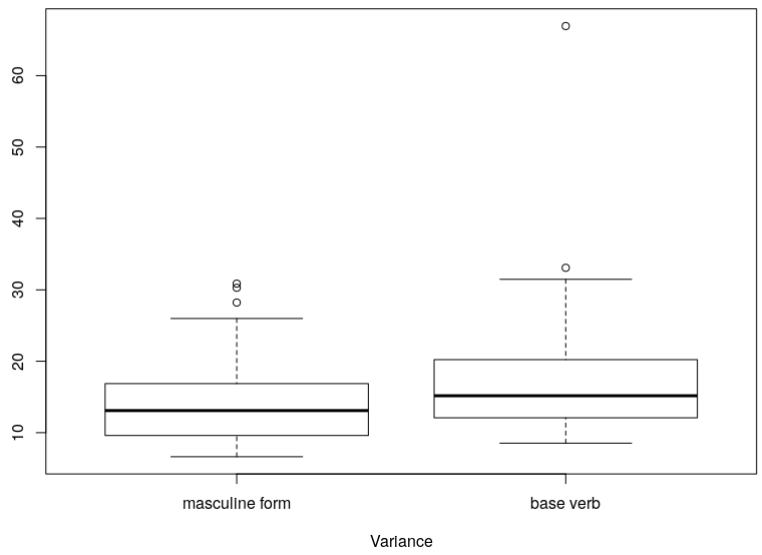


Figure 10: Compared distributions of sets C_{1_D} and C_{2_D}

shift than verb forms. The mean, the quartiles, and the deciles in C_{1_D} are all lower than their respective counterparts in C_{2_D}. What is more, the set C_{2_D} shows a significantly deviant outlier and values greater than the mean are more broadly distributed, which suggests that data is more scattered in C_{2_D} than in C_{1_D}. This is coherent with the proposed hypothesis that French human nouns constitute a paradigm that seems more regular — and therefore more inflectional — than general derivation. If vectors are to be thought as a representation of word senses then the less scattered are the vectors, the more predictible are the word senses.

The second figure to be studied plots $C_{1_{MC}}$ against $C_{2_{MC}}$. It is therefore a measure of the angle formed by the shift for a given pair of vectors and the mean shift for the whole set of pairs embodying the process. Figure 11 shows that the angles are more varied when it comes to masculine than when it comes to verbs. This result seems to contradict the previous interpretations. In fact, one would expect the opposite results:

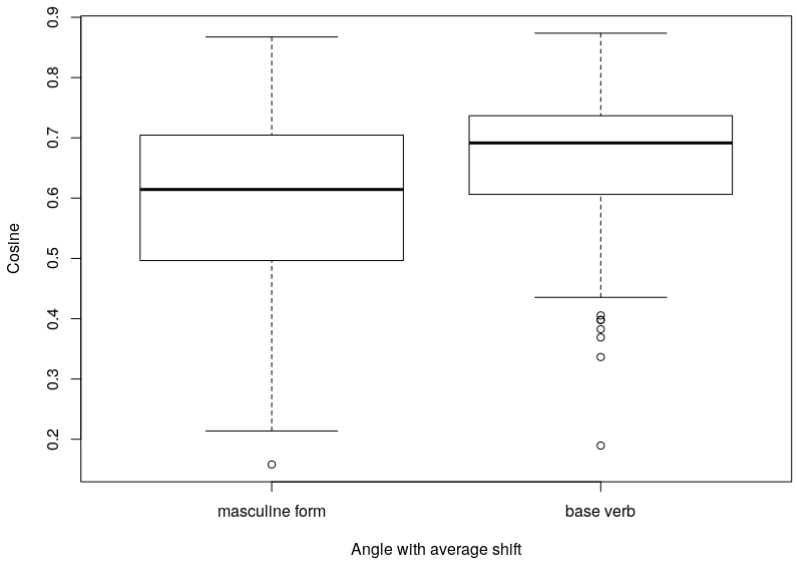


Figure 11: Compared distribution of sets $C_{1_{MC}}$ and $C_{2_{MC}}$

if the process is inflectional in nature, the shift should be more systematic, so the masculine vs. feminine should present a distribution much more restricted around the mean, and the mean itself (being a cosine) be much nearer to 1.

This may perhaps be explained by the fact that feminine forms are in general less frequent, and that therefore the associated vectors are of less quality than masculine forms or base verbs. In this experiment, stricter conditions were imposed on base verbs and masculine forms than on feminine forms — the latter only need occur more than 50 times each. Feminine vectors may therefore be more noisy since they were less controlled. If these vectors are noisy and the shift from masculine forms to feminine forms itself is small — and that is exactly what Figures 10 and 12 (see below) show– then one would expect the cosine relative to the mean to be greatly affected by the noise of vectors associated to the feminine forms.

Lastly, Figure 12 compares box charts for $C_{1_{PC}}$ and $C_{2_{PC}}$. It gives a visual representation of how similar the vectors are to their paired word — either the feminine form for the masculine, or the deverbal noun for the base verb. As previously shown in Equation 11, the greater the cosine, the more similar two vectors are. The figure concords generally with the assumptions made here. On average, the vectors for masculine vs. feminine forms are more similar than the vectors for base verb vs. deverbal noun. This shows that masculine and feminine forms tend to be more similar in meaning than nouns and verbs. The set $C_{1_{PC}}$ is slightly more closely centered around the mean than the set $C_{2_{PC}}$. This is coherent with the analysis that the shift from a masculine to a feminine form is more regular than from a verb base to the associated deverbal noun, since it applies that there is less variation compared to the mean angle of paired vectors — *id est*, that there are less pairs that stray far off of average when it comes to human nouns than when it comes to agent derivation.

The average cosine being higher for set $C_{1_{PC}}$ can be partially explained by the fact that the set $C_{1_{PC}}$ quantifies shift between nouns, whereas the set $C_{2_{PC}}$ describes shifts from a verb to a noun: a transcategorial process such as agent noun derivation involves not only changes in meaning, but also in syntax; therefore the shift of context in a transcategorial process might be more drastic than that of a category-preserving process, in the sense that the same syntactic constraints that hold for the starting and ending points of the latter do not hold when it comes to the former.

The syntactic context being similar for set $C_{1_{PC}}$ might therefore have an influence on the cosine. However, as the vectors were derived from data containing only implicit syntactic relations, the higher cosine is probably not entirely due to the shared part of speech: the lexical relationships between a word and its context should be all the more important that little emphasis was put on the syntactic relationship when computing the model. On the other hand, the distribution being more compact for set $C_{1_{PC}}$ is co-

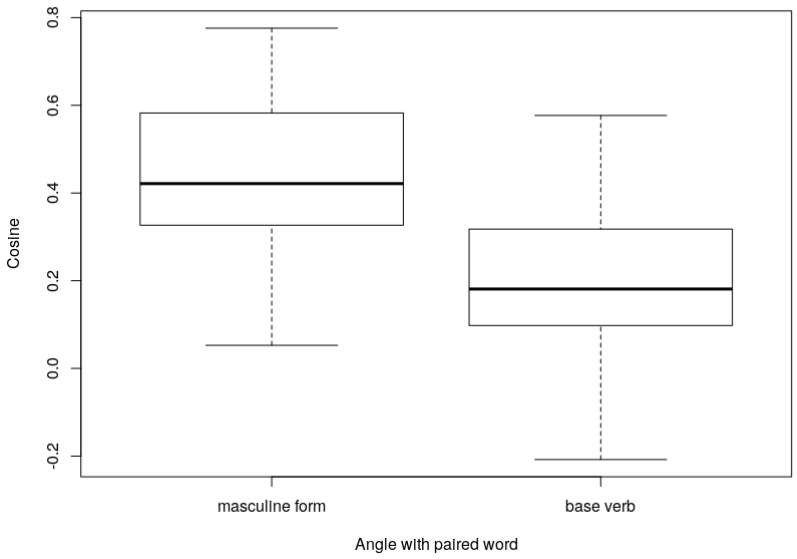


Figure 12: Compared distribution of sets $C_{1_{PC}}$ and $C_{2_{PC}}$

herent with the hypothesis that French human noun paradigms are more inflectional than deverbal nouns derivation, since one expects a more systematic vector for an inflectional process than for a derivational process.

4.2 Second experiment : Teasing apart vector spaces mechanism and morphological effects

The specific issue that Figure 11 highlights is counterintuitive enough and begs for another explanatory experiment. As the cosine measure of the deviation compared to the mean shift is precisely the opposite of what is expected, this specific measure could falsify the general hypothesis of the study. As such, it seems important to separate external factors from the core assumptions. External factors may include the frequencies of the paired items — if one of the vectors is based on a fairly infrequent word, its representation may be poorer and lead to a less trustworthy measure —, as well as the size of the shift between the two paired vectors — the angle deviation to a shift vector may be starker with a smaller shift vector, as the vectors themselves are smaller. On the other hand, the core assumption to be tested is that the kind of morphological process plays a significant role when it comes to angle deviation, that is to say, verbs should deviate more from their related deverbal agent nouns than masculine human nouns from their feminine counterparts.

The general hypothesis of this experiment is that lower frequencies and smaller shifts entail more statistical noise, and that therefore the regularity of g-gender alternation is obfuscated by these. In that respect, the experiment shall consist in finding a model to predict the cosine deviation of the shift based on the morphological process embodied by the paired items, the morphological family shared by the two items, the frequencies of both compared items, and the size of the shift vector from one item to the other. After controlling for the effects of frequency and shift size, if an effect can be found for the morphological process such that g-gender alternation is shown to yield higher cosines, then this hypothesis will be confirmed.

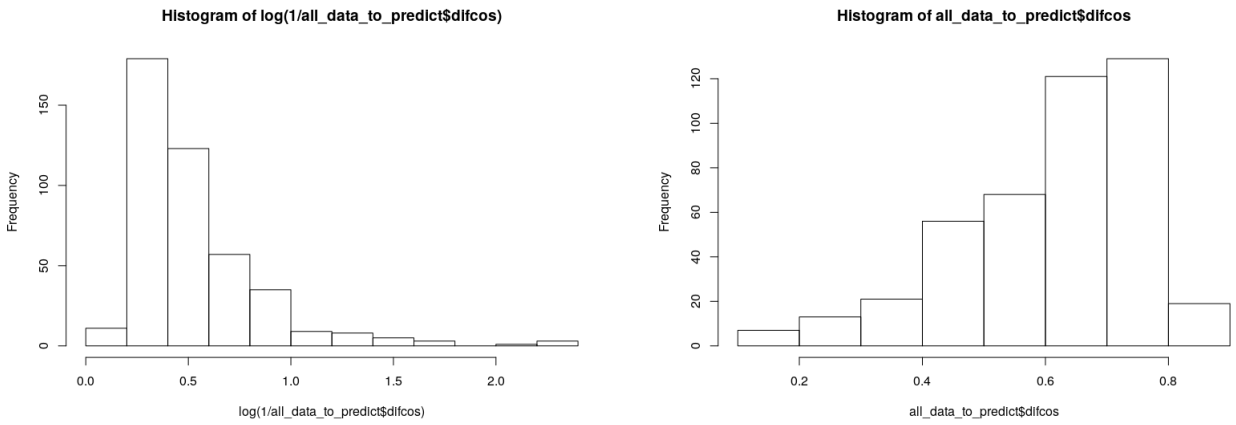
4.2.1 Setup

Such an analysis is technically possible using a generalized linear mixed-effects model (GLMM), which corresponds to a statistical analysis of how predictible a measure (formally, the **dependent variable**) is when taking into account a set of observations (the **predictors**). Statistical models can be conceived as the functional approach to modelisation: the model aims at expressing the dependant variable under scrutiny

— in this case, the cosine measure of the deviation — as a combination of the different predictors. More precisely, since the model that will be used is a GLMM, it expresses the dependent variable as a linear combination of (possibly transformed) linear or random effects. Linear effects are used to model constant factors applying in the same manner to ever datapoint, whereas random effects are meant to represent uneven factors weighing differently depending on the datapoint. Therefore, the model can be conceived as a function associating to every set of values attested for the predictors the resulting dependent variable. In the case of GLMMs, the function can be computed using matrix calculus, and is defined by the general formula : $y = P\beta + Ru + \epsilon$, where y corresponds to the dependent variable, P to the linear effects of the predictors, R to the random effects of the predictors, β and u correspond to coefficients associated to P and R, and ϵ is the residual, *id est* the part that the model failed to generalize using such a function.[7]

The dependent variable, in this case, is the angle deviation as measured by the cosine of the shift vector for a pair and the mean shift for the associated set. The predictors selected for this analysis were the following: the morphological process embodied by the paired items (either g-gender alternation or agent noun derivation), the morphological family shared by the two items (therefore both the cosine deviation of the pair *chanteuse, chanteur* and the cosine deviation of the pair *chanteuse, chantait* are associated to the family *chant-*), the frequencies of both compared items (*id est* the count of occurrences in the corpus for each word in the pair), and the size of the shift vector from one item to the other (computed as the norm of the shift vector). The morphological family was treated as a random effect. Random effects allow for a greater range of possible representations; as it is not certain that all morphological families have the same influence on agent noun derivation or g-gender alternation, the exact fashion in which this influence manifests itself is not easily explained. In particular, this predictor should partially define the vectors themselves, *id est* it should partially capture the semantics of the words themselves: therefore the associated factor is assumed to vary on a per–observation basis. By contrast, the other predictors are all expected to have a more straightforward impact on the dependent variable, and they were therefore represented as fixed effects.

The data for the model needs not be controled in the same way as previously explained in section 4.1.2. In particular, as this analysis takes into account the frequencies of both items, neither the constraint that the ratio of frequencies of compered items be within 1 and $1/5$ nor the need of a median of 1 for these frequencies ratios are needed. However, keeping an absolute frequency threshold of 50 is still relevant, since infrequent words tend to produce unreliable vectors. Adhering only to the frequency threshold constraint, a total of 217 triples can be constituted — which is equivalent to 434 distinct observations of paired vectors embodying either one of the morphological process.



(a) Distribution of the dependent variable log $\frac{1}{\text{cosdev}}$

(b) Original distribution of cosdev

Figure 13: Distribution of the dependant variable compared to the original distribution of cosdev

The data required some manipulations so as to obtain a convergent model. Frequencies were log-

[7]Should there be n observations, based on i linear predictor and j random predictors, y would be of shape $(n\times)$, P of shape $(n \times i)$, β of shape $(i \times 1)$, R of shape $(n \times j)$, u of the shape $(j \times 1)$ and ϵ of shape $(n \times 1)$. As such P and R correspond to the observed values for the predictors, and β and u correspond to the weight each of the i linear effect or j random effect carries.

transformed. Likewise, the dependent variable is not the cosine measure of the angle deviation itself, but rather the log of its multiplicative inverse. This manipulation, illustrated in Figure 13, allows the dependent variable to follow a gamma distribution (*cf.* 13a), rather than a heavily skewed, non–normal distribution (*cf.* 13b) — which in turn allows the model to converge. The log scaling of the invert allows for a sharper contrast. The measurements are originally scaled between 0 and 1; hence their invert is higher when the original measurement tends towards zero. However the log provides a smoothing effect for high elements: therefore the log of the invert groups the lowest original measurements together again. The resulting distribution corresponds to a gamma distribution that can be matched by GLMMs — whereas the original distribution couldn't.

The morphological process type is encoded as a sign function where -1 indicates deverbal agent noun formations, and 1 g-gender alternation. As the dependent variable is modeled using the multiplicative inverse of the cosine deviation, the morphological process type and the shift size predictors are expected to have a negative effect on the dependent variable (a greater shift entails a higher cosine compared to the mean shift, and hence a lower inverse of the cosine), whereas the frequency predictors should have a positive effect.

4.2.2 Results

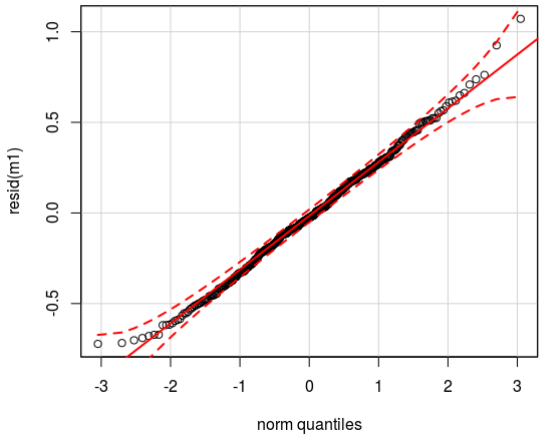


Figure 14: Quantile-quantile plot of residuals for the cosine deviation model

Predictor	Estimate	Std. error	t-stat.	p-values
intercept	−1.061 43	0.280 29	−3.787 00	0.000 15
$log(freq_p)$	0.102 32	0.036 86	2.776 00	0.005 50
$log(freq_c)$	0.558 26	0.040 76	13.697 00	$< 2 \cdot 10^{-16}$
$\lvert p - c \rvert$	−0.078 01	0.009 16	−8.514 00	$< 2 \cdot 10^{-16}$
process embodied	−0.407 07	0.034 53	−11.788 00	$< 2 \cdot 10^{-16}$

Table 2: Fixed effects for the cosine deviation model

The results are summarized in Table 2 . Similar to expectations, the observed effects of shift size and process type are negative, while the effects of the two frequency predictors are positive. As the dependent variable of the model has an inverse growth compared to the measured cosine, this entails that the frequencies of the two items are detrimental to the shift cosine measure, whereas shifts with greater lengths produce more regular cosine measures. These three predictors can be grouped together as intrinsic to the mechanisms of the vector space underlying the DSM, in that they are due to the relative positions of the vectors — for instance, the longer the shift of a given pair of vectors is, the sharper its angle with the mean shift will be. A probable explanation of this phenomenon lies in statistical noise: the shorter a shift between two vectors, the more sensitive to variations will its angle with respect to the mean shift be.

On the other hand, the last row of Table 2 also shows that the morphological process that the paired vectors embody contributes to higher shift cosine measures. This effect is not intrinsic to a vector space itself, but rather reflects the fact that there is a difference of regularity between the two processes compared.

The residual analysis presented in Figure 14 suggests that errors of prediction are within statistically acceptable bounds, and that therefore this model is fit to describe the data. All effects are significant according to the computed p-values; crucially, the hypothesis that gender alternation of human nouns leads to more regular shift than deverbal agent nouns is coherent with the data at hand.

	intercept	$log(freq_p)$	$log(freq_c)$	$\lvert p-c \rvert$
$log(freq_p)$	$-0.767\,00$			
$log(freq_c)$	$-0.717\,00$	$0.239\,00$		
$\lvert p-c \rvert$	$0.622\,00$	$-0.415\,00$	$-0.858\,00$	
process type	$0.330\,00$	$-0.132\,00$	$-0.554\,00$	$0.570\,00$

Table 3: Correlation of fixed effects for first experiment

Another interesting fact is that the predictors are highly correlated with one another, as shown in Table 3. Although high covariance between predictors is often assumed to compromise mixed-effects models, here, all predictors are statistically significant (*cf.* Table 2) and the residual analysis (*cf.* Figure 14) highlights that theGLMM correctly models the data. As such, the high correlation of predictors must be taken as a characteristic of the data itself, rather than a misconception of which factors are relevant to the representation of the studied phenomena. In all, this intercorrelation does not affect the predictors; it can therefore be said that the set of selected predictors explain naturally the variation of the dependent variable.

All in all, the suggestion that morphological process type has an effect on the regularity of paired items shift vectors holds when item frequencies and shift size are taken into account. Moreover, the hypothesis that gender alternation of French human nouns is more regular than deverbal noun derivations is consistent with the proposed GLMM. The model emphasizes that the apparent inverse effect of relation type on cosine is caused by confounding factors, and the actual effect, as observed through the associated predictor, is as expected: human noun g-gender alternation yields higher cosines than deverbal agent noun derivation.

4.3 Partial Conclusions based on the first two experiments

The first experiment follows the expected direction — that the paradigm of French human nouns is more akin to an inflectional paradigm than usual derivational process, such as deverbal formation of agent nouns. However, this first experiment is not sufficient to prove that human nouns g-gender alternation is inflectional: it only proves that human nouns g-gender alternation is more regular than deverbal agent noun derivation — by no means does it indicate where the threshold in terms of regularity should be set so as to tease apart inflectional processes from derivational processes.

Although testing the angle relative to the mean shift in particular showed more puzzling results than other measures, this can be credited to two facts. One is for the most part external to the hypothesis described here: feminine agent nouns being less frequent than their masculine counterparts, the quality of their associated vectors is lower. This first fact leads to two remarks: firstly that removal of bias from the DSM might be of use when studying agent nouns, secondly that results abstracted from mathematical models as DSM or statistical analysis derived of them might be more significant when controlling the potential effects of frequency. Both of these points have already been addressed in the scientific literature (Bolukbasi et al. (2016), Schnabel et al. (2015), and Faruqui et al. (2016) amongst others). The former is due, for the most part, to the fact that human nouns are more often than not very prominently marked when it comes to social gender, all the more so in the rapidly evolving French current trend of marking s-gender on human nouns. As for the latter, it has also been shown that frequency is relevant when it comes to meaning — the naive view that more frequent words are less specialized is not entirely baseless.

The second fact which can partially explain thee puzzling results of the mean shift cosine measures is more intrinsic to the problem of relevance here: the variation seems all the more important when the shift is small. But it has been shown that this variation is a mechanical consequence of the size of the shift

and the frequency of the items, and that, if these are taken into account, g-gender alternation in and of itself produces less important variation. This in turn credits even more the idea that inflection is a more fitting description for French gendered alternating human nouns. The shift vectors being smaller can be conceived as the meaning being more closely related, and it is often remarked in the literature that the meaning of an inflected form is more predictible than that of a derived form.

These two intuitions regarding the importance of frequency and shift size when measuring the angle relative to the mean shift have been statistically verified in the second experiment. What is more, this analysis also highlights that the morphological process embodied by a given pair of vectors contributes significantly to the shift cosine measure. The two other measures from the first experiment consolidate such analysis, since the t-tests show a clear divergence of all these measures for the two morphological processes of deverbal agent noun derivation and noun g-gender alternation. That the distance to the mean shift is smaller when it comes to noun g-gender alternation than when it comes to deverbal agent noun derivation corroborates the relatedness of human noun g-gender alternate forms. That the angle the feminine agent noun forms with the masculine is more consistent than the one it forms with the verb also shows that the g-gender alternation process is more regular — and therefore all the more similar to inflection than derivation — than the deverbal agent noun formation.

These two first experiments could lead to a number of complementary topics of research, which for different reasons fall outside of the scope of the present study. As was previously noted, the transcategorial nature of the deverbal agent noun derivation process might be of importance when comparing the degree of regularity of a canonical derivational process to the g-gender alternation of human nouns. To study this, one would need to reproduce the previous experiments using human nouns derived from a nominal base to constitute triples, such as fermier ('farmer, man'), *fermière* ('farmer, woman'), *ferme* ('farm'). However, to our knowledge, there was no lexicon available that listed enough noun triples of this required form; constituting the necessary dataset would not necessarily prove to be a trivial task.

A second fact which should be studied more in depth concerns the direction of the morphological process. The experiments presented here are based on deverbal agent nouns, and therefore the verbs are systematically the morphological base of the human nouns. It would be of interest to reproduce the current experiments in a setting where the human nouns are the morphological base for the compared derivational process. The production of denominal adjectives can perhaps constitute an interesting point of study. The Dénom lexicon (Strnadová, 2014) does provide a list of denominal adjectives; however very few of these are based on human nouns as extracted from GLAWI, which leads, once more, to the non-trivial issue of constituting a corresponding resource.

What is more, in this specific case g-gender alternating human nouns might complicate a bit the experiment at hand. If one considers for instance the *-eur* and *-euse* suffixes, there is very little data suggesting that these affixes do not provide some sort of morphological closure in the sense of Stump (1998): in the whole 14-billion words corpus used for the study there is only one occurrence of a verb based on an agent noun composed with *-eur* and the classical denominal verb derivative suffix *-iser*, *doctoriser* ('??to doctorize'). What is more, there is no intuitive derivation based on the *-euse* suffix, which would mean that any study focusing on the derivational processes based on *-eur* nouns would be faced either with a lack of complementary data for the *-euse* suffix or with the presupposition that, following the criterion from Stump (1998), the *-eur*, *-euse* alternation is inflectional. Both alternatives would complicate — either theoretically or materially — the reproduction of the experiments presented here. As the reproduced experiments would need to consider a denominal word based on human nouns, the fact that they seem to bring morphological closure would entail that data would be difficult to find, especially for feminine human nouns, and one would therefore have either to assume that feminine and masculine human nouns are two forms of the same word — which is a fault of logic, considering the experiment aims precisely at proving this assumption — or to set aside feminine human nouns from possible bases for derivation due to the lack of relevant data.

One last possible experiment which doesn't lie in the scope of the present study would be to compare human noun g-gender alternation to the inflectional process of human number alternation. Constituting triplets would not prove to be so difficult a task, and would provide an interesting point of comparison. Although the experiments at hand show that g-gender alternation of human nouns is more regular than other typical derivational processes, they say nothing of whether this g-gender alternation should be treated as inflectional, as there is no data that rules out a gradation of regularity amongst derivational processes. These two experiments do not indicate where the domain of inflection ends and where that of derivation begins; and therefore, in and of themselves, they do not answer whether human noun g-gender

alternation should be conceived as inflectional or derivational. Studying this limit requires comparing the g-gender alternation human nouns to an inflectional process — be it nominal number alternation or another morphological process.

5 G-gender alternation in human nouns and adjectives

The previous experiments provided a means of comparing g-gender alternation in human nouns in French to a known derivational process — that of deverbal agent noun formation. Although comparing g-gender alternation to a canonically derivational proces did highlight some significant statistical differences, this comparison isn't sufficient to decide whether human nouns g-gender alternation is inflectional or derivational. One would need to compare human nouns g-gender alternation to another inflectional process since the differences previously found could be conceived as gradation of regularity amongst derivational processes.

When considering which inflectional process to study, French adjectives constitute an obvious choice: they also exhibit g-gender variation and are a textbook example of inflection, as can be seen for instance in Stump (1998). What is more, there is a formal similarity between human nouns g-gender alternation and adjectives g-gender alternation (Bonami and Boyé, forthcoming). These structural and formal similarities bring to the fore the question whether a third semantic similarity can be found. To study how similar these two g-gender alternations are, one should therefore assess both their respective degrees of regularity and their interchangeability with one another.

This in turn questions a specific aspect of the adjective paradigm. One should consider how finely grained the regularity of the adjectival inflectional paradigm is. Because of how intricately entwined g-gender and s-gender are, it is possible that there is some sort of effect when they come to coincide, as is the case with adjectives modifying nouns referring to humans (human qualifying adjectives, HQA). As the g-gender of their human referent should be in line with their s-gender, the alternation between the masculine and feminine form of an HQA should be sensitive to some degree to s-gender and therefore social bias.

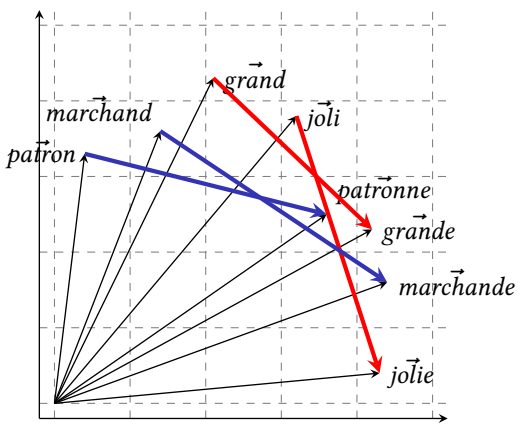
The experiments that this section describes therefore address three questions: firstly, how similar the g-gender alternation of human nouns is to the g-gender alternation of adjectives? Secondly, how constant is the regularity inside a given inflectional paradigm? Thirdly, is there a sub-group of adjectives whose g-gender alternation is more similar to that of human nouns?

5.1 Methodology

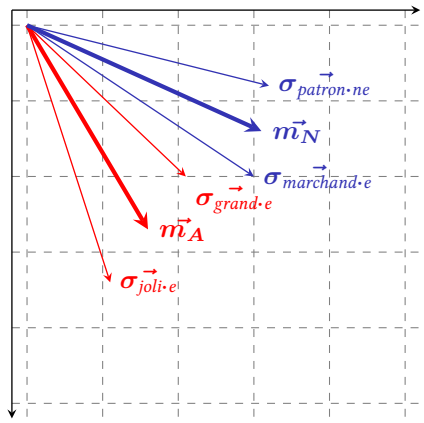
One of the major drawbacks of the methodology of Bonami and Paperno (forthcoming) (*cf.* subsubsection 4.1.1), is that it requires its data to be in the form of triples, *id est* contrastive pairs. However, in the case of g-gender alternation, no common term between adjective g-gender alternation and human noun g-gender alternation exists: the former compares two adjectives, the second compares two nouns, and therefore no pivot can be found. Therefore, unlike previously, there is a need to compare the processes directly, rather than by relying on a common pivot term. This can be done using predictive tasks.When a morphological process is regular, one form is predictable from the other. Hence comparing two processes in terms of predictability informs us on their relative regularity.

A predictive task can be defined as predicting, given a set of inputs and a **predictive function**, a set of target forms. The performance of the predictive function can therefore be studied through the difference between the output of the predictive form (*id est*, the **prediction**) for a given input and the **target** paired to the input.

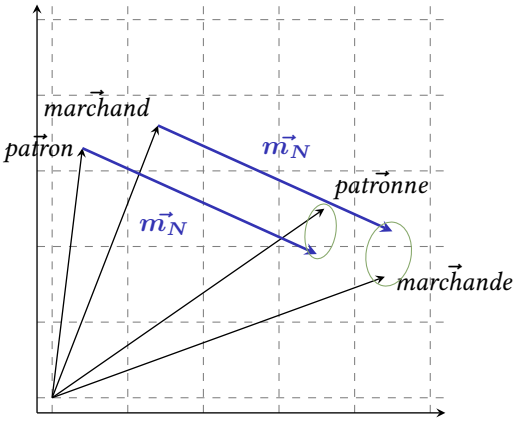
In the following experiments, two predictive tasks were compared, which correspond to the **intrinsic** and the **extrinsic** predictions. Figure 15 provides an illustration of the two predictive functions contrasted. The topleft subfigure represents two morphological processes: g-gender alternation for human nouns (viz. *marchand, marchande*, 'salesman, saleswoman' and *patron, patronne*, 'male boss, female boss'), instantiated by shift vectors in blue, and g-gender alternation for adjectives (viz. *joli, jolie* 'pretty' and *grand, grande* 'big'), instantiated by shift vectors in red. The prediction intrinsic to human nouns is based on information abstracted from the set of paired vectors for human nouns. This function is equivalent to the average shift from the vector associated to one g-gender (eg. $\vec{marchand}$ or $\vec{patron}$ of masculine g-gender) to the vector for the other (eg. $\vec{marchande}$ or $\vec{patronne}$, which are feminine), *id est*, the mean shift $\vec{m_N}$ as computed in the top right subfigure. The intrinsic predictive task itself is illustrated with the bottom left subfigure: the predicted vectors $\vec{marchand} + \vec{m_N}$ or $\vec{patron} + \vec{m_N}$ are those given by the mean shift ($\vec{m_N}$ in blue) added to the inputs ($\vec{marchand}$ or $\vec{patron}$), and the discrepancy with the target vectors (resp., $\vec{marchande}$ and $\vec{patronne}$) is visualised by the green circles.



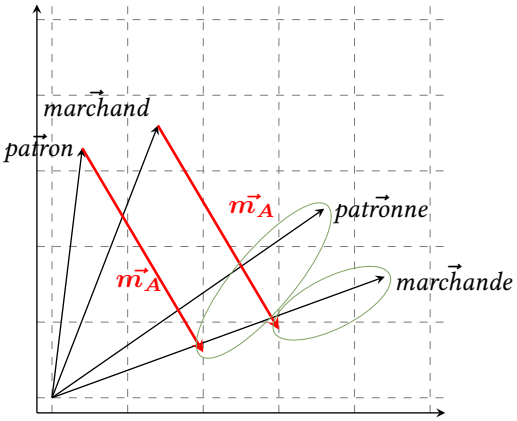
(a) Illustration of the two g-gender alternations



(b) Computation of mean shifts for the two processes



(c) Intrinsic predictions for human nouns g-gender alternation using the mean shift of nouns $\vec{m_N}$



(d) Extrinsic predictions for human nouns g-gender alternation using the mean shift of adjectives $\vec{m_A}$

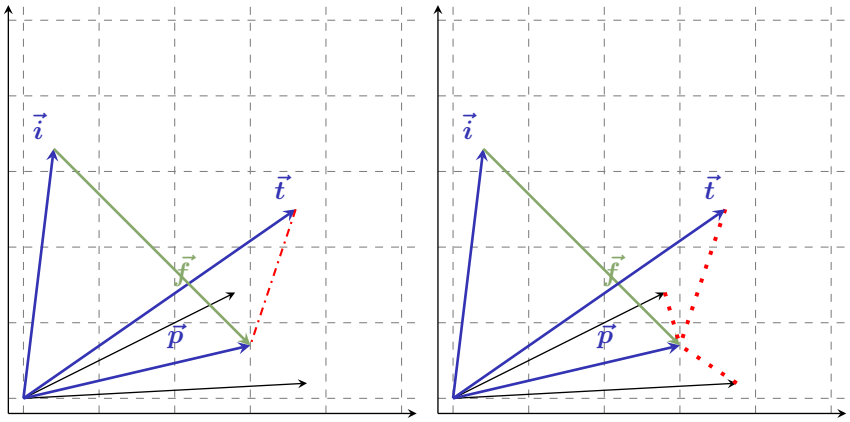
Figure 15: Intrinsic and extrinsic predictions illustrated

An extrinsic prediction, on the other hand, is based on information abstracted from a set of paired vectors embodying a different process — in the example here, the adjectives' g-gender alternation. Similarly to intrinsic predictions, the predictive function is computed as the average translation from the vector of an adjective of a given gender (viz. *joli*, *grand*) to its counterpart (resp. *jolie*, *grande*) which is the mean shift $\vec{m_A}$ as computed in the top right subfigure. What differs is that the set it is computed from, in this case adjectives' g-gender alternation, differs from the set it is applied to, *id est* human nouns' g-gender alternation, therefore this prediction is said to be extrinsic to the morphological process of human noun g-gender alternation. The predictive task derived from that extrinsic mean shift $\vec{m_A}$ (in red) is illustrated in the bottom right subfigure, where the discrepancy between the target vectors (eg. *marchande*, *patronne*) and the predicted forms (resp. $marchand + \vec{m_A}$ and $patron + \vec{m_A}$) is highlighted by the green circles.

In this example, the two processes differ significantly from one another, which leads to poorer performance when it comes to the extrinsic prediction based on adjectives than for the prediction intrinsic to human nouns. However, if the two processes have similar enough semantic effects, the performances could be equivalent in both predictive tasks.

Figure 16 illustrates two different measures for assessing one specific prediction. On the left subfigure, the predicted vector $\vec{p}$ in blue is computed from the input $\vec{i}$ in blue and the predictive function $\vec{f}$ in green, and the quality of that prediction is measured as the distance (illustrated by the red dash-dotted line) between the predicted vector $\vec{p}$ and the target vector $\vec{t}$ in blue. In this experiment, the distance is computed as the Euclidean distance as was previously defined in equation 14.

The right subfigure exhibits the same configuration of input $\vec{i}$, predicted $\vec{p}$ and target $\vec{t}$ vectors using the same predictive function $\vec{f}$. The measure represented however is that of distance ranking, or rank; it is



(a) Measuring the distance between the predicted vector and the target

(b) Measuring the rank of proximity to the target prediction for the predicted vector

Figure 16: Distance & rank measures

equivalent to assessing to what extent the predicted vector is the best approximation of the target vector, knowing the other vectors populating the vector space. In this figure, the measured rank of the prediction is 3, as there are two other vectors (in black) which are closer to the predicted vector than the target. This can be mathematically defined by the following formula:

$$rank(\vec{t}, \vec{p}) = \#\{\vec{v} \mid dist(\vec{p}, \vec{v}) < dist(\vec{p}, \vec{t}) \ \wedge \ \vec{v} \in V\} + 1 \tag{29}$$

The use of both a rank measure and a distance measure can be seen as avoiding any assumptions regarding the evenness of the distribution of the vectors in the DSM. Distance measurements fail to capture that a predicted vector far off from its target might be best approximated by the target — *id est*, that the nearest neighbour of the predicted vector is the target. On the other hand, clusters of closely grouped vectors tend to deteriorate rank measurements. Studying both measures conjointly allows for a more nuanced account of the performances of the predictive functions.

In all, intrinsic predictive tasks resemble somehow the methodology of Bonami and Paperno (forthcoming), presented in subsubsection 4.1.1. Bonami and Paperno (forthcoming) defined three sets of measures: two based on a comparison of a pair's shift to the mean shift of its group, the third by directly comparing paired vectors. Here, in intrinsic predictive tasks, the variation within the process is assessed by comparing the predicted vector based on the mean shift to the target vector. In that respect, it is similar to comparing a pair's shift to the mean shift; what differs most are the measures applied to assess this variation: for instance, instead of cosine deviation, rank is used.

Extrinsic predictive tasks, on the other hand, aim at assessing a different aspect that the methodology of Bonami and Paperno (forthcoming) does not take into account: to what degree the semantic effects of the two processes are similar. As these predictions are akin to comparing a pair's shift to the mean shift of the *other* process, it entails a different pont of view than what was implicit in the methodology of Bonami and Paperno (forthcoming): these author are interested in assessing the variation within a process — here, we study the similarity of processes with one another.

5.2 Third experiment: Comparing g-gender alternation in human nouns and adjectives

In order to assess how similar the g-gender alternation of human nouns is to the g-gender alternation of adjectives, the methodology described in was used to define intrinsic and predictive tasks pertaining to human noun g-gender alternation and adjective g-gender alternation. There were therefore four different tasks: an intrinsic predictive task for human noun g-gender alternation, an intrinsic predictive task for adjective g-gender alternation, an extrinsic prediction task for human noun g-gender alternation based on the adjective mean shift, and an extrinsic task for adjective g-gender alternation based on the mean shift for human nouns.

The intrinsic prediction should give an intuition of how regular a process is. However this intuition only makes sense when provided with a point of comparison. The intrinsic predictions therefore give the relative degrees of regularity of the two processes, and do not constitute an absolute measure. It therefore makes sense to compare intrinsic predictions with one another.

The study of extrinsic prediction is here justified by the formal and structural similarities previously underscored. In the hypothetical case that the same semantic effects are to be found in human nouns and adjective g-gender alternation the extrinsic and intrinsic predictions should yield the same results. The extrinsic prediction for a given process should therefore be compared to the intrinsic prediction for the same process.

This third experiment will therefore focus on three comparisons: comparing the intrinsic prediction of adjectives g-gender alternation to the intrinsic prediction of human nouns g-gender alternation, comparing the intrinsic and the extrinsic predictions for adjectives, and comparing the intrinsic and the extrinsic predictions for human nouns. Intuitively, the two last comparisons should express two facets of the same coin — that is, the sameness or the difference of the semantic effects of the two processes of g-gender alternation discussed here.

5.2.1 Setup

The third experiment consists in intrinsic and extrinsic predictive tasks for both human nouns and adjectives. Following the methodology defined in subsection 5.1, the performances for the four different tasks were assessed with both distance and rank measures, for a total of eight measurements.

The lexicon of human nouns computed for the previous experiment (subsection 3.4 and Figure 5) provided the basis for the selection of human nouns. Contrary to what was done in the first experiment, there is no need to restrict the selected data to human nouns suffixed with *-eur*: therefore any pair of human nouns that were extracted from the GLAWI closure was used. The only filter applied to human nouns was that as was noted before, the data on which the experiment is run must be controled for frequency. The same filter was of course applied to adjectives as well, thus in this experiment all items had a frequency between 100 and 1000. These constraints resulted in 120 human noun pairs and 4874 adjective pairs, therefore the sets are large enough for statistical observations to be significant.

Finally, the DSM used was a word2vec model (CBOW architecture, negative sampling, window of 5) computed over the corpus previously described (subsection 3.2), with feminine and masculine homographs as well as nouns and adjectives homographs disambiguated — which resulted in different vectors for possibly ambiguous forms, for instance for the word *communiste* all the following possible senses received a distinct representation:

- 'communist man' (masculine noun)
- 'communist woman' (feminine noun)
- 'that is communist' (masculine adjective)
- 'that is communist' (feminine adjective)

Therefore none of the contrasts conjointly observed — masculine and feminine grammatical gender on the one hand, and noun and adjective words on the other — overlap in any way. This first model shall be referred to as $\mathcal{M}_1$.

5.2.2 Results

Two different factors were assessed by this $\mathcal{M}_1$ model: firstly, the relative degrees of regularity of adjectival and nominal g-gender alternation, and secondly, the similarity of the semantic effects of these morphological processes.

The first point can be studied by a statistical analysis of whether the two intrinsic predictions yield similar measurements. Contrary to the first experiment, the classic paired t-test could not be applied, due to the fact that the processes sets weren't paired by morphological families. The predictions were

therefore compared using an independent t-test, also known as an unpaired t-test.[8]

Measure	t-statistics	p-values
Euclidean distance	2.882 4	0.004 7
Rank	−1.098 0	0.272 9
Log rank	1.109 5	0.269 4

Table 4: T-test results for intrinsic predictions in the $\mathcal{M}_1$ model

The results of the Welch t-tests are recapitulated in Table 4. For each specific measure detailed in the previous section, the two processes of nominal g-gender alternation and adjectival g-gender alternation were compared. The Euclidean distance indicates that there is a significant difference between human nouns and adjectives, in the sense that adjective predictions are closer to their respective targets than human noun predictions. On the other hand, the rank measure did not prove useful to the distinction of adjectives and human nouns, as can be shown by their large p-values — to be precise, the t-tests is statistically inconclusive due to these large p-value. As can be seen from the log scale rank measurements, this issue is not resolved by re-scaling the measures.

Studying more closely the distribution of rank and distance measurements might highlight interesting facts. Table 5 summarizes the descriptive statistics of the measurements of the intrinsic predictive tasks conducted in the $\mathcal{M}_1$ model. The first column indicates which process is studied, the second gives the measure used to assess the discrepancy between prediction and target, and the six reamining columns display descriptive statistics (minima, first quartiles, medians, means, third quartiles, maxima). In this

Prediction using	Measure	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Nouns	Euclidean distance	0.078 0	0.236 4	0.303 7	0.331 7	0.418 7	0.739 6
	Rank	1	2	10.5	2597	94.25	156400
Adjectives	Euclidean distance	0.025 7	0.210 5	0.277 3	0.296 3	0.362 3	1.051 0
	Rank	1	3	8	4728	51	3467000

Table 5: Descriptive statistics for the intrinsic predictions in the $\mathcal{M}_1$ model

specific case, it is best to look at the median and the quartiles, rather than the mean and the extrema for two reasons: firstly, outliers tend to strongly influence the maximum and the mean; secondly, if the measures here presented are consistent with the observations of Mikolov, Yih, and Zweig (2013), then one could expect that most predictions should fall near their targets and that therefore the distributions of the measurements be skewed towards the minimum.

What can be seen from these descriptive statistics is that adjectives, in general, yield lower measurements. The difference is more clearly pronounced with the distance measurements. This was to be expected from the t-tests.

When it comes to noun distance measurements, the interval between the first quartile and the median spans only about half of the interval between the median and the third quartile, whereas the interval between the third quartile and the maximum spans even longer than the interval between the minimum and the median. The smallest interquartile interval is between the first quartile and the median.

Looking now at adjectives distance measurements, the same sorts of remarks can be made, although the measurements are in general lower and the intervals smaller. The interval between the third quartile and the maximum is however much larger than what it was for nouns, spanning roughly twice the interval between the minimum and the third quartile. The smallest interquartile interval is once again between the first and second quartiles. The distribution presents the similar trend of being heavily skewed towards the first quartile.

[8] More specifically, the applied algorithm was a Welch t-test, where the t-statistic is given by the formula :

$$t = \frac{\bar{m_1} - \bar{m_2}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \tag{30}$$

where $\bar{m_i}$ is the mean of a given set, n_i the size of a set and $\hat{\sigma}_i^2$ the variance of a set, *id est*, the square of its standard deviation (*cf.* equation 27). Only the computation of the t-statistic vary, and the remainder of the algorithm is fairly similar to that which was described before.

Rank measurements, however, express a different trend. A quarter of the human noun predictions are amongst the top two nearest neighbors, and half of them are in the top ten. On the other hand, the third quartile is set at about nine times the median. The rank measurement distribution ressemble more closely a powerlaw instead on being denser between the first quartile and the median, as was the case for distance measurements.

The case is still different when it comes to adjectives: quartiles are generally similar, except for the third quartile, which is only around half of what was observed for human nouns. Finally, adjective pairs constituting a larger set, the maximum value is higher, and therefore the mean is lower. However this implies that the distribution in both cases strictly follows a power law — and this is coherent with the t-tests that show no significant statistical difference for these measurements.

As a side remark, since the maxima are in each case very far from the third quartiles, a manual account of the most salient outliers should bring about interesting facts relating to irregular adjectives and human nouns.

The second point which this experiment addresses is that of how similar the semantic effects of the two processes are — or in other words, whether the two processes yield similar outputs for the same input.

Measure	t-statistics	p-values
Euclidean distance	$-5.772\,0$	$2.512\,0 \cdot 10^{-08}$
rank	$-2.019\,9$	$0.045\,5$
logrank	$-6.244\,6$	$1.999\,0 \cdot 10^{-09}$

(a) T-test results for human nouns intrinsic and extrinsic prediction in the $\mathcal{M}_1$ model

Measure	t-statistics	p-values
Euclidean distance	$-39.328\,0$	$< 2 \cdot 10^{-16}$
rank	$-1.911\,8$	$0.055\,9$
logrank	$-33.169\,0$	$< 2 \cdot 10^{-16}$

(b) T-test results for adjectives intrinsic and extrinsic prediction in the $\mathcal{M}_1$ model

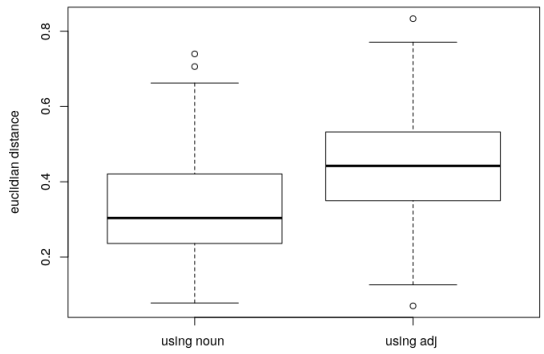
Table 6: T-test results for extrinsic predictions in the $\mathcal{M}_1$ model

This can be done by comparing the intrinsic and extrinsic predictions for a process. Results are summarized in Table 6. The sub- 6a presents the t-test results for the human nouns, and likewise the sub- 6b those for the adjectives. All tests highlight a significant statistical difference between the intrinsic and the extrinsic predictions, which always favours lower measurements for the intrinsic predictions; the only exception being the adjective rank measurements, with a p-value slightly over 0.05. However, the same measurements pitted on a log scale are unequivocal.

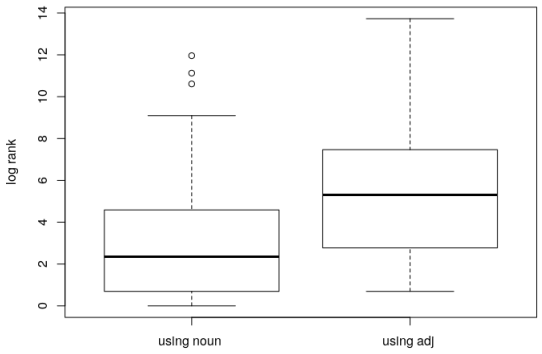
Prediction using	Measure	Min.	1^{st} Qu.	Median	Mean	3^{rd} Qu.	Max.
Nouns with adjective shift	Euclidean distance	$0.070\,3$	$0.349\,8$	$0.442\,3$	$0.441\,2$	$0.531\,4$	$0.833\,2$
	rank	2	16	201.5	20300	1690	916000
Adjectives with noun shift	Euclidean distance	$0.099\,8$	$0.264\,9$	$0.382\,2$	$0.394\,0$	$0.496\,7$	$0.823\,3$
	rank	1	6	50	5623	788	194500

Table 7: Descriptive statistics for the extrinsic predictions in the $\mathcal{M}_1$ model

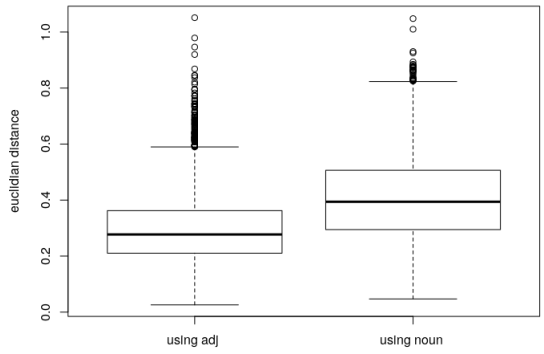
Comparing the descriptive statistics of the extrinsic predictive tasks as recapitulated in Table 7 to the same statistics for the intrinsic predictive tasks (as were presented in Table 5) gives an indication of the dissimilarity of the semantic effects of the two processes. For instance the rank medians, which were respectively of 10.5 and 8 for the intrinsic predictions of nouns and adjectives, are much higher — respectively 50 and 201.5. The distributions for distance measurements of nouns and adjectives extrinsic prediction are less skewed than it was, and more centered around the median. Ranks still follow a power law however it is even steeper than what was attested for the intrinsic predictions. The same conclusion can be made from the observation of Figure 17, comparing intrinsic and extrinsic predictions measurements. Subplot 17a compares distance measurements of noun instrinsic and extrinsic predictions, whereas 17b compares the log scaled rank measurements for these two predictions. Likewise, subplot 17c compares distance measurements for adjective intrinsic and extrinsic predictions, and subplot 17d their log scaled rank measurements. The same observation can be made for each of the four cases regrouped in Figure 17: intrinsic predictions yield lower measurements than extrinsic predictions, and the semantic effects af nominal and adjectival g-gender alternation therefore differ.



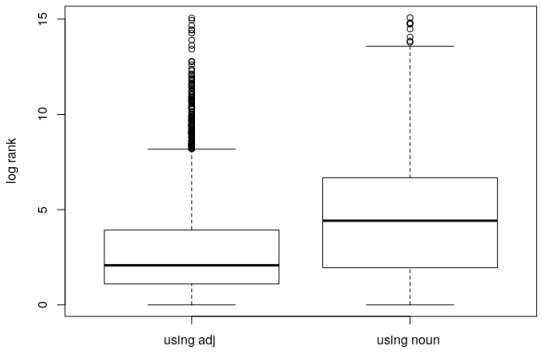
(a) Comparison of distance measures of noun intrinsic and extrinsic predictions



(b) Comparison of log-scaled rank measures of noun intrinsic and extrinsic prediction



(c) Comparison of distance measures of adjective intrinsic and extrinsic predictions



(d) Comparison of log-scaled rank measures of adjective intrinsic and extrinsic prediction

Figure 17: Comparisons of extrinsic and intrinsic predictions in $\mathcal{M}_1$ model

5.2.3 Discussion

This third experiment provides interesting observations relevant to the two questions previously mentioned — that of the difference in regularity between nominal and adjectival g-gender alternation, and that of the difference in semantic effects between these two processes.

The statistically lower measurements for adjectival g-gender alternation indicate that this process is less subject to variation than the nominal g-gender alternation. In other terms, adjectives, as a whole, are more regular than human nouns. However, rank measurements do not coincide with distance measurements, in that they exhibit no statistically significant difference between the adjectives and the nouns. That no such difference can be found using rank measurements provides an intuition as to how the vector space is structured, hinting that the word vectors might be highly clustered. This might be due to the fact that the corpus over which the model was computed being rather large, so is the vocabulary (*id est*, the set of word types) — and the disambiguation of homographs (as described previously in subsubsection 5.2.1) further accentuates this fact. Testing this hypothesis is however not necessary to the present study, as the distance measurement has proven to be reliable. What matters chiefly is the fact that adjectives constitute a group more regular than nouns when it comes to the operationalisation of formal analogy that Mikolov, Yih, and Zweig (2013) described.

The second point that this set of measures enquired was the similarity of the semantic effects of nominal and adjectival g-gender alternation. It is not exactly clear whether one should expect g-gender alter-

nation in nouns to have the same effects as g-gender alternation in adjectives. From a formal perspective, these processes share much of their morphological material and syntactic characteristics. However this formal similarity do not logically imply a semantic similarity; and the ranks and distance measurements described above tend towards proving the opposite. This downgrade in predictability when comparing extrinsic and intrinsic predictive tasks can be summarized as a divergence in semantic effects — to use a masculine human noun rather than a feminine semantically differ from using a masculine adjective rather than a feminine one.

This particular observation leads to two remarks. Firstly, one possible reason why these two processes are not semantically equivalent might be that the g-gender in human nouns is inherent to the lexeme and independent of the syntactic context, whereas it is not inherent in the case of adjectives and dependent on the syntactic context. This explanation, which is similar to that of Booij (1995), would however imply that the g-gender alternation of human nouns is inflectional.

A second point concerns the regularity inherent to the predictive tasks themselves, which depends on the regularity of the semantic effects. Although there is a clear case that adjectives are most of the time more regular than human nouns when it comes to g-gender alternation, the mean and maximum for the distance measurements of intrinsic predictive task of adjectives are noticeably higher than that of nouns. This can be explained by two facts — neither of which excludes the other. Firstly, the adjectives form a set almost 50 times larger than nouns, which does impact the probability of including atypical adjectives. This would in turn explain the higher number of outliers as well as the higher maximum value and therefore the higher mean value for adjectives.

Another possible explanation is that the internal consistency of the adjectives is less than that of the nouns. In other words, the adjective set might cover a wider range of semantic effects — most of which would be fairly similar to the average effect as expressed by the mean shift, and some others which would be atypical and therefore not well described by the same average shift. The extrinsic predictions for adjectives using the mean noun shift tend to produce worse results and yet less outliers: perhaps this smoother distribution can be conceived as a more homogeneous representation of the wider coverage of semantics effects for adjectives — which would then imply that there is a subregularity of semantic effects inside the adjective set that in some manner or another would be comparable to the effects of nominal g-gender alternation.

5.3 Fourth experiment: Consistency of semantic effects

The previous experiment led to the hypothesis that a subregularity of semantic effects might exist in adjectival g-gender alternation, and that some of them might exhibit a behaviour more similar to that of human nouns.

Adjectives describing primarily humans — such as *talentueux, talentueuse* ('gifted') — might be more similar to human nouns than other adjectives, as they should share the same context for the most part. The semantic effects for the g-gender alternation of adjectives describing primarily humans would then be similar to the semantic effects of human noun g-gender alternation. If it is so, the more an adjective pair is used to qualify human nouns rather than other nouns, the more similar the shift from one item of the pair to the other will be to the mean shift of human nouns.

Rather than constituting a subclass of atypical adjectives, the subregularity would be expressed on every adjective, but with different levels of intensity. If it is confirmed, the adjective shift should express a continuous effect, from adjectives primarily used to describe humans (eg. *talentueux, talentueuse*) to those not necessarily describing humans (eg. *grand, grande*, 'tall'), and to those (almost) never describing human referents (eg. *nucléaire*, 'nuclear').

5.3.1 Setup

So as to compare each pair of adjectives to the mean human noun shift, first, the mean human noun shift had to be computed.

The set of human nouns was once again computed using the method described in subsection 3.4, pairs were computed using concordance information from GLAWI paradigms, Lexeur, or when the items listed in GLAWI were common nouns — in which case the feminine and masculine forms were disambiguated, therefore constituting a pair of distinct human nouns. As to be consistent with the definition of the $\mathcal{M}_1$ model, nouns were disambiguated from adjectives. The only frequency constraint that was deemed

necessary was that human nouns should be attested at least 50 times in the corpus, so as to ensure that they do encode their distributional contexts. The mean shift from masculine human nouns to feminine human nouns was therefore computed from this set.

The hypothesis tested here is that the type of noun — human or not — that the g-gender alternate adjective forms usually refer to impacts the shift from one of them to the other; therefore a mean to discriminate what type of noun the pair refers to was necessary. Since the parser from Coavoux (2017) which was used to uniformize the corpus annotations outputs a CONLL format representation of each sentence, it was possible to match adjective tokens to the nouns they modify: either the noun token indexed directly by the HEAD column of the adjective, or, when the adjective token's HEAD column was a verb and its syntactic category was either object or subject attribute (indicated by an ATO or ATS value in the REL column), to the noun object or subject of the verb.

This matching naturally defines a ratio of usage as qualifying a human noun, computed as the number of times the paired adjective forms were matched to a human noun as extracted from GLAWI divided by the total number of occurrences of the adjective. Since only human nouns with a frequency equal or greater to 50 were taken into account to compute the mean noun shift, likewise, adjective pairs for which any item occurred less than 50 times were not considered for this section of the second experiment. These constraints resulted in a total of 15624 adjective pairs.

The hypothesis itself can be tested using a mixed-effects model. In this case, the dependent variable consists of a similarity measure between the adjective pair shift and the mean noun shift. As was stated before, cosine measurements are a measure of such similarity. Predictors would include on the one hand frequency and shift size, and the human qualification ratio computed as described above.[9] Finally, the identity of the morphological family of the adjective pairs was used as random effects.

5.3.2 Results

The model was ran using the R-Studio LME4 library (Bates et al., 2015). The model converged to the

predictor	estimate	std. error	t-stat.	p-values
intercept	$2.29 \cdot 10^{-01}$	$6.66 \cdot 10^{-03}$	34.35	$< 2 \cdot 10^{-16}$
log freq	$5.78 \cdot 10^{-02}$	$1.36 \cdot 10^{-03}$	42.65	$< 2 \cdot 10^{-16}$
ratio	$1.66 \cdot 10^{-01}$	$1.62 \cdot 10^{-02}$	10.20	$< 2 \cdot 10^{-16}$
shift size	$-1.40 \cdot 10^{-02}$	$3.47 \cdot 10^{-04}$	-40.38	$< 2 \cdot 10^{-16}$

Table 8: Fixed effects for human qualification ratio model

results described in Table 8. As the dependent variable is monotonous with respect to the cosine similarity measure — that is, the dependent variable grows all the more the cosine nears 1 —, predictors with positive effects should be taken as improving similarity. What is more, as the dependent variable was scaled between 0 and 1, the estimates of the predictors can also be compared to one another.

All predictors were deemed significant. Moreover, the residual analysis displayed in Figure 18 shows that the model is sound and provides an accurate description of the data. A significant effect can be found in particular for the frequency and shift size predictors, which confirms once more the fact that the regularity of a process is affected by the frequency of the items that embodied it and their relative positions in the vector space.

The quantitatively most important effect is the one associated with the ratio and contributes to higher cosine values. This indicates that the type of noun that the adjective pair tends to qualify has a significant impact on the shift from one adjective form to the other, precisely in that adjectives qualifying human nouns have a shift more similar to the average noun shift.

5.3.3 Discussion

The model highlights the importance of the type of nouns that the adjective qualifies to the semantic effect of the g-gender alternation. This model shows that inflectional processes are not as systematic as

[9] To be perfectly precise, so as to obtain a normal distribution, the dependent variable was computed as $dv_i = -\log(\log(\frac{1}{\cos(\vec{A^i_f} - \vec{A^i_m}, \vec{m_N})}))$ — where $\vec{A^i_f} - \vec{A^i_m}$ is the shift for a given adjective and $\vec{m_N}$ the mean noun shift — and further rescaled between 0 and 1 so as to obtain easily interpretable results. Likewise the frequency was first log-transformed.

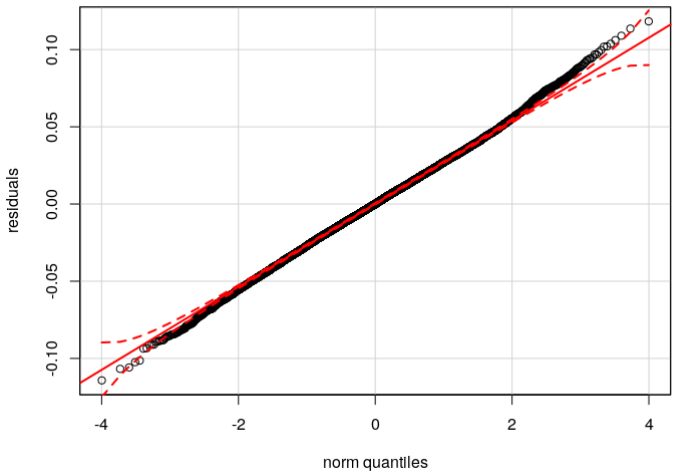


Figure 18: Residual analysis for human qualification ratio model

what would be expected from Mikolov, Yih, and Zweig (2013) (*cf.* Figure 4), and that the regularity of an inflectional process is not as absolute as what transpired from the criteria of Stump (1998). This model stresses the fact that regularity, expressed either in terms of formal analogy as in Mikolov, Yih, and Zweig (2013) or in terms of predictable semantic effects as in Stump (1998) does not hold in an absolute fashion. That such a subregularity can be found when scrutinizing an inflectional process indicates that inflection cannot be as systematic as it is generally assumed, as it implies that the semantic effect of an inflectional process is modulated by semantic characteristics of the base.

What is more, the model also shows that frequency matters in such comparisons, further implying that not all cases of a same inflectional process behave the same under any and all measurements. This in turn naturally questions whether inflection is as cleanly distinct from derivation as what Stump (1998) argued: frequency, as well as subregularity, undermines such assumption.

The precise nature of the subregularity indicates that adjective g-gender alternation in some cases resembles that of nouns. In other words, there are some adjectives for which g-gender alternation is more similar to that of nouns than adjectives in general; and these adjectives are those that primarily qualify human nouns. The fact that they tend to appear in contexts to that of human nouns need not follow, from a logical standpoint, their definition: the human qualification ratio that was computed here was defined in syntactic and not contextual terms. Therefore, that an adjective qualify a human noun does not imply that they cooccur in a fixed window, nor does it imply that the non-coincident window of occurrences of these two would not obscure the similarity of their contexts. Considering the sentence *Un beau garçon, assurément : beau de corps à cause de sa force visible, et beau de visage à cause de ses traits nets et de ses yeux téméraires…*[10] ('A handsome boy, certainly: handsome due to the visible strength of his body, and handsome due to the sharp features of his face and his daring eyes…') shows that the context of the human noun (*garçon*, 'boy') does not necessarily coincide with that of the adjectives that modify it (here, the three occurrences of *beau*, 'handsome'). Although the first occurrence of *beau* is right next to the human noun it modifies, the third occurrence of *beau* does not overlap at all unless the size of the window was set to 15 words or more. Moreover the DSM representations of any paired adjective and human noun are abstracted for a wider variety of contexts than just those where they co-occur — if the effect that this model highlights was purely random, it would have been expected that the more observations are made, the less this randomness would be expressed.

This provides an objective basis to disambiguate adjectives according to their usage. That some adjective behave differently because they tend to qualify human nouns implies that these effects would be more sharply displayed when restricted to only contexts where adjectives qualify human nouns. To reproduce the previous measurements in a model in which this distinction is pertinent should prove highly

[10] L. Hémon, 1916, *Maria Chapdelaine*, quoted in the CNRTL.

informative of the regularity of adjectival g-gender alternation, and how comparable that process is to nominal g-gender alternation.

5.4 Fifth experiment: Comparing g-gender alternation in human nouns and HQA

The third experiment highlighted a clear difference between human nouns and adjectives g-gender alternation. However, the GLMM of the fourth experiment proved that the shift of an adjective pair tend to ressemble more closely the average shift of nouns the more they are used to qualify these human nouns. It is therefore likely that the g-gender alternation of adjectives only used to qualify human nouns (HQA) exhibit a different behavior from the general case that was previously studied.

5.4.1 Setup

So as to contrast adjectives' usages — *id est*, distinguish HQA from non-HQA — a second DSM $\mathcal{M}_2$ was computed (as a word2vec CBOW architecture using negative sampling and a 5 words window), where common gender nouns are disambiguated as well as homonymous adjectives and nouns (like in subsection 5.2), and adjectives tokens are disambiguated between those that modify a human noun according to the dependency parse (*cf.* subsection 5.3) and those that don't. The model therefore distinguishes all cases relevant for the analysis, and for instance the word-form *communiste*, which would be ambiguous between six possible cases, is properly distinguished between:

- 'communist man' (masculine noun)
- 'communist woman' (feminine noun)
- 'that is communist' (masculine adjective, not referring to a noun classified as human)
- 'that is communist' (feminine adjective, not referring to a noun classified as human)
- 'that is communist' (masculine adjective, referring to a noun classified as human)
- 'that is communist' (feminine adjective, referring to a noun classified as human)

The subsequent measurements could therefore contrast three groups: HQA, non-HQA, and human nouns. However as the non-HQA do not form a natural class, this experiment will only compare HQA and human nouns. Each vector of every pair embodying either of the two processes was controlled to ensure its frequency was betwen 100 and 1000. As in subsection 5.2, any pair of human nouns could be selected, provided that both item weren't filtered out by the frequency constraint. This resulted in 121 human nouns and 481 HQA pairs.

As these measurements attempt to reproduce what was described previously in subsection 5.2 while contrasting HQAs and human nouns, the same measures of rank and distance were used. The predictive tasks were once again either intrinsic — using the average shift from a given set to predict g-gender alternate members from this set, *cf.* 15c — or extrinsic — using another set to predict g-gender alternate members from a given set, *cf.* 15d.

For each of these predictive tasks, and so as to be consistent with the previous measurements, two measures were used in this $\mathcal{M}_2$ model : Euclidean distance and rank — resulting in eight distinct measurements over four sets of predictions.

5.4.2 Results

The similarity of regularity between HQA and human nouns was assessed using Welch t-tests over the corresponding intrinsic predictions measurements. Results are presented in Table 9, showing t-tests results

Measure	t-statistics	p-values
Euclidean distance	$0.556\,4$	$0.578\,7$
Rank	$-0.462\,4$	$0.644\,1$
Log rank	$-1.232\,7$	$0.219\,3$

Table 9: t-test results for intrinsic predictions in the $\mathcal{M}_2$ model

for distance, rank, and log rank.

Similarly to what was observed in subsection 5.2, the rank measurements aren't statistically conclusive, even when log scaled. On the other hand, the Euclidean distance measurements yield different results in the human nouns vs. HQA comparisons of the $\mathcal{M}_2$ model and the human nouns vs. adjectives comparisons of the $\mathcal{M}_1$ model. This highlights that although a difference in regularity can be found when looking at adjectives in general, no such conclusion can be made, from a statistical standpoint, when focusing on HQA.

The results from a t-test indicate a general trend and do not tell much about finer details of the distribution, which needs to be studied more closely. Table 10 regroups the descriptive statistics for the intrinsic

Prediction	Measure	Min.	1^{st} Qu.	Median	Mean	3^{rd} Qu.	Max.
Nouns	Euclidean distance	0.098 86	0.213 90	0.293 10	0.324 60	0.405 20	0.756 30
	Rank	1	2	7	3406	103	228500
HQA	Euclidean distance	0.068 68	0.231 50	0.294 30	0.316 70	0.382 50	0.848 90
	Rank	1	3	12	4769	178	704500

Table 10: Descriptive statistics for the intrinsic predictions in the $\mathcal{M}_2$ model

predictive tasks in the $\mathcal{M}_2$ model. The two top rows detail the distribution of the noun intrinsic predictions measurements, whereas the two bottom rows concern HQA intrinsic prediction. Looking at distance measurements, although HQAs have a higher first quartile, they also have a lower third quartile, and their means are almost equal. This was to be expected from the t-tests. When it comes to rank measurements, the same observation as in the third experiment can be made: the distribution in both cases seems very similar to a power law, and therefore the higher quartiles for HQA can be seen as a consequence of the larger number of HQA pair samples.

Measure	t-statistics	p-values
Euclidean distance	$-5.053\,3$	$8.632\,0 \cdot 10^{-07}$
Rank	$-1.169\,2$	$0.244\,3$
Log rank	$-5.236\,5$	$3.577\,0 \cdot 10^{-07}$

(a) T-test results for human nouns intrinsic and extrinsic prediction in the $\mathcal{M}_2$ model

Measure	t-statistics	p-values
Euclidean distance	$-9.341\,5$	$< 2 \cdot 10^{-16}$
Rank	$1.316\,6$	$0.188\,5$
Log rank	$-6.347\,7$	$3.365\,0 \cdot 10^{-10}$

(b) T-test results for adjectives intrinsic and extrinsic prediction in the $\mathcal{M}_2$ model

Table 11: T-test results for extrinsic predictions in the $\mathcal{M}_2$ model

Table 11 summarizes the t-tests results for the extrinsic predictions measurements in the $\mathcal{M}_2$ model. Euclidian distances measurements consistently highlight a statistically significant difference between the instrinsic prediction and the extrinsic predictions, for both g-gender alternation processes. Rank measurements, on the other hand, do not indicate any difference unless they are log-transformed first. This differs from the third experiment, where the highest p-value was slightly above 0.05: here rank measurements have a p-value around 0.24 when it comes to nouns, and 0.19 when it comes to HQA. However, when log-transformed, a statistically significant difference can be observed.

As a side remark, although no definitive conclusion can be drawn from a comparison accross two different vector spaces, the contrast between human nouns and HQA differs from what could be seen in the third experiment. The t-statistics for the adjective distance and log-scaled rank measurements were around five time that of nouns; in this fifth experiment however the human nouns t-statistics are never below half what is computed for the HQA. What is more, the sign of the adjective rank measurements t-value is opposite to what is expected, which isn't the case for any t-statistics of the third or the fifth experiment.

Descriptive statistics for the extrinsic predictions in the $\mathcal{M}_2$ model can be found in Table 12. Similarly to what was observed in the third experiment, the measurements reported for the extrinsic predictions are generally higher than what is observed for the intrinsic predictions. On the other hand, the extrinsic predictions do not reshape the distribution observed for distance measurements: contrary to what was seen in the third experiment, in both intrinsic and extrinsic predictions and for both processes, the smallest interquartile interval observed for distance measurements is between the first quartile and the median.

Prediction	Measure	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Nouns with HQA shift	Euclidean distance	0.093 07	0.330 60	0.401 50	0.419 80	0.529 10	0.826 60
	rank	2	16	99	11660	939	791700
HQA with noun shift	Euclidean distance	0.128 40	0.303 40	0.392 40	0.392 60	0.472 10	0.819 40
	rank	1	6	50	5623	788	194500

Table 12: Descriptive statistics for the extrinsic predictions in the $\mathcal{M}_2$ model

A more visual representation of the fact that extrinsic predictions for both HQA and human nouns perform systematically less well than the corresponding intrinsic predictions can be seen on Figure 19, which mirrors in $\mathcal{M}_2$ what Figure 17 showed in $\mathcal{M}_1$. Subfigure 19a compares the distance measurements of intrinsic (on the left) and extrinsic (on the right) predictions of human nouns, subfigure 19b compares the log-scaled rank measurements for intrinsic and extrinsic predictions of human nouns, subfigure 19c compareds the distance measurements of intrinsic (on the left) and extrinsic (on the right) predictions of HQA, and finally subfigure 19d compares the log-scaled rank measurements for intrinsic and extrinsic predictions of HQA. The same facts observations can be made for the third and the current experiment : the distribution is systematically lower for the intrinsic predictions, which implies that the semantic effects of human nouns and HQA g-gender alternation are different.

5.4.3 Discussion

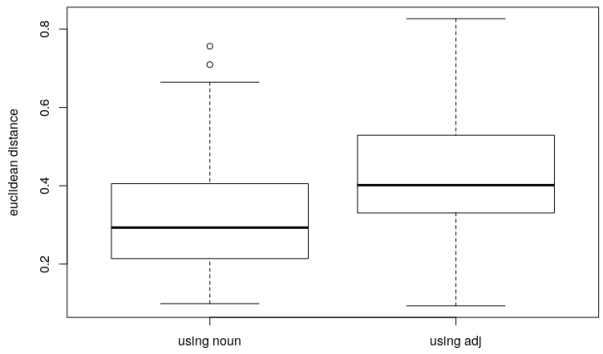
This final set of measurements shows that the difference studied in section 5.2 is not as clear-cut as it seems when results from section 5.3 are taken into account. No difference in terms of predictability can be found between HQA and human nouns. If regularity does entail predictability, these results indicate that the same degree of regularity can be observed in human nouns g-gender alternation and HQA g-gender alternation.

Therefore, the assumption that nominal g-gender alternation is derivational in nature is in contradiction with the criterion from Stump (1998) claiming that inflection is more regular than derivation. To assume that nominal g-gender alternation is inflectional, on the other hand, either goes against the traditional assumption that gender is intrinsic to the noun's lexeme or requires that human nouns in French are neither of feminine nor of masculine g-gender, but rather, following the proposal of Bonami and Boyé (forthcoming), that they vary inflectionally in gender.[11] However, as these authors underscore, this tentative explanation would have many consequences in terms of paradigm uniformity.

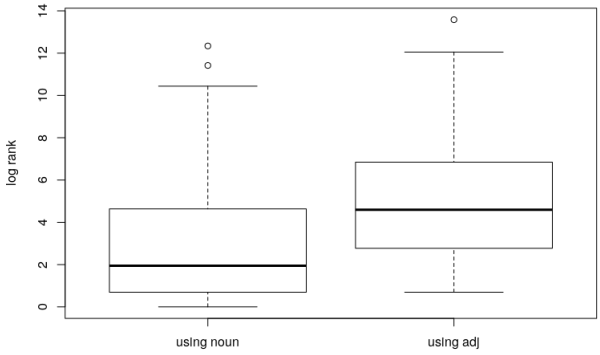
A second point that emerges from this fifth experiment is that the semantic effects of human nouns g-gender alternation aren't equivalent to those of adjectives, both in general and in particular those of HQA. The generally poorer results from extrinsic predictions show that there is no semantic equivalence betwen nominal and adjectival g-gender alternation, even when taking into account the subregularity mentionned in subsection 5.3. One can conclude from such observations that the g-gender alternation of human nouns and that of adjectives do not play the same semantic role. This consolidates the claim that there is a difference between adjective g-gender and nominal g-gender — although the exact distinction adopted from Booij (1995) and generally admitted in the literature between contextual g-gender of adjectives and inherent g-gender of nouns is still to be proved. It is noteworthy that the distinction between contextual and inherent processes as defined in Booij (1995) is conceived as a distinction amongst inflectional processes — thus adopting this explanation would imply treating nominal g-gender variation as an inflectional process.

Another interesting fact to be highlighted is that distinguishing HQA from adjectives in general does not fully resorb the variation in semantics effects, as can be attested from Table 10 describing the distribution of the intrinsic predictions. The interval between the third quartile and the maximum for the distance measurements of HQA intrinsic predictions summarized in this table indicates that there is an

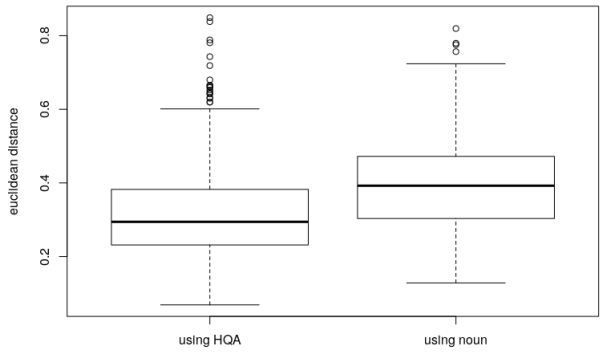
[11] Strenghtening this claim from a purely morphological perspective is the fact that there exist some nouns such as *photocopieur, photocopieuse* ("copy machine") or *ravin, ravine* ("ravine") which exhibit the same pattern of possessing two g-genders — although since these nouns do not denote humans, their alternative forms have for the most very little to do with s-gender; in fact, it is very difficult to pinpoint a consistent semantic difference between the feminine and the masculine forms of these words. However, what matters chiefly is that the same pattern can be found in these nouns, and this entails that the four-cell inflectional paradigm is not merely an illusion stemming from a highly regular derivational process — words like *photocopieur, photocopieuse* have nothing to do with s-gender differentiation, and still possess a four-cell paradigm.



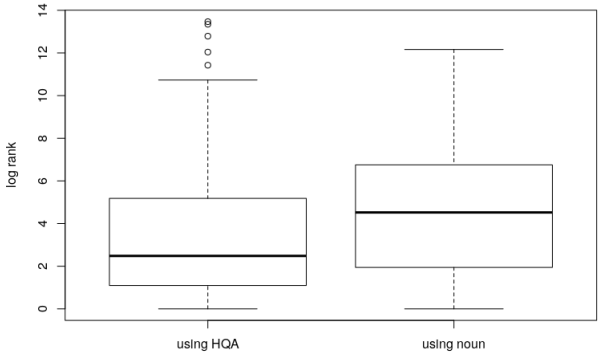
(a) Comparison of distance measures of noun intrinsic and extrinsic predictions



(b) Comparison of log-scaled rank measures of noun intrinsic and extrinsic prediction



(c) Comparison of distance measures of HQA intrinsic and extrinsic predictions



(d) Comparison of log-scaled rank measures of HQA intrinsic and extrinsic prediction

Figure 19: Comparisons of extrinsic and intrinsic predictions in $\mathcal{M}_2$ model

important group of adjectives amongst HQA which are not well represented by the average HQA shift. One would expect, if the human qualification ratio was the only important subregularity, that the distance measurements follow a normal distribution. The general tendency of the distribution therefore indicates that some other effect than the mean shift is necessary to compute the actual shift of a pair of words — this is in fact true regardless of the morphological process that the pair embodies. Section 5.3 implies that frequency and shift size could play this role, although further research that lies outside of the scope of the present study would be necessary to establish this fact with certainty.

A last point of interest is that this experiment compares human nouns to a specifically restricted subset of adjectives — which were selected so as to share syntactic context with human nouns. For the sake of completeness, comparing HQA to another sub-class of adjectives should provide interesting facts to study. So as to provide as stark as possible a contrast, the selected sub-class should be that of adjectives qualifying a non-human noun. This class will be noted as NHQA in the following. It's important to note that NHQAs constitute a natural class: their syntactic context follow the same sort of restriction than what was applied to select HQA, and they never modify any word that may have both human and non-human senses, like *allumeur* ('person in charge of lighting up', 'male flirt', or 'ignition device'). Therefore both the semantics and the syntax of NHQAs are homogeneous, which entails that NHQAs form a natural class.

5.5 Sixth experiment: Comparing g-gender alternation in nouns, HQA and NHQA

The fifth experiment underscored a similarity between HQA and human nouns g-gender alternation. However this experiment did not explicitly show how different HQA were from adjectives in the general sense; moreover it tested two different factors at the same time: the effect of selecting a sub-group of the whole set of adjectives pairs, and the effect of qualifying human nouns on the meaning of a given adjective pair. However either of these two factors do not necessary entail the consequences of the other.

Therefore the following experiment will aim at studying g-gender alternation within three groups: HQA, NHQA and human nouns. The expectations are that g-gender alternation of human nouns is more similar to that of HQA than to the alternation of NHQA. Moreover, as NHQA refer to inanimate or abstract nouns, which therefore should not possess a s-gender, one would expect to observe a greater regularity within NHQA than within HQA.

5.5.1 Setup

The sixth experiment will focus on studying semantic regularity of human nouns and of two classes of adjectives: that of HQA and of NHQA. These classes are defined using the information mined from the GLAWI lexicon; which distinguishes nouns having only human senses, nouns having no human senses, and nouns having both human and non-human senses. Therefore, if HQA are related to nouns having only human senses, and NHQA to nouns having no human senses, there are still a number of adjectives qualifying nouns which are neither purely human or non-human. What is more, the two classes of adjectives are defined through their syntactic contexts; in particular many adjectives do not qualify any nouns, but rather pronouns, verbs, or other, and therefore are not part of either class.

Thus NHQAs do not constitute the complementary set of HQA among the set of all adjectives, and three categories must be distinguished : adjectives qualifying human nouns only — *id est*, HQA —, adjectives qualifying non-human nouns only — NHQA — and the remaining adjectives which either do not qualify a noun, or a qualify a noun with mixed human and non-human referents — noted ?HQA. Only HQA and NHQA constitute natural classes, and ?HQA should only be understood as a mathematical notation to denote the remaining adjectives not listed in HQA nor NHQA.

So as to provide a distinct representation for human nouns, HQA and NHQA, it was necessary to compute another model — which will be referred to as the $\mathcal{M}_3$ model. In this $\mathcal{M}_3$ model, feminine gender is distinguished from masculine gender, and the class to which a token belongs to is unequivocal — *id est*, nouns, adjectives in general, HQA and NHQA receive different representations. As such, the word-form *communiste* in this model is properly distinguished between eight different possible ambiguous meanings:

- 'communist man' (masculine noun)
- 'communist woman' (feminine noun)
- 'that is communist' (masculine adjective, referring to a noun classified as non-human)
- 'that is communist' (feminine adjective, referring to a noun classified as non-human)
- 'that is communist' (masculine adjective, referring to a noun classified as human)
- 'that is communist' (feminine adjective, referring to a noun classified as human)
- 'that is communist' (masculine adjective, not referring to a noun, or referring to a noun neither classified as human or non-human)
- 'that is communist' (feminine adjective, not referring to a noun, or referring to a noun neither classified as human or non-human)

The model itself was computed as a a word2vec model(CBOW architecture, negative sampling, window of 5).

In order to be consistent with the previously defined methodology, only pairs where both items were attested between 100 and 1000 times were taken into account. However, this constraint resulted in sets of very different sizes: 118 noun pairs, 481 HQA pairs and 5074 NHQA pairs. As such imbalanced set size might impact the results due to a greater number of outliers in the more numerous set, the NHQA were subjected to one more restriction.

One way to test this is to compute pairwise distances within each groups of vectors — resulting in six sets of measurements: one for feminine human nouns, one for masculine human nouns, one for masculine HQA, one for feminine HQA, one for feminine NHQA and one for masculine NHQA — and study whether the process a pair belongs to influences the computed distance. This analysis can be done using an ANOVA first, to see if the process entails significant variation of the measurements. As a second optional step, a Tukey's Honest Significant Difference (HSD) test[12] can be performed so as to estimate the factor on distance and of each processes. Results of this analysis are presented in Table 13. As can be seen from the results of

Measurement	F-value	$Pr(>F)$
Process type	115473	$< 2.2e-16$

(a) ANOVA results for pairwise distance variation per process

Processes compared		diff.	lower	upper	adj. p-value
Nouns	HQA	0.024 48	0.020 20	0.028 77	0.000 00
Nouns	NHQA	−0.180 71	−0.184 88	−0.176 54	0.000 00
NHQA	HQA	0.205 19	0.204 17	0.206 22	0.000 00

(b) Tukey HSD test results for pairwise distance variation per process

Table 13: Evaluation of pairwise distance variation per process

the ANOVA presented in 13a, pairwise distances vary significantly ($p < 2.2e-16$) according to which morphological process is embodied by the pair. The results of the HSD test shown in 13b indicate that any two proces yield different measurements, however the estimated variation introduced by comparing NHQAs to human nouns or HQA is about ten times what is observed from comparing human nouns to HQAs. This confirms the intuition that the number of samples correlates with more variation within the sampled items. In that respect, as NHQA pairs are about ten times as numerous as human noun pairs, it was necessary to select the most cohesive group of NHQA. This can be done using clustering techniques, such as unweighted pair group method with arithmetic mean (UPGMA).[13]

The pair-wise Euclidean distance matrix between all feminine NHQA — here noted $\mathcal{D}_f$ — and between all masculine NHQA — here noted $\mathcal{D}_m$ — were computed, then added to one another, in order to provide a distance matrix for all NHQA pairs — noted $\mathcal{D} = \mathcal{D}_f + \mathcal{D}_m$. A hierarchical, single-link UPGMA clustering was computed on the basis of this distance matrix $\mathcal{D}$. From this clustering, the bottom-most cluster covering at least 100 elements was selected. This clustering constraint resulted in a set of 101 NHQA pairs.

As this experiment is concerned with reproducing the previous observations, while taking into account both HQA and NHQA, the same sort of predictive tasks were devised. The tasks were once again divided between intrinsic predictions (as was illustrated in 15c) for each of the three groups — human nouns, HQA and NHQA — and extrinsic predictions (*cf.* 15d) based on all possible combination of two processes — therefore predictions on human nouns based on information abstracted from HQA, predictions on human nouns based on NHQA, predictions on HQA based on human nouns and based on NHQA, predictions for NHQA based on human nouns and based on HQA. For all three intrinsic predictions and all six extrinsic predictions, results were measured in terms of Euclidean distance and rank, as previously defined using Figure 16, resulting in a total of eighteen different sets of measurements.

5.5.2 Results

As this experiment involved three morphological processes at a time, rather than two as previously, the methodology that was adopted up to this point wasn't adapted to the subsequent analysis. In fact, t-tests — be they parametric as in the case of Student's or non-parametric in the case of Welch's — do not provide

[12] Tukey HSD test is an inferential statistic test for comparing more than two groups at once. It compares the means of all pairs of groups, using the formula:

$$\text{HSD} = \frac{\bar{m_1} - \bar{m_2}}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{31}$$

where $\bar{m_1}$ and $\bar{m_2}$ correspond to the means of the two groups, and $\frac{\hat{\sigma}}{\sqrt{n}}$ to the standard error *for the entirety of the measurements*. It closely resembles t-tests; in fact, Tukey's HSD test can be thought of as a t-test accounting for family-wise error rate. This test is therefore used when comparing multiple related groups pairwise, as multiplying t-tests would entail a greater chance of false positive findings.

[13] UPGMA is a hierarchical clustering algorithm that merges clusters two by two. At initialization, all elements to be clustered are represented as singleton clusters, which constitute a set of clusters to be merged. The algorithm then retrieves the two clusters nearest to one another, removes them from the set of clusters to be merged, and merges them into a new cluster which is added to the set of clusters to be merged. Since at every step two clusters are removed and one is added, the algorithm finishes when there is only one cluster remaining to be merged. This last cluster corresponds to the top-most cluster.

any means of comparing three items at a time; and although it is technically feasible to multiply t-tests to study the items pairwise, doing so entails a greater chance of false positive findings.

So as to avoid this, the following methodology was adopted to analyze the measurements: an analysis of variance (ANOVA) was conducted first to see if the factors under scrutiny were significant; when they were, a Tukey's Honest Significant Difference (HSD) test was applied so as to provide estimated factor and adjusted probabilities for each pair of processes. In the particular case of this experiment, the ANOVA consisted in testing whether the processes could be discriminated from one another on the basis of a given measure; whenever the analysis of variance indicated such a variation in measurements, a Tukey HSD test was applied to compare processes pairwise on the basis of said measure.

Measurement	F-value	$Pr(> F)$
Euclidean distance	9,444 8	$8{,}97 \cdot 10^{-05}$
rank	0,659 1	0,517 6
log rank	15,087	$3{,}85 \cdot 10^{-07}$

Table 14: ANOVA results for intrinsic predictions in the $\mathcal{M}_3$ model

Table 14 presents the results for the analysis of variance of the three intrinsic prediction tasks. The assigned probabilities underscore a variation amongst what the intrinsic predictions of the morphological processes yield, both in terms distance measurements and — when log scaled — rank measurements.

Pred.	compared	diff.	lower	upper	adj. p-value
nouns	HQA	−0.007 83	−0.036 40	0.020 74	0.795 89
nouns	NHQA	0.048 46	0.010 76	0.086 16	0.007 41
NHQA	HQA	−0.056 29	−0.086 73	−0.025 85	0.000 05

(a) Tukey HSD test results for distance measurements

Pred.	compared	diff.	lower	upper	adj. p-value
nouns	HQA	−0.620 57	−1.234 41	−0.006 72	0.046 85
nouns	NHQA	0.857 21	0.047 22	1.667 20	0.035 10
NHQA	HQA	−1.477 78	−2.131 79	−0.823 77	0.000 00

(b) Tukey HSD test results for log rank measurements

Table 15: Tukey HSD tests results for measurements of intrinsic predictions in the $\mathcal{M}_3$ model

A Tukey HSD test is required so as to assess what this difference consists of. The corresponding results are displayed in Table 15. Looking at the adjusted p-values for distance, in 15a, underscores no significant difference between nouns and HQA, but NHQA are shown to yield lower measurements than HQA and human nouns, as both p-values are under 0.05. This therefore implies — when considering distance measurements — that NHQA are more regular than nouns and HQA, and that nouns and HQA cannot be really distinguished from one another in terms of regularity.

Rank measurements displayed in 15b accordingly highlight the same difference between NHQA on the one hand and nouns and HQA on the other. Moreover, adjusted p-values suggest that there is a significant difference in measurements not only when comparing nouns and HQA to NHQA, but also when comparing nouns to HQA. More precisely, nouns are shown to yield lower measurements than HQA — although the actual computed difference is less than what is computed when comparing to NHQA. This therefore implies that NHQA embody the more regular process, followed by human nouns, and that HQA correspond to the least regular process.

Extrinsic predictions might provide an interesting insight regarding the similarity of the degrees of regularity of nouns and HQA. Because the results of two different extrinsic prediction tasks — one based on information from the set of HQA pairs, the other based on information abstracted from the set of NHQA — are to be compared to the results of the intrinsic prediction task for human nouns, following the previously defined methodology, an ANOVA is first required to see if the different prediction tasks yield different measurements. This analysis is summarized in Table 16. Much like what was observed for

Measurement	F-value	$Pr(> F)$
Euclidean distance	21,37	$1{,}746 \cdot 10^{-09}$
rank	1,841 1	0,160 2
log rank	37,611	$1{,}586 \cdot 10^{-15}$

Table 16: ANOVA results for extrinsic predictions of human nouns in the $\mathcal{M}_3$ model

intrinsic predictions, both distance and log-scaled rank measurements highlight a variation amongst the different predictive tasks.

To study more precisely what this difference entails, a Tukey HSD test is performed for distance and log rank. The Tukey HSD test also provides a comparison of the two extrinsic predictive tasks with one another, therefore assessing whether either one of the two processes is more similar — in terms of semantic effects — to the g-gender alternation of human nouns than the other. Results of this HSD test can be found in

Pred.	compared	diff.	lower	upper	adj. p-value
intr. pred.	HQA based	−0.101 39	−0.145 68	−0.057 10	0.000 00
intr. pred.	NHQA based	−0.111 02	−0.155 31	−0.066 73	0.000 00
NHQA based	HQA based	0.009 63	−0.034 66	0.053 92	0.865 65

(a) Tukey HSD test results for distance measurements

Pred.	compared	diff.	lower	upper	adj. p-value
intr. pred.	HQA based	−2.001 57	−2.877 08	−1.126 06	0.000 00
intr. pred.	NHQA based	−3.191 96	−4.067 48	−2.316 45	0.000 00
NHQA based	HQA based	1.190 39	0.314 88	2.065 90	0.004 27

(b) Tukey HSD test results for log rank measurements

Table 17: Tukey HSD tests results for measurements of extrinsic prediction of human nouns in the $\mathcal{M}_3$ model

Table 17. Studying distance measurements as reported in 17a, using information from either HQA g-gender alternation or NHQA g-gender alternation entails a clear deterioration of the measurements, indicating that the semantic effects of these two processes differ from that of human nouns g-gender alternation. Moreover, as no significant statistical effect is attested when comparing the two extrinsic predictive tasks, both processes seem equally removed from human nouns in terms of semantic effects.

When studying log rank measurements — summarized in 17b —, the observation can be made that extrinsic predictions lead to poorer results. However, log rank measurements also account for a significant difference when comparing the two extrinsic predictive tasks. NHQA yield even higher measurements than HQA, suggesting that the semantic effects of HQA g-gender alternation are more similar to the semantic effects of human nouns g-gender alternation than the semantic effects of NHQA.

The second group of extrinsic predictive tasks concerns HQA. First, like previously, an ANOVA is performed to assess whether a difference can be observed. Results for each measure are listed under Table 18.

Measurement	F-value	$Pr(> F)$
Euclidean distance	66,853	$< 2 \cdot 10^{-16}$
rank	0,571 3	0,564 9
log rank	132,33	$< 2 \cdot 10^{-16}$

Table 18: ANOVA results for extrinsic prediction of HQAs in the $\mathcal{M}_3$ model

As previously, a significant variation amongst processes is observed for distance and log-scaled rank measures. A Tukey HSD test can therefore be conducted to study the exact nature of this variation. Results for the HSD are to be found in Table 19.

Similar to what was found for human nouns, HQA extrinsic prediction, either based on human nouns information or on NHQA information, yields significantly lower distance measurements than intrinsic prediction. Moreover, the adjusted p-value for the comparison of the two extrinsic predictive tasks with one another indicate that there is no sufficient grounds to distinguish the measurements of the two processes. The same conclusion can therefore be made, that, when looking at distance measurements, the semantic effects of HQA g-gender alternation are equally not well captured by either of the two other processes.

Log-scaled rank measurements, however, consistently with what was found before, clearly indicate that noun-based extrinsic predictions are a better fit than NHQA based extrinsic predictions, as the computed difference shows. The results also underscore once more that extrinsic predictive tasks yield poorer results than the intrinsic predictive task. From this, a possible conclusion — consistent with what was said

Pred.	compared	diff.	lower	upper	adj. p-value
intr. pred.	noun-based	−0.073 83	−0.092 89	−0.054 76	0.000 00
intr. pred.	NHQA based	−0.087 25	−0.106 31	−0.068 18	0.000 00
noun-based	NHQA based	−0.013 42	−0.032 49	0.005 64	0.224 33

(a) Tukey HSD test results for distance measurements

Pred.	compared	diff.	lower	upper	adj. p-value
intr. pred.	noun-based	−0.994 66	−1.396 95	−0.592 36	0.000 00
intr. pred.	NHQA based	−2.754 37	−3.156 66	−2.352 07	0.000 00
noun-based	NHQA based	−1.759 71	−2.162 01	−1.357 42	0.000 00

(b) Tukey HSD test results for log rank measurements

Table 19: Tukey HSD tests results for measurements extrinsic prediction of HQAs in the $\mathcal{M}_3$ model

for human nouns extrinsic predictions — is that there is a gradation of semantic effects similarity: human nouns would be more similar to HQA than NHQA.

The third and last group of extrinsic predictions are those for NHQA. As always, an ANOVA is necessary

Measurement	F-value	$Pr(> F)$
Euclidean distance	532,38	$< 2 \cdot 10^{-16}$
rank	16,213	$2{,}06 \cdot 10^{-07}$
log rank	268,55	$< 2 \cdot 10^{-16}$

Table 20: ANOVA results for extrinsic prediction of HQAs in the $\mathcal{M}_3$ model

to show whether variation amongst processes is attested. Table 20 displays the results of this analysis. An effect can be observed for distance, rank, and log-scaled rank measurements. Therefore, in this case, it will be necessary to test rank measurements as well as distance and log-scaled rank measurements. As previously, a Tukey HSD test is conducted for each significantly varying measurement; the results are summarized in Table 21.

Results for the distance measure, recapitulated in 21a, are consistent with what was previously found: intrinsic prediction yields better measurements than extrinsic predictions. What is more, as can be seen from the last row, the extrinsic prediction based on human nouns yields better measurements than the one based on HQA. This would imply that the semantic effects of NHQA are more similar to that of human nouns than to that of HQA.

Results for rank measurements are listed in 21b. Surprisingly, the comparison of the intrinsic predictions of NHQA to the extrinsic predictions based on human nouns does not bring about a statistically significant difference in the case of this measure. It is however important to note that raw rank measurements throughout the previous experiments have proven to be not as reliable as Euclidean distance or log-scaled rank, and that the adjusted p-value in this case does not reflect a negative conclusion, but rather a lack of conclusive evidence. A significant difference can nonetheless be observed when comparing the intrinsic prediction to the extrinsic HQA based prediction, highlighting once again that the semantic effects of these two processes differ. The last row of this table underscores that the comparison of the two extrinsic predictions highlights that the semantic effects of human nouns g-gender alternation differ less than that of HQA, consistently with what was found for distance measurements.

Finally, the log-scaled measurements are shown to be consistent with the previous measurements in 21c. The intrinsic prediction is systematically better than any of the extrinsic predictions. Moreover, a statistic difference can be shown when comparing the two extrinsic predictions, which would imply once more that the human noun based extrinsic prediction deviates less from the extrinsic prediction than the HQA based extrinsic prediction.

Pred. compared		diff.	lower	upper	adj. p-value
intr. pred.	noun-based	$-0.343\,83$	$-0.375\,84$	$-0.311\,82$	$0.000\,00$
intr. pred.	HQA based	$-0.414\,43$	$-0.446\,44$	$-0.382\,43$	$0.000\,00$
noun-based	HQA based	$-0.070\,61$	$-0.102\,62$	$-0.038\,60$	$1.100\,00 \cdot 10^{-06}$

(a) Tukey HSD test results for distance measurements

Sets compared		diff.	lower	upper	adj. p-value
intr. pred.	noun-based	$-25\,978.380\,00$	$-65\,958.980\,00$	$14\,002.220\,00$	$0.278\,11$
intr. pred.	HQA based	$-93\,616.310\,00$	$-133\,596.910\,00$	$-53\,635.710\,00$	$0.000\,00$
noun-based	HQA based	$-67\,637.930\,00$	$-107\,618.530\,00$	$-27\,657.330\,00$	$0.000\,25$

(b) Tukey HSD test results for rank measurements

Sets compared		diff.	lower	upper	adj. p-value
intr. pred.	noun-based	$-5.495\,77$	$-6.263\,94$	$-4.727\,60$	$0.000\,00$
intr. pred.	HQA based	$-7.241\,53$	$-8.009\,70$	$-6.473\,35$	$0.000\,00$
noun-based	HQA based	$-1.745\,75$	$-2.513\,93$	$-0.977\,58$	$5.000\,00 \cdot 10^{-07}$

(c) Tukey HSD test results for log rank measurements

Table 21: Tukey HSD tests results for measurements extrinsic prediction of NHQAs in the $\mathcal{M}_3$ model

5.5.3 Discussion

The hypotheses that this experiment tested were that HQA g-gender alternation was more similar to human nouns g-gender alternation than NHQA, and that NHQA pairs exhibited more regular shifts than HQA pairs.

The comparison of the intrinsic predictions seems to confirm this second point: a statistically significant difference amongst the two processes was found, showing that NHQA would yield lower distance measurements, and lower log-scaled rank measurements than HQA. It is however noteworthy that the actual difference between the measurements for two processes is very small, highlighting that the distinction it introduces is very nuanced.

Another point of interest is that human noun pairs exhibit more regularity than HQA, but less than NHQA. In that respect, the criterion of regularity is as much satisfied when it comes to human nouns than when it comes to some sub-classes of adjectives. This strengthens the claim that the assumptions that regularity distinguishes inflection from derivation on the one hand, and that human nouns g-gender alternation is derivational on the other contradict each other. Removing the criterion of regularity from the criteria listed in Stump (1998) may however prove to be more adequate than the alternative: as Bonami and Boyé (forthcoming) highlighted, treating human nouns g-gender alternation as inflectional opens the debate of paradigm uniformity.

The higher degree of regularity of human nouns can be explained from at least two facts, which do not necessary exclude each other. The first would be mathematical in nature : HQA pairs, in this experiment, were more numerous than human nouns and NHQA. As a greater number of observations leads to a greater number of outliers, it is possible that redoing the experiment based on a smaller subset of HQA might bring forth different conclusions. However the analysis of pairwise distances conducted prior to this experiment (*cf.* Table 13) showed that the difference between HQA and human nouns was negligible compared to the the full sets of NHQA. Therefore,it is not guaranteed that result would vary. A second tentative explanation would be that it may be adjectives, rather than human nouns, that convey most of the social behaviour of the speaker. As human nouns often refer to objective facts — for instance, that the person the speaker is mentioning be a cashier, a president, or a German —, their use is to a certain extent constrained by the very message that the speaker utters. On the other hand, qualifying adjectives are *in fine* chosen by the speaker — she or he has the freedom to judge said cashier, or president, or German, either gifted, smart, handsome, or other. This degree of freedom in the possible way to describe someone would be where social factors, like s-gender, might come into play — thus leading to more variation amongst adjectives. This second explanation however cannot be tested as directly as the first.

Other than the relative degrees of regularity of HQA and NHQA, one question that was addressed by this experiment regarded the relative degrees of similarity of the semantic effects of the three processes.

The extrinsic predictions have shown that the semantic effects of HQA g-gender alternation were more similar to that of human nouns than the semantic effects of NHQA g-gender alternation, and conversely that the semantic effects of human nouns g-gender alternation were more similar to that of HQA g-gender alternation than the semantic effects of NHQA g-gender alternation. In that respect, human nouns g-gender alternation and HQA g-gender alternation form a more coherent group, and NHQA g-gender alternation differs from these.

Another interesting fact is that the semantic effects of human nouns g-gender alternation are shown to be more similar to that of NHQA g-gender alternation than the semantic effects of HQA g-gender alternation. This questions a point that was mentioned previously: while discussing the results of the fifth experiment, it was suggested that the attested difference between HQA g-gender alternation and human nouns g-gender alternation might be related to the distinction of inherent and contextual inflection, as explained in Booij (1995). This however is not consistent with what was measured for NHQA g-gender alternation in the $\mathcal{M}_3$ model: adjectival g-gender alternation is by definition contextual inflection, however the semantic effects of NHQA g-gender alternation — which is contextual — were shown to be more similar to the semantic effects of nominal g-gender alternation — which would arguably be inherent — than to the semantic effects of HQA g-gender alternation — which is contextual. Therefore the distinction between these processes cannot be entirely summarized by a distinction between inherent and contextual inflectional processes.

In brief, this sixth experiment has brought factual quantitative observations questioning the nature of semantic regularity and its validity as a criterion to tease apart inflection and derivation. Sub-regularities within an inflectional process can lead to different behaviours for different sub-groups: in this particular example, NHQA and HQA are in many cases more different from one another than they are from human nouns.

5.6 Partial conclusions

Regardless of their actual results, the experiments presented in this section have also highlighted how DSM could provide insights on morphological processes. In particular, the methodology defined here is not specific to the exact morphological processes under study: although the fourth experiment was more specifically tailored to study g-gender alternation in adjectives and human nouns, the third, fifth and sixth experiments presented a general method to compare morphological processes with one another, both in terms of regularity — which was studied through the comparison of intrinsic predictions — and in terms of semantic effects — for which extrinsic predictions proved to be insighful.

One general caveat concerns the measures used to inspect these predictions. Rank measurements have proven significantly less adequate than distance measurements: rank often fails to capture distinctions that Euclidean distance can show. As was mentioned previously, it is conceivable that word vectors are grouped in clusters: in which case the accuracy of rank measurement would also be dependent on the position of the target relative to the rest of its cluster. Log-scaling rank measurements was shown to yield contrasted results. In the third and fourth experiments where human nouns were compared respectively to adjectives in general and to HQA, log-scaled measurements did not add anything to what could be shown through distance. On the other hand, in the sixth experiment, where human nouns, HQA and NHQA were simultaneously studied, log-scaled measurements were able to provide more fine-grained distinctions than what distance measurements alone highlighted. In that respect, studying the general layout of vectors across the DSM constitues a topic of research which lies outside the scope of the present study, but would certainly bring about results of crucial importance to the methodology presented here.

This methodology, applied to human noun and adjective g-gender alternation, brought about results with rather far-reaching consequences, as adjective g-gender alternation is taken to be a typically inflectional process. Comparing human nouns g-gender alternation to that of adjectives yields conclusions less univoqual than what the first two experiments hinted at. The first and second experiments underscored a clean distinction between human nouns g-gender alternation and deverbal agent noun derivation: the alternation between g-genders was shown to be much more regular than agent noun derivation. The third experiment, which compared human nouns g-gender alternation to that of adjectives in general, highlighted a similar divide between adjectives and human nouns: adjectives were shown to be more regular than nouns.

However, the subsequent experiments have shown that this apparent difference did not cover all the facts. The fourth experiment showed that there existed subregularities in adjectival g-gender alternation:

the more frequently an adjective pair was used to qualify a human nouns, the more its shift was similar to the mean shift for human nouns pairs. The fifth experiment underscored how regularity for the whole of adjectives did not imply regularity for each subclass of adjectives: when comparing human nouns and human qualifying adjectives, no difference in terms of regularity could be found. The sixth experiment even showed that there could be more variation of regularity within the category of adjectives than when comparing HQA to human nouns.

These findings concern regularity. As semantic regularity is linked to the semantic effects of a process, they also apply to the description of the semantic effects of g-gender alternation in human nouns and in adjectives. In particular, results from the fourth and sixth experiments clearly indicate how, even in the case of inflection, semantic effects are not completely defined by the process itself. The fourth experiment highlighted that there existed a class of adjectives whose semantic effects strayed off from what was expected for adjectives. The sixth experiment stressed that the semantic effects of that class of adjectives were more similar to the semantic effects of human nouns g-gender alternation than to NHQA. In the light of these findings, it is difficult to assume both.

Inflection does not exhibit clockwork regularity, and the semantic regularity criterion from Stump (1998) does not take into account the wide range of semantic effects that can appear in a single group. These experiments on adjective g-gender alternation also strongly suggest that either nouns ought to be treated as inflectional, or that inflection is not necessarily more regular than derivation. The consequences of either choice have already been explained: either French human nouns constitute a case of paradigm non-uniformity and the French g-gender system ought to be reanalysed, or the distinction between inflection and derivation ought to be revised.

6 General Conclusions

Perhaps the most important point of this study is that it provides strong evidence suggesting human nouns g-gender alternation is, in fact, inflectional, *id est* that the inflectional analysis of g-gender alternation (*cf.* subsection 2.2) advocated by Bonami and Boyé (forthcoming) is correlated by objective facts and the derivational analysis is inconsistent with the data at hand. Section 5 clearly concludes that, although a difference in terms of regularity can be attested when comparing human nouns to adjectives as a whole (*cf.* subsection 5.2), this difference vanishes when comparing them to human-qualifying adjectives (see subsection 5.4, subsection 5.5). On the other hand, in section 4, human noun g-gender alternation was shown to be more regular than deverbal agent noun derivation using the methodology of Bonami and Paperno (forthcoming), and this distinction can be systematically made as long as statistical noise is taken into account (subsection 4.2).

From this it stems that human noun g-gender alternation is more regular than deverbal agent nouns, as regular as some adjectives' g-gender alternation, but less regular than adjectives as a whole. This conclusion highlights two facts: one, that derivation and inflection most certainly form a continuum — it has been shown here from the perspective of regularity —, two, that the semantic regularity of inflectional processes is to be nuanced, as different classes of adjectives attribute different effects to g-gender alternation (subsection 5.3, subsection 5.5).

Moreover, this study also proposed a methodology to compare morphological processes in DSMs with one another (*cf.* subsection 5.1), as well as a way to use semantic networks to automatically compute a lexicon of human nouns (subsection 3.4). Both the methodology and the lexicon extraction algorithm can be easily adapted to other studies: subsection 5.2, subsection 5.4 and subsection 5.5 illustrate how to adapt the methodology to different sample sizes and number of processes studied. The human noun lexicon was computed using a set of manually selected seeds empirically chosen so that they refer to humans in the context of a definition; therefore, by selecting other seeds referring to other semantic categories, other lexica can be extracted using the same algorithm.

The various conclusions stemming from this study all rely on accepting the distributional hypothesis as true; moreover, as was discussed in section 2, they are not free from theoretical consequences. This study was concerned with three questions :

- Is distributional semantics a sound model of theoretically abstruse morphological processes?
- Is the grammatical gender alternation of French human noun more akin to inflection or derivation?
- What role does social gender play in grammatical gender assignment in French?

The various experiments conducted in sections 4 and 5 provided some elements to answer these three questions.

As was discussed in section 5, distributional semantics models have been able to highlight subtle differences between the g-gender alternation in human nouns and in adjectives In particular, the experiments underscored how fit distributional semantics models were to study different sub-classes within a process. These models also proved how semantically dissimilar human nouns and adjective g-gender alternations were — although the two processes are extremely similar from a formal point of view, their usage are clearly different.

Another relevant point concerns the different methodologies used throughout this study. This study only used vector shift to represent morphological processes. However, more sophisticated approaches exist — including, for instance, functional representations, *id est* representations of composition by means of multilinear functions. Therefore, the possible methodologies to research morphology within a distributional semantics frameworks were far from exhaustively explored in this limited study. In all, what matters chiefly is the capacity of distributional semantics to mathematically compute similarity and difference between words. Assigning a mathematically defined similarity allows for an objective, data-driven research to be computed. In that respect, distributional semantics is well fitted to studying complex morphological phenomena, as they provide means of quantifying them.

Aside from methodological considerations, the experiments conducted in this study also highlighted how difficult it was to classify g-gender alternation of human nouns as inflectional or derivational. The first experiment — which compared g-gender alternation in agent nouns to deverbal agent noun formation — and the third experiment — which studied the g-gender alternation of adjectives as a whole — indicate that human nouns g-gender alternation seems to lie somewhere in between inflection and derivation. However,

the assumption that inflection exhibit clockwork regularity was challenged by subsequent experiments: human nouns were shown to be more similar to some adjective classes than others, and adjective g-gender alternation was shown to be subjected to subregularities. In all, the experiments seem to indicate that inflection and derivation are not two distinct categories, but rather the two extrema of a continuum, ranging from extremely regular morphological processes to loosely related words. On this continuum, human nouns g-gender alternation would be rather close to g-gender alternation in human qualifying adjectives, but less regular than that of adjectives qualifying non-human nouns.

In that respect, an important finding of this study is that the criteria of Stump (1998) are not consistent with the traditional assumption of treating French human nouns g-gender alternation as derivational. The criterion of semantic regularity being challenged by the observations made here implies that perhaps the inflection-derivation distinction is closer to what Štekauer (2014) advocates. On the other hand, should the inflection-derivation distinction in its current form of two cleanly divided categories of phenomena hold, then it would be necessary to treat human g-gender alternation as inflection (or to reject the g-gender alternation of some adjectives as inflectional). This in turn would question what we know of grammatical gender: the traditional definition of g-gender as a noun classifying system would need to take into account that some French nominal lexemes can be feminine or masculine, depending on the speaker's choice. To sum up, the present study has stressed how complex this issue is: the morphological status of g-gender alternation in human nouns is linked to several debates, concerning for instance the status of g-gender or the nature of the inflection-derivation distinction. The present study provided some objective facts to the debate, but is by no means sufficient in and of itself to resolve it.

Another aspect that complicates this question is that, in this case, g-gender is linked to s-gender. As for the role of s-gender on g-gender assignment in French, two different observations can be made from this study. First, the s-gender of human nouns referents tend to be reflected in the g-gender of the human noun — although a masculine noun may sometimes refer to a woman, within pairs of morphologically-related human nouns, the feminine variant cannot refer to a man. As such, if g-gender alternation of human nouns is deemed to be inflectional, it therefore leads to an important case of paradigm non-uniformity. Second, in the case of adjectives, semantic effects of g-gender alternation are influenced by whether the g-gender and s-gender of the referent are aligned: less regularity can be found for adjectives modifying human nouns. From this, one can surmise that external social factors influence the choice of what adjective to pick to describe a human referent.

To conclude on these theoretical considerations, this study has succesfully shown that the framework of distributional semantics is fit to study morphological processes and to compare them with another; it has stressed that some of the theorical assumptions traditionally adopted were not consistent with objective observations, and it highlights the complexity of the relation between s-gender and g-gender. This study is in itself incomplete, and calls for further research on this domain.

References

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli (2014). "Frege in Space: A Program for Compositional Distributional Semantics Marco Baroni,1 Raffaella Bernardi1 and Roberto Zamparelli". In: *LiLT Volume 9 Perspectives on Semantic Representations for Textual Inference*.

Baroni, Marco et al. (2009). "The WaCky Wide Web:A Collection of Very Large Linguistically Processed Web-Crawled Corpora". In: *Language Resources and Evaluation*. DOI: 10.1007/s10579-009-9081-4. URL: https://doi.org/10.1007/s10579-009-9081-4.

Baroni, Marco et al. (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bolukbasi, Tolga et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., pp. 4349–4357. URL: http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

Bonami, Olivier and Gilles Boyé (forthcoming). "Paradigm uniformity and the French gender system". In: *Morphological Perspectives*. Ed. by Matthew Baerman, Oliver Bond, and Andrew Hippisley. Edinburgh University Press. Forthcoming.

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

Booij, Geert (1995). "Yearbook of Morphology 1995". In: *Yearbook of Morphology 1995 G.E. Booij and J.v. Marle, Editors, 1 - 16 (1995)*. Chap. Inherent versus contextual inflection and the split morphology hypothesis, pp. 1–16.

Burnett, Heather and Olivier Bonami (submitted). "Linguistic Prescription, Ideological Structure and the Actuation of Linguistic Changes: Grammatical Gender in French Parliamentary Debates". In: submitted.

Coavoux, Maximin (2017). "Discontinuous Constituency Parsing of Morphologically Rich Languages". PhD thesis. Univ Paris Diderot, Sorbonne Paris Cité.

Corbett, Greville G. (1991). *Gender*. Cambridge: Cambridge University Press.

Eckert, Penelope and Sally McConnel-Ginnet (2003). *Language and Gender*. Ed. by Cambridge. Cambridge University Press.

F. Hockett, Charles (1958). "A Course in Modern Linguistics". In: 8.

Faaß, G. and K. Eckart (2013). "SdeWaC - A Corpus of Parsable Sentences from the Web." In: *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.

Fabre, Cécile, 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.

Faruqui, Manaal et al. (2016). "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks." In: *RepEval@ACL*. Association for Computational Linguistics, pp. 30–35. ISBN: 978-1-945626-14-2. URL: http://dblp.uni-trier.de/db/conf/repeval/repeval2016.html#FaruquiTRD16.

Firth, J.R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press. URL: https://books.google.fr/books?id=jDu3AAAAIAAJ.

Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum (2007). "Topics in semantic representation." In: *Psychological review* 114 2, pp. 211–44.

Hathout, Nabil and Franck Sajous (2016). "Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Hathout, Nabil, Franck Sajous, and Basilio Calderone (2014a). "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*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 65–74. URL: http://www.aclweb.org/anthology/W14-5809.

Hathout, Nabil, Franck Sajous, and Basilio Calderone (2014b). "GLÀFF, a Large Versatile French Lexicon". English. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland. ISBN: 978-2-9517408-8-4.

Herbelot, Aurélie and Eva Maria Vecchi (2015). "Building a shared world: mapping distributional to model-theoretic semantic spaces". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 22–32. DOI: 10.18653/v1/D15-1003. URL: http://www.aclweb.org/anthology/D15-1003.

Kiela, Douwe and Stephen Clark (2015). "Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception". In: *EMNLP*.

Kruszewski, German, Denis Paperno, and Marco Baroni (2015). "Deriving Boolean structures from distributional vectors". In: *Transactions of the Association for Computational Linguistics, vol. 3*.

Kruszewski, Germán et al. (2016). "There is No Logical Negation Here, but There Are Alternatives: Modeling Conversational Negation with Distributional Semantics". In: *Comput. Linguist.* 42.4, pp. 637–660. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00262. URL: https://doi.org/10.1162/COLI_a_00262.

Landauer, Thomas K and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". In: *Psychological Review 1997, Vol. 104*.

Lazaridou, Angeliki, Elia Bruni, and Marco Baroni (2014). "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world". In: *ACL*.

Lenci, Alessandro (2008). "Distributional semantics in linguistic and cognitive research". In: *talian Journal of Linguistics, 20*.

Louwerse, Max M. and Nick Benesh (2012). "Representing Spatial Structure Through Maps and Language: Lord of the Rings Encodes the Spatial Structure of Middle Earth". In: *Cognitive Science* 36.8, pp. 1556–1569. DOI: 10.1111/cogs.12000.

Louwerse, Max M. and Rolf A. Zwaan (2009). "Language Encodes Geographical Information". In: *Cognitive Science 33 (2009) 51–73*.

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

Michelson, Karin (2015). "Gender Across Languages: Volume 4". In: *Gender Across Languages: Volume 4*. Chap. Gender in Oneida, pp. 277–301.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations." In: *HLT-NAACL*, pp. 746–751.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781.

Mikolov, Tomas et al. (2017). "Advances in Pre-Training Distributed Word Representations". arXiv preprint arXiv:1706.00286 (2017.

Mitchell, Jeff and Mirella Lapata (2008). "Vector-based models of semantic composition". In: *In Proceedings of ACL-08: HLT*, pp. 236–244.

Muller, Philippe, Nabil Hathout, and Bruno Gaume (2006). "Synonym Extraction Using a Semantic Distance on a Dictionary". In: *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. TextGraphs-1. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 65–72. URL: http://dl.acm.org/citation.cfm?id=1654758.1654773.

Paperno, Denis et al. (2014). "A practical and linguistically-motivated approach to compositional distributional semantics". In: *ACL*.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Quine, William Van Ormann (1960). *Word And Object*. MIT Press.

Roller, Stephen, Katrin Erk, and Gemma Boleda (2014). "Inclusive yet selective: Supervised distributional hypernymy detection". English (US). In: *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*. Association for Computational Linguistics, ACL Anthology, pp. 1025–1036.

Sajous, Franck and Nabil Hathout (2015). "GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary". English. In: *Proceedings of the of the eLex 2015 conference*. Herstmonceux, England, pp. 405–426.

Sajous, Franck, Nabil Hathout, and Basilio Calderone (2013). "GLÁFF, un Gros Lexique Á tout Faire du Français". In: *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*. Les Sables d'Olonne, France, pp. 285–298.

— (2014). "Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif." In: *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*. Berlin, pp. 663–680. DOI: http://dx.doi.org/10.1051/shsconf/20140801216.

Santus, Enrico et al. (2014). "Chasing Hypernyms in Vector Spaces with Entropy". In: *EACL2014-SP 2014*.

Santus, Enrico et al. (2016). "ROOT13: Spotting Hypernyms, Co-Hyponyms and Randoms". In: *Association for the Advancement of Artificial Intelligence*.

Schnabel, Tobias et al. (2015). "Evaluation methods for unsupervised word embeddings." In: *EMNLP*. Ed. by Lluís Màrquez et al. The Association for Computational Linguistics, pp. 298–307. ISBN: 978-1-941643-32-7. URL: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2015.html#SchnabelLMJ15.

Schäfer, Roland (2015). "Processing and querying large web corpora with the COW14 architecture". In: *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Ed. by Piotr Bański et al. UCREL. Lancaster: IDS. URL: http://rolandschaefer.net/?p=749.

Štekauer, P. (2014). "The Oxford Handbook of Derivational Morphology". In: *The Oxford Handbook of Derivational Morphology . Oxford: Oxford University Press*. Ed. by R. Lieber and P. Štekauer. Oxford University Press. Chap. Derivational Paradigms, pp. 354–369.

Strnadová, Jana (2014). "Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français". PhD thesis. Université Paris Diderot et Univerzita Karlova V Praze.

Stump, Gregory (1998). "The handbook of morphology". In: ed. by A. Spencer & A. M. Zwicky. Oxford: Blackwell. Chap. Inflection. Pp. 13–43.

Varvara, Rossella (2017). "Verbs as nouns: empirical investigations on event-denoting nominalizations". PhD thesis. Università degli Studi di Trento.

Wauquier, Marine (2017). "Explorations méthodologiques pour la caractérisation distributionnelle de dérivés morphologiques". MA thesis. Université Toulouse Jean Jaurès.

Wittgenstein, Ludwig (1921). *Tractatus Logico-Philosophicus*. Ed. by Wilhelm Ostwald. Annalen der Naturphilosophie, 14.

Zwanenburg, Wiecher (1988). "Aspects de linguistique française." In: *Yearbook of Morphology 1995 G.E. Booij and J.v. Marle, Editors, 1 - 16 (1995)*. Chap. Flexion et dérivation: le féminin en français, 191–208.