Antoine Chambaz* and Guillaume Desagulier

Predicting Is Not Explaining: Targeted Learning of the Dative Alternation

DOI 10.1515/jci-2014-0037

Abstract: Corpus linguists dig into large-scale collections of texts to better understand the rules governing a given language. We advocate for ambitious corpus linguistics drawing inspiration from the latest developments of semiparametrics for a modern targeted learning. Transgressing discipline-specific borders, we adapt an approach that has proven successful in biostatistics and apply it to the well-travelled case study of the dative alternation in English. A dative alternation is characterized by sentence pairs with the same verb, but different syntactic patterns, *e.g. I gave a book to him* (prepositional dative) and *I gave him a book* (double-object dative). Our aim is to explain how native speakers of English choose a pattern over another in any given context. The essence of the approach hinges on causal analysis and targeted minimum loss estimation (TMLE). Through causal analysis, we operationalize the set of scientific questions that we wish to address regarding the dative alternation. Drawing on the philosophy of TMLE, we answer these questions by targeting some versatile machine learners. We derive estimates and confidence regions for well-defined parameters that can be interpreted as the influence of each contextual variable on the outcome of the alternation (prepositional *vs.* double-object), all other things being equal.

Keywords: causal analysis, TMLE, semiparametric inference, dative alternation

1 Introduction

A *corpus* (plural *corpora*) is a large-scale collection of texts sampled from genuine linguistic productions by native speakers. From a statistical viewpoint, a corpus is a sample drawn from the true, unknown law of a given language. Corpus linguistics consists in digging into corpora to better understand the rules governing the language under study. This is why Gries [1] describes corpus linguistics as a "distributional science", a science that infers knowledge from data. Often, corpus linguists focus on the frequencies of occurrence of various elements in corpora, their dispersion, and their co-occurrence properties. Baayen [2] argues that "corpus linguistics should be more ambitious". Focusing on a classification problem, he compares the performances of different classifiers based either on the principle of parametric regression or on more data-adaptive algorithms gathered under the banner of machine learning, both in terms of accuracy of prediction and of quality of the underlying models for human learning. Following Baayen [2], we also advocate for ambitious corpus linguistics drawing inspiration from the latest developments of semiparametrics for a modern targeted learning.

We break free from artificial discipline-specific boundaries, as we benefit from the lessons of state-of-the-art causal analysis and biostatistics to address a long-standing issue in linguistics. Our guiding principle is the following: predicting is not explaining. It conveys the idea that one should always carefully cast the questions at stake as statistical parameters of the true, unknown law of the data. Once this is done, we suggest the two-step procedure known as targeted minimum loss estimation (TMLE [3, 4]). The first step takes advantage of the power of machine learning, while acknowledging its limits in terms of inference. To overcome these limits, the second step consists in bending the initial estimators by targeting them toward the parameters they are meant to capture.

***Corresponding author: Antoine Chambaz,** Modal'X – Université Paris Ouest Nanterre La Défense, 200 av de la République, Nanterre 92001, France, E-mail: achambaz@u-paris10.fr
Guillaume Desagulier, MoDyCo – Université Paris 8, CNRS, Université Paris Ouest Nanterre La Défense, Nanterre, France

For the paper to be accessible to non-linguists, Section 2 introduces keys notions and issues in linguistics. In Section 3, we briefly introduce the dative alternation, the theoretical issues it raises, and a summary of recent corpus-based, statistics driven investigations. In Section 4, we lay out our plan for the prediction and explanation of the dative alternation based on corpus data. We claim that these two tasks differ substantially. Our approach is motivated by causal considerations. Section 5 is a concise presentation of the statistical apparatus that we elaborate to tackle the statistical problems defined in Section 4. We present and comment on the results in Section 6. Additional material is gathered in the appendix. In particular, details on the machine learning and on TMLE procedures are given in Sections A.2 and A.3, respectively. These are the most technical parts of the article.

2 A brief introduction to linguistics

2.1 What linguistics is about

Like biologists studying the structure, function, growth, evolution, distribution, and taxonomy of living cells and organisms, linguists study language. In this respect, doing linguistics means investigating the cognitive system which we identify as the knowledge of a language. This knowledge takes the form of a mental grammar. Therefore, understanding what it means to know a language is to understand the nature of such a grammar.

Despite aiming at an objective description of language, linguists have their theoretical preferences, depending on what they believe the true essence of mental grammar is. In this regard, two competing theories have shaped contemporary linguistics.

As described by Chomsky [5–7], transformational-generative grammar (henceforth "Chomskyan grammar/linguistics") is based on the assumption that, like formal languages, the grammars of natural languages consist of *(a)* a set of abstract algebraic rules and *(b)* a "lexicon" that contains meaningful linguistic elements. The algebra is the innate core of grammar. It constitutes what Chomskyans call the "universal grammar", common to all natural languages. It is therefore what Chomskyans truly look for. The lexicon is relegated to the periphery of grammar, along with what makes a language idiosyncratic (*e.g.* inter-speaker variation, cultural connotations, stylistic mannerism, non-standard usage, etc.). Central to Chomskyan grammar is the opposition between *deep structure* and *surface structure*. This opposition hinges on syntax, *i.e.* the way in which words are combined to form larger constituents such as phrases, clauses, or sentences. The deep structure of a sentence is its abstract syntactic representation. The surface structure of a sentence is its final syntactic representation in speech or text. For instance, the sentence *students hate annoying professors* has one surface structure but two alternative interpretations at the level of the deep structure: *(a)* students hate to annoy professors and *(b)* students hate professors who are annoying. *Derivation* is the process whereby a sentence is generated from abstract operations in the deep structure to a string of words in the surface structure. To sum up, Chomskyan grammar is a "top-down" approach to language: linguists study how algebraic rules "at the top" generate an infinite number of sentences "at the bottom". One major problem with this approach is that by focusing on the top, linguists tend to look down upon the bottom.

Conversely, usage-based linguistics is a "bottom-up" approach. Its tenet is that actual language usage shapes the structure of language [8–11]. From a usage-based viewpoint, grammar is the product of usage varying from speaker to speaker and there are no hard and fast rules. Grammar is therefore derivative, not generative. It does not have a core and a periphery. It is instead a structured inventory of symbolic units. Not only does the inventory of a native speaker of English contain highly schematic constructions (the past tense, the ditransitive construction, the active construction, etc.), concrete words or phrases such as ritualized or formulaic expressions (*double whammy*, *hang in there!*), idioms (*that didn't go down well with the editor*, *he kicked the bucket*, etc.), or non-canonical phrasal collocations (*you're getting to me these*

days), but also mixed constructions having both abstract and concrete elements (*the more you drink, the smarter you think you are*). We adopt the viewpoint of usage-based linguistics because *(a)* we believe it offers a psychologically realistic view of grammar and *(b)* such a view, unlike Chomskyan grammar, can be operationalized.

2.2 Corpus linguistics

Linguists must rely on the native speakers of a language acting as informants and providing data such as sentences. On the basis of such data, linguists test their hypotheses about the cognitive systems of native speakers. Chomskyans have relied heavily on introspecting judgments for data collection. Their motivation may date back to de Saussure, the father of modern linguistics, who delimited the object of study (*langue*) as a structured system disconnected from the vagaries of place and time wherein it is deployed. However, the method has been called into question because linguists' intuitions are not always consonant with what they observe in the data.

With the rise of massive digital collections of texts, linguists who have been dissatisfied with the practice of using themselves as informants have found corpora to be far more useful than introspective judgments to test their hypotheses. Unsurprisingly, these linguists, who consider that genuine language use in all its complexity should be at the center of linguistic research, share the main tenets of the usage-based model.

All present-day digital corpora such as The Brown University Standard Corpus of Present-Day American English, the Switchboard corpus, or the British National Corpus conform to the following definition:

> A corpus is a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of texts is compiled with the intention *(a)* to be representative and balanced with respect to a particular linguistic variety or register or genre and *(b)* to be analyzed linguistically [12].

Note that the above definition covers general corpora as opposed to, say, language acquisition corpora or convenience corpora such as the *Guardian* data, as pointed out by one anonymous reviewer. Arguably, a combination of the usage-based methodology and corpus-driven research has led to a paradigm shift in linguistics.

Corpora have their limitations. One of the most frequent criticisms leveled against corpus linguistics by Chomskyans is that corpora do not indicate whether a given expression or use of an expression is impossible. From our usage-based perspective, the response to this critique is simple: because grammatical rules are mere generalizations about actual usage, negative evidence is of limited importance. A more serious criticism is the following: no corpus – however large and balanced – can ever hope to be representative of a speaker, let alone of a language. Supporters of introspective linguistics might argue that a corpus should not be a basis for linguistic studies if it cannot represent language in its full richness and complexity. Most corpus linguists rightly counter that they do not aim to explain all of a language in every study [13]. In this paper, we take a slightly different stance. While we acknowledge that no corpus can provide access to the true, unknown law of a language, we firmly believe that a corpus is a sample drawn from this law. We consider that ambitious corpus linguistics consists in bridging the gap between what we can observe and measure from a corpus, and what we do not know about a language. To achieve this, ambitious statistics is needed.

2.3 Why alternations matter to linguists

Verbs express events. Events involve participants. Participants occupy "places" in the clauses that verbs control. These places are called *arguments*. The verb *drink*, for example, is a two-argument verb because in

a drinking event there is at least a drinker and a liquid that is drunk. In the active voice, the drinker occupies the subject position and the liquid the object position as in (i):

(i) Bob is drinking whiskey.
Subject Object

Suppose Bob wants to brush the whiskey off his breath before setting off to the linguistics lab. He is now holding his toothbrush in one hand and a tube of toothpaste in the other. Four participants are involved: Bob, his teeth, his toothbrush, and the toothpaste. In sentence (ii), three participants are assigned a *semantic role* by the verb: *Bob* is the *agent* (the initiator of the event), *toothpaste* is the *theme* (an inanimate entity undergoing a change of location from the brush to the teeth), and *his teeth* is the *goal* (the end-point of a motion). The toothbrush is not syntactically realized. The ensuing scene can be described using either (ii a) or (ii b):

(ii) a. Bob is brushing his teeth with toothpaste.
AGENT GOAL THEME
b. Bob is brushing toothpaste onto his teeth.
AGENT THEME GOAL

In (ii a), *his teeth* is the object of the verb and *toothpaste* is the object of the preposition *with*. In (ii b), *toothpaste* is the object and *his teeth* is the object of the preposition *onto*. This means that our attention is brought to *his teeth* in (ii a) and to *toothpaste* in (ii b). The phenomenon of a verb exhibiting variation in its syntactic realization is called an *alternation*. Each alternating form is called an *alternant*.

Linguists have long believed that alternations were conditioned by verb meaning, two verbs with identical or similar meaning displaying similar alternating behavior [14]. To illustrate this, suppose now that Bob proceeds to shaving. The verb *slather* involves three participants: Bob, his face, and the shaving cream. Again, the scene can be described using either alternant in (iii):

(iii) a. Bob is slathering his face with shaving cream.
b. Bob is slathering shaving cream onto his face.

Given that *brush* and *shave* are close in meaning (both associated events imply that a 'thick mass' is transferred from a container to a body part), we should not be surprised to see that they display the same alternation. Linguists might be enticed to include *brush* and *slather* in the same typological verb class, *e.g.* the class of "grooming verbs". However, they are somehow different. For example, even though neither verb allows the theme argument to stand alone, as shown in (iv), only *brush* allows the goal to stand alone, as evidenced in (v):[1]

(iv) a. *Bob is brushing toothpaste.
b. *Bob is slathering shaving cream.

(v) a. Bob is brushing his teeth.
b. *Bob is slathering his face.

The question remains open as to whether one should *(a)* include both verbs in the same class, acknowledging that the class is heterogeneous, or *(b)* assign them to related but distinct classes, at the risk of demultiplying classes and blunting Ockham's razor.

1 The asterisk "*" marks ungrammatical sentences.

Such puzzles are central in an area of linguistics called *argument realization*, *i.e.* the study of the syntactic patterns that the arguments of a verb may enter [15]. In (ii) and (iii), the challenge is to explain why two apparently similar verbs display diverging behaviors and why the divergences takes these forms. Addressing these issues matters because of their far-reaching implications for our understanding of language. From a cognitive perspective, one may wonder how speakers classify store verbs in their mental inventories of linguistic units.

3 The dative alternation

Well known to linguists is the dative alternation, which consists of the prepositional dative (henceforth PD) and the ditransitive constructions (or double-object construction, henceforth DO), as exemplified in (vi) and (vii) respectively:

(vi) John gave the book to Mary. (PD)
S_{AGENT} V O_{THEME} $O_{RECIPIENT}$

(vii) John gave Mary the book. (DO)
S_{AGENT} V $O_{RECIPIENT}$ O_{THEME}

The dative event involves three participants: a giver (John), someone who receives something (Mary), and an entity transferred from the giver to the recipient (the book). In terms of semantic roles, the giver is an *agent*, the participant receiving something is the *recipient*, and the entity transferred from the agent to the recipient is a *theme*. What alternates in this case is the realization of the recipient and the theme, one of which must be an object while the other can be either a direct object or a prepositional object. Levin and Rappaport Hovav [15] describe the dative alternation as a case of object alternation.

To account for the dative alternation, linguists have relied on either intuition or corpus-based quantitative methods. We review each trend in Sections 3.1 and 3.2.

3.1 Theoretical issues

The dative alternation has been a fruitful research topic in many different theories. Substantial accounts of past research can be found for instance in Levin [16], Krifka [14] and Levin and Rappaport Hovav [15, chapter 7].

Chomsky [5, 6] suggests that an alternating verb has a single lexical entry for both forms. These forms have the same deep syntactic structure. Differences visible at the sentence level are explained by the fact that the surface structure of the basic form is a direct projection of the deep structure, whereas the surface structure of the derived form is the product of a transformation.

Subsequent Chomskyan studies holding a distinction between deep and surface structures debate over which variant of the dative alternation is transformationally derived from the basic argument realization. Conclusions differ. On the one hand, Fillmore [17], Hall [18], and Emons [19] contend that PD is basic whereas DO is derived. On the other hand, Burt [20] and Aoun and Li [21] argue for the opposite pattern of transformation: DO is basic whereas PD is derived.

Semantic restrictions to the dative alternation have challenged Chomskyan accounts. One restriction is that certain verbs alternate while others readily enter only one variant:[2]

2 A question mark "?" indicates that the example is relatively unacceptable. Two question marks "??" indicate that the example is definitely unacceptable.

(viii) a. Anthony gave $100 to charity.
b. Anthony gave charity $100.

(ix) a. Anthony donated $100 to charity.
b. ?Anthony donated charity $100.

(x) a. ??The bank denied a checking account to me.
b. The bank denied me a checking account.

Proponents of the Localist Hypothesis [22], according to whom locative expressions are seen as the source from which all more abstract expressions derive, construe the recipient as a spatial goal. They further argue that DO is possible in (xi b) if London refers metonymically to a person or an institution, in which case it differs from (xi a) where London is clearly a place:

(xi) a. She sent a parcel to London.
b. She sent London a parcel.

A second restriction is the frequent lack of semantic equivalence between alternating forms in cases where the verb readily enters both variants [23, 24], as in (xii):

(xii) a. Will taught linguistics to the students.
b. Will taught the students linguistics.

DO conveys a sense of completion in such a way that the teaching is successful in (xii b). Example (xii a) is more neutral in this respect. However, more recent studies warn that these semantic differences are intuitive and may be subject to contextual modulation [15, 25, 26].

Despite continuous efforts to maintain that alternating verbs have a single meaning underlying both formal variants [27–29], there is now cross-theoretical consensus that the two variants of the dative alternation have distinct semantic representations. According to Pinker [30] and Rappaport Hovav and Levin [31], caused motion underlies PD, whereas caused possession underlies DO, as schematized in (xiii):

(xiii) a. John gave the book to Mary.
X cause Z to be at Y (CAUSED MOTION, Y is a goal)
'John causes the book to go to Mary'
b. John gave Mary the book.
X cause Y to have Z (CAUSED POSSESSION, Y is a recipient)
'John causes Mary to have the book'

In a similar fashion, Speas [32, pp. 88–89] schematizes the semantic representations of both variants as follows:

(xiv) a. X cause [Y to be at (possession) Z] (PD)
b. X cause [Z to come to be in STATE (of possession)] by means of [X cause [Y to come to be at (poss) Z]] (DO)

In the Construction Grammar framework, Goldberg [10] posits that PD is a subtype of the more general caused-motion construction (*cf*. the Localist Hypothesis), whereas DO expresses a transfer of possession:

(xv) a. X cause Y to move Z (PD)
b. X cause Y to receive Z (DO)

The above finds empirical support in Gries and Stefanowitsch [33].

Given that the distribution of verbs across the dative variants is semantically constrained, and given the frequent lack of semantic equivalence between PD and DO for a given verb, a set of semantic factors have been recognized to influence the choice of PD *vs*. DO. Among the known lexical semantic restrictions applying to verbs in the dative alternation are the following:

- Movement (PD) *vs*. possession (DO): in PD, the theme undergoes movement (literal or figurative) from an origin to a goal, whereas in DO the agent possesses the theme via the verb event.
- Affectedness: as seen in (xii), the recipient of a dative verb is more likely to receive an affected interpretation when expressed as the first object in DO than in PD;
- Continuous imparting of force: in PD, the verb can express a continuous imparting of force (*e.g*. *haul*, *pull*, *push*). DO shows a dispreference for such verbs (??*Will pushed Anthony the biscuits*). Under certain conditions, exceptions occur [34].
- Communication verbs: as opposed to speech act verbs (*tell*, *read*, *write*, *cite*, etc.) and verbs derived from nouns expressing communication means (*fax*, *email*, *phone*), which can occur in PD or DO, verbs that denote a manner of speaking (*shout*, *yell*, *scream*, *whisper*, etc.) have a strong dispreference for DO. Exceptions are listed in Gropen et al. [35].
- Verbs of impeded possession: such verbs (*deny*, *spare*, *cost*) have a preference for DO.
- Latinate verbs: due to their morphophonology, such verbs (*donate*, *explain*, *recite*, *illustrate*, etc.) disprefer DO, except when they express a future possession (*guarantee*, *assign*, *offer*, *promise*), as pointed out by Pinker [30, p. 216].

Lexical semantic restrictions are sometimes overridden by information-structure factors (interalia [36–38]). Such factors have to do with how information is formally packaged in a sentence. The first factor is discourse givenness, that is to say the fact that the reference of an expression is present in the minds of speakers. In general, given material precedes new material. PD is expected when the theme is more given than the recipient, as in (xvi a), whereas DO is more likely when the recipient is more given than the theme, as in (xvii b):

(xvi) a. Will gave his manuscript to a first-year student. (PD)
b. ??Will gave a first-year student his manuscript. (DO)

(xvii) a. ??Will gave a manuscript to his best student. (PD)
b. Will gave his best student a manuscript. (DO)

The second factor is a corollary of the first: because recipients are typically human and themes typically inanimate, they are more likely to be given and thus to occur before themes. In this respect, DO is more frequent than PD. Bresnan and Nikitina [39] find empirical support for this, but they also find exceptions such as (xviii a):

(xviii) a. It would cost nothing to the government. (PD)
b. It would cost the government nothing. (DO)

Although peripheral, the third factor, heaviness, is correlated with information-structure considerations. Heaviness is characterized by the complexity and/or length of sentence constituents. Heavy material comes last, as exemplified below:

(xix) a. ??Anthony gave a bottle of his favorite red wine to Will. (PD)
b. Anthony gave Will a bottle of his favorite red wine. (DO)

Because given material is generally shorter than non-given material (*e.g*. given recipients will generally occur in the form of pronouns), DO is the preferred realization of the dative alternation due to the last two factors.

Which factor(s) take(s) precedence over the other(s) is still theoretically unclear. Snyder [40] claims that information-structure factors are more important than heaviness, whereas Arnold et al. [36] treat all factors on equal footing. What is clearer is that what determines the dative alternation is a multifactorial problem whose full understanding is best resolved empirically. This is why we now turn to recent corpus-based, statistics-driven investigations of the dative alternation.

3.2 Corpus-based answers

Since Williams [41], the dative alternation has become a model construction for benchmarking predictive methods [2, 36, 42–44]. Focusing on DO, Williams [41] uses the logistic procedure to test on a two-part but limited data set (original data set, sample size is 168; aggregate data set, sample size is 59). The model construction includes 8 variables: syntactic class of verb, register, modality, givenness of goal, prosodic length of goal *vs*. theme, definiteness of goal, animacy of goal, and specificity of goal. Williams [41] finds that not all independent variables are predictors of the position of the goal. Only three reach a relatively high level of significance in the model: the prosodic length of goal *vs*. theme (the length of the goal is shorter than the length of the theme), syntactic class of verb (ditransitive), and register (informal).

Arnold et al. [36] investigate the effects of newness and heaviness on word order in the dative alternation. Their data consists of debate transcriptions from the Canadian parliament (the Aligned-Hansard corpus). Utterances are manually annotated for: constituent order (non-shifted *vs*. shifted; prepositional *vs*. double object), heaviness (three categories of relative length measured as follows: number of words in the theme minus number of words in the recipient), and newness (given, inferable, or new). Arnold et al. [36] conclude that heaviness and newness are significantly correlated with constituent order. DO is preferred when the theme is *(a)* newer and *(b)* heavier than the goal.

Gries [43] uses linear discriminant analysis to investigate the effect of multiple variables on the choice of PD *vs*. DO. in the British National Corpus. Gries observes that all properties of NPgoal along with morphosyntactic variables have the highest discriminatory power. However, *(a)* discriminant analysis makes distributional assumptions that are seldom satisfied by the data, and *(b)* Gries [43] concedes that the data set is limited: being part of a larger project, it consists of only 117 instances of the dative alternation.

To circumvent assumptions about the data distribution and to control for the influence of multiple variables on a binary response, Bresnan et al. [42] use (mixed-effects) logistic regression, like Williams [41] and Arnold et al. [36]. Unlike those previous works, Bresnan et al. [42] predicting's data set is relatively large, consisting of 2,360 dative observations from the 3M-word Switchboard collection of recorded telephone conversations. More importantly, the authors also address the question of circular correlations, which are largely ignored in former statistical models, *e.g.*:

- personal pronouns are short, definite and have animate, discourse-given referents;
- animate, discourse-given nominals are often realized as personal pronouns, which are short and definite.

Such correlations trick researchers into believing that the dative alternation can be explained with one or two variables.

Bresnan et al. [42] 's dative data set is annotated for 14 explanatory variables whose influence on the choice of the dative variants is considered likely: modality, verb, semantic class of verb use, and length, animacy,[3] definiteness,[4] pronominality, and accessibility of recipient/theme; see also Section 4.1. One of their logistic regression models predicts which variant of the dative alternation is used with high accuracy.

3 If the referent of a noun is sentient or alive, it is animate, otherwise it is inanimate.

4 A noun phrase is definite when its referent is identified or identifiable in context. It is indefinite otherwise.

Using Bresnan et al.'s data set, Baayen [2] tests naive discriminative learning (henceforth NDL) on the dative alternation. Baayen compares NDL to other well-established statistical classifiers such as logistic regression [42, 45], memory-based learning [44, 46], analogical modeling of language [47], support vector machines [48], and random forests [49]. He addresses two questions:

- how can statistical models faithfully reflect a speaker's knowledge without underestimating or overestimating what a native speaker has internalized?;
- how do occurrence and co-occurrence frequencies in human classification compare to such frequencies in machine classification?

NDL is based on supervised learning, namely the equilibrium equations for the Rescorla-Wagner model [50]. According to the Wagner-Rescorla equations [51], learners predict an outcome from cues available in their environment if such cues have a value in terms of outcome prediction, information gain, and statistical association. When the learner predicts an outcome correctly on the basis of the available cue, the association strength between cue and outcome is weighted in such a way that prediction accuracy improves in subsequent trials. Whereas the Rescorla-Wagner equations are particularly useful in the study of language acquisition [52, 53], the equilibrium equations for the Rescorla-Wagner model apply to adult-learner states (*i.e.* when weights from cues to outcomes do not change as much). NDL estimates the probability of a given outcome independently from the other outcomes.

Like memory-based learning, NDL stands out because it reflects human performance. Unlike parametric regression models, it is unaffected by collinearity issues. When two or more predicting variables are highly correlated, multiple regression models may indicate how well a group of variables predicts an outcome variable, but may not detect *(a)* which individual predictor(s) improve the model, and *(b)* which predictors are redundant. Unlike memory-based learning however, NDL does not need to store exemplars in memory to capture the constraint networks that shape linguistic behavior. Such exemplars are merged into the weights [54, p. 320].

Baayen [54] corpus fits a NDL model with the following predictors: verb, semantic class of verb use, and length, animacy, definiteness, accessibility, and pronominality of recipient and theme. NDL provides a very good fit to the dative data set, which compares well to predictions obtained with other classifiers such as memory-based learning, mixed-effects logistic regression and support vector machine.

The prediction of the dative alternation is now a well-travelled path in quantitative linguistics, as evidenced by the high accuracy of the most recent methods. Yet, the community is in midstream. There is far more to the dative alternation than its prediction, since *predicting* is not *explaining*. We believe that this distinction is worth maintaining both at the conceptual and the operational levels. This idea is the backbone of our article.

4 Targeting the dative alternation in English

4.1 Data

We used the dative data set available in the languageR package [42, 55]. It contains 3,263 observations consisting of 15 variables. The variables divide into:

- speaker, a categorical variable with 424 levels, including NAs;
- modality, a categorical variable with 2 levels: spoken *vs.* written;
- verb, a categorical variable with 75 levels: *e.g. accord*, *afford*, *give*, *etc.*;
- semantic class, a categorical variable with 5 levels: abstract (*e.g. give* in *give it some thought*), transfer of possession (*e.g. send*), future transfer of possession (*e.g. owe*), prevention of possession (*e.g. deny*), and communication (*e.g. tell*);
- length in words of recipient, an integer valued variable;

- animacy of recipient, a categorical variable with 2 levels: animate *vs*. inanimate;
- definiteness of recipient, a categorical variable with 2 levels: definite *vs*. indefinite;
- pronominality of recipient, a categorical variable with 2 levels: pronominal *vs*. nonpronominal;
- length in words of theme, an integer valued variable;
- animacy of theme, a categorical variable with 2 levels: animate *vs*. inanimate;
- definiteness of theme, a categorical variable with 2 levels: definite *vs*. indefinite;
- pronominality of theme, a categorical variable with 2 levels: pronominal *vs*. nonpronominal;
- realization of recipient, a categorical variable with 2 levels: PD *vs*. DO;
- accessibility of recipient, a categorical variable with 3 levels: accessible, given, new;
- accessibility of theme, a categorical variable with 3 levels: accessible, given, new.

We considered speakers coded NA as mutually independent speakers, also independent from the set of identified speakers. About 80% of the identified speakers contribute more than one construction. This is a source of dependency between observations.

The approach we develop below takes this dependency into account. For the sake of clarity, we describe our approach in the context of independent observations. However, our results were obtained considering dependency.

4.2 Predicting and explaining the dative alternation

Our goal is to both *predict* and *explain* the dative alternation in English. In the next two subsections, we rephrase these two challenges in statistical terms. In a unifying probabilistic framework reflecting subject-matter knowledge, we specifically elaborate two statistical parameters *targeted* toward the above two goals. By "subject-matter knowledge" we mean what has been operationalized from what linguists know about the dative alternation and, more specifically, our data set. The parameters differ substantially because the two goals are radically different.

4.2.1 Predicting

Predicting the dative alternation in English means building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation. The objective could be to deceive a native English speaker sitting in front of a computer and trying to figure out whether his or her interlocutor is also a native English speaker. To do so, the player can only rely on limited information, namely a transcribed construction involving a dative alternation with contextual information. The algorithm does not need to tell us how the dative alternation works. Telling us how the alternation works falls within the scope of explaining it. It is the topic of Section 4.2.2.

For us to learn how to build such an algorithm based on experimental data, a random experiment ideally follows these steps:

1. randomly sample a generic member from the population of native English speakers;
2. observe her until she formulates either in thoughts, orally, or in writing, a construction that involves a dative alternation;
3. record the construction with all the available contextual information;
4. repeat the three above steps a large number of times.

Of course, realizations of this ideal experiment are out of reach. A less idealized, surrogate random experiment, say P_0 (P stands for "*p*robability", and 0 for "truth"), could go as follows: in an immense library gathering all spoken and written English documents produced by native English speakers during a period of interest:

1. randomly sample a document that contains at least one dative alternation;
2. randomly sample a dative alternation from it;
3. record the specific construction with all the available contextual information;
4. repeat the three above steps a large number of times.

We posit that the data set described in Section 4.1 is a set of realizations of a similar random experiment.

The random experiment P_0 is a complex byproduct of the English language seen itself as a probability distribution, or law. We invite the reader to think of P_0 as the quintessential law of the dative alternation. One might dispute this representation. We shall not go down the route of counter-arguing. We see random variation and change as inherent to natural phenomena. They are not errors. This conception of randomness is the byproduct of what Hacking [56] calls the "erosion of determinism". Thus it is legitimate, if not inescapable, for a scientific approach to reality in general, and to language in particular, to place variation and change at the core of the representation, not at its periphery (see Section 2.1).

The law P_0 fully describes the random production of an observation O that decomposes as $O = (W, Y)$. Here, $W \in \mathcal{W} \subset \mathbb{R}^d$ is the contextual information attached to the random construction summarized by O. As for $Y \in \{0,1\}$, it encodes the corresponding form taken by the dative alternation, say 0 for DO and 1 for PD, without loss of generality. Predicting the dative alternation in English requires that we learn a specific feature of P_0 that we call a statistical parameter. The statistician will first define a loss function to unequivocally identify which feature of P_0 she wants to unveil to predict the alternation. A loss function operationalizes the cost of a wrong prediction. The loss function underlies the definition of a statistical parameter.

One may want to minimize the overall probability to wrongly predict the dative alternation. In this case, one may choose the loss function ℓ whose cost is 1 if the prediction is incorrect and 0 otherwise. The construction of the predicting algorithm that we referred to at the very beginning of this section may involve ℓ at some point. Formally, ℓ maps any function f from $\mathcal{W}$ to $\{0,1\}$ and O to

$$\ell(f, O) = \mathbf{1}\{Y \neq f(W)\} = \begin{cases} 1 \text{ if } Y \neq f(W) \\ 0 \text{ otherwise} \end{cases}.$$

Indeed, the risk $R^{\ell}_{P_0}(f)$ of f which is, by definition, the mean value of the loss, satisfies $R^{\ell}_{P_0}(f) = E_{P_0}\{\ell(f, O)\} = P_0\{Y \neq f(W)\}$. Statisticians know well that $f \mapsto R^{\ell}_{P_0}(f)$ is minimized at the statistical parameter $f = \Phi(P_0)$ characterized by

$$\Phi(P_0)(W) = \mathbf{1}\{P_0(Y = 1|W) \geq 0.5\} \tag{1}$$

(see for instance Theorem 2.1 Devroye et al. [57]. Equality (1) means this: the optimal classification rule from the point of view of the loss ℓ is the so-called Bayes classifier which predicts PD if and only if PD is more likely to occur than DO in the current context.

The second statistical parameter $Q(P_0)$ characterized by

$$Q(P_0)(W) = P_0(Y = 1|W) \tag{2}$$

plays a crucial role in the prediction since knowing $Q(P_0)$ implies knowing $\Phi(P_0)$. Note that the reverse is false. In particular, eq. (1) suggests that if q is close to $Q(P_0)$ then f given by $f(W) = \{q(W) \geq 0.5\}$ should be close to $\Phi(P_0)$. We deduce that a predictor can be conveniently built by *(a)* approaching $Q(P_0)$ with a function q mapping $\mathcal{W}$ onto $[0,1]$, and *(b)* deriving by substitution the related classifier f given by $f(W) = \mathbf{1}\{q(W) \geq 0.5\}$. Another loss function is at play in this two-step procedure, namely, L which maps any function q from $\mathcal{W}$ to $[0,1]$ and O to $L(q, O) = (Y - q(W))^2$. Just like $\Phi(P_0)$ minimizes the risk $R^{\ell}_{P_0}$, $Q(P_0)$ minimizes the risk $R^{L}_{P_0}$ attached to L and characterized by $R^{L}_{P_0}(q) = E_{P_0}\{L(q, O)\}$.

It is important now to emphasize what the notation only suggests. The statistical parameters $\Phi(P_0)$ and $Q(P_0)$ are actually the values at P_0 of two functionals Φ and Q. These functionals map the set $\mathcal{M}$ of all laws compatible with the definition of O to the set of functions mapping $\mathcal{W}$ to $\{0,1\}$ and to the set of functions mapping $\mathcal{W}$ to $[0,1]$, respectively. Constraints on $\mathcal{M}$ must only reflect what the linguist knows for sure

about P_0. The linguist may know for instance that the first component of W is binary whereas its second and third components are categorical with three levels and integer values, respectively. In any case, the current state of the art on the dative alternation does not guarantee that $\mathcal{M}$ is parametric. Hence $\Phi(P_0)$ and $Q(P_0)$ do not belong to specific parametric models already known to us.

4.2.2 Explaining

In contrast, explaining the dative alternation in English means uncovering what drives the choice of one dative form over the other. This is certainly a multi-faceted challenge, one that cannot be exhausted and yet is worth being taken up for itself through a specifically designed analysis. To the best of our knowledge, however, such a targeted approach has not yet been carried out. It is indeed through the back-door that explanations have been sought so far, typically by *(a)* predicting the dative alternation, and *(b)* extracting features of the resulting estimator $\hat{\Phi}$ of $\Phi(P_0)$. For instance, Baayen [2] assesses non-parametrically the variable importance of the jth component W^j of the contextual information W on Y by comparing how well the predictor behaves when the information conveyed by W^j is either conserved or blurred. Specifically, a predictor $\hat{\Phi}$ is built based on the original data set. Then the observed values of W^j which the construction relies on are randomly permuted in order to break its potential relation with Y and a second predictor $\hat{\Phi}'$ is built. The greater the decrease in prediction performances of $\hat{\Phi}'$ is with respect to those of $\hat{\Phi}$, the greater the importance of W^j. Of course, resulting variable importance depends heavily on the prediction algorithm. Yet, a sensible variable importance should be defined universally. Let us see how we can define sound variable importance measures universally.

In Section 4.2.1, we imagined an ideal random experiment for the sake of learning to predict the dative alternation. What could an ideal experiment be for the sake of explaining it? More precisely, what could such an experiment be to assess the effect of each component of the contextual information on the dative alternation? We draw our inspiration from a common reasoning in the design and statistical analysis of randomized clinical trials for the sake of evaluating the effect of a drug on a disease. The interested reader will find an accessible review on this topic, presented as a trialogue between a philosopher, a medical doctor and a statistician, in (see Sections 3, 8, and 9 in particular [58]. We consider in turn how to proceed with a categorical component as opposed to a non-categorical component.

4.2.3 Assessing the effect of a categorical contextual variable on the dative alternation

First, let us clarify what we mean by the importance of W^j on Y, with $j \in J$, the set of indices of the categorical components of W (there are many ways of doing it). To keep things simple, we consider a categorical variable, say W^1, with two levels only, *e.g.* the animacy of recipient with its levels animate and inanimate. We denote the levels by 0 and 1, without loss of generality. An ideal random experiment could go along these lines:

1. randomly sample a generic member from the population of native English speakers;
2. randomly sample some contextual information W, and a message to convey;
3. give her all this information except W^1, some partial contextual information, which we denote W^{-1};
4. ask her to formulate a construction involving a dative alternation to convey the message under the constraint $W^1 = 0$;
5. record the resulting form of the alternation, which we denote Y_0^1;
6. take her back in time and ask her to formulate a construction involving a dative alternation to convey the message under the constraint $W^1 = 1$;
7. record the resulting form of the alternation, which we denote Y_1^1;
8. repeat the seven above steps a large number of times.

Here and henceforth, the superscript "1" refers to the fact that we intervene on W^1 while the subscripts "0" and "1" refer to the fact that W^1 is set to 0 and 1, respectively. The two forms of the dative alternation Y_0^1 and Y_1^1 are obtained *ceteris paribus sic standibus*, *i.e.* all other things being equal. Within this conceptual framework, the form of the dative alternation that would have been observed had the speaker been given all the contextual information W (and not W^{-1} and an additional constraint on W^1) would have been $Y = Y^1_{W^1}$, *i.e.* $Y = Y_0^1$ if $W^1 = 0$ and $Y = Y_1^1$ if $W^1 = 1$. The variables Y_0^1 and Y_1^1 are called counterfactuals in causal analysis [59].

If we denote $\mathbb{P}_0^1$ the law of the above ideal random experiment, then the difference $E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\} = \mathbb{P}_0^1(Y_1^1 = 1) - \mathbb{P}_0^1(Y_0^1 = 1)$ can be interpreted as an "effect" of W^1 on Y all other things being equal. Note that this is a parameter of $\mathbb{P}_0^1$. Moreover, if we could indeed sample data from $\mathbb{P}_0^1$ (time travel is not a realistic option yet), then the statistical inference of the latter parameter would be child's play based on the trivial estimator $(1/n)\sum_{i=1}^n (Y_{i,1}^1 - Y_{i,0}^1)$, with n the sample size and $(Y_{i,0}^1, Y_{i,1}^1)$ the ith counterfactual outcome.

It turns out that $\mathbb{P}_0^1$ and the less idealized, surrogate random experiment P_0 that we introduced in Section 4.2.1 can be modeled altogether by means of a non-parametric system of structural equations, a notion which originates in the works of Wright [60], Haavelmo [61] and was brought up-to-date by Pearl [59].

Let us now describe a system of structural equations that encapsulates both $\mathbb{P}_0^1$ and P_0. We characterize the variable importance of W^1 on Y as a parameter of $\mathbb{P}_0^1$. Unfortunately, it is not possible to sample observations from $\mathbb{P}_0^1$, so that one might be tempted to give up on estimating this parameter. Fortunately, the system of structural equations that links $\mathbb{P}_0^1$ and P_0 offers the opportunity to see the apparently inaccessible parameter of $\mathbb{P}_0^1$ as a parameter of P_0 that we can estimate based on data sampled from P_0.

Assume that there exist two deterministic functions F and f, taking their values in $\mathcal{W}$ and $\{0,1\}$, respectively, and a source of randomness (U,V) such that sampling $O = (W,Y)$ from P_0 is equivalent to *(a)* sampling (U,V) from its law and *(b)* computing, deterministically given (U,V),

$$\begin{cases} W &= F(U) \\ Y &= f(W,V) \end{cases}. \tag{3}$$

Model (3) is our first system of structural equations. It is quite general. In particular, taking F equal to the identity (*i.e.* $F(w) = w$ for all $w \in \mathcal{W}$) and $U = W$ yields that a model of the form (3) for P_0 exists whenever Y can be written as an implicit function of W and additional terms, at the exception of Y itself, gathered in a variable that we call V. Necessarily, eq. (3) can be rewritten under the equivalent form

$$\begin{cases} W^j &= F^j(U^j), \quad j = 1,\dots,d \\ Y &= f((W^1,\dots,W^d),V) \end{cases} \tag{4}$$

for some deterministic functions $F^1,\dots,F^d$ derived from F, the same f as in eq. (3), and some source of randomness $(U^1,\dots,U^d,V)$. Now, note that eq. (4) allows us to define the following system

$$\begin{cases} W^j &= F^j(U^j), \quad j = 1,\dots,d \\ Y_0^1 &= f((0,W^2,\dots,W^d),V) \\ Y_1^1 &= f((1,W^2,\dots,W^d),V) \\ Y &= Y^1_{W^1} \end{cases}, \tag{5}$$

provided that the second and third equations always make sense. What is changed there is the value of the first component of the first argument of f. We substitute either 0 or 1 for W^1. Model (5) gives us a joint model for $\mathbb{P}_0^1$ and P_0. Furthermore, eq. (5) allows to define a counterpart to $E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\}$ characterized as a statistical parameter of P_0.

Let us now introduce the functional which maps the set $\mathcal{M}$ to $[-1,1]$ and is given at any $P \in \mathcal{M}$ by

$$\begin{aligned} \Psi^1(P) &= E_P\{P(Y=1|W^1=1,W^{-1}) - P(Y=1|W^1=0,W^{-1})\} \\ &= E_P\{Q(P)(1,W^{-1}) - Q(P)(0,W^{-1})\}, \end{aligned} \tag{6}$$

because $Q(P)(W) = P(Y = 1|W)$ (see eq. (2) for the case $P = P_0$). It is well-known to statisticians that under suitable, untestable assumptions, $\Psi^1(P_0) = E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\}$. We state this result formally and give its simple proof in Section A.1. The equality grants Ψ^1 a causal interpretation.

The fact that W^1 takes only two different values plays a minor role in the above argument. Say that W^2 takes $(K+1)$ different values with $K \geq 1$ and denote these values by $0, \ldots, K$. In addition to eqs (5), (4) also yields the following system

$$\begin{cases} W^j &= F^j(U^j), \quad j = 1, \ldots, d \\ Y_k^2 &= f((W^1, k, W^3, \ldots, W^d), V), \quad k = 0, \ldots, K, \\ Y &= Y_{W^2}^2 \end{cases} \tag{7}$$

provided that the second equation always makes sense. What is changed there is the value of the second component of the first argument of f. We substitute $0, \ldots, K$ for W^2. Model (7) gives us a joint model for P_0 and $\mathbb{P}_0^2$, the law of the ideal random experiment where we intervene on W^2 instead of W^1. The counterpart to the parameter of $\mathbb{P}_0^1$ that we introduced earlier is merely the collection of parameters $(E_{\mathbb{P}_0^2}\{Y_k^2\} - E_{\mathbb{P}_0^2}\{Y_0^2\} = \mathbb{P}_0^2(Y_k^2 = 1) - \mathbb{P}_0^2(Y_0^2 = 1) : k = 1, \ldots, K)$, where $W^2 = 0$ serves as a reference level. As for the related statistical parameter of P_0, it is the value at P_0 of the functional Ψ^2 which maps the set $\mathcal{M}$ to $[-1, 1]^K$ and is given at any $P \in \mathcal{M}$ by $\Psi^2(P) = (\Psi_k^2(P) : 1 \leq k \leq K)$ with

$$\begin{aligned} \Psi_k^2(P) &= E_P\{P(Y = 1|W^2 = k, W^{-2}) - P(Y = 1|W^2 = 0, W^{-2})\} \\ &= E_P\{Q(P)(W^1, k, W^3, \ldots, W^d) - Q(P)(W^1, 0, W^3, \ldots, W^d)\}, \end{aligned} \tag{8}$$

where W^{-2} equals W deprived from its second component W^2. One can also endow Ψ^2 with a causal interpretation under suitable, untestable assumptions.

4.2.4 Assessing the effect of an integer valued contextual variable on the dative alternation

We now turn to the elaboration of a notion of the importance of W^j on Y, with $j \notin J$, *i.e.* W^j is an integer valued contextual variable. Say that $W^3 \in \mathbb{N}$ is such a variable. Drawing inspiration from the way we defined the importance of W^2 based on the definition of the importance of W^1, one might think of treating W^3 like a categorical contextual variable that can take many different values. This option has several drawbacks. First, we would lose the inherent information provided by the ordering of integers. Second, we might have to infer many different statistical parameters if W^3 does take many different values. The proliferation of statistical parameters makes it less likely to extract significant results from our analysis due to an unavoidable, more stringent multiple testing procedure. To circumvent this, we define a statistical parameter of a different kind.

We rely again on eq. (4) to carve out a new system similar to systems (5) and (7). The resulting statistical parameter is tailored to the fact that the importance we wish to quantify is that of a non-categorical variable. Let $\mathcal{W}^3 \subset \mathbb{N}$ be the set of values that W^3 can take. The new system is

$$\begin{cases} W^j &= F^j(U^j), \quad j = 1, \ldots, d \\ Y_w^3 &= f((W^1, W^2, w, W^4, \ldots, W^d), V), \quad \text{all } w \in \mathcal{W}^3, \\ Y &= Y_{W^3}^3 \end{cases} \tag{9}$$

provided that the second equation always makes sense. Among other things, system (9) induces a model for $\mathbb{P}_0^3$, the law of the ideal random experiment where we intervene on W^3 instead of W^1 or W^2. Based on systems (5) and (7), we introduced $\mathbb{P}_0^1$, $\mathbb{P}_0^2$, and some parameters of the latter which are interpretable as importance measures. In the present situation, though, we cannot yet introduce our parameter of $\mathbb{P}_0^3$ that will serve as an importance measure of W^3. We still need two more ingredients to reduce the dimensionality of the problem at stake.

The first ingredient is a so-called marginal structural model, a statistical model for the function $w \mapsto E_{\mathbb{P}_0^3}\{Y_w^3\}$ which maps $\mathcal{W}^3$ to $[0,1]$, *i.e.* a parametric set $\mathcal{F} = \{w \mapsto f_\theta(w) : \theta \in \Theta\}$ of functions mapping $\mathcal{W}^3$ to $[0,1]$, indexed by a finite-dimensional parameter $\theta \in \Theta$. The second ingredient is merely a weight function h mapping $\mathcal{W}^3$ to $\mathbb{R}_+$ such that $\sum_{w\in\mathcal{W}^3} h(w) < \infty$. Based on $\mathcal{F}$ and h, we can now propose the following parameter of $\mathbb{P}_0^3$ as a measure of the importance of W^3 on Y:

$$\arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\Big(E_{\mathbb{P}_0^3}\{Y_w^3\}, f_\theta(w)\Big) = \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\big(\mathbb{P}_0^3(Y_w^3 = 1), f_\theta(w)\big), \tag{10}$$

where we use the notation $\Lambda(p,p') = p\log(p') + (1-p)\log(1-p')$ for all $p \in [0,1]$ and $p' \in]0,1[$. Robins [62] first introduced marginal structural models in causal analysis. Robins et al. [63] discuss their use in epidemiology. More recently, Rosenblum et al. [64] use them to define and estimate the impact of adherence to antiretroviral therapy on virologic failure in HIV infected patients.

As opposed to the previous parameters $E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\}$ on the one hand and $(E_{\mathbb{P}_0^2}\{Y_k^2\} - E_{\mathbb{P}_0^2}\{Y_0^2\} : k = 1,\dots,K)$ on the other hand, eq. (10) has no closed-form explicit expression in terms of $\mathbb{P}_0^3$ in general. However, its implicit characterization gives us a direct interpretation. Parameter (10) is a specific $\theta \in \Theta$ such that f_θ is closer to $w \mapsto E_{\mathbb{P}_0^3}\{Y_w^3\}$ than every other $f_{\theta'}$, where the gap between two functions f, f' mapping $\mathcal{W}^3$ to $[0,1]$ is measured by

$$\begin{aligned}&\sum_{w\in\mathcal{W}^3} h(w)[f(w)log(f/f'(w)) + (1-f(w))\log((1-f)/(1-f')(w))]\\ &= -\sum_{w\in\mathcal{W}^3} h(w)\Lambda(f(w), f'(w)) + \sum_{w\in\mathcal{W}^3} h(w)[f(w)\log(f(w)) + (1-f(w))log((1-f)(w))].\end{aligned}$$

The above is a so-called integrated Kullback-Leibler divergence. The minus sign before the first term in the RHS of the above display explains why eq. (10) involves an arg max and not an arg min. In particular, if $w \mapsto E_{\mathbb{P}_0^3}\{Y_w^3\}$ coincides with f_θ for some $\theta \in \Theta$ and if the weight function h only takes positive values then eq. (10) equals θ. This is very unlikely. If, on the contrary, no f_θ equals $w \mapsto E_{\mathbb{P}_0^3}\{Y_w^3\}$ then eq. (10) can still be interpreted as the projection of the latter onto $\mathcal{F}$.

Often, users of logistic regression models take for granted that their model assumptions are met by the true, unknown law of their data. They are unaware of the precautionary measures required when assessing the results of a fit. This is especially true for the interpretation of the pointwise estimates, and for the reliability of the confidence intervals, which comes at a high price in terms of untestable assumptions about the true, unknown law of the data. We refer the reader to the discussion about the effect of definiteness of theme in Section 6 to hammer home this important point.

Because the set $\mathcal{F}$ is a tool that does not contain the truth, it is often referred to as a "working model". It is selected so as to retrieve information on how $E_{\mathbb{P}_0^3}\{Y_w^3\}$ depends upon w. For technical reasons, $\mathcal{F}$ must be identifiable, *i.e.* such that $f_\theta = f_{\theta'}$ implies $\theta = \theta'$. Recall that expit and logit are two reciprocal functions characterized on $\mathbb{R}$ and $[0,1]$ by $\mathrm{expit}(q) = 1/(1+e^{-q})$ and $\mathrm{logit}(p) = \log(p/(1-p))$, respectively. In this article, we consider the set

$$\mathcal{F} = \big\{w \mapsto \mathrm{expit}\big(\theta_0 + \theta_1 w + \theta_2 w^2\big) : \theta = (\theta_0,\theta_1,\theta_2) \in \Theta = \mathbb{R}^3\big\}, \tag{11}$$

and assume that eq. (10) uniquely defines a single element of Θ for this specific choice of $\mathcal{F}$, an assumption that cannot be tested on data. Thus, parameter (10) should be understood as the best second-order polynomial approximation to $w \mapsto \mathrm{logit}\Big(E_{\mathbb{P}_0^3}\{Y_w^3\}\Big)$ with respect to the aforementioned gap.

By analogy, it is now time to characterize a statistical parameter of P_0 which is a good proxy to eq. (10) in the sense that *(a)* under appropriate assumptions it is equal to eq. (10) and *(b)* it can be inferred from data sampled from P_0. Let Ψ^3 be defined as the function mapping $\mathcal{M}$ to Θ such that, for any $P \in \mathcal{M}$,

$$\begin{aligned}\Psi^3(P) &= \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\big(E_P\{P(Y=1|W^3=w,W^{-3})\}, f_\theta(w)\big)\\ &= \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\big(E_P\{Q(P)(W^1,W^2,w,W^4,\dots,W^d)\}, f_\theta(w)\big),\end{aligned} \tag{12}$$

where W^{-3} equals W deprived from its third component W^3. Here too, a lemma similar to Lemma 1 may guarantee that $\Psi^3(P_0)$ coincides with eq. (10) under suitable, untestable assumptions.

5 Statistical apparatus

Now that the parameters we wish to infer are specified, we turn to their targeted estimation. The targeted estimation relies on machine learning prediction, see Section 5.1, followed by targeted minimum loss explanation, see Section 5.2.

5.1 Machine learning prediction

We consider first the inference of $Q(P_0)$ as defined in eq. (2). The literature on the topic of classification, both from the theoretical and applied points of view, is too vast to select a handful of outstanding references. Instead of choosing one particular approach, we advocate for considering all our favorite approaches, seen as a library of algorithms, and combining them into a meta-algorithm drawing data-adaptively the best from each of them. Many methods have been proposed in this spirit, now gathered under the name of "ensemble learners" (see to cite only a few seminal works, with an emphasis on methods using the cross-validation principle [65–69]). Specifically, we choose to rely on the super-learning methodology [3, 70].

We now give a nutshell description of the super-learning methodology. Say that we have n independent observations $O_1 = (W_1, Y_1), \ldots, O_n = (W_n, Y_n)$ drawn from P_0 and an arbitrarily chosen partition of $\{1, \ldots, n\}$, *i.e.* a collection of sets $\{T(v) \subset \{1, \ldots, n\} : 1 \leq v \leq V\}$ such that $\cup_{v=1}^{V} T(v) = \{1, \ldots, n\}$ (their union covers $\{1, \ldots, n\}$) and for each $1 \leq v_1 \neq v_2 \leq V$, $T(v_1) \cap T(v_2) = \emptyset$ (the sets are pairwise disjoint). For convenience, we introduce the notation $P_{n,S}$ to represent the subset $\{O_i : i \in S\}$ of the complete data set, represented by P_n, corresponding to these observations index by $i \in S \subset \{1, \ldots, n\}$. We use the data to infer the best combination of K algorithms $\hat{Q}_1, \ldots, \hat{Q}_K$ which map any subset of the data set to a function from $\mathcal{W}$ to $[0, 1]$. For instance, $\hat{Q}_1(P_{n,T(2)})(W)$ is the predicted conditional probability that $Y = 1$ given W according to the first algorithm trained on $\{O_i : i \in T(2)\}$. Among a variety of possible ways to combine $\hat{Q}_1, \ldots, \hat{Q}_K$ we decide to resort to convex combinations: thus, for each $\alpha \in \mathcal{A} = \{a \in \mathbb{R}_+^K : \sum_{k=1}^K a_k = 1\}$, we define $\hat{Q}_\alpha = \sum_{k=1}^K \alpha_k \hat{Q}_k$, the meta-algorithm mapping any subset $P_{n,S}$ of the data set to the function $\hat{Q}_\alpha(P_{n,S}) = \sum_{k=1}^K \alpha_k \hat{Q}_k(P_{n,S})$ from $\mathcal{W}$ to $[0, 1]$. Note that if every $\hat{Q}_k$ produces functions mapping $\mathcal{W}$ to $[0, 1]$ then so does $\hat{Q}_\alpha$ for any $\alpha \in \mathcal{A}$.

Recall that the risk $R^L_{P_0}(\hat{Q}_\alpha(P_{n,S})) = E_{P_0}\{L(\hat{Q}_\alpha(P_{n,S}), O)\}$ quantifies how close $\hat{Q}_\alpha(P_{n,S})$ is to $Q(P_0)$, the parameter of P_0 that we wish to target. Of course, we cannot compute $R^L_{P_0}(\hat{Q}_\alpha(P_{n,S}))$ in general because we do not know P_0. Its estimator

$$\begin{aligned} R^L_{P_{n,S}}(\hat{Q}_\alpha(P_{n,S})) &= E_{P_{n,S}}\{L(\hat{Q}_\alpha(P_{n,S}), O)\} \\ &= \frac{\sum_{i \in S} L(\hat{Q}_\alpha(P_{n,S}), O_i)}{\mathrm{card}(S)} \end{aligned}$$

is known to be over-optimistic, since the same data are involved in the construction of $\hat{Q}_\alpha(P_{n,S})$ and in the evaluation of how well it performs. Cross-validation offers a powerful way to circumvent this: the cross-validated estimator

$$R^L_{P_n}(\hat{Q}_\alpha) = \frac{1}{V} \sum_{v=1}^{V} \frac{\sum_{i \in T(v)^c} L(\hat{Q}_\alpha(P_{n,T(v)}), O_i)}{\mathrm{card}(T(v)^c)} \tag{13}$$

(we slightly abuse notation) accurately evaluates how good are the estimators of $Q(P_0)$ produced by the α-indexed meta-algorithm $\hat{Q}_\alpha$. They key is that in each term of the RHS of eq. (13), the subset of data used to

"train" $\hat{Q}_\alpha$, represented by $P_{n,T(v)}$, and the subset used to evaluate its performances, represented by $P_{n,T(v)^c}$, are disjoint. This motivates the introduction of

$$\alpha_n = \arg\min_{\alpha\in\mathcal{A}} R^L_{P_n}(\hat{Q}_\alpha), \tag{14}$$

the minimizer of the cross-validated risk, which finally yields the super-learner

$$\hat{Q}_{\alpha_n}(P_n)$$

by training $\hat{Q}_{\alpha_n}$ on the complete data set. It can be shown that, if every $\hat{Q}_k$ produces functions mapping $\mathcal{W}$ to $[0,1]$ then the super-learner performs almost as well as the so-called "oracle" (since it cannot be inferred without knowing the true law P_0) best algorithm in the library. We refer the reader to Section A.2.1 for a more accurate mathematical statement of this remarkable fact.

5.2 Targeted minimum loss explanation

We now turn to the estimation of $\Psi^1(P_0)$, $\Psi^2(P_0)$ and $\Psi^3(P_0)$ as defined in eqs (6), (8) and (12). We take the route of TMLE, a paradigm of inference based on semiparametrics and estimating functions (see Chapter 25, for recent and comprehensive introductions [71, 72]). Introduced by van der Laan and Rubin [4], TMLE has been studied and applied in a variety of contexts since then (we refer to for an overview [3]). An accessible introduction to TMLE is given in (Sections 12, 13 and 14 [58]).

It is apparent in eqs (6), (8) and (12) that the parameters $\Psi^1(P_0)$, $\Psi^2(P_0)$ and $\Psi^3(P_0)$ all depend on $Q(P_0)$. Let us assume that we have already built an estimator of $Q(P_0)$, which we denote by Q^{init}_n– that could be, for instance, the super-learner $\hat{Q}_{\alpha_n}(P_n)$ whose construction we described in Section 5.1. Here, the superscript "init" indicates that we think of Q^{init}_n as an initial estimator of $Q(P_0)$ built for the sake of predicting, not explaining.

Taking a closer look at eqs (6), (8) and (12), we see that it is easy to estimate $\Psi^1(P_0)$, $\Psi^2(P_0)$ and $\Psi^3(P_0)$ by relying on Q^{init}_n. Consider eq. (6): if we substitute Q^{init}_n for $Q(P)$ in the formula, then only the marginal law of W^{-1} is left unspecified. The simplest way to estimate the latter, which can be shown to be the most efficient too, is to use its empirical counterpart. That means estimating the marginal law of W^{-1} by the empirical law under which $W^{-1} = W_i^{-1}$, the ith observed value of W^{-1} in the data set, with probability $1/n$. Substituting the empirical marginal law of W^{-1} for its counterpart under P in eq. (6) yields an initial estimator of $\Psi^1(P_0)$, say $\psi^{1,\text{init}}_n$, writing as

$$\begin{aligned}\psi^{1,\text{init}}_n &= E_{P_n}\{Q^{\text{init}}_n(1,W^{-1}) - Q^{\text{init}}_n(0,W^{-1})\}\\ &= \frac{1}{n}\sum_{i=1}^n [Q^{\text{init}}_n(1,W_i^{-1}) - Q^{\text{init}}_n(0,W_i^{-1})].\end{aligned}$$

[Correction added after online publication 11 December 2015: "Substituting the empirical marginal law of.." should read "Substituting the empirical marginal law of W^{-1}"]

Likewise, the parameter $\Psi^2(P_0)$ can be simply estimated by $\psi^{2,\text{init}}_n = (\psi^{2,\text{init}}_{k,n} : 1\le k\le K)$ with

$$\begin{aligned}\psi^{2,\text{init}}_{k,n} &= E_{P_n}\{Q^{\text{init}}_n(W^1,k,W^3,\dots,W^d) - Q^{\text{init}}_n(W^1,0,W^3,\dots,W^d)\}\\ &= \frac{1}{n}\sum_{i=1}^n [Q^{\text{init}}_n(W^1_i,k,W^3_i,\dots,W^d_i) - Q^{\text{init}}_n(W^1_i,0,W^3_i,\dots,W^d_i)]\end{aligned}$$

while the parameter $\Psi^3(P_0)$ can be estimated by

$$\psi^{3,\text{init}}_n = \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\big(E_{P_n}\{Q^{\text{init}}_n(W^1,W^2,w,W^4,\dots,W^d)\}, f_\theta(w)\big) \tag{15}$$

$$= \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\left(\frac{1}{n}\sum_{i=1}^n Q^{\text{init}}_n(W^1_i,W^2_i,w,W^4_i,\dots,W^d_i), f_\theta(w)\right). \tag{16}$$

Interestingly, the optimization problem eq. (15) can be solved easily, see Section A.3.3.

Arguably, $\psi_n^{1,\text{init}}$, $\psi_n^{2,\text{init}}$ and $\psi_n^{3,\text{init}}$ are not targeted toward $\Psi^1(P_0)$, $\Psi^2(P_0)$ and $\Psi^3(P_0)$ in the sense that, although they are obtained by substitution, the key estimator Q_n^{init} which plays a crucial role in their definitions was built for the sake of prediction and not specifically tailored for estimating either $\Psi^1(P_0)$, $\Psi^2(P_0)$ or $\Psi^3(P_0)$. In this respect, the targeting step of TMLE can be presented as a general statistical methodology to derive new substitution estimators from such initial estimators so that the updated ones really target what they aim at.

Targeting is made possible because Ψ^1, Ψ^2 and Ψ^3, seen as functions mapping $\mathcal{M}$ to $[-1,1]$, $[-1,1]^K$ and Θ, respectively, are differentiable, see Section A.3.1. In these three cases, the resulting gradients (derivatives), denoted by $\nabla\Psi^1$, $\nabla\Psi^2$ and $\nabla\Psi^3$, drive our choices of estimating functions. Targeting the parameter of interest consists in *(a)* designing a collection $\{Q_{n,\varepsilon}^{\text{init}} : \varepsilon \in \mathcal{E}\}$ of functions mapping $\mathcal{W}^3$ to $[0,1]$ conceived as fluctuations of $Q_n^{\text{init}} = Q_{n,\varepsilon}^{\text{init}}\Big|_{\varepsilon=0}$ in the direction of the parameter of interest, and *(b)* identifying that specific element of the collection which better targets the parameter of interest, see Section A.3.2. Let us denote by $Q_n^{1,\text{targ}} = Q_{n,\varepsilon_n^1}^{\text{init}}$, $Q_n^{2,\text{targ}} = Q_{n,\varepsilon_n^2}^{\text{init}}$ and $Q_n^{3,\text{targ}} = Q_{n,\varepsilon_n^3}^{\text{init}}$ the three a priori different fluctuations of Q_n^{init} that respectively target $\Psi^1(P_0)$, $\Psi^2(P_0)$, and $\Psi^3(P_0)$. They finally yield, by substitution, the three estimators

$$\psi_n^{1,\text{targ}} = \frac{1}{n}\sum_{i=1}^{n}\left[Q_n^{1,\text{targ}}(1, W_i^{-1}) - Q_n^{1,\text{targ}}(0, W_i^{-1})\right], \tag{17}$$

$$\psi_n^{2,\text{targ}} = (\psi_{k,n}^{2,\text{targ}} : 1 \le k \le K) \qquad \text{where, for each } 1 \le k \le K, \tag{18}$$

$$\psi_{k,n}^{2,\text{targ}} = \tfrac{1}{n}\sum_{i=1}^{n}\left[Q_n^{2,\text{targ}}(W_i^1, k, W_i^3, \ldots, W_i^d) - Q_n^{2,\text{targ}}(W_i^1, 0, W_i^3, \ldots, W_i^d)\right],$$

$$\begin{aligned}\psi_n^{3,\text{targ}} &= \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3} h(w)\Lambda\left(\frac{1}{n}\sum_{i=1}^{n} Q_n^{3,\text{targ}}(W_i^1, W_i^2, w, W_i^4, \ldots, W_i^d), f_\theta(w)\right)\\ &= \arg\max_{\theta\in\Theta} \sum_{w\in\mathcal{W}^3}\sum_{i=1}^{n} h(w)\Lambda\left(Q_n^{3,\text{targ}}(W_i^1, W_i^2, w, W_i^4, \ldots, W_i^d), f_\theta(w)\right).\end{aligned} \tag{19}$$

The optimization problem (19) can be solved easily just like eq. (15), see Section A.3.3.

The above estimators satisfy $\psi_n^{1,\text{targ}} = \Psi^1(P_n^{1,\text{targ}})$, $\psi_{k,n}^{2,\text{targ}} = \Psi_k^2(P_n^{2,\text{targ}})$, $\psi_n^{3,\text{targ}} = \Psi^3(P_n^{3,\text{targ}})$ for three empirical laws $P_n^{1,\text{targ}}, P_n^{2,\text{targ}}, P_n^{3,\text{targ}} \in \mathcal{M}$. They are targeted in the sense that they satisfy $E_{P_n}\{\nabla\Psi^1(P_n^{1,\text{targ}})(O)\} = 0$, $E_{P_n}\{\nabla\Psi^2(P_n^{2,\text{targ}})(O)\} = 0$, $E_{P_n}\{\nabla\Psi^3(P_n^{3,\text{targ}})(O)\} = 0$, three equalities which are the core of the theoretical study of their asymptotic properties. The two main properties concern the consistency of the estimators and the construction of asymptotic confidence intervals. An estimator is consistent if it converges to the truth when the sample size goes to infinity. The targeted estimators defined in eqs (17), (18) and (19) are double-robust: the stronger requirement for them to be consistent is that *either* the corresponding targeted estimator of $Q(P_0)$, say Q_n^{targ}, converge to $Q(P_0)$ *or* the conditional law of the variable whose importance is sought given the other components of W, say $g(P_0)$, be consistently estimated by, say, g_n. Furthermore, the stronger requirement to make it possible to build asymptotically conservative confidence intervals is that the product of the rates of convergence of Q_n^{targ} to $Q(P_0)$ and of g_n to $g(P_0)$ be faster than $1/\sqrt{n}$. Finally, we wish to acknowledge that it is possible to target all parameters with a single, specifically designed, richer collection of fluctuations. Targeting all parameters at once enables the construction of simultaneous confidence regions that better take the mutual dependency of the estimators into account. In a problem with higher stakes, we would have gone that bumpier route.

6 Application

We consider in turn every component of the contextual information variable W and estimate its effect on the dative alternation as defined in Section 4.2.2 along the lines presented in Section 5. We systematically report

95%-confidence intervals and p-values when testing whether the parameter is equal to 0 or not. We emphasize that these are not simultaneous 95%-confidence intervals. It is possible, however, to use the p-values to carry out a multiple testing procedure, controlling a user-supplied type-I error rate such as the familywise error rate.

As explained in Section 4.1, the forthcoming results are obtained with consideration for speaker-related dependency, see Section A.3.4.

6.1 Categorical contextual information variables

Let us now comment on the results of Table 1. We disregard the estimates whose p-values are large, because they correspond to insignificant results. We arbitrarily set our p-value threshold to 1%. An estimate ψ_n of the effect of setting $W = w_1$ as opposed to setting $W = w_0$ can be interpreted as follows: all other things being equal, the probability of obtaining a PD construction increases/decreases additively by ψ_n when W is set to w_1 as opposed to. Ranked by decreasing magnitude of the estimates, we obtain:

- a 38.24% decrease when accessibility of recipient switches from accessible to new;
- a 16.57% increase when semantic class switches from abstract to communication meaning;
- a 14.71% decrease when semantic class switches from abstract to future transfer of possession meaning;
- a 13.98% decrease when pronominality of recipient switches from nonpronominal to pronominal;
- a 11.68% decrease when pronominality of theme switches from nonpronominal to pronominal, see examples (xxii) and (xxiii);
- a 11.52% increase when semantic class switches from abstract to transfer meaning;
- a 9.38% increase when animacy of recipient switches from animate to inanimate, see example (xx);
- a 9.28% decrease when semantic class switches from abstract to prevention of possession meaning;
- a 8.43% increase when animacy of theme switches from animate to inanimate;
- a 7.82% decrease when accessibility of theme switches from accessible to new;
- a 5.68% decrease when definiteness of theme switches from definite to indefinite, see example (xxi);
- a 3.95% increase when definiteness of recipient switches from definite to indefinite.

Table 1: Estimated effects of the categorical information variables.

Variable	*Versus*	Estimate	CI	*p*-Value
Modality	written%spoken	0.0277	[−0.0031,0.0585]	0.0776
AnimacyOfRec	inanimate%animate	0.0938	[0.0549,0.1327]	0.0000
DefinOfRec	indefinite%definite	0.0395	[0.0102,0.0688]	0.0083
PronomOfRec	pronominal%nonpronominal	−0.1398	[−0.2171,−0.0624]	0.0004
AnimacyOfTheme	inanimate%animate	0.0843	[0.0337,0.1348]	0.0011
DefinOfTheme	indefinite%definite	−0.0568	[−0.0865,−0.0272]	0.0002
PronomOfTheme	pronominal%nonpronominal	−0.1168	[−0.1377,−0.0959]	0.0000
AccessOfRec	new%accessible	−0.3824	[−0.5458,−0.2189]	0.0000
	given%accessible	0.0411	[−0.0149,0.0971]	0.1506
AccessOfTheme	new%accessible	−0.0782	[−0.1100,−0.0463]	0.0000
	given%accessible	−0.0415	[−0.0673,−0.0157]	0.0016
SemanticClass	t%a	0.1152	[0.0548,0.1755]	0.0002
	p%a	−0.0928	[−0.1532,−0.0324]	0.0026
	f%a	−0.1471	[−0.1946,−0.0997]	0.0000
	c%a	0.1657	[0.1238,0.2077]	0.0000

Source: For each such contextual information (named in the first column) and each comparison (possibly several, identified in the second column), we report the corresponding estimated effect(s), 95%-confidence interval(s) and p-value(s) when testing whether the parameter is equal to 0 or not (in the third, fourth and fifth columns, respectively).

As we go down the list, differences in acceptability are less striking. This reflects the fact that the corresponding estimates get smaller. Let us comment on the above findings about the importance of animacy of recipient, definiteness of theme, and pronominality of theme. We deliberately follow the steps of the thought experiment process designed in Section 4.2.2.

Consider for instance example (xx): under the constraint "set the animacy of recipient to inanimate", the speaker selects either (xx a) or (xx b); under the constraint "set the animacy of recipient to animate", she selects either (xx c) or (xx d). What matters is the extent to which the probability to select the PD construction is altered when one switches from one constraint to the other. Even if linguists might find (xx d) slightly more natural than (xx c), (xx a) is undoubtedly more natural than (xx b). This is consonant with our result, which states that the probability of the PD construction increases when the animacy of recipient is set from animate to inanimate.

(xx) a. Anthony gave $100 to charity.
b. Anthony gave charity $100.
c. Anthony gave $100 to Will.
d. Anthony gave Will $100.

Illustrating the inferred statement about the effect of definiteness of theme is challenging. We see this as a welcome opportunity to emphasize the singularity of our statistical approach. To produce a convincing example, we have to choose a longer theme than before. Indeed, linguists know for a fact that when the theme is long, PD is dispreferred. In example (xxi), one can conceive that the preference of (xxi d) over (xxi c) is slightly stronger than that of (xxi b) over (xxi a). This is consonant with our result, which states that the probability of the PD construction decreases slightly when the definiteness of theme is set from definite to indefinite.

(xxi) a. Anthony bought the incredibly good cake for Will.
b. Anthony bought Will the incredibly good cake.
c. Anthony bought an incredibly good cake for Will.
d. Anthony bought Will an incredibly good cake.

Example (xxi) is clearly counterintuitive to linguists used to interpreting results from logistic regression models. This is a common pitfall. It is due to the belief that the interpretation of a fitted logistic regression still holds even when the true law does not belong to the logistic model. This is never the case. From a mathematical point of view, the parameter matching definiteness of theme in a logistic regression model is a very awkward function of the true law. No matter how awkward the function is, no sensible interpretation can be built without it. In contrast, the parameter we define and estimate to assess the effect of definiteness of theme is a rather simple function of the true law. Moreover, its simple statistical interpretation is buttressed by a causal interpretation, at the cost of untestable assumptions. The above lines epitomize the approach defended in this article.

How do statisticians intuit then? Denote W^1 the definiteness of theme ($W^1 = 1$ for indefinite and $W^1 = 0$ for definite), W^2 the length of theme, and consider this baby model, tweaked for demonstration purposes. Say, contrary to facts, that the true difference $P_0(Y = 1|W^1 = 1, W^{-1}) - P_0(Y = 1|W^1 = 0, W^{-1})$ depends on W^{-1} only through a thresholded version of W^2. More precisely, say that

$$P_0(Y = 1|W^1 = 1, W^{-1}) - P_0(Y = 1|W^1 = 0, W^{-1}) = \begin{cases} 1.00\% & \text{if} \quad W^2 \leq 2 \\ -8.54\% & \text{if} \quad W^2 \geq 3 \end{cases}. \tag{20}$$

Here, for a given context, PD is 1% more likely to occur when definiteness is switched from definite to indefinite and when the theme is short. Concomitantly, PD is 8.54% less likely to occur when definiteness is switched from definite to indefinite and when the theme is long. In addition, assume that

$P_0(W^2 \leq 2) = 30\%$, hence $P_0(W^2 \geq 3) = 70\%$. These are the actual empirical probabilities computed from the data set. Then

$$\Psi^1(P_0) = 30\% \times 1.00\% - 70\% \times 8.54\% \approx -5.68\%.$$

We fine-tuned the values in eq. (20) so that the above coincide with our estimate of the effect of definiteness of theme based on eq. (6).

Now that the reader is more familiar with the statistical reasoning underlying our approach, let us consider one last example. Intuitively, when the theme is pronominal, PD is largely preferred:

(xxii) a. Anthony sent it to you.
b. ??Anthony sent you it.

Yet, Table 1 shows a 11.68% decrease of the probability of obtaining a PD construction when pronominality of theme switches from nonpronominal to pronominal. This is a consequence of averaging out the context, which is reminiscent of what happens with definiteness of theme. Indeed, the intuition at work in example (xxii) holds when the theme is indefinite. If the theme is definite, then the preference for PD is not so marked anymore:

(xxiii) a. Anthony sent this to you.
b. Anthony sent you this.

A reader can only be surprised by our finding if she is lulled into believing that examples such as (xxii) are as a rule more frequent in the data set than those such as (xxiii). It is immensely difficult to apprehend the variety of contexts where speakers choose to use a pronominal theme as opposed to a nonpronominal one, even in the limited context of our data set. We do not embark on this impossible task. We leave that to our method, through the definition of the effect of pronominality of theme and the power of our statistical apparatus.

6.2 Simpson's paradox

Because we are concerned with the difference between predicting and explaining the outcomes of dative alternations, one of the reviewers rightly points out that the article should benefit from a realistic linguistic example where the effect of a contextual variable on the dative alternation is confounded.

We already argued in the first paragraph of Section 4.2.2 that predictions do not readily lend themselves to explanations. As discussed when commenting on example (xxi), even if we had relied solely on a logistic regression model to make predictions (we chose to rely on machine learning prediction), the estimated parameters could not have been interpreted as measures of the effects of the contextual information variables on the dative alternation. Therefore, we shall not illustrate confusion by opposing numerical predictions to numerical explanations.

Instead, confusion can simply be assessed by comparing estimates of naive measures of statistical association to estimates of the parameters introduced in Section 4.2.2. For simplicity, we focus on the effect of a binary contextual information variable. Among other choices, we oppose $\Psi^1(P_0)$ given in eq. (6) to $\mathrm{ER}(P_0) = P_0(Y = 1|W^1 = 1) - P_0(Y = 1|W^1 = 0)$, the excess risk parameter. The latter compares the probabilities to obtain a PD construction in sentences with either $W^1 = 1$ or $W^1 = 0$, neglecting the remaining information summarized by W^{-1}. On the contrary, $\Psi^1(P_0)$ takes W^{-1} into account and averages it out.

Let us resume the discussion closing Section 6.1 on the definiteness of theme, denoted by W^1 ($W^1 = 1$ for indefinite and $W^1 = 0$ for definite). Table 2 summarizes the sentence counts in all strata of (W^1, Y) focusing, for simplicity, on a data set where each identified speaker contributes only one sentence (we

Table 2: Contingency table summarizing the count of each stratum of (W^1, Y) when W^1 is the definiteness of theme, focusing on a data set where each identified speaker contributes only one sentence (we choose the first one for each identified speaker). This selection ensures Independence.

	$W^1 = 1$	$W^1 = 0$
$Y = 1$	63	378
$Y = 0$	28	858

systematically keep the first one). This ensures independence. Let $P_n{}'$ be the corresponding empirical measure. The substitution estimator of $\mathrm{ER}(P_0)$ is $\mathrm{ER}(P_n{}')$ given by

$$\mathrm{ER}(P_n{}') = \frac{63}{63+28} - \frac{378}{378+858} \approx 38.65\%.$$

If we set

$$v_n = \frac{63 \times 28}{(63+28)^3} + \frac{378 \times 858}{(378+858)^3},$$

then the interval

$$[\mathrm{ER}(P_n{}') \pm 1.96 \times \sqrt{v_n}\,] \approx [28.82\%, 48.48\%]$$

contains $\mathrm{ER}(P_0)$ with 95%-probability by the central limit theorem. The estimator $\mathrm{ER}(P'_n)$ differs dramatically from our estimator of $\Psi^1(P_0)$, which equals -5.68% with a 95%-confidence interval of $[-8.65\%, -2.72\%]$. This shows numerically that confusion is at play.

[Correction added after online publication 11 December 2015: "The estimator.." should read "The estimator $\mathrm{ER}(P'_n)$" and "-11.68% (...) $[-13.77\%, -9.59\%]$" should read "-5.68%(...) $[-8.65\%, -2.72\%]$"]

The above illustrates Simpson's paradox. Well-known to epidemiologists and statisticians, it states that a trend appearing in different data groups may disappear or even reverse once these groups are combined.

6.3 Integer valued contextual information variables

Just like Ψ^1 and Ψ^2 differ from Ψ^3 (only Ψ^3 involves a working model), Table 3 is different in nature from Table 1. Instead of commenting on the values in Table 3, we comment on Figure 1.

The left panel represents the effect of length of theme on the alternation. It shows how the probability of PD (y-axis) evolves as a function of w when length of theme (x-axis) is set to w, all other things being equal. The weight values are the values of the function h appearing in eq. (12) when evaluated at the integers $1, \ldots, 10$. The vertical bars are *simultaneous* 95%-confidence intervals for the probabilities. We observe a decreasing trend, with significant differences between the smallest and the largest values of

Table 3: Estimated effects of the integer valued information variables.

Variable	Component	Estimate	CI	p-Value
`LengthOfRecipient`	1	−0.9781	[−1.3324,−0.6238]	0.0000
	w	0.1297	[−0.0659,0.3253]	0.1937
	w^2	0.0011	[−0.0209,0.0231]	0.9237
`LengthOfTheme`	1	0.1457	[−0.3658,0.6571]	0.5767
	w	−0.2133	[−0.3287,−0.0979]	0.0003
	w^2	0.0054	[0.0007,0.0101]	0.0248

Source: For each such contextual information (named in the first column) and each component of the related parameter (identified in the second column), we report the corresponding estimated effect(s), 95%-confidence interval(s) and p-value(s) when testing whether the parameter is equal to 0 or not (in the third, fourth and fifth columns, respectively).

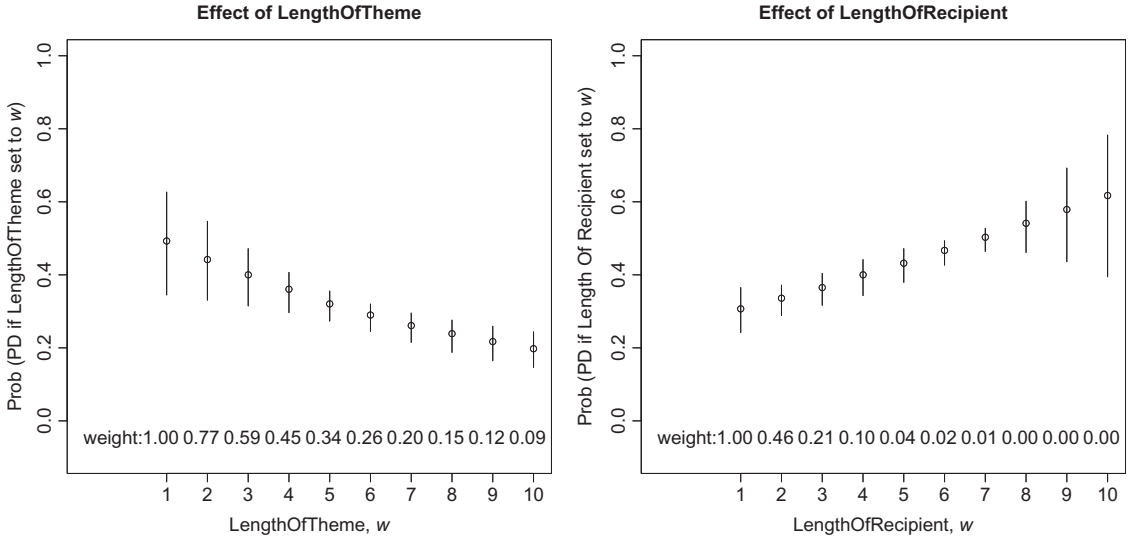

Figure 1: Representing the effects of the integer valued information variables.

length of theme, as evidenced by non-overlapping confidence intervals. From a linguistic point of view, this comes as no surprise because of the following information-structure consideration: a long theme is heavy material, and heavy material comes last, see example (xix).

The right panel represents the effect of length of recipient on the alternation. It shows how the probability of PD (y-axis) evolves as a function of w when length of recipient (x-axis) is set to w, all other things being equal. Again, the weight values are the values of the function h appearing in eq. (12) when evaluated at the integers $1, \ldots, 10$. Here too, the vertical bars are simultaneous 95%-confidence intervals for the probabilities. This time, we observe an increasing trend, with even more significant differences as we go along the x-axis. From a linguistic point of view, this comes as no surprise either for the same reason as above.

7 Discussion

If any, the lessons of this article are about crafting parameters to capture the essence of what one looks for, the merits of scaffolding a thought experiment yielding the ideal data one would have liked to work on, and targeting the above parameters. Using a well-travelled case-study in linguistics, we have adapted and benchmarked a combination of concepts and methods that has already proven its worth in biostatistics.

What is the take-away message on the dative alternation? We cannot answer this question by providing a fitted prediction model, as linguists would expect from a typical statistical study involving, for instance, logistic regression or naive discriminative learning. Prediction is at the core of our approach, but only as a means to an end. Our answer is two-fold: *(a)* we framed our account of the dative alternation in a causal model, as opposed to a prediction model, and *(b)* we investigated the effect of each available, contextual information variable on the choice of PD over DO, resulting in a table of estimates, confidence intervals, and p-values. In comparison with past findings, we found surprising results. For instance, we observed a significant decrease of the probability of obtaining PD when the theme is switched from nonpronominal to pronominal. A crude measure of statistical association such as the excess risk would have indicated a significant increase. This is an illustration of Simpson's paradox.

We showed how to operationalize the effect of any given element of context on the dative alternation as a functional evaluated at the true, unknown law P_0 of the data. We also showed how to estimate this effect in a targeted way, under the form of that functional evaluated at an empirical law built specifically to estimate the corresponding effect. We consider models as useful tools. One of these models is the backbone

of the definition of the effect of an integer valued element of context. Yet, we do not assume that this model reflects the true nature of P_0. The remaining models are at the core of algorithms used by us to build reliable predictors of features of P_0 that are involved in the estimation methodology. The combined power of these algorithms is harnessed by ambitious machine learning. Based on cross-validation, machine learning estimators are reliable but not meant for drawing statistical inference. The targeting step bends them so that valid confidence intervals can be drawn from them. Although we must assume that at least some of these models reflect some aspects of the true nature of P_0, we try to restrict the number of such untestable, unrealistic assumptions to guarantee the validity of inference.

Our method can be applied to an array of linguistic topics. In particular, all case-studies involving alternations such as

- the choice of the predeterminer *vs*. preadjectival position of intensifiers (*e.g.*, *quite* and *rather*),
- the choice of one word over a near-synonym (*e.g.*, *almost/nearly*, *big/large*, *broad/wide*, *freedom/liberty*),

can be handled at no extra cost based on this article. Drawing on the demonstrated versatility of targeted learning, more remote topics of investigation could be dealt with in a similar methodological framework, possibly without invoking causality.

As pointed out by one anonymous reviewer, the causal methods discussed here were pioneered for observational epidemiological studies where randomized interventions would be unethical or infeasible. In contrast, there are very few ethical or logistical challenges to designing randomized experiments to approximate the steps of the ideal study discussed in Section 4.2.2. For instance, a group of volunteers could perfectly read a corrupted, randomly selected corpus sample, with the contextual information in a sentence randomized to $W^1 = 1$ or $W^1 = 0$, and then decide whether the form of the dative alteration Y should be PD or DO. We showed how to infer causally defined statistical parameters based on observational data. We concede that the causal interpretability of our findings still rely on untestable statistical and causal assumptions. Our article provides further support to those linguists who call for better experimental design in the field beside intuitive and corpus-based studies.

We acknowledge that the reasoning underlying the approach advocated in this article is demanding. However, linguistics is at a quantitative turn in its history. Graduate programs throughout the world dramatically improve their offer in statistical training. Junior researchers are more eager than ever for statistics. Massive data sets are piling up. To achieve far reaching results, the discipline needs state-of-the-art theoretical statistics and robust statistical tools. We believe that after the heyday of logistic regression, linguists are now ready to embrace the distinction between predicting and explaining.

Acknowledgements: The authors wish to express their deep gratitude to the anonymous reviewers for their insightful comments. Special thanks also go to the editor, Michael Rosenblum. His suggestions helped greatly improve the manuscript. The authors gratefully acknowledge that this research was partially supported by the French National Center for Scientific Research (CNRS) through the interdisciplinary PEPS-HuMaIn-2013 initiative.

Funding: This work was funded by the CNRS (grant/award number: PEPS-HuMaIn-2013 initiative).

Appendix A

A.1 A lemma

We claimed in Section 4.2.2 that $\Psi^1(P_0) = E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\}$. Formally, the following result holds:

Lemma 1 *Assume that eq. (4) can be extended to eq. (5). Assume moreover that* U^1 *is conditionally independent from* V *given* $(U^2, \ldots, U^d)$. *The first assumption is met for instance if* $P_0(W^1 = 1|W^{-1}) \in]0, 1[$

almost surely, i.e. if W^1 takes both the values 0 and 1 with positive conditional probability given W^{-1}, for almost every W^{-1}. This can be tested on data sampled from whereas the second assumption, dubbed the "randomization assumption", cannot. Then $\Psi^1(P_0) = E_{\mathbb{P}^1_0}\{Y^1_1\} - E_{\mathbb{P}^1_0}\{Y^1_0\}$.

Proof. The conditional independence of U^1 and V given $(U^2, \ldots, U^d)$ implies the conditional independence of W^1 and (Y^1_0, Y^1_1) given W^{-1} under $\mathbb{P}^1_0$. This justifies the second equality below:

$$E_{P_0}\{P_0(Y=1|W^1=1,W^{-1})\} = E_{\mathbb{P}^1_0}\{\mathbb{P}^1_0(Y^1_1=1|W^1=1,W^{-1})\} = E_{\mathbb{P}^1_0}\{\mathbb{P}^1_0(Y^1_1=1|W^{-1})\}.$$

Now, the tower rule (which states that $E(E(A|B)) = E(A)$ for any pair of random variables (A, B)) and the fact that $\mathbb{P}^1_0(Y^1_1 = 1|W^{-1}) = E_{\mathbb{P}^1_0}(Y^1_1|W^{-1})$ imply the equality $E_{P_0}\{P_0(Y=1|W^1=1,W^{-1})\} = E_{\mathbb{P}^1_0}\{Y^1_1\}$. By symmetry, we also have $E_{P_0}\{P_0(Y=1|W^1=0,W^{-1})\} = E_{\mathbb{P}^1_0}\{Y^1_0\}$. Combining these two equalities yields the claimed result. □

A.2 A few details on the super-learner

A.2.1 The super-learner performs almost as well as the best algorithm in the library

The theoretical study of the super-learner's performances is easier when using the loss L characterized by $L(q, O) = (Y - q(W))^2$, when the algorithms $\hat{Q}_1, \ldots, \hat{Q}_K$ produce functions mapping $\mathcal{W}$ to $[0, 1]$, and when the meta-learner is sought under the form of a convex combination. Formally, for every $\delta > 0$, there exists a constant $C(\delta)$ such that

$$E_{P_0}\left\{\frac{1}{V}\sum_{v=1}^{V}\left[R^L_{P_0}(\hat{Q}_{a_n}(P_{n,T(v)})) - R^L_{P_0}(Q(P_0))\right]\right\}$$

$$\leq (1+\delta)E_{P_0}\left\{\min_{\alpha\in\mathcal{A}}\frac{1}{V}\sum_{v=1}^{V}\left[R^L_{P_0}(\hat{Q}_{\alpha}(P_{n,T(v)})) - R^L_{P_0}(Q(P_0))\right]\right\} + C(\delta)\frac{V\log(n)}{n}.$$

In the above display, the outer expectations $E_{P_0}\{(\ldots)\}$ apply to $O_1, \ldots, O_n$. In words, the super-learner performs as well as the oracle best algorithm in the library, up to a factor $(1 + \delta)$ and to the additional term $C(\delta)V\log(n)/n$, which quickly goes to 0 as n grows.

A.2.2 Specifics of our super-learner

The inference of $Q(P_0)$ is carried out by super-learning, as presented in Section 5.1. This is made easy thanks to the `SuperLearner` package [73] for the statistical programming language `R` and to the statistical community as a whole for many contributed packages. The library of algorithms that we rely on consists of estimation procedures based on generalized linear models (`glm` function), classification and regression trees (package `rpart` by Therneau et al. [74]), random forests (package `randomForest` by Liaw and Wiener [75]), multivariate adaptive polynomial spline regression (`polymars` function from the package `polspline` by Kooperberg [76], and the NDL predicting methodology (`ndlClassify` function from the package `ndl` by Antti Arppe et al. [77]).

Incidentally, the minimizer α_n of the cross-validated risk eq. (14) assigns 22% mass on the `glm` algorithm with main terms only, 38% mass on the `randomForest` algorithm with main terms only, and 40% on the `polymars` algorithm with main terms only. The mass assigned to the other algorithms is essentially zero.

A.3 A few details on TMLE

A.3.1 Differentiability of the parameters

Let us consider Ψ^1 as an example. Heuristically, for each $P \in \mathcal{M}$ there exists a function $\nabla\Psi^1(P)$ mapping $\mathcal{W} \times \{0,1\}$ to $\mathbb{R}$ such that, if the law P_ε approaches P from direction s as the real number ε goes to 0, then the $\mathbb{R} \to \mathbb{R}$ function $\varepsilon \mapsto \Psi^1(P_\varepsilon)$ is (classically) differentiable at $\varepsilon = 0$ with a derivative equal to $E_P\{\nabla\Psi^1(P)(O) \times s(O)\}$. Here, s can be (basically almost) any real valued, bounded function defined on $\mathcal{W} \times \{0,1\}$, and "approaching from direction s" means that the *log*-likelihood function under P_ε, $\epsilon \mapsto \log P_\varepsilon(O)$, is a real valued function differentiable at $\varepsilon = 0$ with a derivative equal to $s(O)$. Similar statements hold for Ψ^2 and Ψ^3. It is known (see Chapter 5, for instance [3] that $\nabla\Psi^1$ is characterized by

$$\begin{aligned}\nabla\Psi^1(P)(O) = {} & Q(P)(1, W^{-1}) - Q(P)(0, W^{-1}) - \Psi^1(P) \\ & + (Y - Q(P)(W))\left(\frac{\mathbf{1}\{W^1 = 1\}}{P(W^1 = 1|W^{-1})} - \frac{\mathbf{1}\{W^1 = 0\}}{P(W^1 = 0|W^{-1})}\right).\end{aligned} \tag{21}$$

Similarly, $\nabla\Psi^2$ is characterized by $\nabla\Psi^2(P)(O) = (\nabla\Psi^2_k(P)(O) : 1 \le k \le K)$ with

$$\begin{aligned}\nabla\Psi^2_k(P)(O) = {} & Q(P)(W^1, k, W^3, \ldots, W^d) - Q(P)(W^1, 0, W^3, \ldots, W^d) - \Psi^2_k(P) \\ & + (Y - Q(P)(W))\left(\frac{\mathbf{1}\{W^2 = k\}}{P(W^2 = k|W^{-2})} - \frac{\mathbf{1}\{W^2 = 0\}}{P(W^2 = 0|W^{-2})}\right).\end{aligned} \tag{22}$$

As for $\nabla\Psi^3$, it is such that $\nabla\Psi^3(P)(O)$ equals a 3×3 (deterministic) normalizing matrix times the (random) vector

$$\begin{aligned}\widetilde{\nabla\Psi^3}(P)(O) = {} & \sum_{w \in \mathcal{W}^3} h(w)\Big(Q(P)(W^1, W^2, w, W^4, \ldots, W^d) - f_{\Psi^3(P)}(w)\Big)(1, w, w^2)^\top \\ & + \sum_{w \in \mathcal{W}^3} h(w)(Y - Q(P)(W))\frac{\mathbf{1}\{W^3 = w\}}{P(W^3 = w|W^{-3})}(1, w, w^2)^\top\end{aligned} \tag{23}$$

[3, 78]. Note that there is actually one single non-zero term in the second sum of the RHS of eq. (23), which is the term corresponding to $w = W^3$.

A.3.2 Fluctuating the initial estimators

Let us first describe here the different fluctuations that we use to target Q_n^{init} toward our parameters of interest. Let $g_n^1(1|W^{-1})$, $g_n^2(k|W^{-2})$ and $g_n^3(w|W^{-3})$ be estimators of $P_0(W^1 = 1|W^{-1})$, $P_0(W^2 = k|W^{-2})$ and $P_0(W^3 = w|W^{-3})$, respectively, for all $0 \le k \le K$, $w \in \mathcal{W}^3$ and $W \in \mathcal{W}$. For our specific application, these estimators are based on logistic and multinomial regression models with main terms only. Their fitting is carried out by using the `glm` and `multinomial` functions in `R`.

The fluctuations for Ψ^1 and Ψ^2 are very much alike. To target $\Psi(P_0)$, we rely on $Q_{n,\varepsilon}^{1,\text{init}}$ characterized, for all $\varepsilon \in \mathbb{R}$, by

$$\text{logit}\Big(Q_{n,\varepsilon}^{1,\text{init}}(W)\Big) = \text{logit}\big(Q_n^{\text{init}}(W)\big) + \varepsilon\left(\frac{\mathbf{1}\{W^1 = 1\}}{g_n^1(1|W^{-1})} - \frac{\mathbf{1}\{W^1 = 0\}}{1 - g_n^1(1|W^{-1})}\right). \tag{24}$$

Likewise, we target $\Psi^2(P_0)$ by relying on $Q_{n,\varepsilon}^{2,\text{init}}$ characterized, for all $\varepsilon \in \mathbb{R}^K$, by

$$\text{logit}\Big(Q_{n,\varepsilon}^{2,\text{init}}(W)\Big) = \text{logit}\big(Q_n^{\text{init}}(W)\big) + \sum_{k=1}^{K} \varepsilon_k\left(\frac{\mathbf{1}\{W^2 = k\}}{g_n^2(k|W^{-2})} - \frac{\mathbf{1}\{W^2 = 0\}}{g_n^2(0|W^{-2})}\right). \tag{25}$$

As for the targeting toward $\Psi^3(P_0)$, we choose to rely on $Q^{3,\text{init}}_{n,\varepsilon}$ characterized, for all $\varepsilon \in \mathbb{R}^3$, by

$$\text{logit}\Big(Q^{3,\text{init}}_{n,\varepsilon}(W)\Big) = \text{logit}\big(Q^{\text{init}}_n(W)\big) + \frac{h(W)}{g^3_n(W^3|W^{-3})}(\varepsilon_1 + \varepsilon_2 W^3 + \varepsilon_3 (W^3)^2). \tag{26}$$

We refer the interested reader to [3, Chapter 5; 79]) for further details.

Let us now turn to the next fundamental issue, which pertains to estimating the specific elements $Q^{1,\text{targ}}_n = Q^{\text{init}}_{n,\varepsilon^1_n}$, $Q^{2,\text{targ}}_n = Q^{\text{init}}_{n,\varepsilon^2_n}$ and $Q^{3,\text{targ}}_n = Q^{\text{init}}_{n,\varepsilon^3_n}$ among these collections that better target, each, the corresponding parameter of interest. This is easy. The optimal parameters can be characterized as the following solutions to three different optimization problems:

$$\varepsilon^1_n = \arg\max_{\varepsilon\in\mathbb{R}} \sum_{i=1}^n \Big[Y_i \log(Q^{1,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{1,\text{init}}_{n,\varepsilon}(W_i))\Big],$$

$$\varepsilon^2_n = \arg\max_{\varepsilon\in\mathbb{R}^K} \sum_{i=1}^n \Big[Y_i \log(Q^{2,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{2,\text{init}}_{n,\varepsilon}(W_i))\Big],$$

$$\varepsilon^3_n = \arg\max_{\varepsilon\in\mathbb{R}^3} \sum_{i=1}^n \Big[Y_i \log(Q^{3,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{3,\text{init}}_{n,\varepsilon}(W_i))\Big].$$

These optimization problems can be solved routinely in R with the glm function for the fitting of generalized linear models on data. Interestingly, the fluctuations (24), (25) and (26) can be coded by defining $\text{logit}(Q^{\text{init}}_n(W))$ as an offset and the factors of each component of ε as covariates upon which to regress Y.

A.3.3 Solving eqs (15) and (19)

The numerical computation of the substitution estimators $\psi^{3,\text{init}}_n$ and its targeted counterpart $\psi^{3,\text{targ}}_n$, see eqs (15) and (19), can also be solved routinely using `R`. Firstly, we create a new data set, each observation O_i contributing $\text{card}(\mathcal{W})$ rows, one for every possible value of W^3_i, where each row consists of three entries. For the ith observation, $w \in \mathcal{W}^\ni$ is associated with $(Q^{3,\text{init}}_n(W^1_i, W^2_i, w, W^4_i, \ldots, W^d_i), w, h(w))$ for the computation of $\psi^{3,\text{init}}_n$ and $(Q^{3,\text{targ}}_n(W^1_i, W^2_i, w, W^4_i, \ldots, W^d_i), w, h(w))$ for the computation of $\psi^{3,\text{targ}}_n$. Secondly, we regress the first column of the data set on $f_\theta(w)$ based on its second column using the `glm` function with `binomial` family, `logit` link, `weights` from the third column, and the `formula` encoding our working model (11). Even though the new "outcome" is not binary, the fact that it takes values in $]0,1[$ guarantees that the `glm` function computes the desired iteratively reweighted least squares solutions, provided that the algorithm converges [79].

A.3.4 Including speaker-related dependency

The key to including speaker-related dependency is weighting.

We attach a weight to each observation. This weight is the inverse of the number of constructions contributed by the same speaker in the data set. The observations that we originally noted $O_1, \ldots, O_n$ are now regrouped in $M = 1327$ bigger observations $O^*_1, \ldots, O^*_M$. Here, M is the number of different speakers and each O^*_m decomposes as $O^*_m = (O_{m,1}, \ldots, O_{m,J_m})$, where every $O_{m,j}$ uniquely coincides with one observation among $O_1, \ldots, O_n$.

We may now assume that $O^*_1, \ldots, O^*_M$ are independently sampled from a distribution P^*_0, and that conditionally on the number J_m of constructions contributed by speaker m, the dependent observations $O_{m,1}, \ldots, O_{m,J_m}$ have the same marginal distribution, which coincides with our P_0. Under this assumption, the weighted version of our method accommodates for dependency.

A.3.5 Confidence intervals

We build our confidence intervals by relying on the assumed asymptotic normality of our targeted estimators and their limit standard deviations inferred as the standard deviations of the corresponding efficient influence curves, see eqs (21)–(23). The theory provides us with a set of mathematical assumptions which guarantee that this approach does yield conservative confidence intervals. Some of them can be checked as they only depend on choices we make, such as the algorithms which join forces in the super-learner, see Section A.2.2. Some of them cannot, as they depend on the true, unknown distribution P_0. Thus, we acknowledge that our confidence intervals are valid if the sample size n is large enough and, for instance, if the parametric models upon which the estimation of the conditional probabilities $P_0(W^j|W^{-j})$ (all $1 \leq j \leq d$) are correctly specified. This condition is quite stringent. It is actually possible to weaken it considerably by adding another layer of targeting, as recently shown by van der Laan [80]. This, however, is beyond the scope of this article.

References

1. Gries ST. Frequency tables: tests, effect sizes, and explorations. In: Glynn D, Robinson J. Polysemy and synonymy: corpus methods and applications in cognitive linguistics. Amsterdam: John Benjamins, 2014.
2. Baayen RH. Corpus linguistics and naive discriminative learning. Rev Bras Ling Apl 2011;11:295–328.
3. van der Laan MJ, Rose S. Targeted learning. New York: Springer, 2011. ISBN 978-1-4419-9781-4. doi: 10.1007/978-1-4419-9782-1.
4. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006;2:Art. 11, 40. ISSN 1557-4679. doi: 10.2202/1557-4679.1043.
5. Chomsky N. Syntactic structures. The Hague: Mouton, 1957.
6. Chomsky N. Aspects of the theory of syntax. Cambridge, MA: MIT Press, 1962.
7. Chomsky N. The minimalist program. Cambridge, MA: MIT Press, 1995.
8. Bybee J. Phonology and language use. Cambridge: Cambridge University Press, 2001.
9. Bybee J. Language, usage, and cognition. Cambridge: Cambridge University Press, 2010.
10. Goldberg AE. Constructions: a construction grammar approach to argument structure. Chicago: University of Chicago Press, 1995.
11. Langacker RW. Foundations of cognitive grammar, vol. 1. Stanford, CA: Stanford University Press, 1987.
12. Gries ST. Quantitative corpus linguistics with R. New York and London: Routledge, 2009.
13. Glynn D. Corpus-driven cognitive semantics: introduction to the field. In: Glynn D, Fischer K, editors. Quantitative methods in cognitive semantics: corpus-driven approaches. Berlin: Mouton de Gruyter, 2010:1–42.
14. Levin B. English verb classes and alternations: a preliminary investigation. Chicago and London: The University of Chicago Press, 1993.
15. Levin B, Rappaport Hovav M. Argument realization. Cambridge: Cambridge University Press, 2005.
16. Krifka M. Semantic and pragmatic conditions for the dative alternation. Korean J English Lang Ling 2004;4:1–32.
17. Fillmore C. Indirect object constructions in English and the ordering of transformations. The Hague: Mouton, 1965.
18. Hall BC Subject and Object in English. PhD thesis, Massachusetts Institute of Technology, 1975.
19. Emons JE. Evidence that indirect-object movement is a structure-preserving rule. Found Lang 1972;8(4):546–61.
20. Burt MK. From deep to surface structure. New York: Harper Row, 1971.
21. Aoun J, Li Y-H. Scope and constituency. Ling Inquiry 1989;20:141–72.
22. Jackendoff RS. Semantics and cognition. Cambridge, MA: MIT Press, 1983.
23. Green G. Semantics and syntactic regularity. Bloomington: Indiana University Press, 1974.
24. Oehrle RT. The grammatical status of the English dative alternation. PhD thesis, Massachusetts Institute of Technology, 1976.
25. Baker MC. Thematic roles and syntactic structure. In: Haegeman L, editor. Elements of grammar. Handbook of generative syntax. Dordrecht: Kluwer, 1997:73–137.
26. Davidse K. Agnates, verb classes and the meaning of construals: the case of ditransivity in English. Leuvense Bijdragen 1998;87:281–313.
27. Bresnan J. Lexical-functional syntax. Oxford: Blackwell, 2001.
28. Dowty DR. Thematic proto-roles and argument selection. Language 1991;67:547–619.
29. Jackendoff RS. Semantic structures. Cambridge, MA: MIT Press, 1990.

30. Pinker S. Learnability and cognition: the acquisition of argument structure. Cambridge: MIT Press, 1989.
31. Rappaport Hovav M, Levin B. The English dative alternation: the case for verb sensitivity. J Ling 2008;44:129–67.
32. Speas MJ. Phrase structure in natural language. Dordrecht: Kluwer, 1990.
33. Gries ST, Stefanowitsch A. Extending collostructional analysis – a corpus-based perspective on 'alternations'. Int J Corpus Ling 2004;9:97–129.
34. Baker MC. Review of S. Pinker, learnability and cognition: the acquisition of argument structure. Language 1992;68:402–13.
35. Gropen J, Pinker S, Hollander M, Goldberg R, Wilson R. The learnability and acquisition of the dative alternation in English. Language 1989;65:203–57.
36. Arnold JE, Wasow T, Losongco A, Ginstrom R. Heaviness vs. Newness: the effects of structural complexity and discourse status on constituent ordering. Language 2000;76:1.
37. Davidse K. Ditransitivity and possession. In: Hasan R, Cloran C, Butt DG, editors. Functional descriptions: theory in practice. Amsterdam: John Benjamins, 1996:85–144.
38. Wasow T. Remarks on grammatical weight. Lang Var Change 2008;9:81–105.
39. Bresnan J, Nikitina T. The gradience of the dative alternation. In: Uyechi L, Wee L-H, editors. Reality exploration and discovery: pattern interaction in language and life. Stanford: CSLI, 2009:161–84.
40. Snyder KM. The relationship between form and function in ditransitive constructions. PhD thesis, University of Pennsylvania, PA, 2003.
41. Williams RS. A statistical analysis of English double object alternation. Issues Appl Ling 1994;5:37–58.
42. Bresnan J, Cueni A, Nikitina T, Baayen RH. Predicting the dative alternation. In: Bouma G, Kramer I, Zwarts J, editors. Cognitive Found Interpret. Amsterdam: Royal Netherlands Academy of Arts and Sciences, 2007:69–94.
43. Gries ST. Towards a corpus-based identification of prototypical instances of constructions. Ann Rev Cogn Ling 2003;1:1–27.
44. Theijssen D, Ten Bosch L, Boves L, Cranen B, van Halteren H. Choosing alternatives: using Bayesian networks and memory-based learning to study the dative alternation. Corpus Ling Ling Theory 2013;9:227–62.
45. Speelman D. Logistic regression: a confirmatory technique for comparisons in corpus linguistics. In: Glynn D, Robinson JA, editors. Corpus methods for semantics: quantitative studies in polysemy and synonymy. Amsterdam: John Benjamins, 2014:487–533.
46. Daelemans W, van den Bosch A. Memory-based language processing. Studies in natural language processing. Cambridge: Cambridge University Press, 2009.
47. Skousen R, Lonsdale D, Parkinson DB, editors. Analogical modeling: an exemplar-based approach to language. Amsterdam: John Benjamins, 2002.
48. Vapnik VN. The nature of statistical learning theory. Berlin: Springer Verlag, 1995.
49. Breiman L. Random forests. Mach Learn 2001;45:5–32.
50. Danks D. Equilibria of the Rescorla-Wagner model. J Math Psychol 2003;47:109–21.
51. Wagner AR, Rescorla RA. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II. New York: Appleton-Century-Crofts, 1972:64–99.
52. Ellis N. Language acquisition as rational contingency learning. Appl Ling 2006;27:1–24.
53. Ellis N, Ferreira-Junior F. Constructions and their acquisition: islands and the distinctiveness of their occupancy. Annu Rev Cogn Ling 2009;7:187–220.
54. Polley EC, Rose S, van der Laan MJ. Super learning. In: van der Laan MJ, Rose S, editors. Targeted learning, Springer series in statistics, Chapter 3. New York: Springer, 2011:43–66.
55. Baayen RH. languageR: Data sets and functions with "Analyzing Linguistic Data: a practical introduction to statistics", 2009. Available at: http://CRAN.R-project.org/package = languageR.
56. Hacking I. The taming of chance, vol. 17. Cambridge: Cambridge University Press, 1990.
57. Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition, volume 31 of applications of mathematics (New York). New York: Springer-Verlag, 1996. ISBN 0-387-94618-7. doi: 10.1007/978-1-4612-0711-5. Available at: http://dx.doi.org/10.1007/978-1-4612-0711-5.
58. Chambaz A, Drouet I, Thalabard J-C. Causality, a trialogue. J Causal Inference 2014;2(2):201–41. Ahead of print.
59. Pearl J. Causality: models, reasoning and inference, volume 29. Cambridge: Cambridge University Press, 2000.
60. Wright S. Correlation and causation. J Agric Res 1921;20:557–85.
61. Haavelmo T. The statistical implications of a system of simultaneous equations. Econometrica 1943;11:1–12.
62. Robins JM. Marginal structural models. Proceedings of the American Statistical Association Section on Bayesian Statistical Science, 1997;1–10.
63. Robins JM, Miguel AH, Babette B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
64. Rosenblum M, Deeks SG, van der Laan MJ, Bangsberg DR. The risk of virologic failure decreases with duration of HIV suppression, at greater than 50% adherence to antiretroviral therapy. PLoS ONE 2009;4(9):e7196. doi: 10.1371/journal.pone.0007196.
65. Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

66. Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197–227.
67. Wolpert DH. Stacked generalization. Neural Networks 1992;5:241–59.
68. Breiman L. Stacked regressions. Mach Learn 1996;24:49–64.
69. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statist Sci 1999;14:382–417. doi: 10.1214/ss/1009212519. Available at: http://dx.doi.org/10.1214/ss/1009212519. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors.
70. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Gene Mol Biol 2007;6:Article 25.
71. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer-Verlag, 2003. ISBN 0-387-95556–9.
72. van der Vaart AW. Asymptotic statistics, volume 3 of Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge University Press, 1998.
73. Polley E, van der Laan MJ. SuperLearner, 2011. Available at: http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-4.
74. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees, 2014. Available at: http://CRAN.R-project.org/package = rpart. R package version 4.1-8.
75. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18–22. Available at: http://CRAN.R-project.org/doc/Rnews/.
76. Kooperberg C polspline: Polynomial spline routines, 2013. Available at: http://CRAN.R-project.org/package=polspline. R package version 1.1.8.
77. Arppe A, Hendrix P, Petar Milin R, Baayen H, Shaoul C. ndl: Naive Discriminative Learning, 2014. Available at: http://CRAN.R-project.org/package = ndl. R package version 0.2.16.
78. Rosenblum M, van der Laan MJ. Targeted maximum likelihood estimation of the parameter of a marginal structural model. Int J Biostat 2010;60:Article 19.
79. Rosenblum M. Marginal structural models. In: van der Laan MJ, Rose S, editors. Targeted learning, Springer series in statistics, chapter 9. New York: Springer, 2011:145–60.
80. van der Laan MJ. Targeted estimation of nuisance parameters to obtain valid statistical inference. Int J Biostat 2014;10:29–57.