



# Journal of Language Modelling

VOLUME 10 ISSUE 2  
DECEMBER 2022



*Institute of Computer Science  
Polish Academy of Sciences  
Warsaw*





# Journal of Language Modelling

VOLUME 10 ISSUE 2  
DECEMBER 2022

## Articles

- Idiosyncratic frequency as a measure of derivation vs. inflection 193  
*Maria Copot, Timothee Mickus, Olivier Bonami*
- Simplicity and learning to distinguish arguments from modifiers 241  
*Leon Bergen, Edward Gibson, Timothy J. O'Donnell*
- Neural heuristics for scaling constructional language processing 287  
*Paul Van Eecke, Jens Nevens, Katrien Beuls*

External reviewers 2019–2022 315



JOURNAL OF  
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)

ISSN 2299-856X (printed version)

<http://jlm.ipipan.waw.pl/>

MANAGING EDITOR

*Adam Przepiórkowski* IPI PAN

SECTION EDITORS

*Elżbieta Hajnicz* IPI PAN

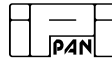
*Małgorzata Marciniak* IPI PAN

*Agnieszka Mykowiecka* IPI PAN

*Marcin Woliński* IPI PAN

STATISTICS EDITOR

*Łukasz Dębowski* IPI PAN



Published by IPI PAN


Institute of Computer Science, Polish Academy of Sciences  
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Circulation: 50 + print on demand

Layout designed by Adam Twardoch.

Typeset in X<sub>Y</sub>L<sup>A</sup>T<sub>E</sub>X using the typefaces: *Playfair*  
by Claus Eggers Sørensen, *Charis SIL* by SIL International,  
*JLM monogram* by Łukasz Dziedzic.

*All content is licensed under  
the Creative Commons Attribution 4.0 International License.*

 <http://creativecommons.org/licenses/by/4.0/>

## EDITORIAL BOARD

*Steven Abney* University of Michigan, USA

*Ash Asudeh* Carleton University, CANADA;  
University of Oxford, UNITED KINGDOM

*Chris Biemann* Technische Universität Darmstadt, GERMANY

*Igor Boguslavsky* Technical University of Madrid, SPAIN;  
Institute for Information Transmission Problems,  
Russian Academy of Sciences, Moscow, RUSSIA

*António Branco* University of Lisbon, PORTUGAL

*David Chiang* University of Southern California, Los Angeles, USA

*Greville Corbett* University of Surrey, UNITED KINGDOM

*Dan Cristea* University of Iași, ROMANIA

*Jan Daciuk* Gdańsk University of Technology, POLAND

*Mary Dalrymple* University of Oxford, UNITED KINGDOM

*Darja Fišer* University of Ljubljana, SLOVENIA

*Anette Frank* Universität Heidelberg, GERMANY

*Claire Gardent* CNRS/LORIA, Nancy, FRANCE

*Jonathan Ginzburg* Université Paris-Diderot, FRANCE

*Stefan Th. Gries* University of California, Santa Barbara, USA

*Heiki-Jaan Kaalep* University of Tartu, ESTONIA

*Laura Kallmeyer* Heinrich-Heine-Universität Düsseldorf, GERMANY

*Jong-Bok Kim* Kyung Hee University, Seoul, KOREA

*Kimmo Koskenniemi* University of Helsinki, FINLAND

*Jonas Kuhn* Universität Stuttgart, GERMANY

*Alessandro Lenci* University of Pisa, ITALY

*Ján Mačutek* Slovak Academy of Sciences, Bratislava;  
Constantine the Philosopher University in Nitra, SLOVAKIA

*Igor Mel'čuk* University of Montreal, CANADA

*Glyn Morrill* Technical University of Catalonia, Barcelona, SPAIN

*Stefan Müller* Humboldt-Universität zu Berlin, GERMANY  
*Mark-Jan Nederhof* University of St Andrews, UNITED KINGDOM  
*Petya Osenova* Sofia University, BULGARIA  
*David Pesetsky* Massachusetts Institute of Technology, USA  
*Maciej Piasecki* Wrocław University of Technology, POLAND  
*Christopher Potts* Stanford University, USA  
*Louisa Sadler* University of Essex, UNITED KINGDOM  
*Agata Savary* Université Paris-Saclay, FRANCE  
*Sabine Schulte im Walde* Universität Stuttgart, GERMANY  
*Stuart M. Shieber* Harvard University, USA  
*Mark Steedman* University of Edinburgh, UNITED KINGDOM  
*Stan Szpakowicz* School of Electrical Engineering  
and Computer Science, University of Ottawa, CANADA  
*Shravan Vasishth* Universität Potsdam, GERMANY  
*Zygmunt Vetulani* Adam Mickiewicz University, Poznań, POLAND  
*Aline Villavicencio* Federal University of Rio Grande do Sul,  
Porto Alegre, BRAZIL  
*Veronika Vincze* University of Szeged, HUNGARY  
*Yorick Wilks* Florida Institute of Human and Machine Cognition, USA  
*Shuly Wintner* University of Haifa, ISRAEL  
*Zdeněk Žabokrtský* Charles University in Prague, CZECH REPUBLIC

# Idiosyncratic frequency as a measure of derivation vs. inflection

*Maria Copot*<sup>1</sup>, *Timothee Mickus*<sup>2</sup>, and *Olivier Bonami*<sup>1</sup>

<sup>1</sup> Université Paris Cité, Laboratoire de linguistique formelle, CNRS

<sup>2</sup> University of Helsinki

## 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 in the literature (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, however, has proved challenging. Linguists often have strong intuitions about whether a process is inflectional or derivational, but there is no single criterion that reliably distinguishes between 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 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 first conjugated verb form instead of second conjugated form of the same verb, 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. In contrast, 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. Derived words are expected to have independent lexical representation and hence, over time, may 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 lexicalized 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 derived from word embedding algorithms (Bonami and Paperno 2018). How accurately one can reconstruct the distributional representation of some target word informs us whether the input 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, known to be relevant to morphological relatedness. 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 the frequency ratio between *quicker* and *quick* to 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. We would not be surprised if the frequency 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 end 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 naïve 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.<sup>1</sup> 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 labour 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 (1986) 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 conceptualization 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, among 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

---

<sup>1</sup>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.

for one class of nouns (and therefore worthy of multiple entries in the dictionary for each set of related items) and as inflectional for a different class (and therefore with each set of related items 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 between 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 behave neither in a typical inflectional nor derivational manner, no matter what specific set of criteria is 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 is generally straightforwardly compositional. However, it can also behave more derivationally: the entity denoted by the plural form may be a different concept from 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 practising 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 a rigorous 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 the 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 characterizing inflection. There is a dearth of attempts to find continuous criteria that characterize the entirety of the gradient.

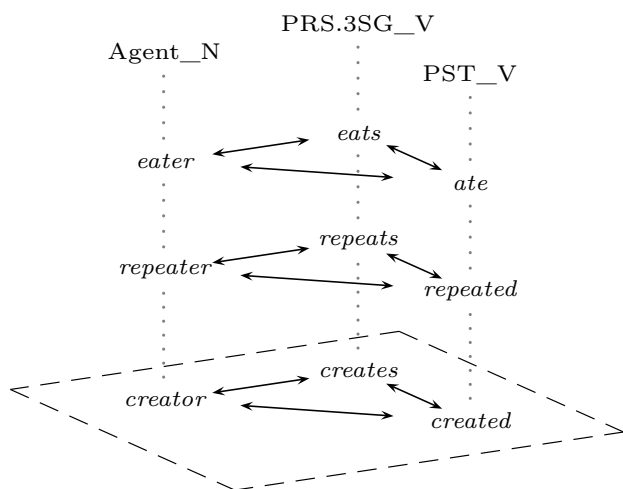


Figure 1:  
A subset of the paradigmatic  
structure of English

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 relationships<sup>2</sup> 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 words are nevertheless in a morphological relationship: their content relationship is instantiated by the same formal means in word pairs like *tooth ~ teeth*. In contrast, 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 is important to note that morphological relationships describe systematic patterns in a way that does not reify the traditional inflection–derivation distinction: *(she) sings* and *singer* are also in a morphological relationship, the same as that 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,*

---

<sup>2</sup>A 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: the amount of information provided on a word's form, meaning, usage, and other characteristics by information about a related word.

*emotion, emotional*. Because the notion of morphological relatedness is agnostic to the inflection-derivation divide, morphological families will group together words that stand in either inflectional or derivational relations in traditional terms, as well as any type of relationship 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 speakers have 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 that of 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 that of 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 unpredictable 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 of 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, with 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 extent to which 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 accurately predicted by the frequency of a related word in a different cell in a systematic fashion, it is likely that the two cells represent 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 and 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.<sup>3</sup> The median frequency ratio varies independently of the inflection–derivation divide,

---

<sup>3</sup>Frequencies 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.



Table 1: Distribution of frequency ratios for three morphological relations

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

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 first and ninth deciles.<sup>4</sup> Here we note very striking differences: for the derivational relation, we witness more than 3 orders of magnitude of variation in the frequency ratios between related words; 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 slightly more than two orders of magnitude of variation. This is strongly suggestive of a gradient quantitative difference that captures the intermediate status of inherent inflection.

A qualitative look at 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

---

<sup>4</sup>We compare the first and ninth quantiles rather than more extreme values 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.

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

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

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 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 perceived as rather affected, 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 that 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.

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

LEXEME	COND.3SG	COND.3PL	Frequency 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>joncherait</i>	<i>joncheraient</i>	9.000
S’ENTRECHOQUER ‘knock against one another’	<i>s’entrechoquerait</i>	<i>s’entrechoqueraient</i>	13.000

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 situations found in derivational and contextually inflectional examples. The frequency ratio is low for 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 of the noun: a type reading for *uranium* (referring to 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.

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

Singular	Plural	Frequency 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> ‘bone’	<i>ossements</i> ‘bones’	29
<i>concitoyen</i> ‘fellow citizen’	<i>concitoyens</i> ‘fellow citizens’	56

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 pluralia 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 pluralia 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.

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

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 are 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. This approach has long been used to quantify 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 should be represented by high-dimensional vectors; these word vectors should be empirically computed from corpus data; the vector space should be continuous. Many algorithms have been proposed 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; Devlin et al. 2019) – whereas most previous DSMS focused on describing word types.<sup>5</sup>

Another theoretical argument in favour 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  $name$ . 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.

---

<sup>5</sup>These 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 (e.g. Mickus et al. 2020; Lenci et al. 2022), often harking back to previous context-specific, exemplar-based approaches (e.g., Erk and Padó 2010).

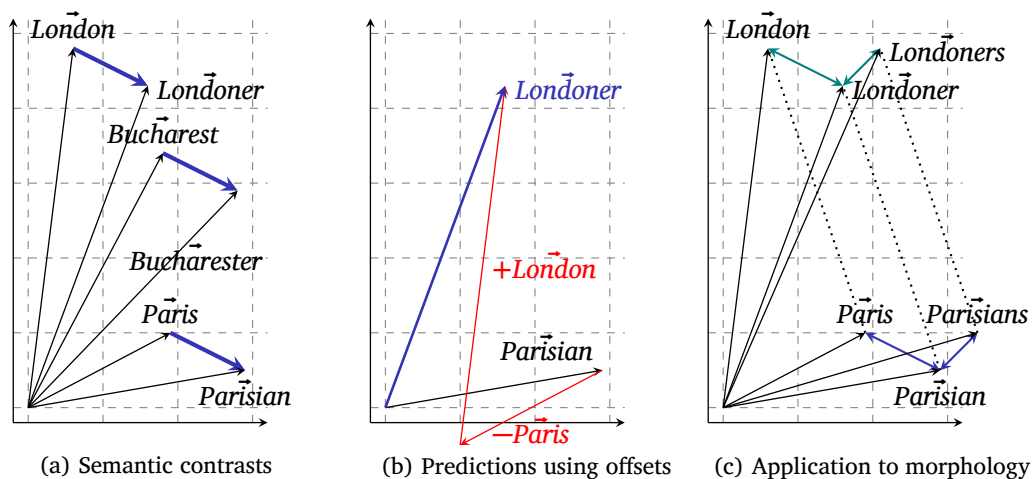


Figure 2: Operationalization of semantic analogy

One DSM architecture in particular has proven to be very popular in such studies: the word2vec model of Mikolov *et al.* (2013a). The chief reason for this popularity 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 2(a) depicts the key insight: stable semantic contrasts, such as the relation between a city and its demonym (e.g. between *Parisian* and *Paris* or *Londoner* and *London*), should translate as a stable vector difference between the two related terms, viz.,  $\vec{Parisian} - \vec{Paris} \approx \vec{Londoner} - \vec{London}$ . Basic vector operations give a predictive force to this observation, as shown in Figure 2(b): given a pair of words that instantiate a semantic contrast (e.g., *Paris* and *Parisian*) and a cue (e.g., *London*), we can infer what the counterpart for this cue word should be (viz. *Londoner*) by means of a simple equation:  $\vec{Londoner} \approx \vec{Parisian} - \vec{Paris} + \vec{London}$ . This ability to make use of stable semantic contrasts is especially useful in paradigm-based morphology, where we can expect pairs of cells in a paradigm to instantiate a stable semantic contrast (see Figure 2(c)).

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 for 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 London* in the previous example) is often one of the most likely predicted outputs. Rogers *et al.* (2017) point out that the distance from the target vector often impacts 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 an incentive to explore other means of using distributional representations to predict morphological properties.

We therefore list the 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 modelled as word-co-occurrences, such as word2vec, but not in term-document models such as LSA,<sup>6</sup> which is why we favour 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 indicated 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).

---

<sup>6</sup> 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 this expectation 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 our 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.<sup>7</sup> 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.

---

<sup>7</sup>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 correspond 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 (e.g. *lirai*) using the frequency of the corresponding reference form (*lire* in this example).

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

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

with  $r_i$  the  $i^{\text{th}}$  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 do not rely on POS tags.

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

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

with  $N(r)$  the semantic neighbourhood of  $r$ , i.e., a set of forms belonging to the same morphological category as the reference form  $r$  and semantically similar to  $r$ . The final term can be seen as an estimate of the shift in frequency we can expect by observing the behaviour of 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 neighbour 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 final 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 neighbourhood 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 neighbour vector  $v_n(r)$ .

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

Throughout all the 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 the POS tags provided with the corpus: the vectors used to find neighbours  $N(r)$  for models of types C and D are based on POS-tagged data,<sup>8</sup> but the 8-dimensional vectors used as predictors in models of types B and D are based on raw word embeddings. All distributional representations correspond to word2vec models (Mikolov *et al.* 2013a) trained with the gensim library implementation (Řehůřek and Sojka 2010) on FRCOW.<sup>9</sup>

It is worth stressing that, by adding different types of predictors to the baseline model structure, models of types B, C and D target lexical semantics in different ways. Our reasoning for using distributional

<sup>8</sup>These vectors are POS-tagged but unlemmatized. Introducing lemmatization would have created asymmetry between inflectional and derivational data.

<sup>9</sup>We use a skip-gram 100-dimensional architecture with a window of 20 tokens, 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).

neighbours instead of the reference form itself in models of types C and D is that we expect similar words in the cell of interest to be better predictors of the behaviour 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 how this lexical semantics interacts with the semantics of the morphological cell. Simply put, it is important to ascertain that differences in prediction error for inflectional and derivational data are not merely the result of differences unaccounted for 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 identify which words are most similar to the reference form, we use the nearest neighbours of the distributional representation of the reference form. Depending on the exact formulation of  $N(r)$ , this can lead to a variable number of neighbours, 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. These 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 would not 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.<sup>10</sup>

---

<sup>10</sup> A truncated SVD reduction corresponds to zeroing out the least important eigenvalues of an 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 set of vectors introduced by the dimensionality reduction.

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 opens up the possibility of considering the inflection-derivation distinction as a gradient rather than a 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, an 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 well-suited both for 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. Therefore RSE is better equipped for comparing model fit both within and between relations than possible alternatives such as  $R^2$  or AIC/BIC.

4

EXPERIMENTS

4.1

*Experiment I*

We trained 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 the francophone wiktionary.<sup>11</sup> 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 (2023), 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 maximum 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.

---

<sup>11</sup> Among others, all these databases are currently being integrated into Démonette version 2 as part of the Demonext project (Namer *et al.* 2019). Unfortunately the enlarged database was not yet available when 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

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 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 entire 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 neighbours inflected/derived in the target cell. To minimize the risk of overfitting, models of type D require roughly 100 data points 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 neighbouring forms – therefore, for a data point to qualify, it needs to fulfil 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 parts of speech: for example, thousands of nouns and adjectives have identical forms. As a result, no precise estimate of the frequency of individual word-forms paired with a morphological category is 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 neighbours. 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 neighbours 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 neighbours can be reasonably informative about the usage of the form of interest). The neighbours of the reference form should have the same PoS as the reference form itself, since the idea behind finding the reference form's neighbours 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 neighbours of the reference form which share a PoS with it: *book (noun)* may be a close neighbour of *read (verb)*, but *book (noun)* cannot 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 neighbour should be discarded. The threshold on the number of usable neighbours 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 neighbour having a disproportionate impact on this average (as each neighbour has its own syntactic/semantic/morphological characteristics which may influence frequency), we set a minimum of 5 neighbours with the desired characteristics in order for the data point to be included. For the same reason that 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 fulfils 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 → agent noun; verb → action noun; verb → agent noun; verb → adjective), one nominal inflection relation (singular noun → 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 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.

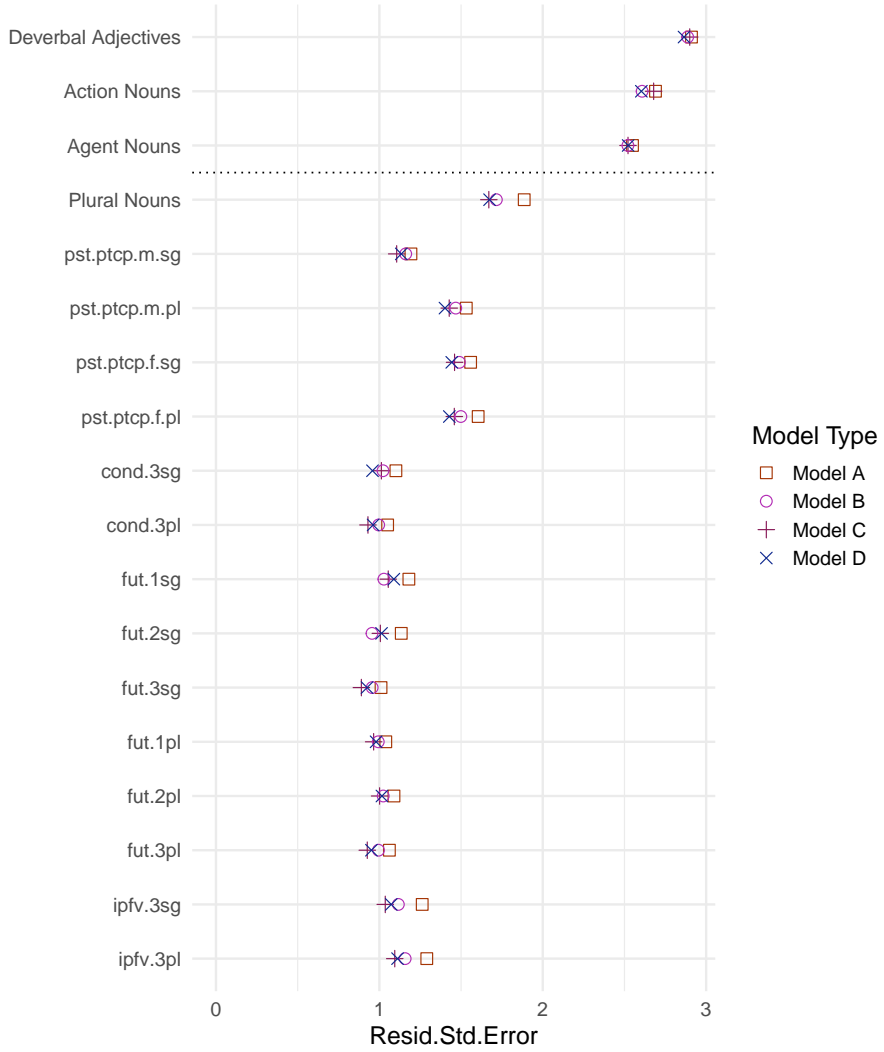


Figure 3: RSE for each model type by paradigmatic relation

*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. Worst  
performing  
model by row  
highlighted  
in red, best  
performing  
model  
highlighted  
in green

There has been much debate about the nature of the inflection-derivation divide. Our results suggest that they are the two ends of a uniformly populated gradient: RSE values do not 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. While there are differences in performance for models within each relation, the RSE for the four different models is very consistent: as

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

	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 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 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 type C is to be imparted on the frequency of the reference form; whereas in type D models, this proportion only reaches 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: they predict the frequency of the plural noun by finding semantic neighbours 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 of 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 behaviour 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 neighbours can take into account distributional properties resulting from complex interactions of semantic values: it does so automatically when selecting neighbours 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).<sup>12</sup>

- (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 arrivera demain.  
send.PST.PTCP.F.SG yesterday DEF.F.SG letter  
arrive.FUT.3SG tomorrow  
‘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 prop-

---

<sup>12</sup>We 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.

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 use  
of the past  
participle  
by construction  
and agreement  
in the UD\_French-  
GSD corpus

erties (Aronoff 1994; Abeillé and Godard 2002). Under this view, each of our four past participle datasets is in fact composed of aggregate data corresponding to two distinct but homophonous paradigm cells, one of which is higher than the other on the inflection–derivation continuum.

How does this relate to 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.

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 neighbours 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 (if not, 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 (if not, 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 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 output 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 data points need to be excluded, under the assumption that the estimate of their relative predictability will be comparable to what could be obtained with models incorporating semantic information.

This strategy allowed 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 data points.<sup>13</sup>

---

<sup>13</sup> 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



Table 9 and Figure 4 confirm the tendency observed in Experiment 1: relations that are traditionally regarded as derivational have higher RSEs than those traditionally regarded as inflectional. Three additional observations are made possible by the presence of more derivational data.

	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. Worst performing model by row highlighted in red, best performing model highlighted in green

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 scrambled the values of the response variable, fit the model, and extracted the AIC – this was repeated 10 times for each relation and model type combination. We then compared the AIC for the target model to that of the models trained on scrambled data. For all relations and model type combinations, the AIC for the target model was more than two standard deviations below the mean of the models fitted to scrambled data, and often many more standard deviations lower. This reassures us that overfitting is not an issue.

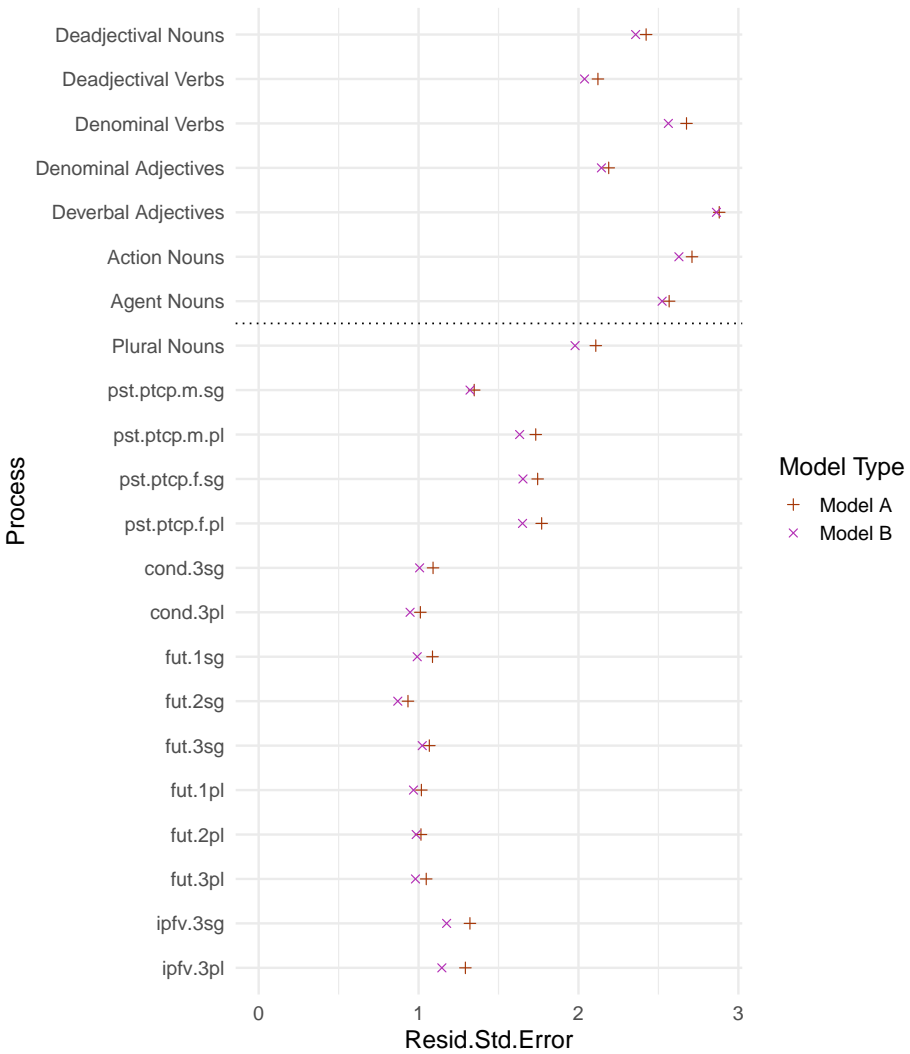


Figure 4: RSE by model type for all relations included in Experiment 2

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 sub-

ject 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.* 2004).

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.

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 dif-

ferent 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 compared 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 attempted 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 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*, Oxford University Press, Oxford.
- Stephen R. ANDERSON (1982), Where's morphology?, *Linguistic Inquiry*, 13:571–612.
- Stephen R. ANDERSON (1992), *A-Morphous Morphology*, Cambridge University Press, Cambridge.
- Mark ARONOFF (1976), *Word Formation in Generative Grammar*, MIT Press, Cambridge, MA.
- Mark ARONOFF (1994), *Morphology by itself: Stems and Inflectional Classes*, MIT Press, Cambridge, MA.
- Charles BALLY (1944), *Linguistique générale et linguistique française*, Francke, Berne.
- Laurie BAUER (2004), The function of word-formation and the inflection-derivation distinction, in Henk AERTSEN, Mike HANNAY, and Rod LYALL, editors, *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.
- Leonard BLOOMFIELD (1933), *Language*, Holt, Rinehart and Winston, Inc., New York, NY.
- Harry BOCHNER (1993), *Simplicity in Generative Morphology*, De Gruyter Mouton, 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.
- 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://www.rivisteweb.it/doi/10.1418/91864>.
- Olivier BONAMI and Jana STRNADOVÁ (2019), Paradigm structure and predictability in derivational morphology, *Morphology*, 29:167–197, 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, <https://scholarlypublications.universiteitleiden.nl/access/item%3A2717489/view>.
- 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, <https://doi.org/10.1037/0033-2909.114.3.542>.
- Joan L. BYBEE (1985), *Morphology: A Study of the Relation between Meaning and Form*, John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Greville G. CORBETT (2010), Canonical derivational morphology, *Word structure*, 3(2):141–155.
- Greville G. CORBETT (2019), Pluralia tantum nouns and the theory of features: a typology of nouns with non-canonical number properties, *Morphology*, 29(1):51–108, <https://doi.org/10.1007/s11525-018-9336-0>.
- William CROFT (1991), *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*, 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, MN, 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, <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, <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 MARTOHARDJONO (1997), *Pace Panini: Towards a Word-based Theory of Morphology*, 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? The structural image of language behind word embeddings, *Philosophy & Technology*, 34(1):149–214, <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 (2023), A distributional assessment of rivalry in word formation, *Word Structure*, 16(1):86–113.
- Zellig HARRIS (1954), Distributional structure, *Word*, 10(23):146–162.
- Frank 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, in G. BOOIJ and J. VAN MARLE, editors, *Yearbook of Morphology 1995*, pp. 43–66, Springer Netherlands, Dordrecht, [https://doi.org/10.1007/978-94-017-3716-6\\_3](https://doi.org/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, <https://aclanthology.org/2014.lilt-11.6/>.
- Nabil HATHOUT, Marc PLÉNAT, and Ludovic TANGUY (2004), Enquête sur les dérivés en *-able*, *Cahiers de grammaire*, 28:49–90, <https://shs.hal.science/halshs-00284612/>.
- Nabil HATHOUT, Franck SAJOUS, and Basilio CALDERONE (2014), GLÀFF, a large versatile French lexicon, in *Proceedings of Conference on Language Resources and Evaluation (LREC), May 2014, Reykjavik, Iceland*, pp. 1007–1012, <https://hal.science/hal-00998467/>.
- Jennifer HAY (2001), Lexical frequency in morphology: Is everything relative?, *Linguistics*, 39(6):1041–1070, <https://doi.org/10.1515/ling.2001.041>.
- 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, [https://doi.org/10.1007/1-4020-2900-4\\_5](https://doi.org/10.1007/1-4020-2900-4_5).

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)*, pp. 3922–3931, 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, Berlin, Germany. Association for Computational Linguistics, <https://aclanthology.org/W16-2015/>.

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, 104:211–240, <https://doi.org/10.1037/0033-295X.104.2.211>.

Alessandro LENCI (2018), Distributional models of word meaning, *Annual Review of Linguistics*, 4(1):151–171, <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*, 56:1269–1313, <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, <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*, 122:485–515.

Peter H. MATTHEWS (1991), *Morphology*, Cambridge University Press, Cambridge, second 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 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, <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, NY, <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, <https://arxiv.org/abs/1301.3781>.

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, GA, <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 L. PAYNE (1986), Inflection versus 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, LA, <https://aclanthology.org/N18-1202>.

Frans PLANK (1991), Inflection and derivation, Noun phrase structure, Theme 7, *EUROTYP Working Papers*, pp. 1–28.

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, <https://www.jstor.org/stable/23741052>.

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, <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, <https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf>.

Ferdinand de SAUSSURE (1916), *Cours de Linguistique Générale*, Payot & Rivage, Paris.

Roland SCHÄFER and Felix BILDHAUER (2012), Building large corpora from the web using a new efficient tool chain, pp. 242–246, Istanbul, Turkey. European Language Resources Association (ELRA).

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, LA, <https://www.aclweb.org/anthology/N18-2039>.

Robert SCHREUDER and R. Harald BAAYEN (1997), How complex simple words can be, *Journal of Memory and Language*, 37:118–139, <https://doi.org/10.1006/jmla.1997.2510>.

Roland SCHÄFER (2015), Processing and querying large web corpora with the COW14 architecture, in Piotr BAŃSKI, Hanno BIBER, Evelyn BREITENEDER, Marc KUPIETZ, Harald LÜNGEN, and Andreas WITT, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, UCREL, IDS, Lancaster, <http://rolandschaefer.net/?p=749>.

Andrew SPENCER (2013), *Lexical Relatedness*, Oxford University Press, Oxford, ISBN 9780199679928, doi:10.1093/acprof:oso/9780199679928.001.0001.

Gilbert W. STEWART (1993), On the early history of the singular value decomposition, *SIAM Review*, 35(4):551–566, <https://doi.org/10.1137/1035134>.

Jana STRNADOVÁ (2014), *Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français*, Ph.D. thesis, Université Paris Diderot et Univerzita Karlova v Praze, <https://theses.hal.science/tel-01536100/>.

Gregory T. STUMP (1998), Inflection, in Andrew SPENCER and Arnold ZWICKY, editors, *The Handbook of Morphology*, pp. 13–43, Blackwell, London.

Gregory T. STUMP and Raphael FINKEL (2013), *Morphological Typology: From Word to Paradigm*, Cambridge University Press, Cambridge.

Pavol ŠTEKAUER (2005), *Meaning Predictability in Word Formation*, John Benjamins, Amsterdam.

Pavol ŠTEKAUER (2014), Derivational paradigms, in Rochelle LIEBER and Pavol ŠTEKAUER, editors, *The Oxford Handbook of Derivational Morphology*, pp. 354–369, Oxford University Press, Oxford.

Pavol ŠTEKAUER (2015), The delimitation of derivation and inflection, in Peter O. MÜLLER, Ingeborg OHNHEISER, Susan OLSEN, and Franz RAINER, editors, *Volume 1 Word-Formation: An International Handbook of the Languages of Europe*, chapter 14, pp. 218–235, De Gruyter Mouton, doi:10.1515/9783110246254-016, <https://doi.org/10.1515/9783110246254-016>.

Delphine TRIBOUT (2010), *Les conversions de nom à verbe et de verbe à nom en français*, Ph.D. thesis, Université Paris Diderot.

Rossella VARVARA (2017), *Verbs as nouns: empirical investigations on event-denoting nominalizations*, Ph.D. thesis, Università degli Studi di Trento, <http://eprints-phd.biblio.unitn.it/2538/>.

Géraldine WALTHER (2013), *De la canonicité en morphologie: perspective empirique, théorique et computationnelle*, Ph.D. thesis, Université Paris Diderot, <https://theses.hal.science/tel-01535976/document>.

Marine WAUQUIER, Nabil HATHOUT, and Cécile FABRE (2020), Semantic discrimination of technicality in French nominalizations, *Zeitschrift für Wortbildung / Journal of Word Formation*, 4:100–119, doi:10.3726/zwjw.2020.02.06, <https://hal.archives-ouvertes.fr/hal-03090796>.

*Maria Copot*

© 0000-0001-5182-9482  
maria.copot@etu.u-paris.fr

Université Paris Cité, Laboratoire  
de linguistique formelle

*Timothee Mickus*

© 0000-0001-9538-7209  
timothee.mickus@helsinki.fi

University of Helsinki

*Olivier Bonami*


© 0000-0003-0688-3855  
olivier.bonami@u-paris.fr

Université Paris Cité, Laboratoire  
de linguistique formelle, CNRS

Maria Copot, Timothee Mickus, and Olivier Bonami (2022), *Idiosyncratic frequency as a measure of derivation vs. inflection*, *Journal of Language Modelling*, 10(2):193–240

doi: <https://dx.doi.org/10.15398/jlm.v10i2.301>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

©  <http://creativecommons.org/licenses/by/4.0/>

# Simplicity and learning to distinguish arguments from modifiers

Leon Bergen<sup>1</sup>, Edward Gibson<sup>2</sup>, and Timothy J. O'Donnell<sup>3</sup>

<sup>1</sup> University of California San Diego

<sup>2</sup> Massachusetts Institute of Technology

<sup>3</sup> McGill University, Canada CIFAR AI Chair, Mila

## ABSTRACT

We present a learnability analysis of the argument-modifier distinction, asking whether there is information in the distribution of English constituents that could allow learners to identify which constituents are arguments and which are modifiers. We first develop a general description of some of the ways in which arguments and modifiers differ in distribution. We then identify two models from the literature that can capture these differences, which we call the argument-only model and the argument-modifier model. We employ these models using a common learning framework based on two simplicity biases which tradeoff against one another. The first bias favors a small lexicon with highly reusable lexical items, and the second, opposing, bias favors simple derivations of individual forms – those using small numbers of lexical items.

Our first empirical study shows that the argument-modifier model is able to recover the argument-modifier status of many individual constituents when evaluated against a gold standard. This provides evidence in favor of our general account of the distributional differences between arguments and modifiers. It also suggests a kind of lower bound on the amount of information that a suitably equipped learner could use to identify which phrases are arguments or modifiers.

*Keywords:*  
*linguistics,*  
*machine learning,*  
*computational*  
*linguistics, syntax,*  
*statistics*

We then present a series of analyses investigating how and why the argument-modifier model is able to recover the argument-modifier status of some constituents. In particular, we show that the argument-modifier model is able to provide a simpler description of the input corpus than the argument-only model, both in terms of lexicon size, and in terms of the complexity of individual derivations. Intuitively, the argument-modifier model is able to do this because it is able to ignore spurious modifier structure when learning the lexicon. These analyses further support our general account of the differences between arguments and modifiers, as well as our simplicity-based approach to learning.

1

INTRODUCTION

The expressivity of natural language is made possible by a division of labor between an inventory of stored items (e.g., morphemes, words, idioms, etc.), known as the *lexicon*, and a set of structure-building operations which combine lexical items to create new expressions, known as the *grammar*.<sup>1</sup> The operation of the grammatical system is highly constrained by requirements imposed by specific lexical items. Consider the verb *put*. In its most basic usage, this verb can only appear in sentences which contain constituents expressing: (i) who is doing the putting, (ii) what is being put, and (iii) the destination of the putting event. The sentence *\*John put the loaf of bread* is incomplete, while the sentence *John put the loaf of bread in his kitchen cupboard* is not. Furthermore, *put* imposes other requirements on sentence structure, such as the requirement that object being put be expressed as a noun phrase. We will refer to such lexically-specified requirements as the *argument structure* of *put*.

---

<sup>1</sup>Note that throughout this paper, we use the term *lexical item* to refer to the elementary units combined by a grammar formalism – whether or not they contain surface words. In the tree-adjoining grammar tradition, which we make use of here, these would more formally be called *elementary trees*. Hence, whenever we use the term *lexical item*, we are referring to what are typically referred to as *elementary trees* in that literature.



Over the last decades, many linguistic theories have adopted a *lexically-driven* view of grammar. Under such an architecture, grammatical computation is performed by small number of structure-building operations (e.g., UNIFY, MERGE, etc.) whose behavior is controlled by the argument-structure specifications of lexical items (Bresnan 2001; Chomsky 1995a,b; Culicover and Jackendoff 2005; Gamut 1991; Gazdar *et al.* 1985; Heim and Kratzer 1998; Huddleston and Pullum 2002; Jackendoff 2002; Johnson and Postal 1980; McConnell-Ginet and Chierchia 2000; Mel'čuk 1988; Moortgat 1997; Pollard and Sag 1994; Sag 2012; Stabler 1997; Steedman 2000).<sup>2</sup> The development of lexically-driven approaches to grammar leads naturally to the suggestion that much of language learning might be reduced to the problem of learning the lexicon (see, e.g., Chomsky 1993).

However, natural language also exhibits constituents that do not appear to be arguments of any lexical item. Consider the sentence *While preparing dinner, John thoughtlessly put the loaf of bread in his kitchen cupboard*. In this sentence, the phrases *while preparing dinner* and *thoughtlessly* specify additional information about the time and manner of the putting event, but they do not seem to be required by any other constituent and the sentence is well-formed and interpretable without them. These phrases also differ in a number of other ways from the core arguments of the verb. For instance, while the argument-phrase specifying the doer of the putting event (i.e., *John*) must appear in the subject position of the sentence (*\*put the loaf of bread John in his kitchen cupboard*), these other phrases can appear in a greater variety of positions (*John thoughtlessly put the loaf of bread in his kitchen cupboard, while preparing dinner*). We will refer to such non-argument phrases as *modifiers*.

The existence of such (apparent) non-argument-driven structure raises a fundamental question. If there are both lexical and non-lexical

---

<sup>2</sup>We note that an alternate tradition of *constructivist* theories argue that argument structure is not associated with particular lexical roots (a position sometimes known as *projectivism*) but rather is a consequence of the functional structure into which roots are inserted during syntactic derivation (see Marantz 2013, for discussion). To the degree that differences between arguments and modifiers in such frameworks still give rise to the distributional differences we discuss below, our results are also consistent with these theories.

modes of composition, how do learners determine when and how each are used? Consider the phrase *in his kitchen*. In the sentence *John put the loaf of bread in his kitchen*, this phrase is an argument, while it is a modifier in the sentence *John made the loaf of bread in his kitchen*. Adult speakers understand these structural differences despite such superficial similarities between the constructions. How do they come by this knowledge?

In this paper, we use computational modeling to address this question. We argue that the statistics of natural language corpora provide evidence that would allow learners to distinguish between argument and non-argument modes of composition in many cases. This evidence is complementary to other forms of evidence available to learners that have been discussed in the context of the argument-modifier distinction in the linguistic literature (such as semantic differences) and can be leveraged by appropriately equipped learners to determine the *argument* or *modifier* status of individual phrases.

In Section 2, we propose that modifiers tend to differ from lexically specified arguments in three ways that have distributional consequences (*inter alia*): **iterability** vs. **finiteness**, **optionality** vs. **obligatoriness**, and **structural flexibility** vs. **structural fixity**. In Section 2.1, we describe two models of lexicon learning designed to minimally capture these differences: the *argument-only model* and the *argument-modifier model*. All formal details can be found in the appendices of the paper.

In Section 3, we describe how lexicon learning under both models can be formulated in terms of a *tradeoff* between two simplicity biases that favor small lexica (*simple-lexicon bias*) and simple derivations (*simple-derivation bias*), respectively. Adopting this tradeoff-based approach, we first show in Section 5.1 that the argument-modifier model is able to recover the argument status of many constituents in a gold-standard corpus, indicating that it captures some aspect of the argument-modifier distinction as discussed in the linguistics literature. We then show in Section 5.2 that the argument-modifier model is able to provide explanations of the input corpus that are more optimal in terms of both the small-lexicon and simple-derivation biases. These results imply that there is clear distributional evidence indicating the argument-modifier status of

many phrases and that this evidence could be leveraged by learners who make use of a tradeoff between derivational and lexical simplicity.

## ARGUMENTS AND MODIFIERS

2

Historically, some distinction between arguments and modifiers (sometimes called *adjuncts*) has been assumed by nearly all theories of syntax and semantics and a number of theoretical mechanisms have been proposed to handle the distinction. Furthermore, many different syntactic and semantic tests have been proposed for distinguishing between the two kinds of phrase (see Bergen *et al.* 2015, for detailed review of this literature). In this paper, we operationalize the argument-modifier distinction by focusing on one particular question: Which constituents in a sentence are there because they were required by some lexical item, and which are not lexically required? In this paper, *argument structure* will refer to any lexically-specified constraint or requirement on constituent co-occurrence. We intend this general notion of argument structure to potentially include many kinds of lexically-specified constraint that have been proposed over the years in different grammatical traditions. Thus, it includes verb-argument structure but, also, the lexical requirements of other categories such as prepositions or nouns.

The difference between arguments and modifiers is often cast in semantic terms. While we do not deny that there are important differences in the way that these types of constituent contribute to the meaning of sentences, in this paper we focus solely on differences between the two types of phrase that affect the distribution of constituents in language.

In lexically-specified grammar formalisms, lexical items list their arguments and (typically) where these arguments appear with respect to the selecting item. This architecture has three critical properties which have important distributional consequences. First, lexical items in such formalisms usually specify only a small number of argument

positions (**finiteness**).<sup>3</sup> Second, lexical arguments are typically obligatory in such systems (**obligatoriness**), though some mechanisms for handling optional arguments are usually provided. Third and finally, particular arguments are required to appear in fixed relationship to the selecting lexical item (**structural fixity**). In languages like English, this typically corresponds to their structural position with respect to their selecting head. In other languages, this may correspond to a grammatical relation which is encoded in other ways (e.g., case).

By contrast, the types of constituents which have been traditionally identified as modifiers differ in each of these three properties. An unbounded number of modifiers can often be added to a constituent (**finiteness** vs. **iterability**); modifiers tend to be optional (**obligatoriness** vs. **optionality**); and modifiers often occur in a greater variety of structural relationships with their head (**structural fixity** vs. **structural flexibility**). These three dimensions of variation summarize a large number of properties and linguistic tests that have been discussed in the literature (Borsley 1999; Comrie 1993; Creissels 2014; Croft 2001; Forker 2014; Haegeman 1994; Haspelmath 2014; Hornstein and Lightfoot 1981; Koenig et al. 2003; Kroeger 2004; Matthews 1981; Przepiórkowski 1999a,b; Radford 1988; Rákosi 2006; Schütze 1995; Schütze and Gibson 1999; Tallerman 2015; Tutunjian and Boland 2008; Vater 1978; Wichmann 2014; Zwicky 1993).<sup>4</sup>

We emphasize that these properties are not definitional and do not represent necessary and sufficient conditions on argumenthood. Instead, they are tendencies: Arguments are sometimes optional (e.g., *John ate/John ate the cake*) and in some cases there is more than one structural realization of the same arguments of some lexical item (e.g., *John gave Mary the book/John gave the book to Mary*). At the same time, there are often strong constraints on the structural position of modifiers (e.g., *John gave Mary the book quickly/\*John gave quickly Mary the book*) and there are constructions in which modifiers are obligatory (e.g., *These ovens clean easily*).<sup>5</sup> Nevertheless, the three properties do roughly summarize a number of linguistic tests for argument-/modifierhood often dis-

---

<sup>3</sup> See Przepiórkowski (2017) for an exception.

<sup>4</sup> See Bergen et al. (2015) for a detailed review of this literature.

<sup>5</sup> We thank an anonymous reviewer for this example.

cussed in the literature. We propose that these statistical tendencies can be used by suitably equipped learning to determine the argument/modifier status of many constituents and, thus, provide a useful source of evidence for lexicon learning that is complementary to other sources of evidence that have been discussed in the literature.

### *Tree-substitution and sister-adjunction grammars*

2.1

In this paper, we use *probabilistic tree-substitution grammars* as our model of lexical argument structure. A tree-substitution grammar formalizes the lexicon as an inventory of stored tree fragments, such as those shown in Figure 1 (Bod 1998; Joshi and Levy 1975; Scha 1990, 1992). This figure shows the inventory of elementary trees that we will use as examples below.<sup>6</sup> Each tree fragment encodes the category and structural position of argument phrases that must be present in a complete sentence which is derived using the fragment. In a tree-substitution grammar, lexical fragments are combined via the *SUBSTITUTE* operation, which replaces a node at the frontier of a derivation with another tree fragment from the lexicon – subject to the condition that the category of the frontier node and the root category of the substituted fragment are identical. The *SUBSTITUTE* operation is applied recursively until no substitutable nodes remain at the frontier, and a complete sentence has been derived.

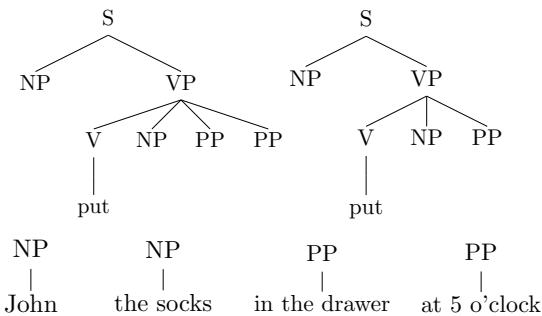


Figure 1:  
Inventory of tree fragments

<sup>6</sup>Note that the internal constituent structure of the noun and prepositional phrases (NP and PP) has been suppressed.

Tree-substitution grammars capture the three core properties of argument structure discussed Section 2. Each lexical fragment can only possess a fixed (and in practice small) number of leaf variables (**finiteness**). All such variables must be filled in a complete derivation (**obligatoriness**); and finally, the position of each argument phrase is fixed relative to the lexical item which selects for it (**structural fixity**).

To model modification, we make use of an extension of tree-substitution grammars which introduces a second structure-building operation, *sister-adjunction* (Chiang 2000; Chiang and Bikel 2002; Rambow *et al.* 1995; Schabes and Shieber 1994). While SUBSTITUTE must be licensed by the presence of an argument node at the frontier, SISTER-ADJOIN can insert a constituent as the sister to any node in an existing tree. The formalism is strongly equivalent to (unlexicalized) tree-insertion grammar and, therefore, has the same weak generative capacity as context-free grammar (Schabes and Waters 1995).<sup>7</sup>

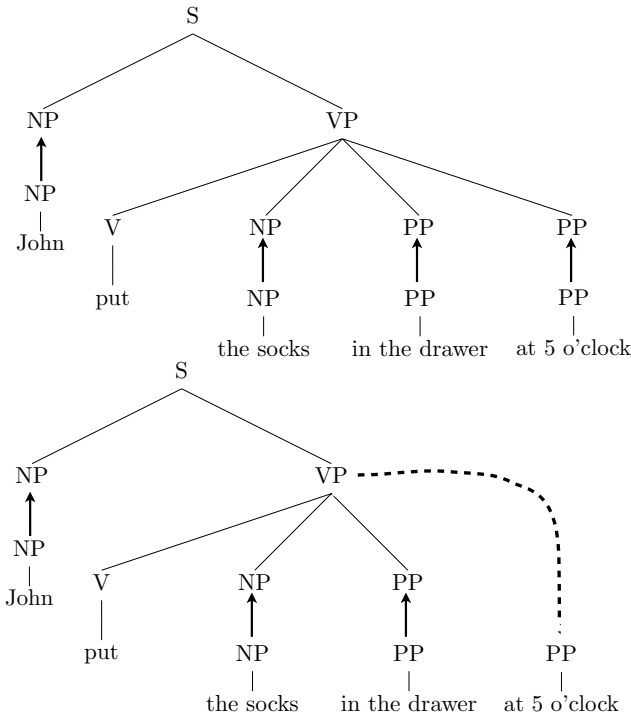
To derive the complete tree for a sentence using a set of fragments such as those shown in Figure 1, the generative process starts from a single nonterminal node of category S (i.e., the start symbol), and then recursively samples arguments and modifiers according to the following procedure. For each node  $f$  with nonterminal category  $A$  on the frontier of our derivation, we perform the following two steps. First, we choose an elementary tree  $t$  with category  $A$  from our lexicon and, for each position before or after a node on the interior of  $t$ , we sister-adjoin zero or more new nonterminal nodes, representing modifier phrases. Second, we substitute  $f$  – now with modifier category nodes – into the derivation at node  $n$  (see discussion of Figure 2 below). This process then repeats on any nonterminal nodes now on the frontier of the tree. In particular, if we have sister-adjoined a modifier node with category  $X$ , its internal structure will be determined recursively by choosing an elementary tree of category  $X$  from the lexicon.

The SISTER-ADJOIN operation formalizes the three core ways in which modifiers differ from arguments: (i) The decision to insert or not insert a modifier does not change the well-formedness of a generated structure with respect to the satisfaction of lexical argument

---

<sup>7</sup>We note that these formalisms have different strong generative capacity, however.

Figure 2:  
TSG versus SAG derivations



requirements (**optionality**) (ii) **SISTER-ADJOIN** can insert any number of modifiers at a position in a derivation (**iterability**), and (iii) **SISTER-ADJOIN** can insert a modifier at any position in a constituent (**structural flexibility**).

Figure 2 illustrates two derivations of the same tree, one in a standard tree-substitution grammar (TSG) without sister-adjunction, and one in the model extended with **SISTER-ADJOIN**, which we term *sister-adjunction grammar* (SAG). The tree-substitution grammar derivation, at the top of the figure, uses an elementary tree with four leaf non-terminals as the backbone for the derivation. The four phrases filling these arguments are then substituted into the elementary tree, as indicated by arrows. Note that in tree-substitution grammars the prepositional phrase, *at 5 o'clock*, which is a temporal modifier, enters the derivation through an argument node. However, the sister-adjunction grammar in the lower part of the figure is able to insert this modifier using **SISTER-ADJOIN** (indicated using dotted lines) and, therefore, uses an elementary tree with only three leaf nonterminals as the back-

bone of this derivation. This difference will mean that tree-substitution grammars will require a greater number of tree fragments in the lexicon to account for variability that could otherwise be accounted for using modification.

### 3 HANDLING UNCERTAINTY: TRADEOFF-BASED LEARNING OF LEXICA

Neither language learners nor linguists have a priori knowledge of the set of lexical items in a language, their particular argument structures, or the argument/modifier status of individual phrases in the input. Rather, the set of lexical argument structures in a language must be learned from linguistic input, and the derivation of particular sentences must be inferred on a case-by-case basis. In this paper, we adopt a probabilistic approach to these problems of learning and inference, specifying prior distributions over lexicons and derivations for both the argument-only model and the argument-modifier model, and using probabilistic conditioning to infer language-specific lexicons and utterance-specific derivations from input data. We give formal definitions of our prior distributions, and algorithms for estimating conditional probabilities in Section 6. In this section, we give an intuitive overview of the ideas behind the framework.

Following earlier work, we propose that lexicon learning is guided by two prior biases for simplicity (especially Brent 1999; De Marcken 1996a,b; Goldwater 2006; Johnson *et al.* 2007; O'Donnell 2011, 2015). The first, the *simple lexicon bias*, provides an a priori measure of the quality of lexicons, favoring those with fewer, more reusable lexical items. The second, the *simple derivation bias*, provides an a priori measure of the quality of the derivations of individual sentences, favoring simpler derivations involving smaller numbers of lexical items, and lexical items with higher probability. These two biases lead to a trade-off: For a fixed set of sentences, if we increase the average reusability of lexical items, then we must also increase the average number of lexical items used in any derivation. Likewise, if we decrease the average number of lexical items used per derivation, we must, on average, increase the size of the lexicon. The inference problem is to jointly find a



set of lexical items and sentence derivations that best explains the distribution of forms in the input data, subject to these two prior biases.

Our two prior biases are a special case of the standard Bayesian prior/likelihood tradeoff applied to the problem of lexical storage. The preference for more reusable lexical items is encoded by the prior over lexical items and the preference for smaller derivations results from the likelihood, which favors derivations in which fewer random choices are made. In the two sections below, we provide additional intuitions about the behavior of our models when applied to input datasets and details about their implementation.

### *Simplicity biases and inference*

3.1

As just discussed, our models encode two simplicity biases. The *simple lexicon bias* favors smaller lexicons containing more reusable lexical items. Following Goldwater (2006), Johnson *et al.* (2007), and others, we formalize this bias using a distribution from Bayesian nonparametric statistics known as the *Pitman-Yor Process* (Pitman and Yor 1995). A Pitman-Yor process  $PYP(G_0, a, b)$  is a distribution over lexical items that is specified with three parameters,  $G_0$ ,  $a$ , and  $b$ . The first parameter,  $G_0$ , is a prior distribution over possible tree fragments that can be stored as lexical items. The other two parameters – known as the concentration parameter  $b$  and discount parameter  $a$  – are real-valued such that  $0 \geq a \geq 1$  and  $b > -a$ .

A Pitman-Yor process operates as follows. The first time we sample from  $PYP(G_0, a, b)$ , a new lexical item will be chosen according to  $G_0$ , stored internally by the Pitman-Yor process, and returned to the caller. On subsequent invocations, either a previously sampled lexical item  $i$  will be returned with probability  $\frac{n_i - a}{N + b}$ , or a new lexical item will be sampled from  $G_0$ , stored, and returned, with probability  $\frac{aK + b}{N + b}$ , where  $n_i$  is the number of times that lexical item  $i$  was previously sampled,  $N$  is the total number of lexical items sampled so far (i.e.,  $N = \sum_j n_j$ ), and  $K$  is the number of distinct lexical items that have been previously sampled (i.e., the number of lexical *types*). Notice that these definitions favor smaller numbers of lexical items and induce a rich-get-richer dynamic whereby lexical items that are used more often are more likely to be reused.

The *simple derivation bias* favors derivations for individual sentences that use small numbers of more probable lexical items. In both the argument-only model and argument-modifier model, this bias is captured by our assumption that the probability of a derivation is the product of the probabilities of the lexical tree fragments used to construct it. Because probabilities must be numbers between 0 and 1, the probability of a derivation decreases quickly (geometrically) as the number of fragments it contains increases. However, this can be mitigated somewhat if the fragments are highly probable (i.e., have probability close to 1).

Applying these two simplicity biases to tree-substitution grammar, we arrive at what we call the argument-only model (see Bod et al. 2003; Cohn et al. 2010; O’Donnell 2011, 2015; Post and Gildea 2013, for related models). To better understand the inferential behavior of the argument-only model, it is useful to consider a toy example. Figure 3 shows three possible solutions to the problem of inferring the correct set of stored tree fragments for a toy corpus consisting of three sentences.

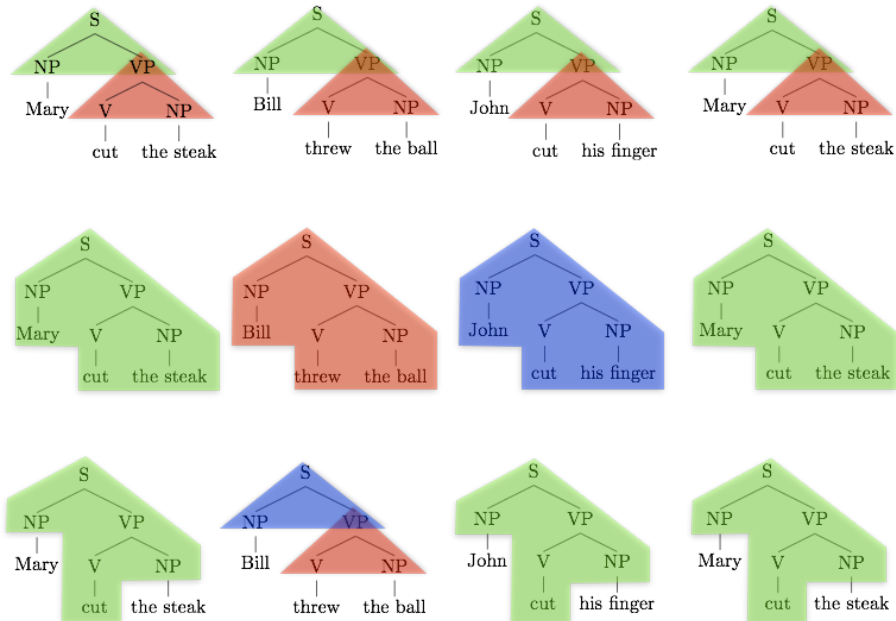


Figure 3: Inference in the argument-only model

Row I of Figure 3 shows the result of storing and using only the smallest, most abstract fragments of sentence structure. In this case, each particular lexical item will be highly reusable, and the lexicon will be maximally compact. However, the derivations of individual sentences will necessarily make use of many lexical fragments and will therefore be more complex. Row II of the figure shows the solution at the other extreme. In this case, every utterance is stored in its entirety. This solution will result in extremely large lexicons with lexical items of limited reusability. However, individual sentences which recur in the data will be derivable with a single lexical item, resulting in potentially low-cost derivations if particular sentences recur in the input. Row III of Figure 3 shows an intermediate solution which is more optimal with respect to this dataset. By storing lexical fragments which express argument structures of intermediate complexity, this solution produces a more compact lexicon than the solution in Row II, and simpler derivations than the solution in Row I, providing a globally better explanation of the input forms. The inference problem solved by the argument-only model is to find such optimal sets of tree fragments given an input corpus.

A similar pair of simplicity biases is used to define the distribution over modifiers. Recall that *SISTER-ADJOIN* inserts modifier category nodes into derivations and that these nodes are then filled using *SUBSTITUTE*. We place a Pitman-Yor process prior over the set of possible modifier node categories. This prior will bias the model towards using a small set of category types when sampling modifiers. For example, the modifier model might prefer to hypothesize that only adjective and adverb phrases are likely to be modifiers rather than adjective, adverb, noun, and verb phrases. A second simplicity bias favors inserting only a small number of modifiers into derivations. This bias is captured by the assumption that the probability of deriving a sequence of modifiers is the product of probabilities of the individual modifiers in this sequence. Because this product drops off geometrically in the number of modifiers, the model will prefer derivations which contain a small number of modifiers.

Applying all of the simplicity biases to sister-adjunction grammar, we arrive at the argument-modifier model. During inference, the argument-modifier model will attempt to find an optimal set of reusable argument-structure fragments by categorizing individual

nodes in input data trees as either (i) internal to a stored tree fragment, (ii) built by substitution of a lexical item at a frontier node, or (iii) built by sister-adjunction. In general, the model will categorize a node as a modifier when doing so will result in a simpler representation of the input corpus, that is, when it allows the input corpus to be explained using a smaller set of lexical items. Intuitively, the SISTER-ADJOIN operation allows the model to prune out constituents when doing so will lead to more compact and generalizable lexical items.

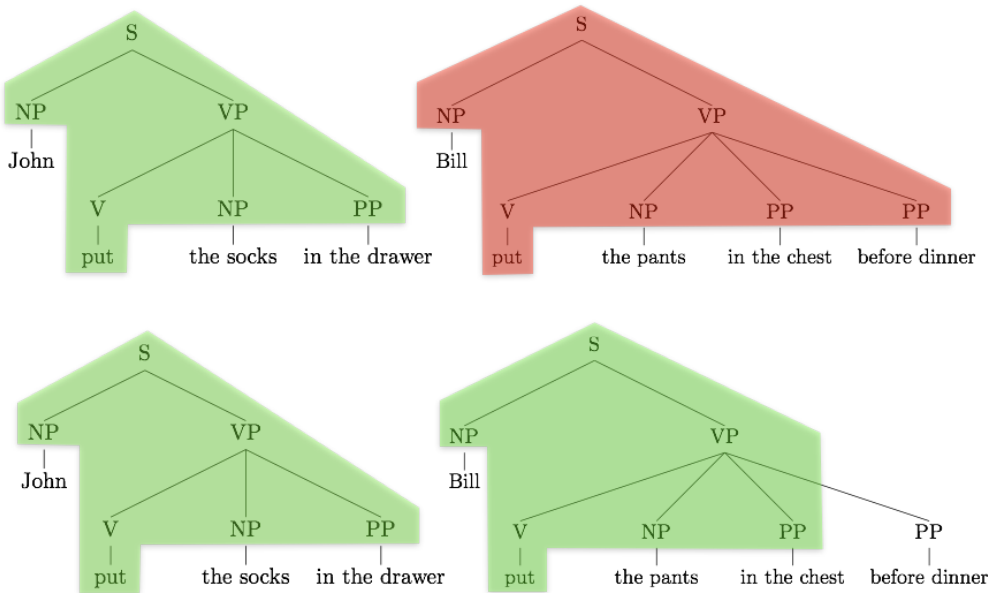


Figure 4: The argument-only model versus the argument-modifier model

Consider Figure 4. If a model posits that there are no modifiers in these sentences, then it will not identify the shared structure between two uses of the verb *put*, and will derive them using distinct sets of elementary trees, as on the top line of Figure 4. On the other hand, if it posits that the PP *before dinner* is a modifier, then it will be able to derive the core structure of these sentences using a single elementary tree, as on the bottom line of Figure 4. Nodes will be identified as modifiers when, like this PP, their removal from the sentence's argument structure leads to simpler derivations of the sentences in the corpus and greater amounts of sharing in the lexicon.

To perform inference, we developed a local Gibbs sampler which generalizes the one proposed by Cohn *et al.* (2010). This sampler jointly explores the space of elementary trees and substitution/adjunction attributions for a corpus consisting of parsed sentences. At each iteration, the sampler determines for each node in the corpus whether (i) the node is internal to an elementary tree, (ii) the node is the root of a tree which was inserted by substitution, or (iii) the node is the root of a tree which was inserted by sister-adjunction. The sampler randomly selects a node in the corpus and resamples its label from the full conditional posterior given the current hypothesis for the rest of the nodes in the corpus and the elementary tree set.

## SIMPLICITY AND EVALUATION METRICS

Before presenting our results in the next section, we make some observations about the relationship between our learning framework and the broader literature. The tradeoff-based approach that we adopt here can be understood as an instantiation of the classical linguistic notion of an *evaluation metric* (Chomsky 1951 [1979], 1955 [1975], 1964).<sup>8</sup> Although we make use of probability theory to capture our two kinds of simplicity, our framework is closely related to other approaches that operationalize simplicity using the idea of description-length or succinctness (e.g., Berwick 1982, 1985; Brent 1997, 1999; Cartwright and Brent 1994; De Marcken 1996a,b; Grünwald 2007; Hsu and Chater 2010; Hsu *et al.* 2011, 2013; Li and Vitányi 2008; Perfors *et al.* 2011; Phillips and Pearl 2014; Rissanen 1978; Stolcke and Omohundro 1994; Wolff 1977, 1980, 1982; Yang and Piantadosi 2022, *inter alia*). Such approaches go back at least to Chomsky (1951 [1979]) in linguistics and have been widely discussed in philosophy, statistics, and cognitive science, both with respect to their normative justification as well as their appropriateness for describing human psychology (see, Li and

---

<sup>8</sup> Also see discussion in Goldsmith (2011) and Rasin and Katzir (2016).

Vitányi 2008, for an overview of many historical threads in statistics, mathematics, and computer science).

Perhaps the most general treatment comes from the theory of Solomonoff induction, which uses a distribution over the set of all possible computer programs to define simplicity preferences related to those used in this work (Grünwald 2007; Li and Vitányi 2008; Rissanen 1978; Solomonoff 1978, 1964a,b). In this framework, theories (i.e. computable distributions over observations) are preferred when they are both simple to describe and provide simple descriptions of the data. It has been proven that this distribution can be used to asymptotically learn any computable theory, given a sufficient amount of data, and as a result it has been proposed as a universal, normative account of learning. The relation between this work and theoretical and empirical problems of language learning are also beginning to be understood in more detail (see, e.g., Hsu and Chater 2010; Hsu *et al.* 2011, 2013, for recent discussion). In cognitive science, there is a large and growing body of work suggesting that human inductive biases are captured by models making use of similar simplicity biases (see, e.g., Feldman 2000; Goodman *et al.* 2008; Piantadosi 2011, 2021, for examples from concept learning).

However, any formal definition of simplicity is dependent on the formalism, representation, or machine model with respect to which it is defined (Li and Vitányi 2008). Therefore, proposals about simplicity are substantive parts of theories of learning and must be evaluated together with other aspects of such theories. There remain several different frameworks implementing the simplicity-based approach – including the Bayesian framework, adopted here, and the minimum description length framework (Grünwald 2007; Rissanen 1978). It remains for future work to achieve a more fine-grained theoretical and empirical understanding of similarities and differences amongst various approaches to learning-via-simplicity.

In this section, we will use the computational models introduced above to evaluate two questions. First, do the statistics of natural language corpora provide evidence for the argument or modifier status of indi-

vidual phrases that is usable by a learner that optimizes a trade-off between lexical and derivational simplicity? Second, why is the argument-modifier model a superior model of the input data under these simplicity biases.

In order to address these questions, we will perform two sets of analyses. In the first, we will look at whether the argument-modifier model learns a distinction between arguments and non-arguments which agrees with a hand-annotated corpus. We will show that the argument-modifier model classifies arguments and non-arguments in a manner that aligns with traditional linguistic assumptions. Thus, we conclude that the argument/modifier status of individual phrases is evidenced in the input. This provides evidence in favor of both our formalization of the distinction and in favor of tradeoff-based learning.

In the second set of analyses, we will examine how the argument-modifier model infers the argument status of constituents using simplicity. We will show that the argument-modifier model learns a simpler representation of the input data than the argument-only model. We illustrate how this arises as a result of the representational and inferential assumptions discussed above.

### *Gold Standard Evaluation*

### 5.1

Our first set of analyses examine the ability of the argument-modifier model to correctly classify constituents as arguments or modifiers. As we discussed above, the model was designed to capture three differences between arguments and modifiers that affect their syntactic distribution: **obligatoriness/optionality**, **finiteness/iterability**, and **fixity/flexibility**. If the argument-modifier model is able to correctly distinguish modifier and argument phrases in the training corpus, we can conclude that these three distributional differences provide a signal to appropriately equipped learners.

We trained the argument-modifier model on sections 2–21 of the Wall Street Journal portion of the Penn Treebank (Marcus *et al.* 1999). The input consisted of approximately 40,000 parsed sentences, without any further annotations for argument or modifier status. Under the Penn Treebank’s tree annotation scheme, arguments and modifiers are not distinguished from each other by their hierarchical relations

in the parse tree (or in any other way). In particular, the arguments and modifiers of a phrase are most often siblings in the tree. Thus, the argument-modifier model could not directly use any information in the input corpus to simply read off each sentence's argument and modifier structure.

In order to evaluate the accuracy of the argument-modifier model classification of arguments and modifiers, we require a gold standard which provides annotations for arguments and modifiers in the Penn Treebank. Unfortunately, no such resource provides a classification of all nodes in the Penn Treebank (or CHILDES, MacWhinney 2000, which we use in our next study). However, for a subset of the phrases in the Penn Treebank, such information is available in the PropBank corpus (Palmer *et al.* 2005) which provides annotations of argument and modifier structure for all of the verbal predicates in the Wall Street Journal portion of the corpus. As noted in Palmer *et al.* (2005), the annotation of modifiers in PropBank is non-standard in certain cases. In particular, NEG and MOD categories are annotated as modifiers. We therefore exclude these categories from our analyses. PropBank does not annotate the arguments or modifiers of expressions which are not verbal predicates. Our model evaluations were performed by running the Gibbs sampler described in Section 3.2 for 100 iterations, and selecting the node labelings which were output on the final iteration.

For the purpose of our analyses, all sister-adjoined nodes are classified as modifiers, and all other nodes (i.e. nodes which are internal to an elementary tree or at the leaf of one) are classified as non-modifiers. We compared the model's labels to those provided by PropBank, on the subset of nodes for which PropBank provides annotations.

To show that differences in the distributions of argument and modifier phrases provide a valuable source of evidence for lexicon acquisition, we first establish that our model is able to correctly classify phrases at a rate which is better than chance. To demonstrate this, we computed the precision (i.e. number of correctly identified modifier nodes divided by the total number of modifier nodes identified by the model) and recall (i.e. number of correctly identified modifier nodes divided by the total number of modifier nodes in the gold-standard) of the model and compared it with two baselines. The first baseline randomly classifies each node as internal to an elementary tree, the leaf of an elementary tree, or a modifier with equal probability. Note



that prior to receiving any training data, the model has no information about which phrase types are likely to be modifiers and which are likely to be arguments. The random baseline therefore represents the model’s knowledge of the argument/modifier distinction prior to training, and any improvement in the model’s classification of modifiers must be attributed to information contained in the input data.

The second baseline treats every node as a modifier. We introduce this baseline in order to illustrate some basic facts about PropBank. Table 1 shows precision and recall in identifying the modifiers of verbal predicates in the corpus. The argument-modifier model is compared to three baselines: an all-modifier baseline, in which every node is labeled as a modifier, a random baseline, and a version of the model that does not use context to predict modifiers. PropBank annotated 179,058 nodes in the corpus for their argument/modifier status. These nodes represent approximately 10% of the total nodes in the corpus. Among the annotated nodes, 45,507 (25%) are modifiers, meaning that 25% of the guesses of the all-modifier baseline are correct.

Precision measures accuracy of modifier-predictions. Table 1 shows that the argument-modifier model is significantly more accurate than the random and the all-modifier baselines, demonstrating that the training data has provided information which allows the model to correctly classify many constituents.

Model	Precision	Recall	Accuracy
all-modifier	0.25	1	0.25
all-argument	N/A	0	0.75
random	0.29	0.23	0.66
<b>SAG</b>	<b>0.66</b>	<b>0.52</b>	<b>0.81</b>

Table 1:  
Precision and recall  
of the argument-modifier model

Recall measures the coverage of gold-standard modifier nodes achieved by the models. Again, the argument-modifier model achieved significantly higher coverage than the random baseline, indicating that the training data contains enough information to increase the number of true modifiers that the model recognizes.

In order to better understand what the argument-modifier model learned about the modifiers of verbal predicates, the evaluations against PropBank were further broken down by the category of the

Table 2:  
Labelings for modifiers  
of VP nodes, broken down  
by child category

VP Parent				
Child category	Model	Precision	Recall	PropBank
ADVP	random	0.95	0.23	12,385
ADVP	SAG	0.95	0.47	12,385
NP	random	0.04	0.23	3,345
NP	SAG	0.47	0.57	3,345
PP	random	0.49	0.22	18,841
PP	SAG	0.56	0.54	18,841
SBAR	random	0.40	0.22	4,552
SBAR	SAG	0.84	0.63	4,552

modifier. Table 2 shows the results for the phrase types which occur most frequently as verbal modifiers: adverb phrases (ADVPs), noun phrases (NPs), prepositional phrases (PPs), and subordinate clauses (SBARs). Together these categories of constituent account for more than 85% of the modifiers in the training corpus.

For the phrase categories of adverb phrases (ADVPs) and prepositional phrases (PPs), the model doubles the recall of the random baseline, and roughly maintains its baseline precision. Adverb phrases are typical modifiers when they appear within a verb phrase (VP). Out of 13,197 ADVPs annotated by PropBank, 12,384 are modifiers. Prepositional phrases are also frequently modifiers when they appear in this context. Out of 38,861 PPs annotated by PropBank, 18,839 are modifiers. The increase in the model's recall therefore indicates that the model learned to correctly classify many of these ADVP and PP modifiers.

In contrast to adverb and prepositional phrases, noun phrases (NPs) which appear within verb phrases are typically arguments to the verb. Out of 92,965 NPs annotated by PropBank, only 3,306 appear as modifiers. Exceptions to this generalization are cases where a noun phrase is used as an adverbial modifier, such as the noun phrase *last night* in *They played the game last night*. The precision of the model increased by a factor of 10 for NPs, indicating that it incorrectly classified many fewer non-modifier NPs. In addition, the model's precision more than doubles the baseline.

Phrases belonging to the category of subordinate clauses SBAR can serve either as arguments or modifiers. For example, in the sentence

*John said that he would be late*, the subordinate clause *that he would be late* is an argument of the verb *said*. By contrast, in the sentence *The woman laughed when she heard the joke*, the clause *when she heard the joke* is a temporal modifier of the verb *laughed*. Out of 13,617 SBAR phrases annotated by PropBank, 4,551 are modifiers. The model's precision and recall on SBAR phrases was more than twice that of the random baseline, showing that the model classified fewer clausal arguments as modifiers, and correctly identified a greater number of clausal modifiers.

As we mentioned above, certain categories of constituents have highly stereotyped argument-modifier status when they appear as children of other categories. For example, adverb phrase (ADVP) children of verb phrases (VP) and adjective phrase (JJ) children of noun phrases (NP) are both typically modifiers of their parent constituents.

PropBank only provides argument-modifier annotations for the children of verb phrases (VP), and therefore we do not have a gold standard for modifiers occurring outside of VPs. Nonetheless, it is possible to use the stereotyped behavior of these categories to examine the model's performance on the children of non-VP nodes. Tables 3–6 show the model's classification of constituents which were children of sentence-level constituents (S), prepositional phrases (PPs), noun phrases (NPs), and subordinate clauses (SBARs), respectively. In each of these cases, the category of child constituents is highly indicative of their argument-modifier status.

For sentence-level (S) constituents, we analyzed three categories of child phrase: noun phrases (NPs), verb phrases (VPs), and (ADVPs). These are the three most common categories which have stereotyped argument/modifier behavior when they appear as children of nodes. Of these three phrase types, noun and verb phrase are typically not modifiers, whereas adverb phrases typically are. For example, in *Usually, John wears a coat*, the adverb *Usually* is a modifier of the sentence while *John* and *wears a coat* are not modifiers. Table 3 shows how often the model labeled the children of sentence-level constituents as modifiers nodes. The model accords with intuition here, most often labeling adverb phrases but not noun or verb phrases as modifiers.

For prepositional phrases (PPs), we considered four categories of child constituent: adverb phrases (ADVPs), noun phrases (NPs), prepositions (INs), and *to* (TOs). Of these phrase types, only adverb phrases

Table 3:  
Labels  
for children  
of S nodes

S parent				
Child category	Model	#Guessed	Corpus total	Typically modifier
ADVP	random	1,393	6,063	Y
ADVP	SAG	2,331	6,063	Y
NP	random	16,654	93,076	N
NP	SAG	1,738	93,076	N
VP	random	16,005	89,984	N
VP	SAG	572	89,984	N

typically modify the parent prepositional phrase. For example, in the prepositional phrase *immediately after the opening*, the adverb phrase *immediately* is a modifier while the prepositional head *after* and noun phrase *the opening* are not. In accord with these intuitions, Table 4 demonstrates that the model classifies most adverb phrase children of prepositional phrases as modifiers, but treats prepositional heads and noun phrases as non-modifiers.

Table 4:  
Labels  
for children  
of PP nodes

PP parent				
Child category	Model	#Guessed	Corpus total	Typically modifier
ADVP	random	216	1,109	Y
ADVP	SAG	547	1,109	Y
IN	random	13,972	83,848	N
IN	SAG	672	83,848	N
NP	random	15,060	88,556	N
NP	SAG	496	88,556	N
TO	random	1,484	8,654	N
TO	SAG	64	8,654	N

We considered four categories of constituents for noun phrases (NPs): determiners (e.g., *the*, *a*; DT), adjectives (JJ), other noun phrases, and prepositional phrases. Determiners are unlikely to modify noun phrases, while adjectives typically do modify them. For example, in the noun phrase *the big chair*, the determiner *the* is not a modifier, while the adjective *big* modifies the noun *chair*. Prepositional phrases are often modifiers (e.g., in *the resort by the sea*, the prepositional phrase *by the sea* modifies the noun *resort*), although in some cases,

NP parent				
Child category	Model	#Guessed	Corpus total	Typically modifier
DT	random	15,791	77,553	N
DT	SAG	1,701	77,553	N
JJ	random	10,544	45,812	Y
JJ	SAG	9,717	45,812	Y
PP	random	7,652	43,420	Y
PP	SAG	3,226	43,420	Y

Table 5:  
Labels  
for children  
of NP nodes

such as deverbal nominalizations, they are typically treated as arguments of the head noun (e.g., in the noun phrase *the destruction of the city*, the prepositional phrase *of the city* is an argument of the head noun; see, e.g., Chomsky 1970).

Table 5 shows the modifier-classification rates of noun phrase children. The model correctly identifies determiners as non-modifiers. However, for adjectives, the most prototypical modifiers of noun phrase, the model's performance is weaker: The number of JJs classified as modifiers is approximately the same as the random baseline. The number of PPs classified as modifiers decreased by more than half relative to the random baseline, though the implications of this are unclear: As discussed above, PPs appear frequently as the modifiers of noun phrases but also as arguments. It should be noted that the Penn treebank is notorious for having many complex noun phrases consisting of long sequences of noun compounds annotated with a single flat structure. This likely affected the ability of the model to distinguish amongst the children of NP nodes.

SBAR parent				
Child category	Model	#Guessed	Corpus total	Typically modifier
S	random	4,873	29396	N
S	SAG	101	29396	N
WHADVP	random	421	2521	N
WHADVP	SAG	38	2521	N
WHNP	random	1,383	8505	N
WHNP	SAG	79	8505	N

Table 6:  
Labels  
for children  
of SBAR nodes

The category SBAR is used to mark subordinate clauses in the Penn treebank. Here we consider the following categories of children: sentence-level constituents (Ss) and wh-expressions (WHADVPs and WHNPs) which are used to introduce subordinate clauses (e.g., the word *when* in the sentence *The woman laughed when she heard the joke*). None of these types of constituent is typically thought of as modifying subordinate clauses. Table 6 shows, consistent with this intuition, that the model treats all three categories as non-modifiers.

### 5.1.1

### Discussion

We have presented two sets of results in this section. First, we have shown that the argument-modifier model's accuracy at classifying arguments and non-arguments substantially improves over a random baseline. Second, we have shown that among phrases that are not labeled in the gold standard (i.e., all phrase types but verb phrases), the argument-modifier model learns an argument/non-argument classification which appears linguistically reasonable for most major phrasal categories.

These results have two consequences for the arguments in this paper. The argument-modifier model is built on the assumption of three distributional differences between lexical argument-structure-derived phrases and modifier phrases: **finiteness** v. **iterability**, **obligatoriness** v. **optionality**, and **structural fixity** v. **structural flexibility**. Since the argument-modifier model made use of these properties in order to classify phrases in the input corpus as arguments or non-arguments, its performance on the gold standard shows that we have captured some linguistically relevant properties of arguments and modifiers using these properties.

The results also show that the distributional information contained in the input corpus is often sufficient for recovering the argumenthood of specific constituents. The argument-modifier model does not have any a priori knowledge about which types of phrases are likely to be arguments, and it leverages only distributional information in order to infer the status of individual phrases. Thus, its performance in categorizing arguments and non-arguments must be attributable to the distributional information contained in the corpus. This distributional information is leveraged by the model by trading

off the simple lexicon and simple derivation biases. We note that our study is an example of an *ideal learner analysis* (Pearl and Goldwater 2016); that is, the model is highly idealized and not intended to veridically represent a child language learner. Therefore, our results do not demonstrate that children use distributional information to identify the argument or modifier status of individual constituents. Instead, they indicate that such information would be available to any learner that made similar assumptions about the relationship between simplicity and learning.

### *Lexicon learning, arguments structure, and simplicity*

5.2

In the previous section, we showed that the argument-modifier model is able to correctly recover the modifier status of many constituents using only the pattern of co-occurrences between constituents in the training set. In this section, we show how this performance is the result of the simplicity biases outlined in Section 3. As discussed in that section, our framework makes use of two competing simplicity biases. The *simple lexicon bias* favors small numbers of highly reusable lexical items and the *simple derivation bias* favors derivations of individual forms using small numbers of lexical items. Typically, these two biases lead to a tradeoff. Smaller, more reusable lexical items mean more complex derivations and vice versa. However in this section, we present simulations demonstrating that compared to the argument-only model, the argument-modifier model learns a more compact generalizable lexicon, while also providing simpler derivations for individual forms. If we fix a particular dataset as well as fix a particular model (argument-only model or argument-modifier model and all parameters), there is at least one optimal (i.e., most probable), solution for that model-dataset combination. Any lexicon/derivation set that increases one kind of simplicity with respect to this optimum will necessarily decrease the other. Thus, our results show that the argument-modifier model is overall a better model for the data since it is able to find a solution which is superior under both measures.

To see how it is possible that the argument-modifier model is able to optimize *both* kinds of simplicity simultaneously consider a verb phrase (VP) headed by a verb like *put*. In simplest form, *put* requires

two VP-internal arguments – a noun phrase (NP) expressing the object which was put somewhere, and a prepositional phrase (PP) expressing the destination – *put his socks in the suitcase*. Across particular uses of this simple *put*-construction, the VP node will often have the following sequence of children: V NP PP. However, because modifiers are optional, iterable, and appear at a variety of positions within a constituent, they can greatly increase the number of different observed sequences of children of the VP node: *put his socks suddenly in the suitcase* [V NP ADVP PP], *put his socks in the suitcase suddenly without warning* [V NP PP ADVP PP], etc. The argument-modifier model is able to explain away the presence of these additional phrases using the SISTER-ADJOIN operation, and is driven to do so because this leads to a lexicon of argument-structure fragments and derivations of individual forms which better optimizes both simplicity biases.

In the analyses in this section, we provide empirical support for this argument. To demonstrate the point, we show that the argument-modifier model can account for the same data as the argument-only model with a more compact lexicon and simpler derivations of each sentence. We show this on both training and holdout data drawn from two corpora: the Wall Street Journal portion of the Penn Treebank, and the Brown (1973) portion of the CHILDES database (MacWhinney 2000). For the WSJ, the model was trained on the 40,000 parsed sentences from sections 2-21 (the same sentences that were used in the gold standard analyses). The CHILDES sections used here consist of approximately 30,000 child-directed utterances which were recorded between ages 1;6 to 5;1. Sentence fragments and *wh*-questions were excluded from our analyses, though the results do not differ substantially when fragments and questions are included.

The training regime was the same as in the gold standard analysis: The models received parse trees for each sentence as input. Because the CHILDES database does not provide parses, we used the corpus of parsed CHILDES sentences developed by Pearl and Sprouse (2013). We include these CHILDES analyses below because it is more likely than newspaper text to be representative of the input received by a typical natural language learner. Note that we did not include the CHILDES corpus in the previous evaluation because we do not have gold-standard annotations of the argument/modifier status of any phrases in this corpus. The differences between the two corpora



can be illustrated by several simple statistics. On average, the sentences in the WSJ corpus contain 25 words, while the sentences in CHILDES contain 6.5 words. The parse trees in the WSJ contain 71 nodes on average, while those in CHILDES contain 19 nodes. Finally, the average maximum depth (i.e., the longest distance from the root node to a leaf) of the parse trees in the WSJ is 10, while the average depth in CHILDES is 5. These statistics show that the sentences in the WSJ are significantly longer and more syntactically complex than those in CHILDES.

In this section, we compare the ways in which the argument-modifier model and the argument-only model represent the input training data for the WSJ and CHILDES corpora. We first examine the bias for reusable lexical items. Figure 5 shows the cumulative frequencies of the 1,000 most often stored tree fragments in the lexicons of the argument-modifier model and argument-only model, as learned on the CHILDES (left) and WSJ (right) training sets. We computed these values by first ranking tree fragments by frequency of occurrence in the lexicon; this resulted in a rank for each type of tree fragment, with lower rank corresponding to greater frequency. Then, for all tree fragments below a given rank (e.g., for the tree fragments below rank 100, corresponding to the 100 most common tree fragments), we computed the sum of the frequencies of these fragments.<sup>9</sup> The figure shows that the most reusable tree fragments learned by the argument-modifier model are used more often across sentences in the training corpus than the most reusable tree fragments learned by the argument-only model. The difference is more pronounced in the WSJ training set. This is likely due to the greater sentence complexity and greater number of modifiers in newspaper text compared to child-directed speech.

---

<sup>9</sup>Tree fragments which were rooted at part-of-speech nodes (pre-terminals) were excluded from this and subsequent analyses. A subtree which is rooted at a part of speech necessarily consists of exactly two nodes (the part of speech and the terminal string which it is a parent of). As a result, there is only one parse for such subtrees, and both models will always parse such a subtree in an identical manner.

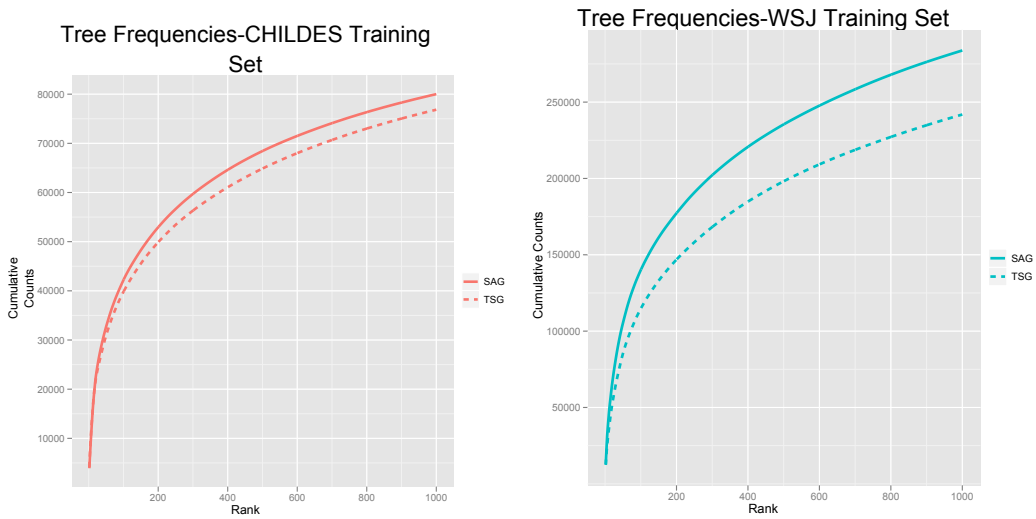


Figure 5: Cumulative Frequencies

We next examine which model was able to provide simpler derivations of individual sentences. One way to measure this is to look at the complexity of stored tree fragments learned by each model. If a model stores tree fragments which are larger (on average), then it must account for each sentence using fewer fragments (on average). Figure 6 shows the cumulative average number of nodes (left) and average depth (right) of the 1,000 most common elementary trees learned by the argument-modifier model and the argument-only model on the CHILDES and WSJ corpora. These figures show that the elementary trees learned by the argument-modifier model are more complex than those learned by the argument-only model, and therefore that the derivation of individual sentences involve fewer lexical items (on average). The difference in tree fragment complexity is greater for the WSJ corpus than for CHILDES, most likely because the parse trees in the WSJ corpus contain a greater number of nodes and have greater depth than those in CHILDES.

Figure 7 shows the cumulative proportion of nodes in the training corpus which are accounted for by the 1,000 most common stored tree fragments learned by the argument-modifier model and the argument-only model. Because it learns both more reusable and larger stored



Figure 6: Complexity of stored tree fragments



Figure 7: Cumulative coverage

tree fragments, the argument-modifier model is able to account for the training data using a smaller number of stored items.

### 5.2.2 Lexical and derivational simplicity in new sentences

The previous analyses demonstrate that the argument-modifier model learns a more parsimonious representation of the input than the argument-only model. An important caveat, however, is that the argument-modifier model is a more complex grammatical formalism than the argument-only model. Whereas the argument-only model only has a single composition operation (SUBSTITUTE), the argument-modifier model has two composition operations (SUBSTITUTE and SISTER-ADJOIN). This means that the model has more degrees of freedom in explaining an input training set. As a result, it is possible that the argument-modifier model's performance is due to *overfitting*. A standard method to diagnose overfitting is to evaluate the model on novel data. If the model is overfit on the training data, then it will have captured spurious regularities in its input, and will therefore generalize poorly to new data.

In order to determine whether the parsimony advantages of the argument-modifier model generalize to novel data, we divided the CHILDES and WSJ corpora into training and test portions. The training portion was used as input to the argument-modifier model and argument-only model, while the test portion was used for evaluating the generalizability of these grammars. For the WSJ corpus, we used the standard split: training on sections 2–21 and testing on section 23. For the CHILDES corpus, we randomly selected 80% of the sentences for training, and used the remaining 20% for test.

Our evaluation of the argument-modifier model follows the method in the previous section. We compare the argument-modifier model to the argument-only model, and conduct similar analyses of fragment reusability and derivation complexity (fragment size). In order to perform these analyses, we applied our sampler to the test portions of the two corpora without incorporating any new tree fragments into the set of learned tree fragments. That is, after training we froze the set of lexical fragments (and associated counts) and did not allow any learning from the test set during inference. Thus each sentence in the test corpus was analyzed as if it were the next observed

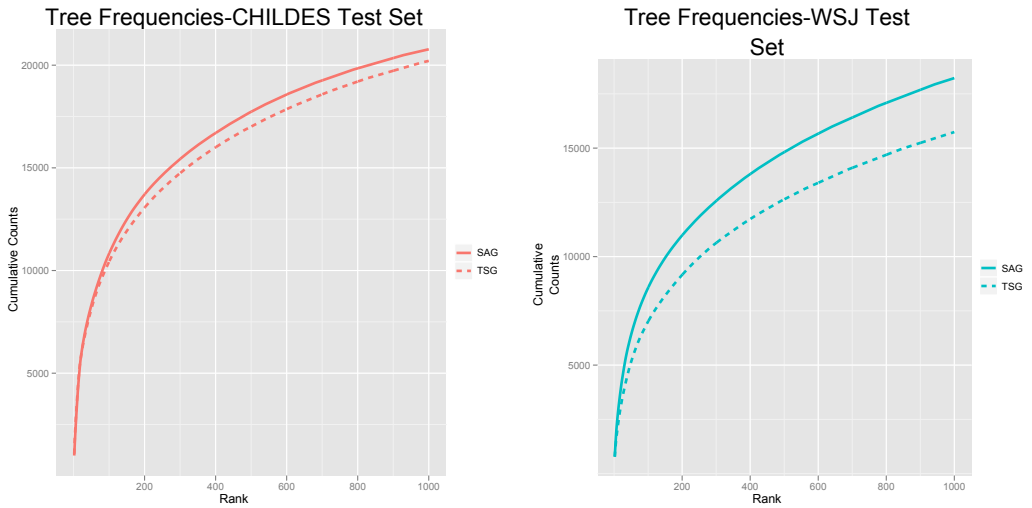


Figure 8: Cumulative frequencies (generalization)

sentence after training. The analyses below are otherwise identical to those in the previous section.

We first examine the bias for reusable lexical items. Figure 8 shows the cumulative frequencies of the 1,000 most common tree fragments from the lexicons of the argument-modifier model and the argument-only model, as inferred on the CHILDES (left) and WSJ (right) test sets. The figure shows that the commonly stored tree fragments learned by the argument-modifier model are used more often across sentences in the test corpus. The difference is again more pronounced in the WSJ test set due the greater sentence complexity and number of modifiers in newspaper text compared to child-directed speech.

We next turn to the bias for simple derivations of individual sentences. As in the previous simplicity analyses, we measure derivation complexity by examining the size of tree fragments used to account for test sentences. Larger tree fragments imply fewer fragments per derivation. Figure 9 shows the cumulative average number of nodes (left) and average depth (right) of the 1,000 most common elementary trees used to account for the new sentences by the argument-modifier model and the argument-only model on the CHILDES and WSJ test corpora. These figures show that the elementary trees learned by the

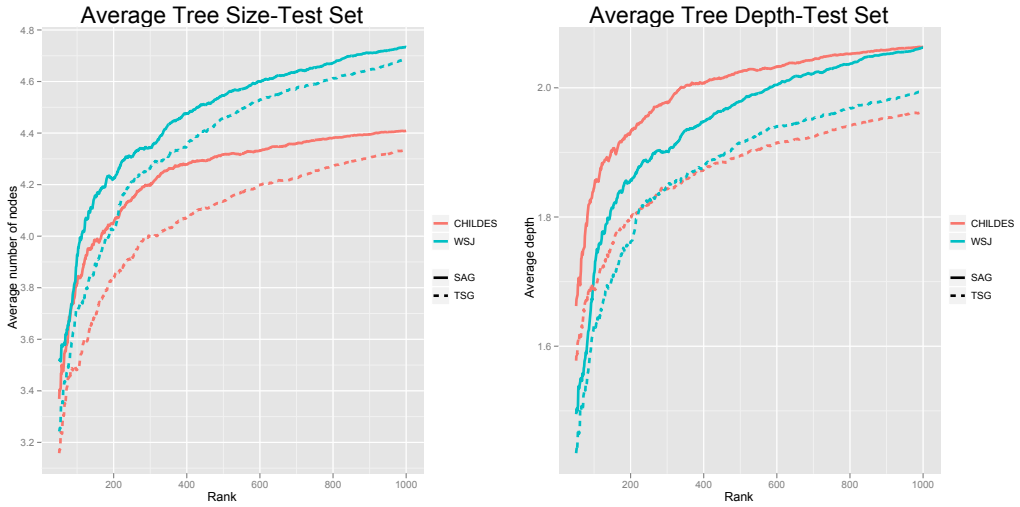


Figure 9: Complexity of stored tree fragments (generalization)

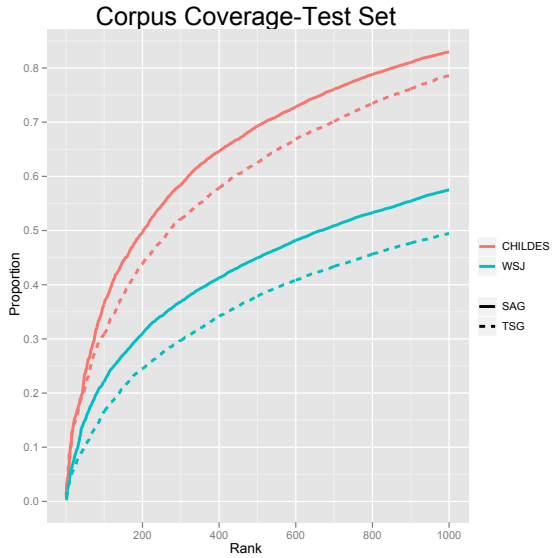


Figure 10: Cumulative coverage (generalization)

argument-modifier model are more complex than those learned by the argument-only model and, therefore, that the derivations of individual sentences are simpler.

Thus, the argument-modifier model's advantage in both kinds of simplicity transfers to the case of new sentences. This is further confirmed in Figure 10 which shows the cumulative proportion of nodes in the training corpus that are accounted for by the 1,000 most common stored tree fragments learned by the two models.

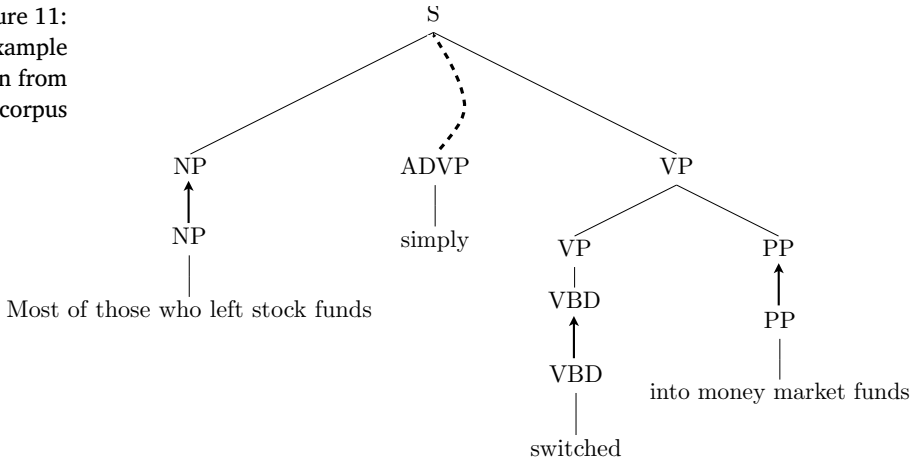
#### Discussion

#### 5.2.3

The preceding analyses indicate that the sister-adjunction model is able to learn both more reusable lexical items, and simpler derivations of each sentence than the tree-substitution model. As we discussed previously, the inference performed in learning the set of lexical fragments for the argument-only model can be understood in terms of a tradeoff. All else being equal, smaller tree fragments are more reusable, leading to smaller lexica. However, larger tree fragments lead to simpler derivations, since fewer are needed per derivation. For a given corpus, if a particular model is at or near an optimum, increasing the reusability of lexical items in an otherwise optimal model will necessarily increase the complexity of derivations, and decreasing the complexity of derivations will necessarily increase the size of the lexicon. Nevertheless, the argument-modifier model is better in both simplicity measures, indicating that it provides a globally superior account of the input data by learning a smaller lexicon of larger tree fragments.

To understand these results better, again consider the example sentences in Figure 4. The argument-modifier model is able to use a single elementary tree (stretching from the root S node to the verb *put*) to derive the core of both sentences. In contrast, as Figure 4 shows, the argument-only model will require two distinct elementary trees, one with three arguments under the VP node (for the first sentence) and one with four arguments (for the second). Thus, because the argument-modifier model can compose an optional PP such as *at 5 o'clock* separately from a sentence's core argument structure, it can re-use the same elementary tree to derive a greater number of sentences in the corpus. This explains how the argument-modifier model can use

Figure 11:  
Example  
derivation from  
the WSJ corpus



its sister-adjunction operation to find more reusable elementary trees than the argument-only model. It is driven to do so by the prior preference for a smaller lexicon.

Figure 11 illustrates a representative example from the WSJ corpus. By using SISTER-ADJOIN to account for the ADVP node separately from the rest of the sentence’s derivation, the argument-modifier model was able to use a common depth-three elementary tree to derive the backbone of the sentence. By contrast, the argument-only model must include the ADVP node in an elementary tree; this elementary tree is much less common in the corpus.

6

CONCLUSION

In this paper, we have studied the learnability of the argument-modifier status of phrases. We have formulated the distinction as one between lexical and non-lexical modes of composition which give rise to three differences between the two types of constituents which have distributional consequences: **iterability v. finiteness**), **optionality v. obligatoriness**, and **structural flexibility v. structural fixity**. We then proposed that the modifier or argument status of individual phrases could be learned based on optimizing a tradeoff between two competing biases: the *simple lexicon bias* and the *simple derivation bias*.



Our first set of gold standard results indicate that our formalization of the distinction accords with the traditional distinction between arguments and modifiers. Furthermore, our results show that linguistic input contains a strong distributional signal to the modifier/argument status of individual phrases – at least for a learner making use of the tradeoff between lexical and derivational simplicity.

Our second set of results illustrate why the argument-modifier model is able to identify the status of individual phrases. The addition of the SISTER-ADJUNCTION operation allows the model to put derivations into a kind of normal form for which the optimal lexicon contains of both more complex and more reusable fragments. Thus, the argument-modifier model achieves a greater degree of lexical and derivational simplicity simultaneously.

Taken together, these results show that there is considerable distributional evidence for the traditional argument-modifier distinction, but that a simplicity-based learner equipped with lexical and extra-lexical modes of composition could make use of this evidence to acquire the pattern of arguments and modifiers in their language. This result is complementary to traditional linguistic argumentation about the distinction. Our formulation of the problem has deliberately ignored any semantic or, in fact, any non-distributional aspects of the argument-modifier distinction. Any such systematically correlated information should only make the learning problem easier.

## APPENDIX

### FORMALIZATION OF THE MODELS

The argument-modifier model extends earlier work on induction of Bayesian TSGs (Cohn *et al.* 2010; O'Donnell 2011, 2015; O'Donnell *et al.* 2011; Post and Gildea 2009). The Pitman-Yor Process allows the complexity of the lexicon to grow with more input sentences, while still enforcing a bias for more compact lexicons (Pitman and Yor 1995). As discussed in Section 3.1, the model has two components: (i) A distribution over elementary trees, similar to earlier models of Bayesian TSG induction, and (ii) a distribution over modifiers.

Algorithm 1 provides pseudocode for the generative model. Note that throughout, we will use the notation  $c_p$  to refer to the nonterminal category of a node  $p$ .

For each node  $p$ , the distribution over elementary trees rooted at that node is given by:

$$(1) \quad G_{c_p} | a_{c_p}, b_{c_p}, P_E \sim \text{PYP}(a_{c_p}, b_{c_p}, P_E(\cdot | c_p))$$

where  $P_E(\cdot | c_p)$  is a context free distribution over elementary trees with root label  $c_p$ . The hyperparameters  $a_{c_p}, b_{c_p}$  are set to  $a_{c_p} = 0, b_{c_p} = 1$  for this paper.<sup>10</sup>

The context-free distribution over elementary trees  $P_E(e|c)$  is defined by:

$$(2) \quad P_E(e|c) = \prod_{i \in I(e)} (1 - s_{c_i}) \prod_{f \in F(e)} s_{c_f} \prod_{c' \rightarrow \alpha \in e} P_{\text{cfg}}(\alpha | c')$$

where  $I(e)$  is the set of internal nodes in  $e$ ,  $F(e)$  is the set of frontier nodes,  $s_c$  is the probability that we stop expanding at a node labeled  $c$ , and  $P_{\text{cfg}}(\alpha | c')$  is the probability of the context-free rule expanding category  $c'$  to the sequence  $\alpha$ ,  $c' \rightarrow \alpha$ . For this paper, the parameters  $s_c$  are set to 0.5. The distribution  $P_{\text{cfg}}(\alpha | c')$  is defined using a distribution that is similar to the Infinite PCFG (Finkel *et al.* 2007; Liang *et al.* 2007),<sup>11</sup> which provides a Dirichlet process prior for PCFG rules.<sup>12</sup>

<sup>10</sup> Given these parameter values, the prior reduces to the model known as a Dirichlet process. Since our implementation allows for other values of  $a$  we present the more general version of the mathematics.

<sup>11</sup> Our base distribution over PCFG rules differs from the Infinite PCFG as presented in Liang *et al.* (2007) in a number of ways. First, rather than being a hierarchical Dirichlet process model, our set of nonterminal categories is fixed to be equal to the set of nonterminal categories in the treebank. Second, our rules are not fixed to be in Chomsky normal form, but rather the length of the right-hand side of each rule is sampled from a geometric distribution, and each child symbol is drawn conditioned on the parent symbol and then entire left context of the symbol, which is backed-off using the scheme of Teh (2006) and Goldwater *et al.* (2006).

<sup>12</sup> We use this nonparametric prior so that in addition to learning a distribution over elementary trees, we can also learn a distribution over context-free rules. The inferred distribution over context-free rules may substantially differ from the maximum-likelihood estimate derived from the corpus, as nodes that the model

$\beta \sim \text{Dir}(1, \dots, 1)$  [draw prior over nonterminals]

for each nonterminal sequence  $c_1, \dots, c_n$ :

$P_{\text{rhs}}(c_1, \dots, c_n) = \frac{1}{2^n} \prod_i \beta_{c_i}$  [define base distribution for pcfg prior]

for each nonterminal  $c$ :

$P_{\text{cfg}}(\cdot|c) \sim \text{DP}(a, P_{\text{rhs}}(\cdot))$  [draw distributions over CF rules]

for each nonterminal  $c$ :

for each elementary tree  $e$  rooted at  $c$ :

$F(e) =$  frontier of  $e$ ,  $I(e) =$  interior nodes of  $e$

$P_E(e|c) = \prod_{i \in I(e)} (1 - s_{c_i}) \prod_{f \in F(e)} s_{c_f} \prod_{c' \rightarrow \alpha \in e} P_{\text{cfg}}(\alpha|c')$

$G_c \sim \text{PYP}(a_c, b_c, P_E(\cdot|c))$  [draw distributions over elementary trees]

$\theta \sim \text{Dir}(1, \dots, 1)$  [draw base distribution over nonterminals]

for each sequence of nonterminals  $C = q_l, \dots, q_1$ : [draw modifier distributions]

if  $\text{length}(C) = 1$

$H_C \sim \text{DP}(\alpha, \text{Multinomial}(\theta))$

else

$H_C \sim \text{DP}(\alpha, H_{C'})$ , where  $C' = q_{l-1}, \dots, q_1$

for each node  $f$  on the frontier of the parse tree:

$e \sim G_{c_f}$  [sample an elementary tree rooted at category  $c_f$ ]

substitute  $e$  at  $f$

for each internal node  $p$  in  $e$ :

for each argument child  $d_i$  of  $p$ :

$j = 1$

$C = c_{d_1}, s_{1,1}, \dots, c_{d_i}, c_p$  [C is the context for  $d_i$ ]

$s_{i,j} \sim H_C$  [draw from the modifier distribution for  $d_i$ ]

while  $s_{i,j} \neq \text{STOP}$  [continue until drawing a STOP symbol]

sister-adjoint a node labeled  $s_{i,j}$  between  $d_i, d_{i+1}$

$j + 1$

$C = c_{d_1}, s_{1,1}, \dots, c_{d_i}, s_{i,1}, \dots, s_{i,j-1}, c_p$  [add sampled modifier to the context]

$s_{i,j} \sim H_C$

A similar base distribution for elementary trees is used in Cohn *et al.* (2010) and Post and Gildea (2009). The base distribution over elementary trees thus will be biased towards small elementary trees which use frequent context-free expansions.

In addition to defining a distribution over elementary trees, we also define a distribution which governs modification via sister-adjunction. To sample a modifier, we first decide whether or not to sister-adjoin into location  $l$  in a tree. Following this step, we sample a modifier category (e.g., a PP) conditioned on the location  $l$ 's context: its parent and left siblings. Because contexts are sparse, we use a backoff scheme based on hierarchical Dirichlet processes similar to the ngram backoff schemes defined in Teh (2006) and Goldwater *et al.* (2006). Let  $e$  be an elementary tree that has been substituted into the parse tree, and let  $p$  be an internal node in  $e$ . The node  $p$  will have  $n \geq 1$  children derived by argument substitution:  $d_1, \dots, d_n$ . In order to sister-adjoin between two of these children  $d_i, d_{i+1}$ , we recursively sample nonterminals  $s_{i,1}, \dots, s_{i,k}$  until we sample a *STOP* symbol:

$$(3) \quad P_a(s_{i,1}, \dots, s_{i,k}, STOP | C_0) = \left( \prod_{j=1}^k P_a(s_{i,j} | C_j) \right) \cdot P_a(STOP | C_{k+1})$$

where  $C_j = c_{d_1}, s_{1,1}, \dots, c_{d_i}, s_{i,1}, \dots, s_{i,j-1}, c_p$  is the context for the  $j$ th modifier between these children. The distribution over sister-adjoined nonterminals is defined using a hierarchical Dirichlet process to implement backoff in a prefix tree over contexts. Given the context  $C = q_l, \dots, q_1$  (where  $l > 1$ ), we define the distribution  $H_C$  over sister-adjoined nonterminals  $s_{i,j}$  by:

$$(4) \quad H_C \sim DP(\alpha, H_{C'}),$$

where  $C' = q_{l-1}, \dots, q_1$ . A sample is drawn from the root of the hierarchy when the context  $C$  is of length 1 (and hence the backed-off context is empty). A Dirichlet-multinomial distribution is used as the

---

labels as modifiers are not included in the derivation of an elementary tree. This approach is also suitable to the unsupervised setting (as in Cohn *et al.* 2010), in which the derived trees in the corpus are not observed.

prior in this case:

$$\theta \sim \text{Dir}(1, \dots, 1)$$

$$H_C \sim \text{DP}(\alpha, \text{Multinomial}(\theta))$$

where  $C = q_1$  and  $\theta$  is a vector with entries for each nonterminal and an entry for the *STOP* symbol. The backoff scheme for sampling modifiers is illustrated in Figure 12.

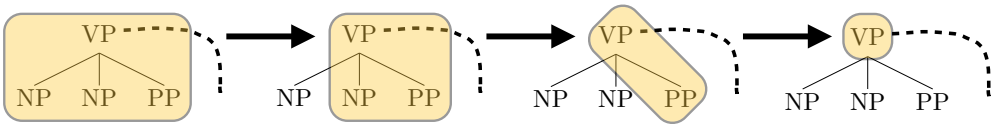


Figure 12: This illustrates the procedure for sampling a modifier at the right edge of a VP. The distribution over modifiers is conditioned on the modifier's context, which contains its VP parent and left siblings, as illustrated on the left of the figure. This distribution is estimated by successively backing off to smaller contexts

## REFERENCES

- Leon BERGEN, Edward GIBSON, and Timothy J. O'DONNELL (2015), A learnability analysis of argument and modifier structure, *lingbuzz*, (lingbuzz/002502).
- Robert C. BERWICK (1982), *Locality principles and the acquisition of syntactic knowledge*, Ph.D. thesis, Massachusetts Institute of Technology.
- Robert C. BERWICK (1985), *The acquisition of syntactic knowledge*, The MIT Press, Cambridge, Massachusetts and London, England.
- Rens BOD (1998), *Beyond grammar: An experience-based theory of language*, CSLI Publications, Palo Alto, CA.
- Rens BOD, Remko SCHA, and Khalil SIMA'AN, editors (2003), *Data-oriented parsing*, CSLI, Palo Alto, CA.
- Robert D. BORSLEY (1999), *Syntactic theory: A unified approach*, Edward Arnold, London, England.
- Michael R. BRENT (1997), Toward a unified model of lexical acquisition and lexical access, *Journal of Psycholinguistic Research*, 26(3):363–375.

Michael R. BRENT (1999), An efficient, probabilistically sound algorithm for segmentation and word discovery, *Machine Learning*, 34:71–105.

Joan BRESNAN (2001), *Lexical functional syntax*, Wiley-Blackwell, Oxford, England.

Roger W. BROWN (1973), *A first language: The early stages*, Harvard University Press, Cambridge, MA.

Timothy A. CARTWRIGHT and Michael R. BRENT (1994), Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition, in *Proceedings of the 16th Annual Meeting of the Cognitive Science Society*.

David CHIANG (2000), Statistical parsing with an automatically-extracted tree adjoining grammar, in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.

David CHIANG and Daniel BIKEL (2002), Recovering latent information in treebanks, in *Proceedings of COLING 2002*.

Noam CHOMSKY (1951 [1979]), *Morphophonemics of modern Hebrew*, Garland Publishing, New York, NY.

Noam CHOMSKY (1955 [1975]), *The logical structure of linguistic theory*, Plenum Press, New York, NY.

Noam CHOMSKY (1964), *Current issues in linguistic theory*, Janua Linguarum: Studia Memoriae Nicolai van Wijk Dedicata, Mouton, The Hague, The Netherlands.

Noam CHOMSKY (1970), Remarks on nominalization, in Roderick JACOBS and Peter ROSENBAUM, editors, *Readings in English Transformational Grammar*, Ginn and Company, Waltham, MA.

Noam CHOMSKY (1993), A minimalist program for linguistic theory, in Kenneth L. HALE and Samuel Jay KEYSER, editors, *The View from Building 20: Essays in Honor of Sylvain Bromberger*, pp. 1–52, The MIT Press, Cambridge, Massachusetts and London, England.

Noam CHOMSKY (1995a), Bare phrase structure, in Gerth WEBELHUTH, editor, *Government and Binding Theory and the Minimalist Program*, pp. 383–349, Blackwell.

Noam CHOMSKY (1995b), *The minimalist program*, The MIT Press, Cambridge, MA.

Trevor COHN, Phil BLUNSOM, and Sharon GOLDWATER (2010), Inducing tree-substitution grammars, *Journal of Machine Learning Research*, 11:3053–3096.

Bernard COMRIE (1993), Argument structure, in Joachim JACOBS, Arnim VON STECHOW, Wolfgang STERNEFELD, and Theo VENNEMAN, editors, *Syntax: An International Handbook*, pp. 905–914, Walter de Gruyter, Berlin, Germany.

- Denis CREISSELS (2014), Cross-linguistic variation in the treatment of beneficiaries and the argument vs. adjunct distinction, *Linguistic Discovery*, 12(2).
- William CROFT (2001), *Radical construction grammar: Syntactic theory in typological perspective*, Oxford University Press, Oxford, England.
- Peter CULICOVER and Ray JACKENDOFF (2005), *Simpler syntax*, Oxford University Press, Oxford, England.
- Carl DE MARCKEN (1996a), The unsupervised acquisition of a lexicon from continuous speech, Technical Report AI-memo-1558, CBCL-memo-129, Massachusetts Institute of Technology – Artificial Intelligence Laboratory.
- Carl DE MARCKEN (1996b), *Unsupervised language acquisition*, Ph.D. thesis, Massachusetts Institute of Technology.
- Jacob FELDMAN (2000), Minimization of Boolean complexity in human concept learning, *Nature*, 407(6804):630–633.
- Jenny Rose FINKEL, Trond GRENAGER, and Christopher D. MANNING (2007), The infinite tree, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.
- Diana FORKER (2014), A canonical approach to the argument/adjunct distinction, *Linguistic Discovery*, 12(2).
- L. T. F. GAMUT (1991), *Logic, language, and meaning volume II: Intensional logic and logical grammar*, University of Chicago Press, Chicago, IL.
- Gerald GAZDAR, Ewan KLEIN, Geoffrey K. PULLUM, and Ivan A. SAG (1985), *Generalized phrase structure grammar*, Harvard University Press, Cambridge, MA.
- John Anton GOLDSMITH (2011), The evaluation metric in generative grammar, in *Proceedings of the 50th Anniversary Celebration of the MIT Department of Linguistics*.
- Sharon GOLDWATER (2006), *Nonparametric bayesian models of lexical acquisition*, Ph.D. thesis, Brown University.
- Sharon GOLDWATER, Thomas L. GRIFFITHS, and Mark JOHNSON (2006), Interpolating between types and tokens by estimating power-law generators, in *Advances in Neural Information Processing Systems 18*.
- Noah D. GOODMAN, Joshua B. TENENBAUM, Jacob FELDMAN, and Thomas L. GRIFFITHS (2008), A rational analysis of rule-based concept learning, *Cognitive Science*, 32(1):108–154.
- Peter D. GRÜNWARD (2007), *The minimum description length principle*, The MIT Press, Cambridge, MA.
- Liliane HAEGEMAN (1994), *Introduction to government and binding theory*, Blackwell, Oxford, England.

Martin HASPELMATH (2014), Arguments and adjuncts as language-particular syntactic categories and as comparative concepts, *Linguistic Discovery*, 12(2).

Irene HEIM and Angelika KRATZER (1998), *Semantics in generative grammar*, Blackwell Publishing, Malden, MA.

Norbert HORNSTEIN and David W. LIGHTFOOT (1981), *Introduction to explanation in linguistics: The logical problem of language acquisition*, Addison Wesley Longman, Upper Saddle River, NJ.

Anne S. HSU and Nick CHATER (2010), The logical problem of language acquisition goes probabilistic: No negative evidence as a window on language acquisition, *Cognitive Science*, 34:972–1016.

Anne S. HSU, Nick CHATER, and Paul M. B. VITÁNYI (2011), The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis, *Cognition*, 120:380–390.

Anne S. HSU, Nick CHATER, and Paul M. B. VITÁNYI (2013), Language learning for positive evidence reconsidered: A simplicity-based approach, *Topics in Cognitive Science*, 5:35–55.

Rodney HUDDLESTON and Geoffrey K. PULLUM (2002), *The cambridge grammar of English language*, Cambridge University Press, Cambridge, England.

Ray JACKENDOFF (2002), *Foundations of language*, Oxford University Press, New York, NY.

David E. JOHNSON and Paul M. POSTAL (1980), *Arc pair grammar*, Princeton University Press, Princeton, NJ.

Mark JOHNSON, Thomas L. GRIFFITHS, and Sharon GOLDWATER (2007), Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models, in *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA.

Aravind K. JOSHI and Leon S. LEVY (1975), Tree adjunct grammars, *Journal of Computer and System Sciences*, 10:136–163.

Jean-Pierre KOENIG, Gail MAUNER, and Breton BIENVENUE (2003), Arguments for adjuncts, *Cognition*, 89:67–103.

Paul R. KROEGER (2004), *Analyzing syntax: A lexical-functional approach*, Cambridge University Press, Cambridge, England.

Ming LI and Paul M. B. VITÁNYI (2008), *An introduction to Kolmogorov complexity and its applications*, Springer, Berlin, Germany, third edition.

Percy LIANG, Slav PETROV, Michael I. JORDAN, and Dan KLEIN (2007), The infinite PCFG using hierarchical Dirichlet processes, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 688–697.



*Arguments and modifiers*

- Brian MACWHINNEY (2000), *The childe project: Tools for analyzing talk*, Lawrence Erlbaum Associates, Mahwah, NJ.
- Alec MARANTZ (2013), Verbal argument structure: Events and participants, *Lingua*, 130:152–168.
- Mitchell P. MARCUS, Beatrice SANTORINI, Mary Ann MARCINKIEWICZ, and Ann TAYLOR (1999), Treebank-3, Technical report, Linguistic Data Consortium, Philadelphia.
- Peter H. MATTHEWS (1981), *Syntax*, Cambridge University Press, Cambridge, England.
- Sally MCCONNELL-GINET and Gennaro CHIERCHIA (2000), *Meaning and grammar: An introduction to semantics*, MIT Press, Cambridge, MA.
- Igor MEL'ČUK (1988), *Dependency syntax : Theory and practice*, The SUNY Press, Albany, NY.
- Michael MOORTGAT (1997), Categorical type logics, in *Handbook of Logic and Language*, pp. 93–177, Elsevier.
- Timothy J. O'DONNELL (2011), *Productivity and reuse in language*, Ph.D. thesis, Harvard University, Cambridge, MA.
- Timothy J. O'DONNELL (2015), *Productivity and reuse in language: A theory of linguistic computation and storage*, The MIT Press, Cambridge, MA.
- Timothy J. O'DONNELL, Jesse SNEDEKER, Joshua B. TENENBAUM, and Noah D. GOODMAN (2011), Productivity and reuse in language, in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Boston, MA.
- Martha PALMER, P. KINGSBURY, and Daniel GILDEA (2005), The proposition bank: An annotated corpus of semantic roles, *Computational Linguistics*, 31(1):71–106.
- Lisa PEARL and Sharon GOLDWATER (2016), Statistical learning, inductive bias, and Bayesian inference in language acquisition, in Jeffrey LIDZ, William SNYDER, and Joe PATER, editors, *The Oxford Handbook of Developmental Linguistics*, Oxford University Press, Oxford, England.
- Lisa PEARL and Jon SPROUSE (2013), Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem, *Language Acquisition*, 20:23–68.
- Amy PERFORs, Joshua B. TENENBAUM, and Terry REGIER (2011), The learnability of abstract syntactic principles, *Cognition*, 118(3):306–338.
- Lawrence PHILLIPS and Lisa PEARL (2014), The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation, manuscript.
- Steven Thomas PIANTADOSI (2011), *Learning and the language of thought*, Ph.D. thesis, Massachusetts Institute of Technology.

- Steven Thomas PIANTADOSI (2021), The computational origin of representation, *Minds and Machines*, 31:1–58.
- Jim PITMAN and Marc YOR (1995), The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, Technical report, Department of Statistics University of California, Berkeley.
- Carl POLLARD and Ivan A. SAG (1994), *Head-driven phrase structure grammar*, University of Chicago Press, Chicago, IL.
- Matt POST and Daniel GILDEA (2009), Bayesian learning of a tree substitution grammar, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.
- Matt POST and Daniel GILDEA (2013), Bayesian tree substitution grammars as a usage-based approach, *Language and Speech*, 56(3):291–308.
- Adam PRZEPIÓRKOWSKI (1999a), *Case assignment and the complement/adjunct dichotomy*, Ph.D. thesis, Neuphilologischen Fakultät der Universität Tübingen, Tübingen.
- Adam PRZEPIÓRKOWSKI (1999b), On case assignment and “adjuncts as complements”, in Gerth WEBELHUTH, Jean-Pierre KOENIG, and A. KATHOL, editors, *Lexical and Constructional Aspects of Linguistic Explanation*, pp. 223–245, CSLI Publications.
- Adam PRZEPIÓRKOWSKI (2017), Hierarchical lexicon and the argument/adjunct distinction, in *Proceedings of the Lexical Functional Grammar 2017 (LFG’17) Conference*, University of Konstanz.
- Andrew RADFORD (1988), *Transformational grammar: A first course*, Cambridge University Press, Cambridge, England.
- György RÁKOSI (2006), *Dative experiencer predicates in Hungarian*, Ph.D. thesis, Universiteit Utrecht.
- Owen RAMBOW, K. VIJAY-SHANKER, and David WEIR (1995), D-tree grammars, in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics.
- Ezer RASIN and Roni KATZIR (2016), On evaluation metrics in optimality theory, *Linguistic Inquiry*, 47(2):235–282.
- Jorma RISSANEN (1978), Modeling by shortest data description, *Automaticata*, 14(5):465–471.
- Ivan A. SAG (2012), Sign-based Construction Grammar: An informal synopsis, in Hans BOAS and Ivan A. SAG, editors, *Sign-Based Construction Grammar*, pp. 101–107, CSLI Publications, Palo Alto, CA.
- Remko SCHA (1990), Taaltheorie en taaltechnologie; competence en performance, in R. DE KORT and G.L.J. LEERDAM, editors, *Computertoepassingen in de Neerlandistiek*, pp. 7–22, Landelijke Vereniging van Neerland.

- Remko SCHA (1992), Virtuele grammatica's en creatieve algoritmes, *Gramma/TTT*, 1(1):57–77.
- Yves SCHABES and Stuart M. SHIEBER (1994), An alternative conception of tree-adjointing derivation, *Computational Linguistics*, 20(1):91–124.
- Yves SCHABES and Richard C. WATERS (1995), Tree insertion grammar: A cubic-time parsable formalism that lexicalizes context-free grammar without changing the trees produced, *Computational Linguistics*, 21(4):479–513.
- Carson T. SCHÜTZE (1995), PP attachment and argumenthood, Technical report, Papers on language processing and acquisition, MIT working papers in linguistics, Cambridge, Ma.
- Carson T. SCHÜTZE and Edward GIBSON (1999), Argumenthood and English prepositional phrase attachment, *Journal of Memory and Language*, 40:409–431.
- Ray SOLOMONOFF (1978), Complexity-based induction systems: comparisons and convergence theorems, *IEEE Transactions on Information Theory*, 24(4):422–432.
- Ray J. SOLOMONOFF (1964a), A formal theory of inductive inference. Part I, *Information and Control*, 7(1):1–22.
- Ray J. SOLOMONOFF (1964b), A formal theory of inductive inference. Part II, *Information and Control*, 7(2):224–254.
- Edward P. STABLER (1997), Derivational minimalism, in *Logical Aspects of Computational Linguistics*, Springer, Berlin, Germany.
- Mark STEEDMAN (2000), *The syntactic process*, MIT Press, Cambridge, MA.
- Andreas STOLCKE and Stephen OMOHUNDRO (1994), Inducing probabilistic grammars by Bayesian model merging, in *Proceedings of the International Conference on Grammatical Inference*.
- Maggie TALLERMAN (2015), *Understanding syntax*, Routledge, London, England, fourth edition.
- Yee Whye TEH (2006), A Bayesian interpretation of interpolated Kneser-Ney, Technical Report TRA2/06, National University of Singapore, School of Computing.
- Damon TUTUNJIAN and Julie E. BOLAND (2008), Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension, *Language and Linguistics Compass*, 2(4):631–646.
- Heinz VATER (1978), On the possibility of distinguishing between complements and adjuncts, in *Valence, semantic case and grammatical relations*, pp. 21–45, John Benjamins.
- Søren WICHMANN (2014), Arguments and adjuncts cross-linguistically: A brief introduction, *Linguistic Discovery*, 12(2).

J. Gerard WOLFF (1977), The discovery of segments in natural language, *British Journal of Psychology*, 68:97–106.

J. Gerard WOLFF (1980), Language acquisition and the discovery of phrase structure, *Language and Speech*, 23(3):255–269.

J. Gerard WOLFF (1982), Language acquisition, data compression, and generalisation, *Language and Communication*, 2(1):57–89.

Yuan YANG and Steven Thomas PIANTADOSI (2022), One model for the learning of language, *Proceedings of the National Academy of Sciences*, 119(5).

Arnold M. ZWICKY (1993), Heads, bases, and functors, in Greville G. CORBETT, Norman M. FRASER, and Scott MCGLASHAN, editors, *Heads in Grammatical Theory*, pp. 292–315, Cambridge University Press, Cambridge, England.

*Leon Bergen*

lbergen@ucsd.edu

Department of Linguistics, University  
of California San Diego, San Diego,  
California

*Edward Gibson*

① 0000-0002-5912-883X

gibson@mit.edu

Department of Brain and Cognitive  
Sciences, Massachusetts Institute of  
Technology, Cambridge,  
Massachusetts

*Timothy J. O'Donnell*

① 0000-0002-5711-977X

timothy.odonnell@mcgill.ca

McGill University,  
Canada CIFAR AI Chair, Mila

Leon Bergen, Edward Gibson, and Timothy J. O'Donnell (2022), *Simplicity and learning to distinguish arguments from modifiers*, *Journal of Language Modelling*, 10(2):241–286

① <https://dx.doi.org/10.15398/jlm.v10i2.263>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

④ <http://creativecommons.org/licenses/by/4.0/>

# Neural heuristics for scaling constructional language processing

Paul Van Eecke<sup>1,2</sup>, Jens Nevens<sup>1</sup>, and Katrien Beuls<sup>3</sup>

<sup>1</sup> Vrije Universiteit Brussel

<sup>2</sup> KU Leuven

<sup>3</sup> Université de Namur

## ABSTRACT

Constructionist approaches to language make use of form-meaning pairings, called constructions, to capture all linguistic knowledge that is necessary for comprehending and producing natural language expressions. Language processing consists then in combining the constructions of a grammar in such a way that they solve a given language comprehension or production problem. Finding such an adequate sequence of constructions constitutes a search problem that is combinatorial in nature and becomes intractable as grammars increase in size. In this paper, we introduce a neural methodology for learning heuristics that substantially optimise the search processes involved in constructional language processing. We validate the methodology in a case study for the CLEVR benchmark dataset. We show that our novel methodology outperforms state-of-the-art techniques in terms of size of the search space and time of computation, most markedly in the production direction. The results reported on in this paper have the potential to overcome the major efficiency obstacle that hinders current efforts in learning large-scale construction grammars, thereby contributing to the development of scalable constructional language processing systems.

*Keywords:*  
*neuro-symbolic*  
*AI, neural*  
*heuristics,*  
*language*  
*processing,*  
*computational*  
*construction*  
*grammar*

Constructionist approaches to language (Goldberg 2003) analyse all linguistic knowledge that is necessary for language comprehension and production in terms of constructions. A construction is defined as a conventionalised pairing between a linguistic form and its meaning (Goldberg 1995; Kay and Fillmore 1999). There exists no restriction in the nature of the form and the meaning that a construction can capture (Fillmore 1988, p. 36). The form pole of a construction can include morphemes and word forms, as well as larger patterns that range from idiomatic expressions (e.g. “Break a leg!”), over partially instantiated structures (e.g. “X takes Y for granted”), to fully abstract schemata (e.g. the ditransitive “X VERB Y Z” as instantiated in “Simon sent his parents a postcard”). The meaning pole of a construction can contain any semantic or pragmatic information that is associated with a particular form, including lexical and phrasal meaning, the assignment of semantic roles, and the composition of logical structures.

According to the constructionist paradigm, the different constructions that constitute a construction grammar can freely combine in order to collaboratively map between a natural language utterance and a representation of its meaning (Goldberg 2006, p. 22). Due to the unrestricted nature of a construction grammar, the non-locality of constructions, and the fact that the application of a construction does not necessarily correspond to a tree-building operation (van Trijp 2016), constructional language processing cannot straightforwardly be implemented in a faithful way using common techniques such as chart parsing and chart generation (see e.g. Pereira and Warren 1983; Shieber 1988; Kay 1996). Instead, current systems implement the process of finding a sequence of constructions that perform an adequate mapping between a linguistic expression and a representation of its meaning as a search process (Bleys *et al.* 2011; Van Eecke and Beuls 2017). This search process is combinatorial in nature and becomes intractable as grammars increase in size. The intractability of construction grammars is a consequential problem as it constitutes a major obstacle that hinders ongoing research in learning large-scale construction grammars. It thereby limits their usability in both usage-based linguistics research and language technology applications.

Previous approaches to overcoming this intractability problem have either only been partially effective, as in the case of priming networks (Wellens and De Beule 2010; Wellens 2011), or have imposed a global order on constructions, which goes against the constructional idea of “*allowing constructions to combine freely as long as there are no conflicts*” (Goldberg 2006, p. 22), as in the case of construction sets (Beuls 2011).

In this paper, we introduce a novel methodology for learning heuristics that substantially optimise the search processes involved in constructional language processing. The heuristics are based on sequence-to-sequence models that are trained to estimate at any point in processing the probability that the application of a particular construction will lead to a solution. We evaluate the methodology on the CLEVR benchmark dataset (Johnson *et al.* 2017) and show that it outperforms state-of-the-art approaches, both in terms of size of the search space and time of computation.

The remainder of this paper is structured as follows. Section 2 precisely defines the search problem involved in constructional language processing, discusses state-of-the-art approaches, and introduces the dataset and grammar that we will use. Section 3 presents our neural methodology for learning heuristics, which constitutes the main contribution of the paper. Section 4 describes the setup of our experiments and presents the evaluation results. Finally, the method and results are discussed in Section 5. An interactive web demonstration accompanying this paper can be consulted at <https://emergent-languages.org/demos/neural-heuristics>. The web demonstration provides examples of the methodology introduced in this paper in full detail.

## PROBLEM DEFINITION

2

We first define constructional language processing as a state-space search problem, which is a class of problems that has a long history in the field of artificial intelligence (Newell and Simon 1956; Nilsson 1971). For doing this, we adopt the terminology that is used in Fluid Construction Grammar (FCG – <https://www.fcg-net.org>) (Steels

2011; van Trijp et al. 2022; Beuls and Van Eecke 2023), the leading computational construction grammar implementation. We then discuss the merits and limitations of state-of-the-art approaches, in particular the use of priming networks and the use of construction sets. Finally, we introduce the CLEVR dataset, which will be used as a benchmark to evaluate our methodology against the state of the art in Section 4.

## 2.1 *Constructional language processing*

Constructional language processing is the process in which the different constructions of a construction grammar combine in order to comprehend or produce natural language expressions. Comprehension refers to the process of mapping a natural language expression to a representation of its meaning, while production refers to the inverse process of mapping a semantic representation to a natural language utterance. Both processes are performed by the same grammar, i.e. the same inventory of constructions. Constructional language processing, as operationalised in the FCG framework, revolves around two basic concepts: ‘transient structures’ and ‘constructions’.

- **Transient structures** A transient structure is a feature structure that represents all that is known about a linguistic expression at a given point during processing. Transient structures correspond to state representations in the classical problem solving paradigm. Before processing has started, the transient structure, which is at that point called ‘initial transient structure’, only contains the input to the comprehension or production process. In comprehension, the input consists of an utterance; while in production, it consists of a semantic representation.
- **Constructions** A construction (CXN) is a feature structure that represents a bidirectional mapping between the formal and the semantic aspects of a linguistic entity. Constructions correspond to operators in the problem solving paradigm and consist of preconditions and postconditions. The preconditions can be ‘matched’ against a transient structure and if matching succeeds, the postconditions can be ‘merged’ into the transient structure. Matching



is a first-order syntactic unification operation that checks the compatibility of two feature structures, whereas merging is a unification operation that combines the information contained in two feature structures. For a formal definition of matching and merging, see Steels and De Beule (2006) and Sierra Santibáñez (2012).

Constructional language processing consists in the sequential application of constructions to a transient structure. Each individual construction application thereby expands the transient structure with new information. Initially, the transient structure only contains the input utterance or input meaning representation, and only constructions that match this information can apply. Through their application, these constructions can contribute additional information to the transient structure, which can in turn satisfy the preconditions of other constructions. Analogous to the use of goal tests in the classical problem solving paradigm, goal tests in constructional language processing verify whether a given transient structure qualifies as a solution to the search problem. Typical goal tests for constructional language processing include (i) checking whether no more constructions can apply, (ii) verifying whether the input utterance or input meaning representation has been fully processed, and (iii) checking whether the meaning comprehended so far consists of a fully connected network of predicates linked through their arguments. When all goal tests succeed for a given transient structure, it qualifies as a solution and the resulting meaning representation (in comprehension) or the resulting utterance (in production) are extracted.

An illustrative example of a construction application process is shown in Figure 1. Note that the constructions used in this example were created for didactic purposes, and do not necessarily correspond to insightful linguistic analyses. From left to right, the figure shows the transient structures and constructions involved in the processing of the utterance *Sam cycles* in comprehension and production. The transient structures shown in the top-left and bottom-left corners (i.e. the green boxes labelled with the number 1) are the initial transient structures in comprehension and production respectively. The initial transient structure in comprehension contains an input unit with a number of predicates representing the utterance. The initial transient structure in production contains an input unit with the meaning representation of

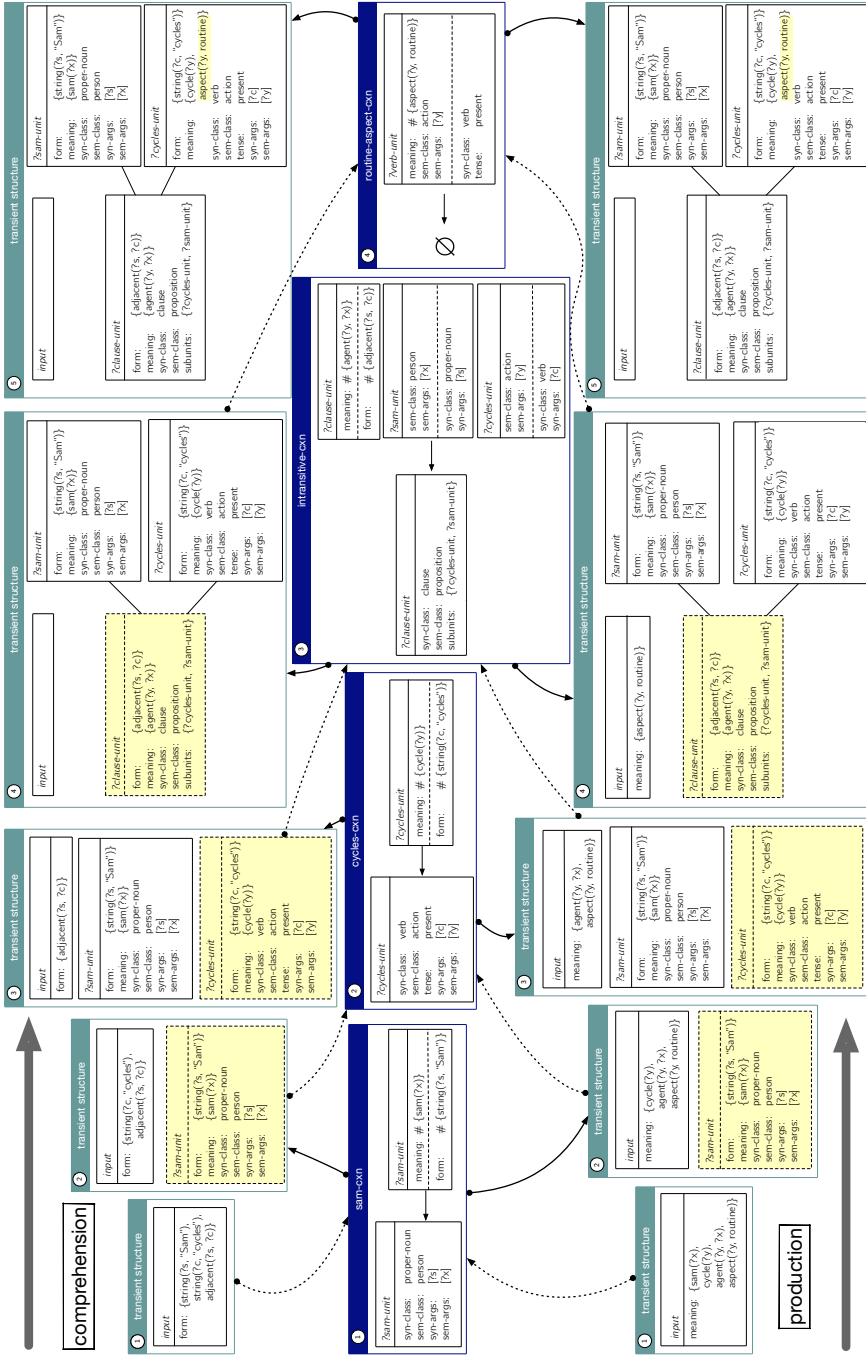


Figure 1: Schematic representation of how different constructions combine in order to comprehend and produce a natural language expression. Here, four constructions collaboratively process the utterance *Sam cycles in comprehension* (top, left-to-right) and *production* (bottom, left-to-right). Dotted arrows represent matching relations and solid arrows represent merging relations. New units added during the merging phase are highlighted using dashed boxes. Symbols preceded by a question mark denote logic variables

the utterance in predicate notation. The middle-left box in the figure represents the SAM-CXN, which matches both initial transient structures. Conventionally, the preconditions and postconditions of a construction are separated by a left-pointing arrow. The preconditions are written on the right-hand side of the arrow, while the postconditions are specified on its left-hand side. As constructions support language processing in both the comprehension and production direction, they contain two sets of preconditions on their right-hand side. The preconditions that are active in comprehension are always specified under a dashed line and the preconditions that are active in production are specified above it. The preconditions of one direction become postconditions in the opposite direction, and are as such treated in the same way as the information specified on the left-hand side. In this case, the construction matches the string *Sam* in comprehension and adds the meaning predicate above the dashed line along with the semantic and syntactic features specified on the left-hand side. The resulting transient structure (labelled with the number 2) is shown just right of the initial transient structure. In production, an analogous process takes place. Here, the construction matches a meaning predicate that is present in the initial transient structure, and contributes a string predicate along with the same semantic and syntactic features as in comprehension.

Next, the CYCLES-CXN applies in the same way to the transient structure that was just created, adding new information related to the string *cycles* in comprehension and to the predicate *cycle(?y)* in production. After that, the INTRANSITIVE-CXN (labelled with the number 3) can apply, as its preconditions are now satisfied by information from the input unit, in combination with information that was contributed by the SAM-CXN and the CYCLES-CXN. The INTRANSITIVE-CXN maps between the adjacency of a proper noun and a verb, and the agentive relation between the person and action they represent. Finally, the ROUTINE-ASPECT-CXN (labelled with the number 4) maps between an action verb in the present tense and a meaning predicate denoting that the aspectual structure of the action corresponds to a routine.

From the final transient structure, shown in the top-right and bottom-right corners of the figure and labelled with the number 5, the result of the construction application process can be extracted.

In comprehension, this is the combination of all ‘meaning’ features in the transient structure, while in production, it is the combination of all ‘form’ features. Note that the construction application process is entirely bidirectional. The output in comprehension is equal to the input in production and vice versa. Moreover, the exact same set of constructions has been applied, in this case even in the same sequential order.

This illustrative example shows how constructions can collaboratively map between an utterance and a representation of its meaning, both in the comprehension and the production direction, with examples of constructions that can apply based on the input only (SAM-CXN and CYCLES-CXN), constructions that build hierarchical structures (INTRANSITIVE-CXN) and constructions that only contribute non-hierarchical information (ROUTINE-ASPECT-CXN). What the example doesn’t show is how a constructional language processing engine can determine that it is exactly this combination of constructions that needs to apply. Construction grammars that exceed the size of these toy examples are immediately faced with constructions that are in competition with each other, and in particular with sequences of constructions that can apply but do not ultimately lead to a solution. This challenge, which is central to the problem solving paradigm, can be solved by backtracking to earlier transient structures in case of failure and possibly, in the worst case, exploring the entire search space, i.e. trying out all possible combinations of construction applications. It is this process of construction application and backtracking that makes constructional language processing intractable for larger grammars.

As can be seen in the figure, the constructions that constitute construction grammars differ in many aspects from the rules that constitute traditional formal grammars. First of all, constructions do not necessarily correspond to tree-building operations, as exemplified by the ROUTINE-ASPECT-CXN and discussed in van Trijp (2016). Constructions are also non-local, in the sense that they can match information that is present anywhere in the transient structure. As a consequence, constructional language processing cannot straightforwardly be optimised using well-known techniques for efficiently processing formal grammars, such as chart parsing and chart generation (see e.g. Pereira and Warren 1983; Shieber 1988; Kay 1996).

In the computational construction grammar literature, a number of techniques for reducing the search space created by all possible construction applications have been proposed. A straightforward optimisation that is almost always used consists in checking whether a new transient structure is different from all other transient structures that already occur in the search tree. If this is not the case, the duplicate transient structure can immediately be pruned away. A second common optimisation consists in hashing constructions that match string predicates or meaning predicates in the input unit, which reduces the search problem to abstract constructions only.

When it comes to the choice of the baseline search strategy, a depth-first search algorithm with backtracking is often chosen. For constructional language processing, depth-first search generally outperforms breadth-first search for two reasons. First, solutions are typically found deep in the search tree (after many constructions have been applied) and there is no inherent preference for shorter solutions, like for example in the case of planning problems. Second, there often exist many correct orders in which constructions can apply, which can lead to a high branching factor and an abundance of duplicate transient structures, some of which can only be detected deep in the search tree.

Two more advanced approaches that go beyond the depth-first search with backtracking, duplicate detection and hashing baseline have been proposed in the literature: ‘construction sets’ and ‘priming networks’.

- **Construction sets** This approach consists in subdividing the construction inventory into (possibly overlapping) sets of constructions. Two global orders of construction sets are specified, one for comprehension and one for production. The basic idea is that constructions of a later set are not applied before constructions of an earlier set have at least been matched against the transient structure (Beuls 2011). The use of construction sets can drastically decrease the size of the search space, but comes with a number of important drawbacks. Construction sets are inherently in disagreement with the constructionist idea that constructions can

freely combine as long as there are no conflicts, which is crucial for supporting open-ended and creative language use (Van Eecke and Beuls 2018; Goldberg 2006, p. 22). Also, the global ordering of construction sets and the allocation of constructions to particular sets is difficult to learn, as it presupposes that the general architecture of a grammar is known beforehand. Finally, scaling grammars that make use of construction sets is even difficult in the case of hand-crafted grammars, as the grammar engineer then does not only need to encode the necessary linguistic knowledge, but also needs to determine the order in which constructions need to be scheduled for matching.

- **Priming networks** Priming networks are inspired by the psychological phenomenon whereby current behaviour is nonconsciously influenced by exposure to past experiences (see e.g. Schacter and Buckner 1998). In the case of computational construction grammar, this approach argues that the application of a construction can prime the application of another construction. In this way, frequent co-occurrences of constructions can be captured in the form of a priming network (Wellens and De Beule 2010; Wellens 2011). Priming links can be learned in a usage-based fashion by extracting the frequency of co-occurrences of constructions from successful branches of the search trees generated during past construction application processes. These links can then be used to guide the search process by always expanding the transient structure created by the construction that was most strongly primed. The priming links are based either on the order of the constructions themselves or on dependencies between a construction's preconditions and the postconditions of other constructions by which they were satisfied. Two priming networks are learned for a construction inventory, one for use in comprehension and the other for use in production. The main advantage of priming networks is that they can be learned in a straightforward way. However, an important disadvantage is that if only local priming links are taken into account, priming is only partially effective, and if longer-distance links are taken into account, the networks are often not efficacious as they suffer from sparsity problems.

Another solution that has been proposed in the literature is to process construction grammars using existing systems for implementing generative grammar formalisms (Müller 2017). A major disadvantage of this approach is that this is only possible for generative grammar formalisms that include constructional properties (e.g. constructional HPSG and SBCG). These formalisms are inherently limited to local constructions that correspond to tree-building operations (van Trijp 2016; van Trijp *et al.* 2022). As a consequence, this approach does not satisfy the methodological needs of the construction grammar community at large.

### *The CLEVR dataset and grammar*

2.3

We use the CLEVR dataset (Johnson *et al.* 2017) and CLEVR construction grammar (Nevens *et al.* 2019) to benchmark the effect of the heuristics that we propose in this paper. This choice is motivated by three main reasons. First of all, with its nearly 1,000,000 utterances, the CLEVR dataset is sufficiently large to train even the most data-intensive heuristics. Second, there exists a computational construction grammar that, given infinite computation time, covers the entire dataset, in both the comprehension and production direction (Nevens *et al.* 2019). This means that this grammar achieves 100% accuracy on the tasks of mapping from utterances to their meaning representation and vice versa. This allows us to evaluate the effect of the proposed heuristics in isolation. Finally, the grammar gives rise to a search space that can only be processed efficiently using powerful heuristics.

The utterances in the CLEVR dataset are synthetically generated English questions about images of scenes depicting different configurations of geometrical figures. Each question is annotated with a semantic representation that captures the logical meaning that underlies it. The question-annotation pairs embrace various aspects of reasoning, including attribute identification (*There is a large cube; what is its color?*), counting (*How many green spheres are there?*), comparison (*Are there an equal number of large cubes and small things?*), spatial relationships (*What size is the cylinder that is right of the yellow shiny thing*

*that is left of the cube?*) and logical operations (*How many objects are either red cubes or yellow cylinders?*). The average length of the questions is 18.4 words with a maximum length of 42 words.

The CLEVR grammar consists of 170 constructions, of which 55 are morphological and lexical constructions. Apart from these, the grammar also contains 115 grammatical constructions that capture phenomena including referential expressions, spatial relations, coordination and subordination structures, and a wide range of interrogative structures. On average, 25 constructions should be applied in order to successfully comprehend or produce an utterance from the dataset. This means that the average solution is found at depth 25 in the search tree.

The size of the search space for an average sentence amounts thus in theory to  $170^{25}$  construction applications. In practice, most of these construction applications are not possible given the dependencies between the preconditions and postconditions of the constructions. Still, when using the baseline depth-first strategy with backtracking, duplicate detection and hashing, the search tree in comprehension includes on average more than 3.5 times the number of construction applications than were needed to find a solution. While this might still be manageable to a certain extent, this number grows to more than 29 in production. In practice, this means that many solutions cannot be found in a reasonable amount of time without the use of suitable heuristics.

We make use of the same splits as the original dataset, with the training, validation and test sets consisting of 699,989 utterances, 149,991 utterances and 149,988 utterances respectively.

### 3

## METHODOLOGY

We will now introduce our novel methodology for learning heuristics that substantially optimise the search processes involved in constructional language processing. These heuristics take the form of neural networks that are trained to estimate at any point in processing the probability that the application of a particular construction will lead to a solution. Our approach is inspired by recent successes obtained using



neural heuristics in other domains that typically employ the problem solving paradigm, in particular games (Mnih *et al.* 2015; Silver *et al.* 2016) and planning (Takahashi *et al.* 2019; Wang *et al.* 2019; Ferber *et al.* 2020).

### *General architecture*

3.1

We design neural heuristics that can be used to assign after each construction application a score to the resulting transient structure. This score, called ‘heuristic value’, reflects how close a transient structure is to a solution. It can then be used to decide on the order in which transient structures are expanded, with the goal of minimizing the average number of construction applications that is needed to reach a solution. Intuitively, this score is influenced by both the input utterance (in comprehension) or meaning representation (in production) and the sequence of constructions that have been applied so far in the same branch of the search tree.

For each direction of processing, this intuitive idea is operationalised using two recurrent neural networks (RNNs) that are organised in an encoder-decoder constellation. Before processing starts, the encoder RNN encodes the input utterance or meaning representation into a context vector. During processing, the decoder RNN is called for each transient structure, just before its expansion. The input to the decoder RNN is the sequence of names of constructions that have been applied so far in that branch of the search tree, along with the output of the encoder RNN (context vector) and its hidden states. The output of the decoder RNN is at each decoding timestep a probability distribution over all constructions in the construction inventory. The constructions are then applied and the heuristic values of the resulting transient structures are computed as the sum of the heuristic value of their parent transient structure and the probability score returned by the decoder RNN. The heuristic values of the transient structures are used in combination with a beam search algorithm. This process is graphically depicted in Figure 2, where the beam size is set to three for clarity reasons.

The choice for an RNN-based encoder-decoder architecture is motivated by two main reasons. First of all, the problem of mapping

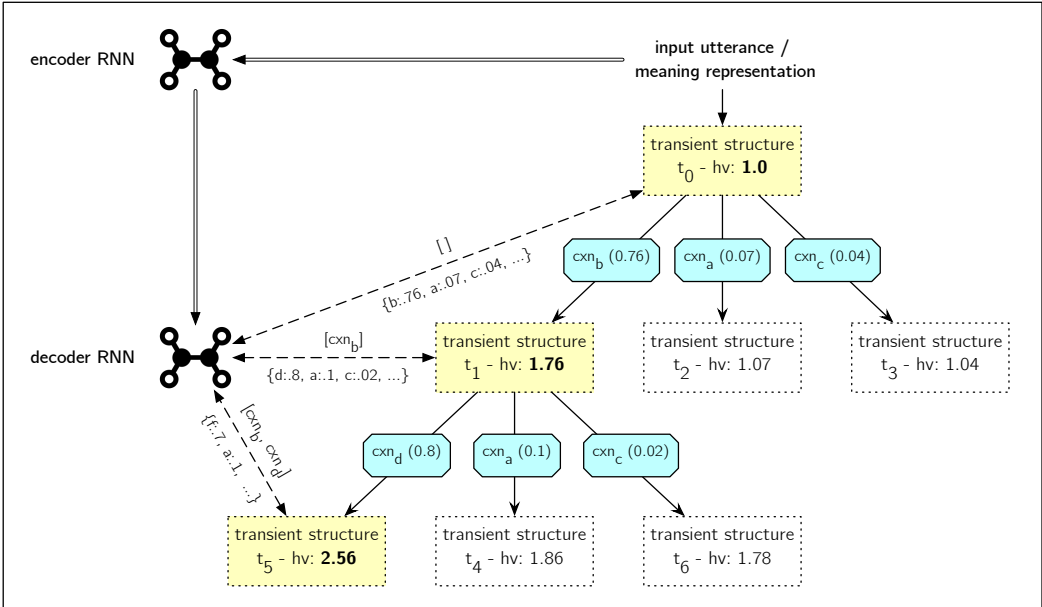


Figure 2: Schematic representation of the integration of the neural heuristics in constructional language processing. Before a node in the search tree is expanded, the encoder-decoder model is queried, and a probability distribution over all constructions of the construction inventory is returned. The heuristic values are then calculated and used by a beam search algorithm, in this case with the beam size set to three. ‘Hv’ stands for ‘heuristic value’

an input utterance or meaning representation to a sequence of constructions can be naturally framed as a sequence-to-sequence problem. RNN-based architectures are typically good at handling this class of problems (Sutskever *et al.* 2014), although also CNN-based (Gehring *et al.* 2017) and transformer-based (Vaswani *et al.* 2017) architectures have more recently been successfully applied to the same class of problems. Second, and most importantly, the sequential nature of the RNN-based architecture allows us to query the decoder RNN while already providing a partial sequence of predictions. This is necessary for integrating the neural architecture as a heuristic in the construction application process, while being able to keep the benefits of the existing search and backtracking facilities. Indeed, the neural networks are used to make the search process created by the grammar more efficient, unlike their use in end-to-end neural semantic parsers, where

they perform the actual mapping from utterances to their semantic representation (see e.g. Jia and Liang 2016; Konstas *et al.* 2017; van Noord *et al.* 2018; Yu and Gildea 2022).

Our neural encoder-decoder architecture is based on the neural machine translation architecture proposed by Bahdanau *et al.* (2015). It consists of an encoder with bidirectional single-layer gated recurrent units (GRUs), a decoder with single-layer GRUs and an attention mechanism that attends over the encoder’s hidden states at every decoder time step. The attention mechanism ensures that the decoder does not need to rely on a single high-dimensional representation of the entire input sequence (the context vector). Instead, the decoder has access to all encoder hidden states and learns to use a subset of these hidden states. Intuitively, the decoder chooses at every timestep to pay attention to specific parts of the input utterance or meaning representation.

The basic idea underlying our methodology is somewhat reminiscent of the use of recurrent neural networks for guiding dependency parsing (Kiperwasser and Goldberg 2016; Dozat and Manning 2017, 2018). In this line of research, RNNs are also used to predict sequences of actions (e.g. transitions) based on utterances and previous actions. The main difference resides in the correspondence between the length of the input and the length of the predictions. In the case of dependency parsing, an action needs to be predicted for each input word. In the case of constructional language processing, however, the number of constructions that needs to be predicted is not tied to the length of the input utterance or meaning representation. Indeed, multiple words or meaning predicates (even organised in non-contiguous patterns) can be covered by the application of a single construction, and single words or meaning predicates can give rise to the application of multiple constructions (see e.g. the ROUTINE-ASPECT-CXN in Figure 1). In order to accommodate for this asymmetry between the length of the input pattern and the length of the output pattern, we have opted for two RNNs in an encoder-decoder constellation instead of directly using an RNN for prediction. This allows us to effectively decouple the length of both sequences.

## 3.2

*Training*

Training the neural encoder-decoder architecture requires a dataset of input utterances (in comprehension) or meaning representations (in production), paired with, for each utterance or meaning representation, a sequence of names of constructions of which the application would lead to a solution. Annotating the original CLEVR dataset, which contains utterances along with a representation of their meaning, in this format is a non-trivial task, as we face at this moment the very search problem that we are aiming to optimise. We therefore adopted a spiral approach. For both comprehension and production, we started processing the data from CLEVR's training and validation splits using the depth-first search strategy with backtracking, duplicate detection and hashing, setting a time limit of 400 seconds. We collected the sequences of construction names for all input utterances or meaning representations that were successfully processed within this time frame. Then, we trained a first version of the sequence-to-sequence heuristic and used it to process more utterances using the same time limit. After three iterations, the entire dataset could be successfully annotated.

The encoder-decoder architecture requires that input utterances or meaning representations are represented as sequences. For utterances, this is naturally done by using sequences of tokens. For meaning representations, this is somewhat more complicated as they come in the form of networks of predicates that share variables. We therefore transformed the predicate networks into sequences notated in reverse Polish notation. In this notation, predicate names follow their arguments. Since the arity of each predicate is known, the notation is unambiguous without the need for variables and their equalities to be explicitly represented.

We trained the encoder-decoder models for 100,000 time steps with a batch size of 64, using the Adam optimisation algorithm with a learning rate of  $5e-4$  and weight decay of  $1e-6$ . We used cross-entropy as the loss function and used a teacher forcing ratio of 1. We included a dropout layer after the embedding layer in both the encoder and the decoder. We ran a hyperparameter optimisation process for the embedding size (100, 200, 300), the hidden layer size (64, 128, 256, 512) and the dropout probability (0.0, 0.1, 0.2, 0.5). We found that

best performance was achieved using the model with an embedding size of 100 in comprehension and 300 in production, a hidden layer size of 512 in comprehension and 256 in production and a dropout probability of 0.2 in comprehension and 0.1 in production. Note that for efficiency reasons, the hyperparameters were optimised based on the gold standard annotation of the dataset, and not based on their performance as a heuristic in FCG.

## EXPERIMENTS

4

In order to benchmark the efficiency of our methodology and compare it against the state of the art, we conducted two experiments that evaluate the use of the proposed neural heuristics in constructional language processing. The first experiment is concerned with the comprehension direction, while the second experiment is concerned with the production direction.

### *Experimental setup*

4.1

Both experiments consist in processing the test split of the CLEVR dataset using three different search strategies. The first strategy makes use of FCG’s standard search algorithm, namely depth-first search with backtracking, duplicate detection and hashing. The second strategy makes use of priming networks as proposed by Wellens and De Beule (2010). The third strategy evaluates the encoder-decoder methodology that we introduced above, with an unrestricted beam size.

The strategies are evaluated in terms of the size of the search space and the time that is required to reach a solution. The size of the search space is defined as the total number of transient structures that were created during processing, divided by the number of transient structures in the branch of the solution. The optimal size of the search space is thus equal to one, indicating that a solution was found without any backtracking taking place. The time of computation is measured in seconds, spanning from the creation of the initial transient structure

until the resulting meaning representation (in comprehension) or utterance (in production) has been extracted from the solution transient structure. In general, the size of the search space is the most accurate measure for gauging the performance of a search strategy, but it does not take into account the computational overhead caused by the heuristic itself. The computation time metric includes both factors, but should be interpreted with extreme caution, as it is also influenced by external factors.

For the purposes of this paper, we have chosen to focus on the fundamental issue of reducing the search space, and have not included any other time-related optimisations. Such optimisations could include the implementation of a more efficient protocol for communication between the FCG engine and the neural networks, deploying the neural networks on GPUs, or not using the neural heuristics for utterances under a maximum number of words. The reason that we include the computation time metric is to show that even without these optimisations, a reduction in the search space already corresponds to a reduction in processing time.

If no solution was found within 400 seconds, the search process was halted and the result was logged as ‘no solution found’.

The evaluation was carried out using computing nodes with  $2 \times 20$ -core Intel Xeon Gold 6148 (Skylake) CPUs and 16GB of RAM.

## 4.2

### *Experimental results: comprehension*

The results of the comprehension experiment are presented in Table 1 and visualised through violin plots in Figure 3. The table provides the mean values, standard deviation and maximum values of the search space size and the computation time for the depth-first, priming and neural strategies. The plots show the probabilistic density of the search space size (Figure 3a) and computation time (Figure 3b) for the three strategies. When it comes to the size of the search space, the results show that the neural strategy greatly outperforms the depth-first and priming strategies. More density mass is situated close to a search space size of one, which is the theoretical minimum. The average size of the search space is 1.16 in the case of the neural strategy, 3.21 in the case of the priming strategy and 3.69 in the case of the depth-first strategy. Importantly, the performance gain obtained through the

Table 1: Performance of the different strategies in the comprehension direction

	Search space size			Computation time (s)			# Timed out > 400 s
	mean	sd	max	mean	sd	max	
Depth-first	3.69	7.09	174.26	0.84	4.42	141.28	0
Priming	3.21	5.98	161.09	0.72	3.73	158.48	0
Neural	1.16	0.19	15.84	0.91	0.80	49.48	0

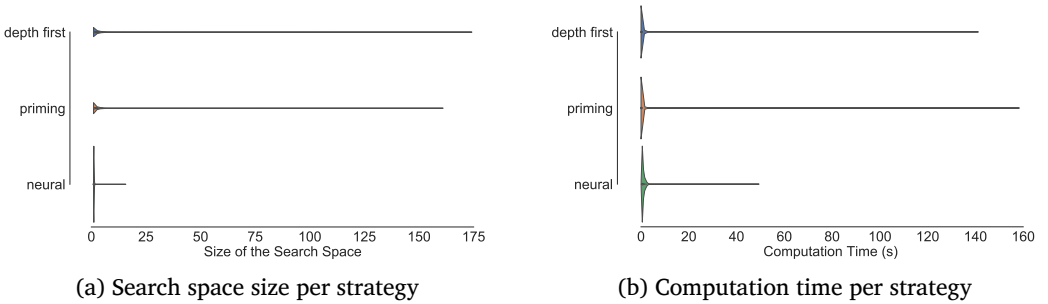


Figure 3: Visualisation of the results of the comprehension experiment

neural strategy also extends to sentences that otherwise require a large search space. The largest search space required by the neural strategy is 15.82, while the depth-first and priming strategies require search space sizes of up to 174.26 and 161.09, respectively. The results obtained through the computation time metric are in line with those obtained through the search space size metric. Even the most difficult sentences take less than 50 seconds using the neural strategy, whereas they take more than 140 seconds using the depth-first and priming strategies. In sum, we can conclude from the comprehension experiment that the neural strategy outperforms the state of the art both in terms of size of the search space and in terms of time of computation. Importantly, the greatest reduction in search space and time of computation is achieved for the most difficult sentences.

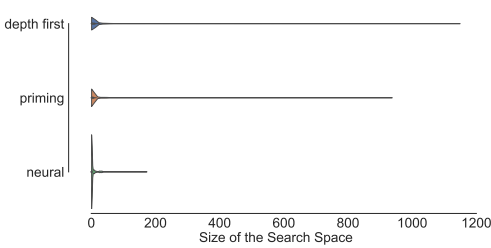
### Experimental results: production

4.3

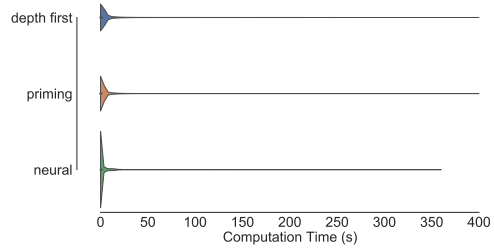
The results of the production experiment are presented in Table 2 and visualised through violin plots in Figure 4. Figure 4a shows the

Table 2: Performance of the different strategies in the production direction

	Search space size			Computation time (s)			# Timed out (> 400 s)
	mean	sd	max	mean	sd	max	
Depth-first	29.08	90.40	1149.74	8.84	37.57	400.00	1325
Priming	20.81	67.78	938.25	6.40	29.38	400.00	475
Neural	6.35	16.85	173.90	3.64	12.32	360.25	0



(a) Search space size per strategy



(b) Computation time per strategy

Figure 4: Visualisation of the results of the production experiment

average size of the search space for each strategy. We can immediately observe that the search problem in production is considerably more difficult than the search problem in comprehension, and that the performance gain that is obtained through the neural strategy is even larger. In the case of the neural strategy, the density mass is concentrated around a lower mean value (6.35) than in the case of the priming (20.81) and depth-first (29.08) strategies. The maximum value is reduced from 1149.74 (depth-first) and 938.25 (priming) to 173.90 (neural). When it comes to computation time (Figure 4b), the results are analogous. The average processing time is reduced from 8.84 (depth-first) and 6.40 (priming) seconds to 3.64 seconds (neural). The maximum processing time that was needed amounts to 360.25 seconds for the neural strategy. For the other two strategies, not all sentences could be produced within the maximum time frame of 400 seconds.

An analysis of the utterances for which the neural strategy could not reduce the search space to under 5 reveals an interesting limitation of the methodology that we have introduced. The decoder RNN takes as input a sequence of constructions that have so far been applied



during the application process and returns as output a probability distribution over all constructions in the construction inventory. In other terms, it makes predictions about which constructions should be applied at which moment in time. However, it does not make any predictions about the way in which the constructions should apply, in particular to which units in the transient structure. As a consequence, only the ambiguity that arises from multiple applicable constructions (i.e. multiple transient structures resulting from the application of different constructions) can be solved, not the ambiguity that arises from multiple ways in which a single construction can apply (i.e. multiple transient structures resulting from the application of a single construction). While this ambiguity is far less substantial than the ambiguity that stems from multiple applicable constructions, it explains why the search space is not consistently reduced to around 1 even if every prediction by the neural network is optimal.

In sum, we can conclude that the production experiment confirms the results obtained in the comprehension experiment. The neural strategy outperforms the state of the art both in terms of size of the search space and in terms of time of computation, especially when it comes to processing the most difficult sentences of the dataset.

## DISCUSSION AND CONCLUSIONS

5

Constructionist approaches to language, as originally laid out by, among others, Fillmore (1988), Goldberg (1995), Kay and Fillmore (1999) and Croft and Cruse (2004), consider form-meaning mappings, called constructions, to be the basic unit of linguistic analysis. Apart from the fact that they constitute form-meaning mappings, constructions are subject to very few restrictions. First of all, constructions do not necessarily correspond to tree-building operations (van Trijp 2016). Second, constructions are non-local in the sense that they can access all information that is known during processing. Third, constructions can involve units of arbitrary size, both on the form and the meaning side. Finally, constructions are not restricted to continuous constituents and are not even required to include word order constraints. As a consequence, constructional language processing cannot

straightforwardly be implemented in a faithful way using common grammar processing techniques, such as chart parsing and generation (see e.g. Pereira and Warren 1983; Shieber 1988; Kay 1996). Instead, faithful computational construction grammar implementations implement constructional language processing as a state-space search problem (Bleys *et al.* 2011; Van Eecke and Beuls 2017).

In order to reliably scale to large problems, state-space search methods rely on heuristics that can estimate the likelihood that a given state will lead to a solution. While certain optimisations have in the past been applied to the case of computational construction grammar, including construction sets (Beuls 2011) and priming networks (Wellens and De Beule 2010; Wellens 2011), a lack of general and powerful heuristics remained a major obstacle to ongoing construction grammar research, in particular to research on representing, processing and learning large-scale construction grammars.

The neural methodology that we have presented in this paper introduces a general and effective way to learn heuristics that substantially optimise the search processes involved in constructional language processing. Analogous to recent successes in many subfields of artificial intelligence, including game playing (Mnih *et al.* 2015; Silver *et al.* 2016) and planning (Takahashi *et al.* 2019; Wang *et al.* 2019; Ferber *et al.* 2020), the methodology combines the predictive strengths of neural networks with the expressive representations, sound logic operations and backtracking abilities of traditional search and unification methods.

An integration of the proposed method in the Fluid Construction Grammar system (Steels 2011; van Trijp *et al.* 2022; Beuls and Van Eecke 2023) and an evaluation of the method using the CLEVR benchmark dataset (Johnson *et al.* 2017) and the CLEVR construction grammar (Neuens *et al.* 2019) show that the neural heuristics indeed outperform the state-of-the-art priming strategy and can substantially reduce the search space and processing time in both the comprehension and the production direction, especially in the case of utterances that otherwise gave rise to a large search space.

We posit that this general methodology for learning neural heuristics that optimise the search processes involved in constructional language processing constitutes a promising contribution towards the scaling of constructionist approaches to language. It thereby has both

theoretical and practical implications. On the theoretical side, scalable processing models will allow construction grammarians to go beyond the study of constructions in isolation, and model the intricate interactions that take place between constructions as part of a larger grammar. On the practical side, the scaling of constructional language processing paves the way for achieving breakthroughs in ongoing research on learning large-scale construction grammars (Nevens *et al.* 2022; Doumen *et al.* 2023), which has in turn major implications on research in usage-based linguistics (Diessel 2015), models of language acquisition (Tomasello 2003) and the use of construction grammar in language technology applications (Willaert *et al.* 2020, 2021; Beuls *et al.* 2021; Verheyen *et al.* 2022).

#### ACKNOWLEDGEMENTS

We would like to thank the three anonymous JLM reviewers for their rigorous yet constructive examination of our paper. The research reported on in this paper was financed by the Research Foundation Flanders (FWO) through a postdoctoral grant awarded to Paul Van Eecke (75929) and the European Union's Horizon 2020 research and innovation programme under grant agreement number 951846 (Meaning and Understanding in Human-centric AI – <https://www.muha.i.org>).

#### CONFLICT OF INTEREST STATEMENT

On behalf of all authors, the corresponding author states that there is no conflict of interest.

#### REFERENCES

- Dzmitry BAHDANAU, Kyunghyun CHO, and Yoshua BENGIO (2015), Neural machine translation by jointly learning to align and translate, in *International Conference on Learning Representations (ICLR 2015)*, pp. 1–15.
- Katrien BEULS (2011), Construction sets and unmarked forms: A case study for Hungarian verbal agreement, in Luc STEELS, editor, *Design Patterns in Fluid Construction Grammar*, pp. 237–264, John Benjamins, Amsterdam, Netherlands.

- Katrien BEULS and Paul VAN EECKE (2023), Fluid Construction Grammar: State of the art and future outlook, in *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs + NLP, GURT/SyntaxFest 2023)*, pp. 41–50, Association for Computational Linguistics, Washington, D.C., USA.
- Katrien BEULS, Paul VAN EECKE, and Vanja Sophie CANGALOVIC (2021), A computational construction grammar approach to semantic frame extraction, *Linguistics Vanguard*, 7(1):20180015.
- Joris BLEYS, Kevin STADLER, and Joachim DE BEULE (2011), Search in linguistic processing, in Luc STEELS, editor, *Design Patterns in Fluid Construction Grammar*, pp. 149–179, John Benjamins, Amsterdam, Netherlands.
- William CROFT and D. Alan CRUSE (2004), *Cognitive linguistics*, Cambridge University Press, Cambridge, United Kingdom.
- Holger DIESEL (2015), Usage-based construction grammar, in Ewa DĄBROWSKA and Dagmar DIVJAK, editors, *Handbook of Cognitive Linguistics*, pp. 295–321, Mouton de Gruyter, Berlin, Germany.
- Jonas DOUMEN, Katrien BEULS, and Paul VAN EECKE (2023), Modelling language acquisition through syntactico-semantic pattern finding, in *Findings of the Association for Computational Linguistics: EAACL 2023*, forthcoming.
- Timothy DOZAT and Christopher D. MANNING (2017), Deep biaffine attention for neural dependency parsing, in *5th International Conference on Learning Representations, ICLR 2017*, pp. 1–8.
- Timothy DOZAT and Christopher D. MANNING (2018), Simpler but more accurate semantic dependency parsing, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 484–490, Association for Computational Linguistics, Melbourne, Australia, doi:10.18653/v1/P18-2077.
- Patrick FERBER, Malte HELMERT, and Jörg HOFFMANN (2020), Neural network heuristics for classical planning: A study of hyperparameter space, in *24th European Conference on Artificial Intelligence*, pp. 2346–2353.
- Charles J. FILLMORE (1988), The mechanisms of “construction grammar”, in *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pp. 35–55.
- Jonas GEHRING, Michael AULI, David GRANGIER, Denis YARATS, and Yann N. DAUPHIN (2017), Convolutional sequence to sequence learning, in *Proceedings of the 34th International Conference on Machine Learning*, pp. 1243–1252.
- Adele E. GOLDBERG (1995), *Constructions: A construction grammar approach to argument structure*, University of Chicago Press, Chicago, IL, USA.
- Adele E. GOLDBERG (2003), Constructions: A new theoretical approach to language, *Trends in Cognitive Sciences*, 7(5):219–224.
- Adele E. GOLDBERG (2006), *Constructions at work: The nature of generalization in language*, Oxford University Press, Oxford, United Kingdom.

Robin JIA and Percy LIANG (2016), Data recombination for neural semantic parsing, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12–22, Association for Computational Linguistics, Berlin, Germany, doi:10.18653/v1/P16-1002.

Justin JOHNSON, Bharath HARIHARAN, Laurens VAN DER MAATEN, Li FEI-FEI, C. LAWRENCE ZITNICK, and Ross GIRSHICK (2017), CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, IEEE Computer Society, Los Alamitos, CA, USA.

Martin KAY (1996), Chart generation, in *34th Annual Meeting of the Association for Computational Linguistics*, pp. 200–204, Association for Computational Linguistics.

Paul KAY and Charles FILLMORE (1999), Grammatical constructions and linguistic generalizations: The what’s x doing y? construction, *Language*, 75(1):1–33.

Eliyahu KIPERWASSER and Yoa GOLDBERG (2016), Simple and accurate dependency parsing using bidirectional LSTM feature representations, *Transactions of the Association for Computational Linguistics*, 4:313–327, doi:10.1162/tacl\_a\_00101, <https://aclanthology.org/Q16-1023>.

Ioannis KONSTAS, Srinivasan IYER, Mark YATSKAR, Yejin CHOI, and Luke ZETTEMAYER (2017), Neural AMR: Sequence-to-sequence models for parsing and generation, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 146–157, Association for Computational Linguistics, Vancouver, Canada, doi:10.18653/v1/P17-1014.

Volodymyr MNIH, Koray KAVUKCUOGLU, David SILVER, Andrei A. RUSU, Joel VENESS, Marc G. BELLEMARE, Alex GRAVES, Martin RIEDMILLER, Andreas K. FIDJELAND, Georg OSTROVSKI, *et al.* (2015), Human-level control through deep reinforcement learning, *Nature*, 518(7540):529–533.

Stefan MÜLLER (2017), Head-Driven Phrase Structure Grammar, Sign-Based Construction Grammar, and Fluid Construction Grammar: Commonalities and differences, *Constructions and Frames*, 9(1):139–173.

Jens NEVENS, Jonas DOUMEN, Paul VAN EECKE, and Katrien BEULS (2022), Language acquisition through intention reading and pattern finding, in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 15–25, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, <https://aclanthology.org/2022.coling-1.2>.

Jens NEVENS, Paul VAN EECKE, and Katrien BEULS (2019), Computational construction grammar for visual question answering, *Linguistics Vanguard*, 5(1):20180070.

Allen NEWELL and Herbert SIMON (1956), The logic theory machine – a complex information processing system, *IRE Transactions on Information Theory*, 2(3):61–79.

Nils NILSSON (1971), *Problem-solving methods in artificial intelligence*, McGraw-Hill Book Company, New York, NY, USA.

Fernando PEREIRA and David WARREN (1983), Parsing as deduction, in *21st Annual Meeting of the Association for Computational Linguistics*, pp. 137–144, Association for Computational Linguistics.

Daniel L. SCHACTER and Randy L. BUCKNER (1998), Priming and the brain, *Neuron*, 20(2):185–195.

Stuart M. SHIEBER (1988), A uniform architecture for parsing and generation, in *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*, pp. 614–619, <https://aclanthology.org/C88-2128>.

Josefina SIERRA SANTIBÁÑEZ (2012), A logic programming approach to parsing and production in Fluid Construction Grammar, in Luc STEELS, editor, *Computational Issues in Fluid Construction Grammar*, volume 7249 of *Lecture Notes in Computer Science*, pp. 239–255, Springer, Berlin, Germany.

David SILVER, Aja HUANG, Chris J. MADDISON, Arthur GUEZ, Laurent SIFRE, George VAN DEN DRIESSCHE, Julian SCHRITTWIESER, Ioannis ANTONOGLU, Veda PANNEERSHELVAM, Marc LANCTOT, et al. (2016), Mastering the game of Go with deep neural networks and tree search, *Nature*, 529(7587):484–489.

Luc STEELS, editor (2011), *Design patterns in Fluid Construction Grammar*, John Benjamins, Amsterdam, Netherlands.

Luc STEELS and Joachim DE BEULE (2006), Unify and merge in Fluid Construction Grammar, in *International Workshop on Emergence and Evolution of Linguistic Communication (EELC 2006)*, pp. 197–223, Rome, Italy.

Ilya SUTSKEVER, Oriol VINYALS, and Quoc V. LE (2014), Sequence to sequence learning with neural networks, in Z. GHARAMANI, M. WELLING, C. CORTES, N. LAWRENCE, and K.Q. WEINBERGER, editors, *Advances in Neural Information Processing Systems*, volume 27, pp. 3104–3112, Curran Associates, Inc., Red Hook, NY, USA.

Takeshi TAKAHASHI, He SUN, Dong TIAN, and Yebin WANG (2019), Learning heuristic functions for mobile robot path planning using deep neural networks, in *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pp. 764–772.

Michael TOMASELLO (2003), *Constructing a language: A usage-based theory of language acquisition*, Harvard University Press, Harvard, MA, USA.

Paul VAN EECKE and Katrien BEULS (2017), Meta-layer problem solving for computational construction grammar, in *The 2017 AAAI Spring Symposium Series*, pp. 258–265, AAAI Press, Palo Alto, CA, USA.

Paul VAN EECKE and Katrien BEULS (2018), Exploring the creative potential of computational construction grammar, *Zeitschrift für Anglistik und Amerikanistik*, 66(3):341–355.

Rik VAN NOORD, Lasha ABZIANIDZE, Antonio TORAL, and Johan BOS (2018), Exploring neural methods for parsing discourse representation structures, *Transactions of the Association for Computational Linguistics*, 6:619–633, doi:10.1162/tacl\_a\_00241.

Remi VAN TRIJP (2016), Chopping down the syntax tree: What constructions can do instead, *Belgian Journal of Linguistics*, 30(1):15–38.

Remi VAN TRIJP, Katrien BEULS, and Paul VAN EECKE (2022), The FCG editor: An innovative environment for engineering computational construction grammars, *PLOS ONE*, 17(6):e0269708, doi:10.1371/journal.pone.0269708.

Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Lukasz KAISER, and Illia POLOSUKHIN (2017), Attention is all you need, in I. GUYON, U. VON LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, and R. GARNETT, editors, *Advances in Neural Information Processing Systems*, volume 30, pp. 6000–6010, Curran Associates, Inc., Red Hook, NY, USA.

Lara VERHEYEN, Jérôme Botoko EKILA, Jens NEVENS, Paul VAN EECKE, and Katrien BEULS (2022), Hybrid procedural semantics for visual dialogue: An interactive web demonstration, in *Workshop on semantic techniques for narrative-based understanding: Workshop at IJCAI-ECAI 2022*, pp. 48–52.

Jingyuan WANG, Ning WU, Wayne Xin ZHAO, Fanzhang PENG, and Xin LIN (2019), Empowering A\* search algorithms with neural networks for personalized route recommendation, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 539–547.

Pieter WELLENS (2011), Organizing constructions in networks, in Luc STEELS, editor, *Design Patterns in Fluid Construction Grammar*, pp. 181–201, John Benjamins, Amsterdam, Netherlands.

Pieter WELLENS and Joachim DE BEULE (2010), Priming through constructional dependencies: a case study in Fluid Construction Grammar, in *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8)*, pp. 344–351, World Scientific.

Tom WILLAERT, Paul VAN EECKE, Katrien BEULS, and Luc STEELS (2020), Building social media observatories for monitoring online opinion dynamics, *Social Media + Society*, 6(2), doi:10.1177/2056305119898778.

Tom WILLAERT, Paul VAN EECKE, Jeroen VAN SOEST, and Katrien BEULS (2021), An opinion facilitator for online news media, *Frontiers in Big Data*, 4:1–10.

Chen YU and Daniel GILDEA (2022), Sequence-to-sequence AMR parsing with ancestor information, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 571–577, Association for Computational Linguistics, Dublin, Ireland, doi:10.18653/v1/2022.acl-short.63.

*Paul Van Eecke*

© 0000-0001-9153-9092  
paul@ai.vub.ac.be

Artificial Intelligence Laboratory,  
Vrije Universiteit Brussel, Pleinlaan 2,  
1050 Brussels, Belgium

KU Leuven, Faculty of Arts, Research  
Unit Linguistics, Blijde Inkomststraat  
21, 3000 Leuven, Belgium

KU Leuven, imec research group itec,  
Etienne Sabbelaan 51, 8500  
Kortrijk, Belgium

*Katrien Beuls*


© 0000-0003-4451-4778  
katrien.beuls@unamur.be

Faculté d'informatique, Université de  
Namur, rue Grandgagnage 21, 5000  
Namur, Belgium

Paul Van Eecke, Jens Nevens, and Katrien Beuls (2022), *Neural heuristics for scaling constructional language processing*, *Journal of Language Modelling*, 10(2):287–314

doi: <https://dx.doi.org/10.15398/jlm.v10i2.318>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

©  <http://creativecommons.org/licenses/by/4.0/>



## EXTERNAL REVIEWERS 2019–2022

The mainstay of any peer-reviewed journal are its reviewers, and JLM is no exception here. Each paper is reviewed by at least 3 carefully selected reviewers, usually including at least one representing the JLM Editorial Board. To increase reviewing anonymity, we do not give the names of the 29 JLM EB reviewers, but we would like to heartily thank them for their hard and timely work. We also express our sincere gratitude to the following 99 external reviewers for papers reviewed during 2019–2022:

*Gulsat Aygen*

Northern Illinois University

*Anne Abeillé*

Université Paris Diderot, Paris 7

*David Adger*

Queen Mary University of London

*Željko Agić*

IT University of Copenhagen, Unity Technologies

*Raquel G. Alhama*

Max Planck Institute for Psycholinguistics

*Afra Alishahi*

Tilburg University

*Jonathan Amith*

Yale University

*Pascal Amsili*

CNRS

*Denis Béchet*

Université de Nantes

*Leon Bergen*

University of California at San Diego

*Ricardo Bermúdez-Otero*

University of Manchester

*Timothée Bernard*

CNRS •

Université Denis Diderot, Paris 7

*Paul Boersma*

University of Amsterdam

*Olivier Bonami*

Université Paris Sorbonne

*Robert Borsley*

University of Essex

*Johan Bos*

University of Groningen

*Gosse Bouma*

University of Groningen

*Gilles Boyé*

Université Bordeaux Montaigne

*Andrew Carnie*

University of Arizona

*Milos Cernak*

Institut Dalle Molle d'Intelligence Artificielle Perceptive

*Lucas Champollion*

New York University

*Jane Chandlee*

Haverford College

- Rui P. Chaves*  
University at Buffalo
- Barbara Citko*  
University of Washington
- Alexander Clark*  
University of Gothenburg
- Robin Cooper*  
University of Gothenburg
- Ryan Cotterell*  
Johns Hopkins University
- Berthold Crysmann*  
CNRS •  
Université Denis Diderot, Paris 7
- Jennifer Culbertson*  
University of Edinburgh
- Tim Van de Cruys*  
Katholieke Universiteit Leuven
- Mark de Vries*  
University of Groningen
- Grażyna Demenko*  
Adam Mickiewicz University  
in Poznań
- Liviu P. Dinu*  
University of Bucharest
- Hossep Dolatian*  
Stony Brook University
- Maciej Eder*  
Institute of Polish Language, Polish  
Academy of Sciences •  
Pedagogical University of National  
Education Commission
- Marina Ermolaeva*  
Lomonosov Moscow State University
- Edward Flemming*  
Massachusetts Institute of Technology
- Danny Fox*  
Massachusetts Institute of Technology
- Richard Futrell*  
University of California at Irvine
- Kim Gerdes*  
Université Sorbonne Nouvelle
- Thomas Graf*  
Stony Brook University
- Gunnar Ólafur Hansson*  
The University of British Columbia
- Nabil Hathout*  
CNRS •  
Université Toulouse II – Jean Jaurès
- Dag Trygve Truslew Haug*  
Universitetet i Oslo
- Jeffrey Heinz*  
Stony Brook University
- Valentin Hofmann*  
University of Oxford
- Adam Jardine*  
Rutgers University
- Szymon Jaroszewicz*  
Institute of Computer Science,  
Polish Academy of Sciences
- Adam Jatowt*  
Universität Innsbruck
- Wojciech Jaworski*
- Roni Katzir*  
Tel Aviv University
- Gerrit Kentner*  
Goethe Universität Frankfurt
- Tracy Holloway King*  
Adobe

*Katarzyna Klessa*  
Adam Mickiewicz University  
in Poznań

*Olga Kolesnikova*  
National Polytechnic Institute  
of Mexico

*Danijel Koržinek*  
Polish-Japanese Academy  
of Information Technology (PJAIT)

*Katarzyna Krasnowska*  
Institute of Computer Science,  
Polish Academy of Sciences

*Diego Gabriel Krivochen*  
University of Oxford

*Lothar Lemnitzer*  
Berlin-Brandenburg Academy  
of Sciences and Humanities

*Jadwiga Linde-Usiekiewicz*  
University of Warsaw

*Tal Linzen*  
Johns Hopkins University

*John Lowe*  
University of Oxford

*Veronika Lux-Pogodalla*  
ATILF (Computer Processing and  
Analysis of the French Language)  
at CNRS

*Antonio Machicao*  
Priemer Humboldt-Universität  
zu Berlin

*Giorgio Magri*  
CNRS • Université Paris 8

*Robert Malouf*  
San Diego State University

*Stella Markantonatou*  
Institute for Language and Speech  
Processing /‘Athena’ RIC

*Roland Meyer*  
Humboldt-Universität zu Berlin

*Günter Neumann*  
DFKI (German Research Center  
for Artificial Intelligence)

*Andrew Nevins*  
University College London

*Garrett Nicolai*  
University of Alberta

*Urpo Nikanne*  
Åbo Akademi University

*Timothy John O’Donnell*  
McGill University

*Hugo Oliveira*  
University of Lisbon

*Timothy John Osborne*  
Zhejiang University

*Lisa Pearl*  
University of California at Irvine

*Katya Pertsova*  
University of Carolina at Chapel Hill

*Omer Preminger*  
University of Maryland

*Ezer Rasin*  
Tel Aviv University

*Jon Rawski*  
San José State University

*Charles Reiss*  
Concordia University Montreal

*Christian Retoré*  
University of Montpellier

*Mehrnoosh Sadrzadeh*  
University College London

*Manfred Sailer*  
Goethe-Universität Frankfurt

*Jacques Savoy*  
University of Neuchatel

*Nathan Schneider*  
Georgetown University

*Raj Singh*  
Carleton University

*Paweł Teisseyre*  
Institute of Computer Science,  
Polish Academy of Sciences

*Amalia Todirascu*  
Université de Strasbourg

*Remi van Trijp*  
Sony Computer Science Laboratory  
Paris

*Leo Wanner*  
Pompeu Fabra University

*Stephen Wechsler*  
University of Texas at Austin

*Alexander Williams*  
University of Maryland

*Colin Wilson*  
Johns Hopkins University

*Grégoire Winterstein*  
Université du Québec à Montréal

*Jan Wiślicki*  
University of Warsaw

*Jacek Witkoś*  
Adam Mickiewicz University  
in Poznań

*Leszek Wroński*  
Jagiellonian University

*Annie Zaenen*  
Stanford University