

Idiosyncratic Frequency as a Measure of Derivation vs. Inflection

*Maria Copot*¹, *Timothee Mickus*², and *Olivier Bonami*¹

¹ Université Paris Cité, Laboratoire de linguistique formelle, CNRS

² University of Helsinki

Final version as accepted on October 2, 2022

ABSTRACT

There is ongoing discussion about how to conceptualize the nature of the distinction between inflection and derivation. A common approach relies on qualitative differences in the semantic relationship between inflectionally versus derivationally related words: inflection yields ways to discuss the same concept in different syntactic contexts, while derivation gives rise to words for related concepts. This differential can be expected to manifest in the predictability of word frequency between words that are related derivationally or inflectionally: predicting the token frequency of a word based on information about its base form or about related words should be easier when the two words are in an inflectional relationship, rather than a derivational one. We compare prediction error magnitude for statistical models of token frequency based on distributional and frequency information of inflectionally or derivationally related words in French. The results conform to expectations: it is easier to predict the frequency of a word from properties of an inflectionally-related word than from those of a derivationally-related word. Prediction error provides a quantitative, continuous method to explore differences between individual processes and differences yielded by employing different predicting information, which in turn can be used to draw conclusions about the nature and manifestation of the inflection–derivation distinction.

Keywords:
Morphology,
Derivation–
Inflection
gradient,
Distributional
Semantics

The theoretical distinction between inflection and derivation is well-defined on paper (Matthews 1991): inflection outputs different forms of the same lexeme (*read, reads, reading*), while derivation outputs related lexemes (*read, reader, readable*). Empirically grounding this binary distinction has however proved challenging. Linguists often have strong intuitions about whether a process is inflectional or derivational, but there is no single criterion that reliably distinguishes the two (Stump 1998). In fact, the distinction appears much more akin to a gradient with two poles (see e.g. Bybee 1985; Dressler 1989). Inflection and derivation both seem to be characterized by loose clusters of features—features that co-occur frequently, but not systematically. This gradient nature suggests that the inflection–derivation distinction ought to be studied from a quantitative and empirical perspective, which is the aim of the present paper.

The theoretical distinction we stated above can be leveraged to make quantitative predictions over different morphological processes. If inflection provides the means of using the same lexeme in different contexts, we can expect that words in inflectional relationships should have stronger relationships of interpredictability. What changes when we use a conjugated verb form or a plural instead of a singular noun is not the concept we wish to name, but merely the syntactic and semantic context in which the word is being employed. On the other hand, derivation is used to fill onomasiological needs (Štekauer 2005): a derived word typically arises because a language user is trying to name a new concept by building on an existing and related word. Because of the imperfect correspondence between language and reality, one cannot assume that there will be a perfect match between the derived meaning and the expectations set by the morphology used to derive it. Relatedly, derived words are expected to have independent lexical representation and hence may over time acquire senses or usages that deviate from those of their base. As a consequence, we expect derivationally related words to have patterns of usage that differ in unpredictable ways—making it in turn harder to predict information pertaining to a word given a derivationally related term. While lexi-

calized differences in usage are also attested for inflectionally related words, one can expect them to be much rarer.

Is this variation in patterns of usage across the inflection–derivation gradient a phenomenon that we can quantify empirically? To do so, we first need to decide how to measure differences in patterns of usage. One approach uses distributional representations as derived from word embedding algorithms (Bonami and Paperno 2018). How accurately one can reconstruct the distributional representation of some target word informs us of whether the input we used is predictive of this target’s patterns of usage. This, in turn, allows one to contrast and compare pairs of morphologically related words depending on where they sit on the inflection–derivation gradient: words in a derivational relation should be less predictive of one another’s patterns of usage, and we should expect the reconstruction to be less accurate. Yet, the sheer diversity of existing architectures and the inherent noisiness of the methods used to derive them raise concerns. Reconstructing a word embedding is tantamount to assuming that the corresponding embedding architecture accurately captures all the relevant distributional characteristics. In the absence of an independent measure of predictability that is both fine-grained enough and applicable at scale, we have no way of establishing that this assumption is warranted. It is therefore relevant to look for other means of characterizing a word’s patterns of usage.

In this paper we focus on frequency as a well-understood, easily obtainable and holistic correlate of word usage, that is known to be relevant to morphological relatedness; for instance, derived words tend to be lower frequency than their bases (Harwood and Wright 1956; Hay 2001), a fact that can be exploited to help establish direction of derivation (Kisselew *et al.* 2016). Two pairs of words that relate to each other in a parallel way should have distributions that contrast in the same way, and hence their frequencies of usage should be related by the same conversion factor. For instance, we expect that the frequency ratio between *quicker* and *quick* be very similar to that between *brighter* and *bright*. On the other hand, where identity of morphological marking does not mean identity of semantic contrast, we have no such expectations. For instance, we would not be surprised if the semantic ratio between *driver* and *drive* is very different from that between *diner* and *dine*. To measure how reliably a given process

causes an identical shift in usage for different lexemes, we measure the variability in frequency ratios between pairs of words linked by the same process: derivationally related words should show higher variation in frequency ratios.

The remainder of this paper is structured as follows: in Section 2, we review the theoretical elements underlying our approach. In particular, we discuss the derivation–inflection gradient in Sections 2.1 and 2.2, and the interface between quantitative morphology and distributional semantics in Sections 2.3 and 2.4. Section 3 outlines the experimental protocol: we train separate linear models for several morphological processes, predicting the frequency of a form in the target cell from various types of information. Section 4 reports the results of two comparable experiments on datasets of different sizes. We finish up with a general summary of our findings and future perspectives for this work in Section 5.

2 THEORETICAL BACKGROUND

2.1 *The Derivation–Inflection Gradient*

The key naive distinction between inflection and derivation is intuitive and easy to grasp: inflection yields forms for talking about the same concept in different syntactic contexts (*I read~she reads*), while derivation yields forms for talking about different but related concepts (*I read~a reader*). Based on such observations, Anderson (1982, 1992) suggests that relevance to syntax is the only criterion necessary to distinguish inflection from derivation. Such a strict, binary categorisation hinging upon a single criterion quickly proves indefensible (Booij 1996). Some inflection is strictly *contextual*, in the sense that the choice of an inflected form is strictly dictated by the syntactic context: this is true, most prominently, of variation in agreement morphology and case. However morphological distinctions within the traditional purview of inflection can also be *inherent*, in the sense that it is the expression of some content. This is the case, for instance, for number

on nouns, or most TAM (tense–aspect–mood) distinctions on verbs.¹ Inherent inflection can thus be semantically potent and irrelevant to syntax: for instance, in many languages, whether a verb is future or past will have no syntactic consequences.

Systematically distinguishing inflection and derivation is thus not a straightforward matter of division of labor between syntax and semantics. Hence linguists have explored many other possible criteria. Bybee (1985) proposes obligatoriness of expression, degree of semantic change to the word, range of applicability; Payne (1985) proposes 8 criteria, among which a variation of Bybee’s, along with additions like presence or absence of category change; Plank (1991) highlights 28 criteria that distinguish at least some cases of inflection and derivation, noting that none of these is either necessary or sufficient to characterize the distinction, but instead these criteria are better conceived of as prototypical properties of two extremes of a gradient.

The conceptualisation of the inflection-derivation distinction is of importance beyond theoretical morphology. Take as an example the use of morphological language data in computational linguistics: large resources such as UniMorph (Kirov *et al.* 2016, 2018; McCarthy *et al.* 2020) have been extensively used to make typological generalisations about the world’s languages, to test linguistic hypotheses on a diverse language sample, and to evaluate the performance of language processing models, amongst other things. Decisions made about the UniMorph tagset and the possible shape of the UniMorph paradigms are dependent on decisions made by editors of the Wiktionary pages for the languages in the resource – deciding where to draw the line between inflection and derivation (or whether to draw a line at all) for an individual language has cascading consequences on all of the uses made of data from Unimorph. For a concrete example, take Malouf *et al.* (2020): contrary to the observation that Navajo noun morphology is fairly straightforward, they find that their method flags Navajo noun paradigms as being particularly unpredictable. This is the outcome of the same paradigmatic pattern being treated as derivational for one class of nouns (and therefore worthy of multiple entries in the

¹ The extent to which phenomena such as sequence of tense and mood selection should be considered contextual or inherent is a fascinating but understudied topic.

dictionary for each set of related items) and as inflectional for a different class (and therefore each set of related items was reported in the same dictionary entry). Insights about the nature of the inflection-derivation distinction could have important consequences for all applications relying on morphological data.

The question of how to distinguish inflection and derivation is a live one (see Spencer 2013 for a recent overview), but few qualitative advances have been made in identifying reliable criteria for distinction since the issue first captured the attention of the field. There is growing agreement that inflection and derivation cannot be characterized as dichotomous or otherwise categorical, and that relatedness between words is a multifactorial and gradient matter (Dressler 1989; Booij 1996; Haspelmath 1996; Bauer 2004; Corbett 2010; Spencer 2013; Štekauer 2015), with some studies arguing that the distinction does not apply in the same way across languages (Bauer and Bauer 2012) or is plainly irrelevant (e.g. Bochner 1993; Ford *et al.* 1997; Haspelmath forthcoming).

There are plenty of morphological processes that don't behave in a typical inflectional nor derivational manner, no matter what specific set of criteria are chosen to characterize the distinction. English noun pluralization is one of many examples that could illustrate this (see among many others Acquaviva 2008; Corbett 2019 for a discussion of its properties). It looks inflectional in many respects: it is a syntactic requirement that plural marking be employed when talking about an entity in a plural syntactic context (*one car~two cars/*two car*), and the resulting semantics are generally straightforwardly compositional. However, it can also behave more derivationally: the entity denoted by the plural form may be a different concept compared to that denoted by the singular form (*spectacle = a show; spectacles = glasses* is an extreme example, but milder cases exist too, such as *practice~practices*, where the singular can denote a habit or the act of practicing a profession, while the plural can mainly denote the habit), and plural marking may not carry plural semantics (*a pair of scissors*). English noun pluralization is not unique in seemingly straddling the inflection-derivation boundary, and an account of the distinction between the two must be informative about such cases.

Continuum approaches to inflection and derivation in
quantitative morphology

2.2

The approaches to the inflection–derivation gradient listed above rely on clustering of dichotomous criteria rather than on a quantitative approach to the difference: in these approaches, a process is considered more inflection-like than another if it ticks more of the boxes of binary criteria that characterize inflection. There is a dearth of attempts to find continuous criteria that characterize the entirety of the gradient.

The quest for such a characterization of the inflection–derivation continuum is a good fit for quantitative paradigmatic approaches to morphology. We adopt Bonami and Strnadová’s (2019) conceptualization of a paradigmatic system as a collection of content-aligned sets of words that instantiate parallel morphological relationships. This is illustrated in Figure 1 with a slice of the paradigm structure of English morphology: morphological families of words are represented on horizontal planes that are aligned based on the content-based contrasts they share. In other words, a paradigmatic system is a set of interpredictability relationships² of form and meaning between words of a language, while an individual paradigm is a morphological family that is structured by a subset of these relationships.

Let us take a closer look at how paradigms can be established under such an operationalization. Two words can be said to be in a morphological relationship if they instantiate a form-meaning correspondence which is also instantiated by other word pairs in the language. So *cake* and *cakes* are in a morphological relationship: their meaning relationship of *one of X ~ more than one of X* is instantiated by the same formal means *X ~ Xs* in other pairs of words in the language such as *squirrel ~ squirrels* or *squid ~ squids*. The pair *foot ~ feet* does not instantiate the same morphological relationship: it shares a content relationship with the words above but not a relationship of form. The two

²A reviewer points out that in the morphological literature, *predictability* mostly refers to the amount of information about a form provided by a related one (see e.g. Ackerman *et al.* 2009; Stump and Finkel 2013). Here, we use *predictability* and *interpredictability* in the broader, statistical sense: it refers to the amount of information provided on a word’s form, meaning, usage, and other characteristics by information about a related word.

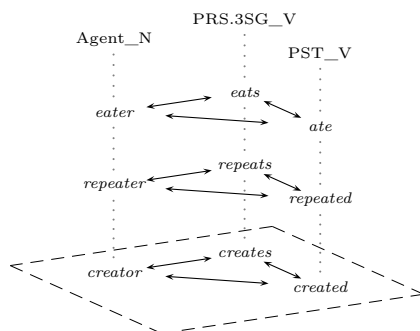


Figure 1: A subset of the paradigmatic structure of English

words are nevertheless in a morphological relationship: their content relationship is instantiated by the same formal means in word pairs like *tooth~teeth*. On the other hand, word pairs like *shingle* (a mass of rounded pebbles) ~ *shingles* (an illness) do not instantiate a morphological relationship: they share a formal relationship with the word pairs above, but there are no other word pairs in the English language with this same form relationship that also share a parallel content relationship. Morphological relationships can also be found within the realm of derivation: *sing* and *singer* have the same relationship of form and meaning as pairs like *read~reader* and *help~helper*. It's important to note that morphological relationships describe systematic patterns in a way that doesn't reify the traditional inflection–derivation distinction: *(she) sings* and *singer* are in a morphological relationship too, the same one which is instantiated by *(she) reads* and *reader*.

Sets of morphologically related words that share a conceptual core are known as *morphological families* (Schreuder and Baayen 1997): *read, reads, reader* constitutes a morphological family, as does *emote, emotion, emotional*. Because morphological relationships are not inherently inflectional nor derivational, morphological families will group together words that are in both traditionally inflectional and traditionally derivational morphological relationships, as well as any type of relationship in between the two extremes.

Paradigmatic structure emerges when morphological families whose members have parallel content relationships are aligned. Under this

particular definition, paradigmatic structure is closely linked to relationships of interpredictability between words, which are exploited by speakers when producing and processing language. If a speaker has knowledge of a partial morphological family and how it fits within the paradigmatic system of the language, they may exploit proportional analogy and probabilistic mapping to generate a new member of said morphological family (Ackerman *et al.* 2009). Knowing that *repeat* (PRS) has a past tense *repeated* will allow a speaker to induce *disembogued* as the past tense of a present form *disembogue*. Encountering the form (*she*) *absquatulated* will likely lead a speaker to identify it as a past tense with a hypothetical present form *absquatulate*, by analogy with the structure established by the previous forms. These relationships of predictability may include morphological relations placed along all parts of the traditional inflection–derivation gradient. The theory makes no assumptions about the reification of such a distinction: as long as there is partial interpredictability of form and meaning, there is paradigmatic structure. As exemplified in Bonami and Strnadová (2019), the probabilistic nature of paradigm structure lends itself well to be investigated with quantitative methods.

Quantitative morphology, frequency and semantics

2.3

The predictability-based view of paradigm structure outlined above invites us to explore explicitly quantitative reflexes of the inflection–derivation continuum. One proposal in that direction is Bonami and Paperno (2018), who use distributional methods to operationalize the idea that inflection relates words in a more semantically transparent fashion than derivation (see e.g. Dressler 1989, 5). Another is Rosa and Žabokrtský (2019), who focus on the idea that word pairs related by inflection tend to be distributionally more similar than pairs related by derivation. In this paper we explore a related but different idea: inflection and derivation differ in how interpredictable the frequencies of morphologically related words are.

Our reasoning is as follows. We start from the basic idea that derivation yields new lexemes, while inflection yields word forms of the same lexeme. Under a gradient understanding of this statement, the output of derivation will tend to be more independent from its

input compared to that of inflection. The more inflectional a morphological relation is, the more the output will be dependent on other members of its paradigm, and have properties that can be more accurately predicted on their basis.

In psycholinguistic terms, words in a derivational relationship are likely to have more independent mental representations. One way that this independence can manifest is in the amount that information about the meaning or usage of one member of the pair can be predictive about the meaning or usage of the other member. An easily measurable correlate of similarity of semantics and usage is frequency. If the frequency of a word in a cell is systematically accurately predicted by the frequency of a related word in a different cell, it is likely that the two cells represent different ways of talking about the same concept in different contexts, and can therefore be said to be in a more inflectional relationship. If related words in two cells are not good predictors of each other's frequency, this points to the relative independence of words belonging to one cell from words belonging to the other, making this a more derivational relationship.

In the remainder of this section we give initial circumstantial evidence pointing to the relevance of this idea. Table 1 provides information on the distribution of frequency ratios between pairs of French words related by one derivational relation, one inherent inflectional relation, and one contextual inflectional relation.³ The median frequency ratio varies independently of the inflection–derivation divide, with the derivational relation standing between the two inflectional relations. This is not really surprising, as the frequency of inflectional paradigm cells is known to be subject to considerable variation. What is of interest to us is the spread of variation in frequency ratios for each morphological relationship, which we can assess by examining the ratio between the ninth and first decile.⁴ Here we witness very striking

³Frequencies are taken from the FRCOW corpus (Schäfer and Bildhauer 2012; Schäfer 2015); derivational relations are extracted from the Démonette database (Hathout and Namer 2014), while inflectional relations are extracted from the GLÀFF inflectional lexicon (Hathout *et al.* 2014). Only pairs of words which both have non-zero frequency in the corpus and each have no homograph documented in the GLÀFF are taken into account.

⁴We compare the first and ninth quantiles rather than more extreme values

Reference form	Target form	Target / Reference frequency ratio			Inter-decile Ratio
		First decile	Median	Ninth decile	
Infinitive verb	-age derived noun	0.003	0.279	6.500	2166.7
Singular noun	Plural noun	0.011	0.207	1.702	155.7
Conditional 3SG	Conditional 3PL	0.136	0.316	1.000	7.4

Table 1: Distribution of frequency ratios for three morphological relations.

differences: for the derivational relation, we witness more than 3 orders of magnitude of variation in the frequency ratios between related words, while for contextual inflection that variability is less than one order of magnitude. This seems to indicate that the frequency of one form is indeed more predictive of that of the other form if the two words are related by contextual inflection. In addition, our example of inherent inflection stands firmly in the middle, with a bit more than two orders of magnitude of variation. This is strongly suggestive of a gradient quantitative difference that matches the intermediate status of inherent inflection.

A qualitative look at sample examples of high and low frequency ratios provides important insights into the likely causes of the observed differences. Table 2 presents examples of denominal verbs in -age. The pair *fixer*~*fixage* is emblematic of the prototypical situation for very low frequency ratio items: the -age derivative is very low frequency because it lost competition with a rival (Aronoff 1976) relying on a different process, here *fixation* (which instantiates most of the expected action noun senses linked to the verb *fixer*). *Fixage* did not disappear but underwent specialization, and is now a rare technical term in chemistry and economics, making it far less frequent than its corresponding infinitive. A comparable but less extreme situation is found with the pair *arriver*~*arrivage*. *Arrivage* is etymologically ‘the act of arriving,’ but has specialized to mean ‘delivery of a large quantity

because the data tends to be noisy at the very end of the distribution, due to errors in the automatically derived linguistic resources we rely on. This is only meant as a preliminary illustrative measure of frequency dispersion, which will be captured in a more principled way in Section 4.

Reference form	Target form	Frequency ratio
<i>fixer</i> ‘to fasten’	<i>fixage</i> ‘fixing’	0.003
<i>arriver</i> ‘to arrive’	<i>arrivage</i> ‘delivery’	0.007
<i>outrer</i> ‘to exaggerate; to cause indignation’	<i>outrage</i> ‘offense’	49
<i>ouvrer</i> ‘to work’	<i>ouvrage</i> ‘work; book’	738

Table 2: Sample frequency ratios for *-age* deverbal nouns

of merchandise.’ The converted past participle *arrivée* is the general event noun corresponding to *arriver*.

At the other end of the spectrum, *ouvrage* acquired an extra sense of ‘book, (artistic) body of work’ in addition to its etymological sense of ‘a work’—this additional sense boosted its frequency of use, since there is now another concept for which the word can be used. More importantly, while the noun *ouvrage* is alive and well in both of its senses, the verb *ouvrer* progressively fell out of usage, displaced by its synonym *travailler*. *Outrer*~*outrage* is a comparable case: although there is a rather transparent semantic relationship between the two words, the verb is rare in contemporary French and felt as rather precious, while the noun has thrived in a legal context.

Let us now turn to examples of the contextual inflectional relationship between the conditional 3SG and 3PL. As exemplified in Table 3, we observe that what variation there is correlates with the syntactic and semantic properties of the underlying lexemes. At the low end of the spectrum we find verbs which are most frequently used in an impersonal construction with 3SG subject *il* or *ça*. At the high end, we find verbs whose subject is semantically constrained to denote a group. While this is not strictly incompatible with singular number, plural number for the subject, and hence agreement on the verb, is much more likely.

Finally, let us examine an example of inherent inflection, by returning to the relationship between singular and plural nouns. As shown in Table 4, we find what looks like a mix of the the situations found in derivational and contextually inflectional examples. Low frequency ratio items include mass terms such as *uranium*, property nouns such as *unanimité*, and names of disciplines such as *géologie*. In all these cases, use of the plural is restricted to some shifted meaning

LEXEME	COND.3SG	COND.3PL	Freq. ratio
ADVENIR ‘happen’	<i>advierait</i>	<i>advieraient</i>	0.0127
ÉTONNER ‘surprise’	<i>étonnerait</i>	<i>étonneraient</i>	0.0156
SEMBLER ‘seem’	<i>semblerait</i>	<i>sembleraient</i>	0.02845
PULLULER ‘swarm’	<i>pullulerait</i>	<i>pulluleraient</i>	8.6667
JONCHER ‘be scattered on’	<i>joncherais</i>	<i>joncheraient</i>	9.000
S’ENTRECHOQUER ‘knock against one another’	<i>s’entrechoquerait</i>	<i>s’entrechoqueraient</i>	13.000

Table 3: Frequency ratio of words in a INF~COND.3SG relationship in French

of the noun: a type reading for *uranium* (we are distinguishing different varieties of uranium), a metonymic sense extension in the case of *unanimité* (an instance of a unanimous vote) or *géologie* (the geological structure of an area). Given that this shifted meaning is much less frequent than the main meaning, but relatively more frequent in the plural, we get a non-zero but small frequency ratio. Arguably then, all these examples exhibit a frequency ratio predictable from lexical semantics.

At the other end of the spectrum, we find items that are nearly pluralia tantum. *Lipide* can be used in the singular to denote a particular type of fat, but the vast majority of uses are in the plural and denote a quantity of fat. *Ossement* was originally an ordinary noun meaning ‘skeleton,’ which then specialized as a plurale tantum denoting specifically bones denuded of flesh. This is the main meaning attested in the corpus, but there is some innovative use in the singular with the same meaning but unambiguously singular reference. *Concitoyen* is nearly always used in the plural with a generic reading; specific readings are possible in both numbers, but rare. Hence the frequency ratio follows from the fact that generic quantification is overwhelmingly expressed in the plural in French. Overall then, we find here effects that are much more similar to what we witnessed in the case of derivation: a high frequency ratio tends to be due to the conventionalization of a plurale tantum use for one of the readings of a noun, a purely lexical property that is not predictable from either the lexical semantics of the noun or the relationship between singular and plural.

Singular	Plural	Freq. ratio
<i>uranium</i> ‘uranium’	<i>uraniums</i> ‘uraniums’	0.001
<i>unanimité</i> ‘unanimity’	<i>unanimités</i> ‘unanimities’	0.001
<i>géologie</i> ‘geology’	<i>géologies</i> ‘geologies’	0.002
<i>lipide</i> ‘lipid’	<i>lipides</i> ‘lipids’	19
<i>ossement</i> ‘bones’	<i>ossements</i> ‘bones’	29
<i>concitoyen</i> ‘fellow citizen’	<i>concitoyens</i> ‘fellow citizens’	56

Table 4: Frequency ratio of words in a SG~PL relationship in French

Given the discussion above, we expect that in general, the frequency of a word is more predictable from that of its inflectional relatives than from that of its derivational relatives. Moreover, we expect this effect to be gradient, with inherent inflection standing somewhere between derivation and contextual inflection. Although we have no specific prediction, we can presume that other cases of morphology that aligns neither with canonical inflection nor with canonical derivation (Corbett 2010) may also exhibit such intermediate behavior.

Finally, we expect the causes of variability in frequency to be different for inflection and derivation, leading to measurably different effects. For all morphological relations, the frequency ratio between pairs of words is modulated by lexical semantics: some lexical meanings lend themselves to higher or lower frequencies in given cells. As a result, we expect the frequency ratio between pairs of morphologically related words to be generally variable, and that variability to be predicted at least in part by lexical semantic information. Where inflection and derivation are expected to differ is in the extent to which the frequency of a word remains unpredictable once the content it shares with other members of its morphological family is known. Within derivation, we expect an additional cause of variability: because derivationally related words are less interdependent than inflectionally related ones, it is more likely that derivationally related words be subject to independent arbitrary semantic shifts, leading to increased unpredictability of their patterns of usage and frequency properties.

This discussion suggests that a proper exploration of the predictability of word frequency should take semantic information into

account. Distributional semantics provides a possible operationalization of this factor.

Distributional Semantics and morphology

2.4

The prevalent method for quantifying semantics in linguistics is through distributional vectors. The tool has long been used to put a number onto the degree of similarity in meaning between words or lexemes. The framework of distributional semantics is based on the hypothesis, first formulated by Harris (1954), that word distribution correlates with word meaning. The core idea is that the meaning of a word influences what we say about it. Given what the word *dog* means, we are more likely to say “*A dog barks*” or “*The dog is wagging its tail*” than “*This dog shares a border with Romania.*” Hence, by virtue of its meaning, the distribution of the word *dog* will be more similar to that of *jackal* or *pug* than that of *Moldova* or *Hungary*. By abductive reasoning, this entails that words with similar distributions should have similar meanings.

The proposal of Harris (1954), taken at face value, implies that any model of word distribution can be understood as a model of word meaning. In practice, computational linguists have adopted a stricter definition of distributional semantics. Lenci (2018) directly begins his review of the field by equating distributional semantics to vector space semantics. Boleda (2020) takes a more nuanced approach, and states that a distributional semantics model (henceforth ‘DSM’) should exhibit the three following characteristics: words ought to be represented by high-dimensional vectors, these word vectors ought to be empirically computed from corpus data, and the vector space ought to be continuous. Many algorithms have been suggested to derive such distributional vectors, from the LSA model of Landauer and Dumais (1997) based on co-occurrence counts and singular value decomposition, to neural networks trained as classifiers such as the word2vec model of Mikolov *et al.* (2013a). A recent trend is the introduction of distributional representations of word tokens (Peters *et al.* 2018; De-

vlin et al. 2019)—whereas most previous DSMs focused on describing word types.⁵

Another theoretical argument in favor of distributional semantics, outlined by Sahlgren (2008), lies in the connections one can make with structuralism (Saussure 1916; Bloomfield 1933). Sahlgren more specifically draws on Saussure’s concept of *value*. The *value* of a sign is a differential conceptualization of meaning: it is characterized both by the allowed positions of the sign on the syntagmatic axis (i.e., the syntactic contexts where this sign may occur) as well as the relations this sign entertains within the paradigmatic axis (i.e., how it differs from other words that could fit in this slot). This concept is framed as *distributional substitutability* in the work of Harris (1954): two words are distributionally substitutable if they can be swapped for one another in any context. In short, we can expect of a DSM that it groups together words that occur in the same contexts—i.e., words with similar semantics and equal morphosyntactic feature values.

On a practical level, the appeal of DSMs in linguistic studies lies in their ability to produce semantic representations for any word attested in their training corpus. They are therefore invaluable to corpus-driven studies of the lexicon, and applications of distributional semantics to morphology have indeed been fruitful. For instance, Marelli and Baroni (2015) propose to model the semantic effects of derivation as a linear transformation of the base form: their proposal amounts to computing the representation of a word such as *nameless* as the application of a transformation \mathcal{L}_{less} on the base word vector \vec{name} . Other studies include Varvara (2017), who compares the semantic stability of deverbal event nominalization processes using an array of metrics, and Wauquier et al. (2020), who study how different French nominalization processes fall into distinct clusters of distributional vectors.

One DSM architecture in particular has proven to be very popular in such studies: the word2vec model of Mikolov et al. (2013a).

⁵These word token models are more often presented as “contextualized” embeddings; it is straightforward to construe a context-specific representation of a word type as a word token representation. Previous studies have also explicitly equated these two characterizations (Mickus et al. 2020; Lenci et al. 2022, a.o.), often harking back to previous context-specific, exemplar-based approaches (e.g. Erk and Padó 2010).

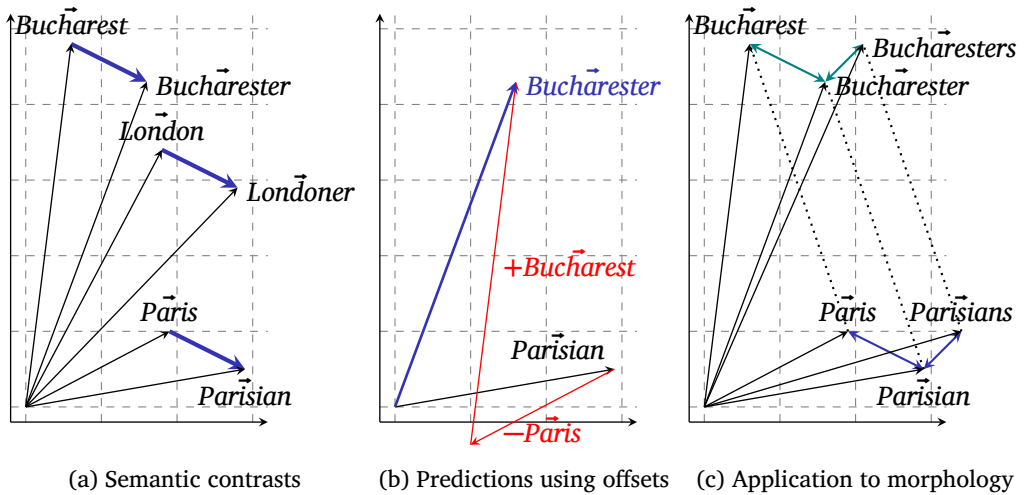


Figure 2: Operationalization of semantic analogy

The chief reason for this fame is that word2vec models arguably encode stable semantic contrasts by means of simple vector offsets. This characteristic was first described by Mikolov *et al.* (2013b); we illustrate it in Figure 2. Figure 2a depicts the key insight: stable semantic contrasts, such as the relation between a city and its demonym (e.g. between *Parisian* and *Paris* or *Bucharester* and *Bucharest*), should translate as a stable vector difference between the two related terms, viz., $\vec{Parisian} - \vec{Paris} \approx \vec{Bucharester} - \vec{Bucharest}$. Basic vector operations give a predictive force to this observation, as shown in Figure 2b: given a pair of words that instantiate a semantic contrast (e.g., *Paris* and *Parisian*) and a cue (e.g., *Bucharest*), we can infer what the counterpart for this cue word should be (viz. *Bucharester*) by means of a simple equation: $\vec{Bucharester} \approx \vec{Parisian} - \vec{Paris} + \vec{Bucharest}$. This ability to make use of stable semantic contrasts is especially worthwhile in paradigm-based morphology, where we can expect pairs of cells in a paradigm to instantiate a stable semantic contrast (see Figure 2c).

A number of works have leveraged this ability to manipulate semantic contrasts to study morphological properties. One approach has been to compare and contrast the stability and predictability of semantic contrasts. Bonami and Paperno (2018) set out to compare the

semantic stability of inflectional and derivational relations, whereas Mickus *et al.* (2019) compare the predictability of grammatical gender variation of different classes of French adjectives.

However concerns have been raised about the validity of this offset method. Linzen (2016) remarks that the terms in an analogy relation tend to be very close to one another—so much so that one of the three cues in an analogy (*viz. Parisian, Paris and Bucharest* in the previous example) is often one of the most likely predicted outputs. Rogers *et al.* (2017) point out that the distance to the target vector often impact results: outliers are much less likely to be retrieved. Schluter (2018) further details how the common practice of normalizing word embeddings before performing vector addition distorts results. We take this criticism as incentive to explore other means of using distributional representations to predict morphological properties.

We can therefore list criteria we require in word embedding architectures before using them in the present study. First, the theoretical argument put forward by Sahlgren (2008) that vector spaces ought to be shaped by structural relations does not hold equally for all models: Sahlgren expects this characteristic to be found in DSMS where context is modeled as word-co-occurrences, such as word2vec, but not in term-document models such as LSA,⁶ which is why we favor the former over the latter. Second, if we wish to study the effects of distributional information and side-step any potential spurious correlations, then we should set aside models that do not rely solely on word-co-occurrences such as the spelling-informed FastText model of Bojanowski *et al.* (2017). Third, as we sketched in the previous section, our interest in the present work lies in the predictability of word frequency: this is a feature we expect word type models to encode more directly than word token representations—hence we will also disregard word token embedding models such as those of Peters *et al.* (2018) or Devlin *et al.* (2019).

⁶ See also Gastaldi (2021) for a discussion.

Section 2.3 outlined why one would expect the frequency of derived lexemes to be subject to more variation than that of inflected forms. We can reframe these expectations in terms of paradigmatic predictability: it is easier to predict the frequency of an inflected word from information about another member of its paradigm than it is to predict the frequency of a derived word from information about its base. Because we are not precommitting to reifying the distinction between inflection and derivation, we shall employ unifying terminology for parallel phenomena in the two domains throughout this paper. We will use the term *reference form* to refer simultaneously to the notions of a base in derivation and the citation form in inflection. Likewise, we call *target form* any form in the inflectional or derivational cell of interest. Our hypothesis can therefore be formulated as follows: the closer the relationship between two words is to canonical inflection, the easier it should be to predict the frequency of the target form from information about its reference form.

To test the hypothesis, we model the frequency of words in the target cell using four sets of predictors. We expect that models of derivational relations will exhibit a higher amount of prediction error than models of inflectional processes; comparing error rates between models and morphological processes will allow us to answer our research question quantitatively. As we focus on comparing error rates, we specifically consider simple models so as to avoid introducing confounding factors. More precisely, we use linear models with no random effects where the dependent variable is the log-transformed frequency of the target cell; our choice is motivated by the overall simplicity of these models.⁷ We consider four sets of predictors:

- (A) Using only the frequency of the reference form.
- (B) Using the frequency of the reference form and the distributional representation of the reference form.

⁷ More complex models, such as neural networks, could be envisioned; we leave those to future work.

- (C) Using the frequency of the reference form and the relative frequency of the word pairs that instantiate the same meaning contrast and are the most semantically similar to the reference form.
- (D) Using the frequency of the reference form and the distributional representations of the words that are the most semantically similar to the reference form.

We therefore establish four types of models according to the set of predictors they use. The models of type A provide a baseline; formally they corresponds to:

$$(1) \quad f(t) \sim f(r)$$

where r and t are the reference and target forms, and $f(\dots)$ measures their frequency. In practice, with this model type, we attempt to predict the frequency of the target form (say, *lirai*) using the frequency of the corresponding reference form (viz., *lire* in this example).

Type B models furthermore include distributional vector components as predictors, or more formally:

$$(2) \quad f(t) \sim f(r) + r_1 + \dots + r_d$$

with r_i the i^{th} component of the d -dimensional vector representation \vec{r} of the word r . Simply put, type B corresponds to predicting the frequency of a target (*lirai*), using the frequency and the distributional vector of the corresponding reference form *lire*. The distributional vectors are raw word embeddings and don't rely on POS tags.

In type C models, we leverage frequency information pooled from the semantic neighborhood of the reference form. Formally, they correspond to:

$$(3) \quad f(t) \sim f(r) + \frac{1}{|N(r)|} \sum_{r' \in N(r)} \frac{t'}{r'}$$

with $N(r)$ the semantic neighborhood of r , i.e., a set of forms belonging to the same morphological category as the reference form r that are semantically similar to r . The final term can be seen as an estimate of the shift in frequency we can expect by observing the behavior or reference and target forms for reference forms that are distributionally similar to r . To give a more concrete example, type C models

try to predict a target form such as *lirai* from the frequency of the reference form *lire* and the average neighbor relative frequency, i.e., $\text{mean}\left(\frac{f(\textit{interpréterai})}{f(\textit{interpréter})}, \frac{f(\textit{déchiffrerai})}{f(\textit{déchiffrer})}, \dots\right)$, as we expect *interpréter* ‘interpret’, *déchiffrer* ‘decipher’ and other semantically similar items to provide helpful insight as to what the target form frequency should be. The last type of models, type D, combines ideas from types B and C. In type D models, we first compute a distributional representation for the semantic neighborhood of the reference form:

$$v_n(r) = \frac{1}{|N(r)|} \sum_{r' \in N(r)} \vec{r}'$$

Simply put, $v_n(r)$ is the average of the word vectors in the neighborhood of r (*interpréter*, *déchiffrer*, etc. in our previous example). We then predict the frequency of the target form (*lirai*) using the frequency of the reference form (*lire*) and the components of this average neighbor vector $v_n(r)$.

$$(4) \quad f(t) \sim f(r) + (v_n(r))_1 + \dots + (v_n(r))_d$$

Throughout all experiments described below, we employ distributional vectors and frequency information computed from the FRCoW corpus (Schäfer and Bildhauer 2012; Schäfer 2015). Where relevant, we employ POS tags provided with the corpus: the vectors used to find neighbours are based on POS-tagged data⁸ (however, the 8-dimensional vectors are based on raw word embeddings). As for distributional representations, we train word2vec models (Mikolov *et al.* 2013a) using the gensim library implementation (Řehůřek and Sojka 2010).⁹

It is worth stressing that models of type B, C and D, by adding different types of predictors to the baseline model structure, target

⁸ these vectors are POS-tagged but unlemmatised. Introducing lemmatisation would have created an asymmetry between inflectional and derivational data

⁹ We use a skip-gram 100-dimensional architecture with a window of 20, 20 negative examples and 10 epochs over the FRCoW corpus. These hyperparameters were selected so as to maximize performance on the French translation of the Google analogy test set (Bojanowski *et al.* 2017).

lexical semantics in different ways. Our reasoning for using distributional neighbors instead of the reference form itself in models of type C and D is that we expect similar words in the cell of interest to be better predictors of the behavior of the target form compared to information about the reference form: similar words in the cell of interest are informative about both the lexical semantics of the data point, and about how the lexical semantics interacts with the semantics of the morphological cell. Simply put, it's important to ascertain that differences in prediction error for inflectional and derivational data aren't merely the result of unaccounted differences in lexical semantics.

Two difficulties arise from our choices of predictors. First, models of types C and D use predictors computed from words that are most semantically similar to the reference form. To qualify which words are most similar to the reference form, we use the nearest neighbors of the distributional representation of the reference form. Depending on the exact formulation of $N(r)$, this can lead to a variable number of neighbors, and hence to a variable number of potential predictors. This issue is why we average distributional representations or frequency information of the most similar words when using them as predictors. The second issue concerns models of types B and D, which include distributional representations as predictors. Said representations consist of high-dimensional vectors: in our case, the representations are originally of 100 dimensions. Including all components as predictors in our models would result in models that are over-specified and possess enough degrees of freedom to encode all the data at our disposal. This would therefore hinder our methodology: we would not be able to compare error rates of such models since they wouldn't have extracted any reasonable generalization from the data but just memorized it. To side-step this issue, we reduce the dimensionality of our embeddings to 8 dimensions when using them as predictors, by applying a truncated SVD dimensionality reduction.¹⁰

¹⁰A truncated SVD reduction corresponds to zeroing out the least important eigenvalues of a SVD factorization. As such, truncating a matrix M to its k largest eigenvalues can be shown to be the optimal approximation to M of rank no greater than k , in that such an approximation \tilde{M} minimizes the difference in Froebenius norm $\|M - \tilde{M}\|_F$ (Eckart and Young 1936; Stewart 1993). Plainly put, using this method guarantees that we minimize the distortion to our entire

To compare the predictability of derivation and inflection, we train models of these four types on data from words instantiating several paradigmatic relations in the French morphological system straddling the inflection-derivation divide as traditionally conceived. We start by collecting examples of word pairs in various paradigmatic relations, such as plural and singular nouns, or agent nouns and their verbal bases. Because of the definition of paradigmatic structure adopted in Section 2.3, which aligns morphological relationships based on their semantic content when building paradigmatic structure, we follow the same practice in our work: formal contrasts that embody the same semantic contrast are treated as realizing the same paradigmatic relation (Gaeta 2007; Štekauer 2014). This is standard in paradigmatic approaches to inflection: words in the same paradigmatic cell are treated as a set with common semantics, regardless of their conjugation or declension class (e.g. French *agiter* and *attendre* are both infinitives, even though their ending is different, since their ending remains the infinitive marker within their class, in the same way that *agitation* and *attente* are both deverbal action nouns, despite their different formal relationship to the base). We then train a model of each type per morphological process. This allows us to compare results on a per-process basis and thus open up the possibility of considering the inflection–derivation distinction as a gradient rather than dichotomy.

We compare the variability of relationships instantiated by each process using residual standard error (RSE) as a metric. This coefficient corresponds to the proportion of the variation in the targets not explained by a model. A model with a lower RSE will be more accurate in its predictions than a model with a higher RSE. In more precise terms, a RSE of x would indicate that predictions with a standard deviation below 1 ought to be accurate to $\pm x$. This measure was chosen because it is at the same time well-suited to comparing prediction accuracy for models of the same process with different predictors, and for comparing accuracy of the same type of model trained on datasets of different sizes. RSE is therefore better equipped for comparing model fit both within and between relations than possible alternatives such as r^2 or AIC/BIC.

set of vectors introduced by the dimensionality reduction.

4 EXPERIMENTS

4.1 *Experiment I*

We train the four model types above for several inflectional and derivational cells in the French morphological system.

4.1.1 Data Selection

Our initial dataset was constructed by compiling information on French (base, derivative) pairs documented in the *Démonette* (Hathout and Namer 2014), *Denom* (Strnadová 2014), *Mordan* (Koehl 2012), and *Converts* (Tribout 2010) databases, and combining it with inflectional information from the *GLÀFF* lexicon (Hathout *et al.* 2014), itself derived from French entries in francophone wiktionary;¹¹ this led to a set of 34 derivational processes and 54 inflectional relations between a citation form and a paradigm cell other than the citation form.

To decide which formal derivational relationship should be treated as semantically equivalent, we look to Guzmán Naranjo and Bonami (under revision), who assess morphosemantic similarity among derivational processes by computing average difference vectors between derived words and their bases and clustering them agglomeratively on the basis of cosine distance. We specifically picked as semantically equivalent collections of processes with the same input and output part of speech and belonging to a cluster with a maximal internal distance of 0.7. The threshold was chosen based on claims in the literature about which formal contrasts have similar semantics, for formal contrasts on which such discussion is available. As a result of this grouping, the 34 processes under examination correspond to 8 paradigmatic relations. Table 5 indicates which processes ended up grouped together, and provides a mnemonic label for each of the groups.

As one of the goals of this research is to compare the effect that different types of predictors have on model accuracy, we wish to train

¹¹ All these databases and more are currently being integrated in *Démonette* version 2 as part of the *Demonext* project (Namer *et al.* 2019). Unfortunately the database was not available yet at the time the present research was conducted.

Idiosyncratic Frequency in Derivation vs. Inflection

Denominal Adjectives	-al:N > A, -aire:N > A, -el:N > A, -ique:N > A, -if:N > A, -eux:N > A, -ier:N > A, -ien:N > A, CONV:N > A
Denominal Verbs	-iser:N > V, CONV:N > V
Deadjectival Verbs	-iser:A > V, -ifier:A > V
Deadjectival Nouns	-té:A > N, -ité:A > N, -itude:A > N, -erie:A > N
Ordinal Adjectives	-ième:Num > A
Deverbal Adjectives	-if:V > A, -ant:V > A, -PST_PART:V > A, -é:V > A, - Vble:V > A
Action Nouns	-erie:V > N, -ance:V > N, -ée:V > N, CONV:V > N, -ure:V > N, -age:V > N, -ment:V > N, -ion:V > N
Agent Nouns	-euse:V > N, -eur:V > N, -rice:V > N

Table 5: Grouping of derivational processes. Processes within the same group are inputs to the same model.

all models for a single paradigmatic relation on the same set of data points. We therefore select the data points for a relation based on the requirements of the most demanding model, and if there are too few data points available to successfully fit the most demanding model, we discard the whole paradigmatic relation from the data.

The most demanding model is type D, which models the frequency of a word in the target cell based on the frequency of its reference form plus each of the dimensions of the 8D average vector of the reference form’s neighbors inflected/derived in the target cell. To minimize the risk of overfitting, models of type D require roughly 100 datapoints per predictor—with 9 predictors (the reference form frequency, together with the eight vector dimensions), the model requires relations with at least 900 data points. Models of type D rely on averaging the vectors of neighboring forms—therefore, for a data point to qualify, it needs to fulfill certain criteria.

French inflection is ripe with syncretisms, some of which are very hard to disambiguate; for instance, regular first conjugation verbs have homographic forms for all three singular forms of the present indicative and subjunctive. Homography also straddles part of speech, with e.g. thousands of nouns and adjectives having identical forms. As a result, precise estimation of the frequency of individual word-forms paired with a morphological category is not currently available. To circumvent that problem, we decided to consider in the model only

words that have no homographs according to the GLÀFF.

The data point should also have a reference form with over 50 occurrences in FRCOW (Schäfer 2015): we wish to employ the distributional vector of the reference form both as a predictor by itself and as a starting point for finding distributional neighbors. Vectors based on few occurrences are unreliable, so data points that rely on vectors derived from too few occurrences should be discarded. We chose 50 occurrences as a threshold for what counts as a reliable vector.

Moreover, the data point should have at least 5 neighbors of the expected PoS, with a cosine similarity of at least 0.7 to the reference form (an arbitrary threshold to ensure the distributional semantic information of the neighbors can be reasonably informative about the usage of the form of interest). The neighbors of the reference form should have the same PoS as the reference form itself, since the idea behind finding the reference form's neighbors is to find semantically similar pairs of forms linked by the same paradigmatic relation as the original pair. If the target form is *reads* and its reference form is *read*, we want semantically similar pairs like *peruses~peruse* or *interprets~interpret*. To find these, we first find the neighbors of the reference form which share a PoS with it: *book (noun)* may be a close neighbor of *read (verb)*, but *book (noun)* can't be inflected in the third person singular in order to get a pair parallel to *read~reads*, so despite being very similar to the reference form, this particular neighbor should be discarded. The threshold on the number of usable neighbors per data point is to do with the fact that some of the predictors are averages: the smaller the number of items going into the average, the more weight each has. To avoid any single neighbor having a disproportionate impact on this average (as each neighbor has their own syntactic/semantic/morphological characteristics which may influence their frequency), we set a minimum of 5 neighbors with the desired characteristics in order for the data point to be included. For the same reason we imposed the 50-token threshold on the reference form, we impose the same threshold on all other distributional vectors we employ in finding word forms, or in the models themselves.

If a data point fulfills all conditions above, it will be included in the dataset for models of type D. If after this filtering the relation still has more than 900 data points available, we fit all four model types to this same set of filtered data points.

For inflection, we also exclude cells such as the past subjunctive and the simple past, which are out of current use or restricted to a specific style of discourse. Usage in these cells is inherently biased for reasons orthogonal to the inflection–derivation debate, introducing noise into any generalizations about how usage in these cells relates to that of a reference form, since the causes for variability would be different.

These filtering conditions leave us with three deverbal derivational relations (verb \rightarrow agent noun; verb \rightarrow action noun; verb \rightarrow agent noun; verb \rightarrow adjective), one nominal inflection relation (singular noun \rightarrow plural noun), and inflectional relations between the infinitive and 15 other verbal paradigm cells. Note that the dataset includes no clear instance of contextual inflection; in particular, because we use the infinitive as the reference form for verbs, the reference and target forms never differ by agreement only.

Results

4.1.2

Full results are presented in tabular format in Table 6, and illustrated graphically in Figure 3. As predicted, the RSE for any derivational targets is higher than the RSE for any inflectional target. This is true both when comparing models of the same type across paradigmatic relations, but also across model types: every model fitted to inflectional data has an RSE that is lower than that of any model fitted to derivational data. Frequency, and therefore patterns of use, appear harder to predict based for derivational relationships compared to inflectional ones. This observation appears to be true regardless of the set of predictors employed. This suggests that there are distinctions in the predictability of usage patterns between processes, which can be captured by our methods, and that traditionally inflectional and traditionally derivational processes pattern together with respect to ease of prediction. Section 2.3 outlined some of the causal factors that we expected would lead to inflectional and derivational relations being distinguished by RSE, all factors ultimately harking back to the fact that inflection normally produces different ways of talking about the same concept in different grammatical contexts.

There has been much debate about the nature of the inflection-derivation divide. Our results suggest the two are ends of a uniformly



Figure 3: RSE for each model type by paradigmatic relation

Idiosyncratic Frequency in Derivation vs. Inflection

	Process	Model A	Model B	Model C	Model D
1	Deverbal adjectives	2.91	2.89	2.90	2.86
2	Action Nouns	2.69	2.61	2.68	2.60
3	Agent Nouns	2.55	2.52	2.52	2.52
4	Plural Nouns	1.89	1.72	1.67	1.67
5	pst.ptcp.m.sg	1.19	1.16	1.11	1.13
6	pst.ptcp.m.pl	1.53	1.47	1.43	1.40
7	pst.ptcp.f.sg	1.56	1.49	1.46	1.44
8	pst.ptcp.f.pl	1.60	1.50	1.46	1.43
9	cond.3sg	1.10	1.02	1.01	0.96
10	cond.3pl	1.05	1.00	0.93	0.96
11	fut.1sg	1.18	1.03	1.05	1.09
12	fut.2sg	1.13	0.95	1.01	1.01
13	fut.3sg	1.01	0.96	0.89	0.92
14	fut.1pl	1.04	0.99	0.96	0.98
15	fut.2pl	1.09	1.02	1.00	1.02
16	fut.3pl	1.06	0.99	0.93	0.95
17	ipfv.3sg	1.26	1.12	1.04	1.07
18	ipfv.3pl	1.29	1.16	1.09	1.11

Table 6: RSE for each model type by paradigmatic relation. Worse performing model by row highlighted in red, best performing model highlighted in green.

populated gradient: RSE values don't pattern in two categorical poles, but span the range between the extremes. The average position of the relations along the gradient patterns well with discussions of their nature in the literature: in the middle, one finds nominal inflection (semantically active) and the past participles (which in French are part verbal and part adjectival, somewhat more independent from the rest of the verbal paradigm compared to other cells).

Within each paradigmatic relation, models of type C or D are generally the best performing, with type A being consistently the worst performer.

While there are differences in performance for models within each relation, the RSE for the four different models is very consistent: as Table 7 shows, there is a very high correlation between RSE values across model types. This suggests that there are properties of the data which make it harder or easier to predict the frequency of words obtained through a given paradigmatic relation, regardless of the exact

	Model A	Model B	Model C	Model D
Model A	1			
Model B	0.997	1		
Model C	0.997	0.998	1	
Model D	0.997	0.998	0.999	1

Table 7: Pearson correlation of RSE for each pair of model types. Values range from 0 to 1. The higher the value, the closer the correlation.

predictors employed.

Focusing solely on the RSE scores, however, leaves out a number of important details. This is apparent if we decompose r^2 coefficients by predictors using dominance analysis (Budescu 1993). According to these analyses, on average 80.3% of the r^2 of type B models and 91.7% of that of of type C is to be imparted on the frequency of the reference form; whereas in type D models, this proportion only tallies up to 50.1%. The fact that different model types lead to converging results while building on a quantitatively different mix of predictors can be construed as confirmation of the robustness of the observed gradient differences between paradigmatic relations.

4.1.3

Discussion

The reason why models C and D appear to be consistently the best performers is probably due to the fact that they integrate information about the target cell and not just about the reference form: it is easier to predict a word’s frequency, which is in part a function of its context of use, if information is available about words that are distributed similarly within that context.

We now discuss those contrasts giving rise to intermediate values for RSE, namely nominal pluralization and the past participles; within that latter set, the masculine singular particularly stands out. These warrant some discussion.

As already discussed, nominal pluralization is semantically active: contexts in which a group of things is talked about may differ from the context in which a singleton thing is talked about. For example, things which in the plural behave as a homogeneous mass (e.g. *crumb~crumbs*) will be biased towards a certain set of contexts in the plural compared to things which in the plural behave as a collective

of individual agents (e.g. *worker~workers*) or as a series of individual objects (e.g. *pie~pies*). This is probably why type C models perform so well compared to the rest for this particular relation: it predicts the frequency of the plural noun by finding semantic neighbors of the singular, and using their average relative frequency in the plural to predict the frequency of the plural form of interest. If we assume that these distinct types of plural classes defined by their semantic properties are an accurate way to describe the data, one might see how semantic information scattered across 8 distributional predictors might perform worse than an estimate for the relative frequency of the plural form for nouns with similar semantics.

To illustrate the mechanism with a simplified case, imagine that establishing the plural subclass of a noun is dependent on properties like agentiveness, mass-like behavior and abstractness, just to give a few examples. These properties are largely orthogonal to one another, and as such might be captured by different dimensions of the word vector. Plural subclasses however might depend on multiple complex interactions between these properties. For instance, we could expect the plural distributions of lexemes to group in four clusters, corresponding to inanimate mass-like lexemes (*crumbs*), inanimate count-like lexemes (*pies*), agentive lexemes with collective tendencies (*workers*) and agentive lexemes without collective tendencies (*CEOs*). Because the model's structure is additive, any features of word usage that are dependent on combinations of properties expressed by different vector dimensions will not be successfully captured. On the other hand, the model based on relative frequency of the neighbors can take into account distributional properties resulting from complex interactions of semantic values: it does so automatically when selecting neighbors in the first place, and aggregates the information about the relative frequency in the plural of words with these properties. By aggregating information, the model type is able to better account for any non-additive relationships between semantic properties.

Past participles have an apparently peculiar distribution as a set: while the masculine singular form gives rise to performance on a par with finite verb forms, the models for masculine plural and feminine forms have higher RSEs, not much lower than those found for noun pluralization. While this is a more subtle point, we argue that this result conforms with our expectations given what we know of usage of

these forms. The French past participle is used in three constructions: in so-called ‘compound tenses,’ where it contributes to the periphrastic expression of TAM and person marking in combination with an auxiliary verb (1); in the passive periphrase, where it expresses passive voice in combination with the auxiliary *être* ‘be’ (2); and finally as the head of an absolute participial modifier (3).¹²

- (1) Paul a envoyé une lettre.
Paul have.PRS.3SG send.PST.PTCP.M.SG IND.F.SG letter
‘Paul sent a letter’
- (2) Une lettre a été envoyée.
IND.F.SG letter have.PRS.3SG be.PST.PTCP.M.SG
send.PST.PTCP.F.SG
‘A letter was sent.’
- (3) Envoyée hier, la lettre
send.PST.PTCP.F.SG yesterday DEF.F.SG letter
arrivera demain.
arrive.FUT.3SG tomorrow
‘As it was sent yesterday, the letter will arrive tomorrow.’

The literature suggests that TAM-expressing uses of the past participle on the one hand, and passive and absolute constructions on the other, do not have the same morphological status: while periphrastic expression of TAM is firmly part of inflection (Bonami 2015), the passive, as a valence-changing operation subject to lexical exceptions, is often argued to belong to derivation (see e.g. Kiparsky 2005; Walther 2013). In a language such as French (or English), where a single form is recruited for the expression of TAM and voice, this entails seeing the past participle as a syncretic form with two discrete functions of a perfect vs. passive participle, with distinct morphological and syntactic properties (Aronoff 1994; Abeillé and Godard 2002). Under this view then, each of our four past participle datasets in fact are composed of aggregate data corresponding to two distinct but homophonous paradigm

¹²We disregard here participles converted to adjectives, as these have been excluded by our data selection strategy, as words having a homograph in a different part of speech.

cells, one of which is higher than the other on the inflection–derivation continuum.

How does this relate with the contrast between RSEs for models of the masculine singular vs. other forms of the participle? As it happens, person and number agreement with the subject is systematic and obligatory for passive and absolute uses of participles, while it is rare for perfect uses. In TAM-expressing uses, the vast majority of verbs use the default masculine singular form in the vast majority of contexts. Only two situations give rise to agreement: transitive verbs agree with a preceding object realized as a weak pronoun or a filler in an unbounded dependency construction, but do not agree in the canonical VO construction; and a minority of intransitive verbs use the auxiliary *être* and agree with their subject.

Construction	M.SG	M.PL	F.SG	F.PL	Total
Non-agreeing TAM	2815	5	6	1	2827
Agreeing TAM	738	236	385	92	1451
Passive	1275	480	803	265	2823
Absolute	2344	630	1241	455	4670
Total	7172	1351	2435	813	11771
Share of TAM	50%	18%	16%	11%	36%

Table 8: Frequency of each use of the past participle by type in the UD_French-GSD corpus

To evaluate the impact of these differences on our data, we queried the UD_French-GSD dependency-parsed corpus (Guillaume *et al.* 2019) and tabulated all combinations of construction type, gender, and number. The results, displayed in Table 8, clearly show that TAM expression makes up a much larger share of the use of masculine singular participles (50%) than the other three gender–number combinations (from 11% to 18%). Hence TAM-expressing uses are over-represented in the pool of masculine singular participle tokens, while conversely the share of passive and absolute tokens uses is over-represented in the three other pools of tokens. Given this, it was to be expected that the masculine singular models have lower RSE, as the share of the data corresponding to more inflection-like uses is higher.

Experiment I showed that the models with information about semantic neighbors within the target cell were the ones that accounted for most variability in target frequency prediction. However, employing such models severely limits the number of paradigmatic relations one can compare: models with semantic information require that enough close neighbors be available for each word form (else, the word form is excluded), and for it to be possible to train a model for a given cell, enough word forms need to have available data (else, the paradigmatic relation is excluded).

Rather than looking at the best absolute fit, let us turn our attention to the relative predictability of the frequency of the output of the different relations. Table 7 indicated that, while that models relying on information about the word form only (models A and B) lead to poorer prediction, their results are highly correlated with those of better performing model types C and D. This suggests that the relative rankings outputted by the method, regardless of which specific model is used, are robust. We can therefore expand the number of morphological processes we are comparing by using models with information about the reference form only, from which fewer datapoints need to be excluded, under the assumption that the estimation of their relative predictability will be comparable to what could be obtained with models incorporating semantic information.

This strategy allows us to obtain data points for 9 other derivational relations, providing a larger set of data points on which to test the prediction that RSE will increase as the relation in question is more extremely derivational in nature. The derivational relation with the smallest number of data points available, given the constraints for models of type A and B, are denominal adjectives in -al (*norme~normal*), with 147 datapoints.¹³

¹³ Given the large number of predictors involved in model type B (reference form frequency + the 8 dimensions of the reference form vector), we should beware of overfitting. To check that the models for these paradigmatic relations are picking up on regularities in the data, we compared the AIC of the target models to the AIC of models for which the values for the dependent variable have been scrambled. If the AIC for the target model is consistently lower than

	Process	Model A	Model B
1	Deadjectival Nouns	2.42	2.36
2	Deadjectival Verbs	2.12	2.04
3	Denominal Verbs	2.68	2.56
4	Denominal Adjectives	2.19	2.14
5	Deverbal Adjectives	2.88	2.86
6	Action Nouns	2.71	2.63
7	Agent Nouns	2.57	2.52
8	Plural Nouns	2.11	1.98
9	pst.ptcp.m.sg	1.35	1.32
10	pst.ptcp.m.pl	1.73	1.63
11	pst.ptcp.f.sg	1.74	1.65
12	pst.ptcp.f.pl	1.77	1.65
13	cond.3sg	1.09	1.01
14	cond.3pl	1.01	0.95
15	fut.1sg	1.09	0.99
16	fut.2sg	0.93	0.87
17	fut.3sg	1.07	1.02
18	fut.1pl	1.02	0.97
19	fut.2pl	1.01	0.99
20	fut.3pl	1.05	0.98
21	ipfv.3sg	1.32	1.18
22	ipfv.3pl	1.29	1.15

Table 9: RSE by model type for all relations included in Experiment 2

Table 9 and Figure 4 confirm the tendency observed in experiment 1: relations that are traditionally regarded as derivational have higher RSE than those traditionally regarded as inflectional. Three

the AIC for the model trained on scrambled data, this suggests that the model is doing more than just memorizing the data and picking up on patterns within it. We scramble the values of the response variable, fit the model, and extract the AIC—this is repeated 10 times for each relation and model type combination. We then compare the AIC for the target model to that of the models trained on scrambled data. For all relations and model types combinations, the AIC for the target model was more than two standard deviations below the mean of the models fitted to scrambled data often many more standard deviations lower. This reassures us that overfitting is not an issue.

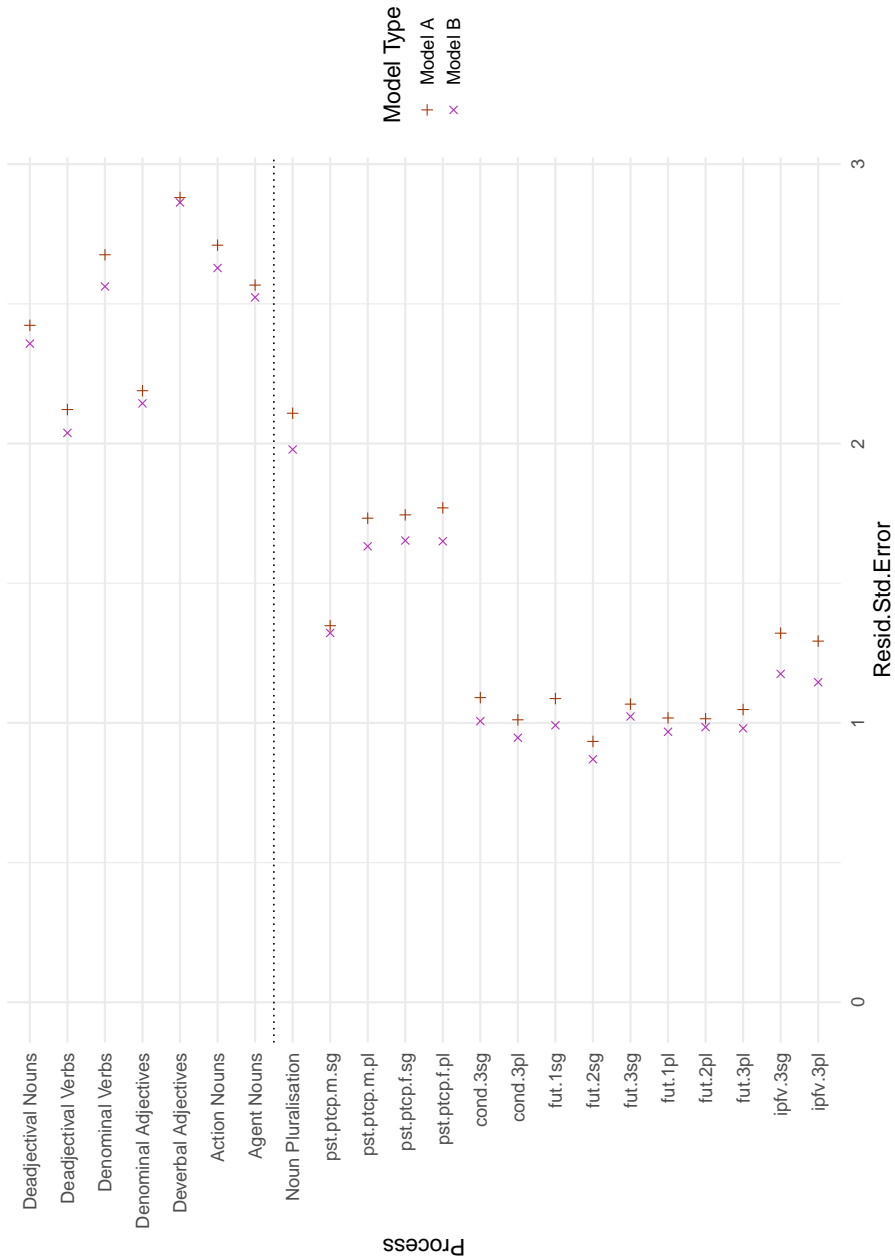


Figure 4: RSE by model type for all relations included in Experiment 2

additional observations are made possible by the presence of more derivational data.

First, some contrasts in predictability among derivational relations match expectations derived from the extant literature. For instance, denominal adjectives are among the most predictable. A considerable proportion of denominal adjectives are so-called ‘relational adjectives’ such as *présidentiel* ‘presidential; of the president’ (Bally 1944). While the characterization of this class of adjectives is the subject of heated debates (McNally and Boleda 2004; Fradin 2007; Rainer 2013; Strnadová 2014), they are generally considered to have very close semantic proximity to their nominal base. At the other end of the spectrum, deverbal adjectives are the most unpredictable. The bulk of these are modal *-able* adjectives, which are notorious for their semantic diversity and unpredictability (Riehemann 1998; Hathout *et al.* 2003).

Second, for other derivational relations, the level of predictability is not readily explained: for instance, there is no immediate explanation for the fact that deadjectival verbs are considerably more predictable than denominal verbs; or for the fact that deadjectival nouns and action nouns, which are often assumed to be minimally different from their bases semantically (Croft 1991; Spencer 2013), lead to contrasting RSEs. These results clearly suggest avenues for future detailed linguistic explorations of the structure of the derived lexicon.

Third, the added data changes the perspective on the inflection-derivation gradient. Based on the smaller sample in experiment 1, we did observe granular differences in predictability within inflectional and derivational relations, but there was still a sharp divide between the two classes: all models for inflectional relations had RSEs below 2, while all models for derivational relations had RSEs above 2.5. In the present experiment, we witness overlap between the two distributions: the least predictable inflectional relation, nominal plural formation, leads to RSEs within the same restricted range (1.95, 2.20) as the two most predictable derivational relations, deadjectival verbs and denominal adjectives. The fact that plural formation has this borderline character is not that surprising: as already hinted at, noun plurals readily gain lexical autonomy as pluralia tantum (cf. e.g. *ciseau* ‘chisel’; *ciseaux* ‘scissors’). However, the general observation strongly suggests that, while derivation is less predictable than inflection on average, the distinction is blurred in some corners of the system; and

hence that no sharp divide can be established between the two.

5

CONCLUSIONS

There has been much discussion concerning the nature of the distinction between inflection and derivation, and how this difference manifests empirically. The paper proposes a quantitative, paradigmatic method to investigate such questions.

The traditional conceptual difference between inflection and derivation is that inflection yields ways of talking about the same concept in different grammatical contexts, while derivation yields ways of talking about different but related concepts. As a consequence, derivationally related words are expected to behave more independently in their patterns of usage than inflectionally related ones for two reasons: first, the relative independence is more likely to enable asymmetric semantic shifts; second, even in the absence of semantic shifts derivationally related words denote different concepts that may have different patterns of usage due to properties of the real world—or more broadly, the semantics of the paradigmatic relation might interact in non-additive ways with the semantics of the base.

If one approaches the lexicon as a series of paradigmatic relationships of interpredictability between words, the difference between inflection and derivation does not need to be reified, but can be emergent from the relative reliability of the paradigmatic relationship in predicting the properties of one form from the other. This would put paradigmatic approaches among those that see inflection and derivation as a gradient.

The paper proposes a method that seeks to compare various morphological relations on the basis of their paradigmatic predictability, to see if this operationalization captures the traditional distinction between inflection and derivation, and whether any interesting patterns emerge either in the relative predictability ranking on different morphological relations or in which types of predictors perform best.

The prediction made by the conceptual distinction between inflection and derivation is effectively one about usage: inflectionally related words will have more interpredictable patterns of usage than

derivationally related words. One easily accessible correlate of patterns of linguistic usage is frequency: if two paradigmatic cells simply constitute ways of talking about the exact same concept in different grammatical contexts (e.g. past vs present) the frequency ratio between members of that paradigmatic relationship should have low variability, since to obtain the frequency of a word in cell B it would suffice to multiply the frequency of the form in cell A by the ratio of contexts that require cell A vs cell B. However, if the two paradigmatic cells link different but related concepts, we expect much more variability in the relationship between the frequencies of two words instantiating said relationship, depending on the semantics of the concept and its real-world properties, the semantics of the morphological relation, and any asymmetrical shift in meaning that might have occurred.

It is therefore expected that the frequency of inflected words would be more accurately predicted than the frequency of derived words, based on comparable information. To establish this, we compare RSEs across models for different relations: RSE provides a normalized, continuous measure for examining differences between relations and model structures. The hypothesis holds up against the data: models predicting the frequency of derived words have consistently higher RSE than models predicting the frequency of inflected words.

We also attempt to fit models containing different kinds of predictors to the same morphological relation. Predictors may include frequency information or distributional information, and they may pertain to a cell of reference within the paradigm or to words obtained by the same relation. We find that it is models which include information about the target cell that tend to provide the best fit for each morphological relation. Nevertheless, all four model structures we employed yielded relatively close RSE estimates for each morphological relation, validating the method: while some information may be more helpful in predicting the frequency of words in a given cell (which information this is for each case is itself informative about the nature of the relation), there appears to be variability that is intrinsic to the data yielded by a given morphological relation.

While comparing the performance of different types of predictors on data from a single relation can give rise to insights about the nature of the relation, the relative consistency in RSE between the

four model types employed for each relation allowed us extend the method to morphological relations with fewer data points available. Given that the relative ranking of relations by their predictability remained constant for each model type, it was possible to use the types of models which required the least amount of data in order to make inferences about a wider range of relations. The larger sample size confirms that the method is capable of capturing differences in predictability of patterns of usage between members of different paradigmatic relationships. Relations traditionally seen as derivational had lower predictability than relations traditionally seen as inflectional. The predictability values did not cluster around the two poles but instead spanned the whole range between the extremes, lending further support to a gradient understanding of the distinction between inflection and derivation, and opening up the possibility that it be seen as emergent from the paradigmatic predictability of the properties of the morphological relation in question.

ACKNOWLEDGMENTS

A previous version of this study was presented at the second Paradigmo workshop (Université Bordeaux Montaigne/Online, June 2021). We thank the audience at the workshop, as well as anonymous reviewers both for the workshop and for this journal, for insightful comments and suggestions.

This work was partially supported by three public grants overseen by the French National Research Agency (ANR): the *Démonext* project (reference: ANR 17-CE23-0005), *Idex Lorraine Université d'Excellence* (reference: ANR-15-IDEX-0004), and *Labex EFL* (reference: ANR-10-LABX-0083), through which it contributes to *Idex Université de Paris* (ANR-18-IDEX-0001). Finally, we acknowledge the support by the *Fo-Tran* project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 771113).

REFERENCES

- Anne ABEILLÉ and Danièle GODARD (2002), The syntactic structure of French auxiliaries, *Language*, 78:404–452.
- Farrell ACKERMAN, James P. BLEVINS, and Robert MALOUF (2009), Parts and wholes: implicative patterns in inflectional paradigms, in James P. BLEVINS and Juliette BLEVINS, editors, *Analogy in Grammar*, pp. 54–82, Oxford University Press, Oxford.
- Paolo ACQUAVIVA (2008), *Lexical Plurals: A Morphosemantic Approach*.
- Stephen R. ANDERSON (1982), Where's morphology?, *Linguistic Inquiry*.
- Stephen R. ANDERSON (1992), *A-Morphous Morphology*, Cambridge University Press, Cambridge.
- Mark ARONOFF (1976), *Word Formation in Generative Grammar*, MIT Press, Cambridge.
- Mark ARONOFF (1994), *Morphology by itself*, MIT Press, Cambridge.
- Charles BALLY (1944), *Linguistique générale et linguistique française*, Francke, Berne.
- Laurie BAUER (2004), The function of word-formation and the inflection-derivation distinction, in *Words in their Places. A Festschrift for J. Lachlan Mackenzie*, pp. 283–292, Vrije Universiteit, Amsterdam.
- Laurie BAUER and Winifred BAUER (2012), The inflection-derivation divide in Māori and its implications, *Te Reo*, 55:3–24.
- L. BLOOMFIELD (1933), *Language*, Holt, Rinehart and Winston, Inc., U.S.A.
- Harry BOCHNER (1993), *Simplicity in Generative Morphology*, Mouton de Gruyter, Berlin.
- Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN, and Tomas MIKOLOV (2017), Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, 5:135–146, doi:10.1162/tacl_a_00051, <https://aclanthology.org/Q17-1010>.
- Gemma BOLEDA (2020), Distributional Semantics and Linguistic Theory, *Annual Review of Linguistics*, 6(1):213–234, doi:10.1146/annurev-linguistics-011619-030303, <https://doi.org/10.1146/annurev-linguistics-011619-030303>.
- Olivier BONAMI (2015), Periphrasis as collocation, *Morphology*, 25(1):63–110.
- Olivier BONAMI and Denis PAPERNO (2018), Inflection vs. derivation in a distributional vector space, *Lingue e Linguaggio*, 17:173–195, <https://halshs.archives-ouvertes.fr/halshs-01957367>.

- Olivier BONAMI and Jana STRNADOVÁ (2019), Paradigm structure and predictability in derivational morphology, *Morphology*, 29, doi:10.1007/s11525-018-9322-6.
- Geert BOOIJ (1996), Inherent versus contextual inflection and the split morphology hypothesis, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of Morphology 1995*, pp. 1–16, Kluwer, Dordrecht.
- David V. BUDESCU (1993), Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression., *Psychological Bulletin*, 114(3):542–551, doi:10.1037/0033-2909.114.3.542, <https://doi.org/10.1037/0033-2909.114.3.542>.
- Joan L. BYBEE (1985), *Morphology: A study of the relation between meaning and form*.
- Greville G CORBETT (2010), Canonical derivational morphology, *Word structure*, 3(2):141–155.
- Greville G. CORBETT (2019), Plurality tantum nouns and the theory of features: a typology of nouns with non-canonical number properties, *Morphology*, 29(1):51–108, doi:10.1007/s11525-018-9336-0, <https://doi.org/10.1007/s11525-018-9336-0>.
- William CROFT (1991), *Syntactic Categories and Grammatical Relations*, The University of Chicago Press, Chicago.
- Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, doi:10.18653/v1/N19-1423, <https://aclanthology.org/N19-1423>.
- Wolfgang U. DRESSLER (1989), Prototypical Differences between Inflection and Derivation, *STUF - Language Typology and Universals*, 42(1):3–10, doi:doi:10.1515/stuf-1989-0102, <https://doi.org/10.1515/stuf-1989-0102>.
- Carl ECKART and Gale YOUNG (1936), The approximation of one matrix by another of lower rank, *Psychometrika*, 1(3):211–218, ISSN 1860-0980, doi:10.1007/BF02288367, <https://doi.org/10.1007/BF02288367>.
- Katrin ERK and Sebastian PADÓ (2010), Exemplar-Based Models for Word Meaning in Context, in *Proceedings of the ACL 2010 Conference Short Papers*, pp. 92–97, Association for Computational Linguistics, Uppsala, Sweden, <https://aclanthology.org/P10-2017>.
- Alan FORD, Rajendra SINGH, and Gita MARTO HARDJONO (1997), *Pace Panini*, Peter Lang, Berlin.

Bernard FRADIN (2007), Three puzzles about denominal adjectives in *-eux*, *Acta Linguistica Hungarica*, 54(1):3–32.

Livio GAETA (2007), On the double nature of productivity in inflectional paradigms, *Morphology*, 17:181–205.

Juan Luis GASTALDI (2021), Why Can Computers Understand Natural Language?, *Philosophy & Technology*, 34(1):149–214, ISSN 2210-5441, doi:10.1007/s13347-020-00393-9, <https://doi.org/10.1007/s13347-020-00393-9>.

Bruno GUILLAUME, Marie-Catherine DE MARNEFFE, and Guy PERRIER (2019), Conversion et améliorations de corpus du français annotés en Universal Dependencies, *TAL*, 60(2):71–95.

Matías GUZMÁN NARANJO and Olivier BONAMI (under revision), A distributional assessment of rivalry in word formation, *Word Structure*.

Zellig HARRIS (1954), Distributional structure, *Word*, 10(23):146–162.

F. W. HARWOOD and Alison M. WRIGHT (1956), Statistical Study of English Word Formation, *Language*, 32(2):260–273.

Martin HASPELMATH (1996), Word-class Changing Inflection and Morphological Theory, doi:10.1007/978-94-017-3716-6_3.

Martin HASPELMATH (forthcoming), Inflection and derivation as traditional comparative concepts, *Linguistics*.

Nabil HATHOUT and Fiammetta NAMER (2014), Démonette, a French derivational morpho-semantic network, *Linguistic Issues in Language Technology*, 11(5):125–168.

Nabil HATHOUT, Marc PLÉNAT, and Ludovic TANGUY (2003), Enquête sur les dérivés en *-able*, *Cahiers de grammaire*, 28:49–90.

Nabil HATHOUT, Franck SAJOUS, and Basilio CALDERONE (2014), GLÀFF, a Large Versatile French Lexicon, in *Proceedings of LREC 2014*.

Jennifer HAY (2001), Lexical frequency in morphology: Is everything relative?, *Linguistics*, 39(6):1041–1070.

Paul KIPARSKY (2005), Blocking and Periphrasis in Inflectional Paradigms, in Geert BOOIJ and Jaap van MARLE, editors, *Yearbook of Morphology 2004*, pp. 113–135, Springer, Dordrecht.

Christo KIROV, Ryan COTTERELL, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sabrina J. MIELKE, Arya MCCARTHY, Sandra KÜBLER, David YAROWSKY, Jason EISNER, and Mans HULDEN (2018), UniMorph 2.0: Universal Morphology, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, <https://aclanthology.org/L18-1293>.

Christo KIROV, John SYLAK-GLASSMAN, Roger QUE, and David YAROWSKY (2016), Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3121–3126, European Language Resources Association (ELRA), Portorož, Slovenia, <https://aclanthology.org/L16-1498>.

Max KISSELEW, Laura RIMELL, Alexis PALMER, and Sebastian PADÓ (2016), Predicting the Direction of Derivation in English Conversion, in *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 93–98.

Aurore KOEHL (2012), *La construction morphologique des noms désadjectivaux suffixés en français*, Ph.D. thesis, Université de Lorraine.

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

Alessandro LENCI (2018), Distributional Models of Word Meaning, *Annual Review of Linguistics*, 4(1):151–171, doi:10.1146/annurev-linguistics-030514-125254, <https://doi.org/10.1146/annurev-linguistics-030514-125254>.

Alessandro LENCI, Magnus SAHLGREN, Patrick JEUNIAUX, Amaru CUBA GYLLENSTEN, and Martina MILIANI (2022), A comparative evaluation and analysis of three generations of Distributional Semantic Models, *Language Resources and Evaluation*, ISSN 1574-0218, doi:10.1007/s10579-021-09575-z, <https://doi.org/10.1007/s10579-021-09575-z>.

Tal LINZEN (2016), Issues in evaluating semantic spaces using word analogies, in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 13–18, Association for Computational Linguistics, Berlin, Germany, doi:10.18653/v1/W16-2503, <https://www.aclweb.org/anthology/W16-2503>.

Robert MALOUF, Farrell ACKERMAN, and Arturs SEMENUKS (2020), Lexical databases for computational analyses: A linguistic perspective, in *Proceedings of the Society for Computation in Linguistics 2020*, pp. 446–456, Association for Computational Linguistics, New York, New York, <https://aclanthology.org/2020.scil-1.52>.

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

P. H. MATTHEWS (1991), *Morphology*, Cambridge University Press, Cambridge, 2nd edition.

Arya D. MCCARTHY, Christo KIROV, Matteo GRELLA, Amrit NIDHI, Patrick XIA, Kyle GORMAN, Ekaterina VYLOMOVA, Sabrina J. MIELKE, Garrett

NICOLAI, Miikka SILFVERBERG, Timofey ARKHANGELSKIY, Nataly KRIZHANOVSKY, Andrew KRIZHANOVSKY, Elena KLYACHKO, Alexey SOROKIN, John MANSFIELD, Valts ERNŠTREITS, Yuval PINTER, Cassandra L. JACOBS, Ryan COTTERELL, Mans HULDEN, and David YAROWSKY (2020), UniMorph 3.0: Universal Morphology, in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, <https://aclanthology.org/2020.lrec-1.483>.

Louise MCNALLY and Gemma BOLEDA (2004), Relational adjectives as properties of kinds, in Olivier BONAMI and Patricia Cabredo HOFHERR, editors, *Empirical issues in formal syntax and semantics*, volume 5, pp. 179–196, <http://www.cssp.cnrs.fr/eiss5>.

Timothee MICKUS, Olivier BONAMI, and Denis PAPERNO (2019), Distributional Effects of Gender Contrasts Across Categories, in *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pp. 174–184, doi:10.7275/g11b-3s25, <https://aclanthology.org/W19-0118>.

Timothee MICKUS, Denis PAPERNO, Mathieu CONSTANT, and Kees VAN DEEMTER (2020), What do you mean, BERT?, in *Proceedings of the Society for Computation in Linguistics 2020*, pp. 279–290, Association for Computational Linguistics, New York, New York, <https://aclanthology.org/2020.scil-1.35>.

Tomas MIKOLOV, Kai CHEN, G.S CORRADO, and Jeffrey DEAN (2013a), Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at ICLR, 2013*.

Tomas MIKOLOV, Wen-tau YIH, and Geoffrey ZWEIG (2013b), Linguistic Regularities in Continuous Space Word Representations, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Association for Computational Linguistics, Atlanta, Georgia, <https://aclanthology.org/N13-1090>.

Fiammetta NAMER, Lucie BARQUE, Olivier BONAMI, Pauline HAAS, Nabil HATHOUT, and Delphine TRIBOUT (2019), Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats, in *Actes de TALN*, Toulouse, France, <https://halshs.archives-ouvertes.fr/halshs-02275652/document>.

Doris PAYNE (1985), Inflection or derivation: is there a difference? , in *Proceedings of the First Annual Meeting of the Pacific Linguistics Conference*, pp. 247–260, U. of Oregon,.

Matthew E. PETERS, Mark NEUMANN, Mohit IYER, Matt GARDNER, Christopher CLARK, Kenton LEE, and Luke ZETTEMAYER (2018), Deep Contextualized Word Representations, in *Proceedings of the 2018 Conference of*

- the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana, doi:10.18653/v1/N18-1202, <https://aclanthology.org/N18-1202>.
- Frans PLANK (1991), Inflection and derivation, *EUROTYP Working Papers. Noun phrase structure*.
- Franz RAINER (2013), Can relational adjectives really express any relation? An onomasiological perspective, *SKASE Journal of Theoretical Linguistics*, 10(1):12–40.
- Radim ŘEHŮŘEK and Petr SOJKA (2010), Software Framework for Topic Modelling with Large Corpora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, ELRA, Valletta, Malta, <http://is.muni.cz/publication/884893/en>.
- Susanne RIEHEMANN (1998), Type-based derivational morphology, *Journal of Comparative Germanic Linguistics*, 2(1):49–77.
- Anna ROGERS, Aleksandr DROZD, and Bofang LI (2017), The (too Many) Problems of Analogical Reasoning with Word Vectors, in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 135–148, Association for Computational Linguistics, Vancouver, Canada, doi:10.18653/v1/S17-1017, <https://www.aclweb.org/anthology/S17-1017>.
- Rudolf ROSA and Zdeněk ŽABOKRTSKÝ (2019), Attempting to separate inflection and derivation using vector space representations, in *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pp. 61–70, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, <https://aclanthology.org/W19-8508>.
- Magnus SAHLGREN (2008), The distributional hypothesis, *Italian Journal of Linguistics*, 20(1):33–54.
- Ferdinand de SAUSSURE (1916), Payot & Rivage, Paris, 1995 edition.
- Roland SCHÄFER (2015), Processing and querying large web corpora with the COW14 architecture.
- Roland SCHÄFER and Felix BILDHAUER (2012), Building Large Corpora from the Web Using a New Efficient Tool Chain, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
- Natalie SCHLUTER (2018), The Word Analogy Testing Caveat, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 242–246, Association for Computational Linguistics, New Orleans, Louisiana, doi:10.18653/v1/N18-2039, <https://www.aclweb.org/anthology/N18-2039>.