Predicting is not explaining: targeted learning of the dative alternation

Antoine Chambaz, Guillaume Desagulier

Predicting is not explaining: targeted learning of the dative alternation

Antoine Chambaz[1] and Guillaume Desagulier[2]

[1]Modal'X – Université Paris Ouest Nanterre La Défense
[2]MoDyCo — Université Paris 8, CNRS, Université Paris Ouest Nanterre La Défense

October 8, 2014

Abstract

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. 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.

1 Introduction

Gries (2014) describes corpus linguistics as a "distributional science" investigating the frequencies of occurrence of various elements in corpora, their dispersion, and their co-occurrence properties. Baayen (2011) argues that, "although this characterization of present-day corpus linguistics is factually correct (...), 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 (2011), 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, van der Laan and Rubin, 2006; van der Laan and Rose, 2011). 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.

In Section 2, we briefly introduce the dative alternation, the theoretical issues it raises, and a summary of recent corpus-based, statistics driven investigations. In Section 3, 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 4 is a concise presentation of the statistical apparatus that we elaborate to tackle the statistical problems defined in Section 3. We present and comment on the results in Section 5. 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 more technical parts of the article.

2 The dative alternation

An argument alternation is characterized by sentence pairs with the same verb, but different syntactic patterns. 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), exemplified in (i) and (ii) respectively:

(i) John gave the book to Mary. (PD)
S_{AG} V O_{THEME} O_{REC}

(ii) John gave Mary the book. (DO)
S_{AG} V O_{REC} O_{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 (2005) describe the dative alternation as a case of object alternation, a subcase of multiple argument realization phenomena.

To account for the dative alternation, linguists have relies on either intuition or corpora and quantitative methods. We review each trend in Sections 2.1 and 2.2.

2.1 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, 1993), (Krifka, 2004) and (Levin and Rappaport Hovav, 2005, chapter 7).

Chomsky (1957, 1962) 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 transformational-generative 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 (1965), Hall (1975), and Emons (1972) contend that PD is basic whereas DO is derived. On the other hand, Burt (1971) and Aoun and Li (1989) argue for the opposite pattern of transformation: DO is basic whereas PD is derived.

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

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

(iv) a. Anthony donated $100 to charity.

b. ?Anthony donated charity $100.

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

Proponents of the Localist Hypothesis (Jackendoff, 1983), who construe the recipient as a spatial goal, might further argue that DO is possible in (vi b) if London refers metonymically to a person or an institution, in which case it differs from (vi a) where London is clearly a place:

(vi) 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 (Green, 1974; Oehrle, 1976), as in (vii):

(vii) 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 (vii b). Example (vii 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 (Baker, 1997; Davidse, 1998; Levin and Rappaport Hovav, 2005).

Despite continuous efforts to maintain that alternating verbs have a single meaning underlying both formal variants (Jackendoff, 1990; Dowty, 1991; Bresnan, 2001), there is now cross-theoretical consensus that the two variants of the dative alternation have distinct semantic representations. According to Pinker (1989) and Rappaport Hovav and Levin (2008), caused motion underlies PD, whereas caused possession underlies DO, as schematized in (viii):

(viii) 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 (1990, pp. 88–89) schematizes the semantic representations of both variants as follows:

(ix) 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 (1995) posits that PD is a subtype of the more general caused-motion construction (cf. the Localist Hypothesis), whereas DO expresses a transfer of possession:

(x) 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, 2004).

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 (vii), the recipient of 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 (Baker, 1992).
- Communication verbs: as opposed to speech act verbs (*tell*, *read*, *write*, *cite*, etc.) and denominal verbs 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.). Exceptions are listed in (Gropen et al., 1989).
- 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 (1989, 216).

Lexical semantic restrictions are sometimes overridden by information-structure factors (Arnold et al., 2000; Davidse, 1996; Wasow, 2008, interalia). The first factor is discourse givenness: given material precedes new material. PD is expected when the theme is more given than the recipient, as in (xi a), whereas DO is more likely when the recipient is more given than the theme, as in (xii b):

(xi) a. Will gave his manuscript to a first-year student. (PD)
b. [??]Will gave a first-year student his book. (DO)

(xii) 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 (2009) find empirical support for this, but they also find exceptions such as (xiii a):

(xiii) 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. Heavy material comes last, as exemplified below:

(xiv) 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.

What is still theoretically unclear is which factor(s) take(s) precedence over the other(s). Snyder (2003) claims that information-structure factors are more important than heaviness, whereas Arnold et al. (2000) 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. We now turn to recent corpus-based, statistics-driven investigations of the dative alternation.

2.2 Corpus-based answers

Since (Williams, 1994), the dative alternation has become a model construction for benchmarking predictive methods (Arnold et al., 2000; Gries, 2003; Bresnan et al., 2007; Baayen, 2011; Theijssen et al., 2013). Focusing on DO, Williams (1994) 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 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. (2000) 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. 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 (2003) 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 NP_{GOAL} 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 (2003) 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. (2007) use (mixed-effects) logistic regression, like Williams (1994) and Arnold et al. (2000). Unlike those previous works, Bresnan et al.'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. (2007)'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, definiteness, pronominality, and accessibility of recipient/theme; see also Section 3.1. One of their logistic regression models predicts which variant of the dative alternation is used with a high accuracy.

Using Bresnan et al.'s data set, Baayen (2011) tests naive discriminative learning (henceforth NDL) on the dative alternation. Baayen compares NDL to other well-established statistical classifiers such as logistic regression (Bresnan et al., 2007; Speelman, 2014), memory-based learning

(Daelemans and van den Bosch, 2009; Theijssen et al., 2013), analogical modeling of language (Skousen et al., 2002), support vector machines (Vapnik, 1995), and random forests (Breiman, 2001). 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 (Danks, 2003). According to the Wagner-Rescorla equations (Wagner and Rescorla, 1972), 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 (Ellis, 2006; Ellis and Ferreira-Junior, 2009), 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 (Baayen, 2011, p. 320).

Baayen 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.

3 Targeting the dative alternation in English

3.1 Data

We used the `dative` data set available in the `languageR` package (Bresnan et al., 2007; Baayen, 2009). It contains 3263 observations consisting of 15 variables. The variables divide into:

- speaker, a categorical variable with 424 levels, including `NA`s;
- 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.

3.2 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.

3.2.1 Predicting. Predicting the dative alternation in English means building an algorithm that poses as a native speaker of English when he or 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 or not. 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 in English works. Telling us how the alternation works falls within the scope of explaining it. It is the topic of Section 3.2.2.

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 3.1 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 (1990) 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.

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 Devroye et al., 1996, Theorem 2.1). 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 a 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, (1) suggests that if q is close to $Q(P_0)$ then f given by $f(W) = \mathbf{1}\{q(W) \geq 0.5\}$ 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 valued, 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.

3.2.2 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 (2011) 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 3.2.1, 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 (Chambaz et al., 2014, see Sections 3, 8 and 9 in particular). We consider in turn how to proceed with a categorical component as opposed to a non-categorical component.

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" refer 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 $Y = Y_{W^1}^1$, *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 (Pearl, 2000).

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 3.2.1 can be modeled altogether by means of a non-parametric system of structural equations, a notion which originates in the works of Wright (1921), Haavelmo (1943) and was brought up-to-date by Pearl (2000).

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),

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

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, (3) can be rewritten under the equivalent form

$$\left\{ \begin{array}{lcl} W^j & = & F^j(U^j), \quad j = 1, \ldots, d \\ Y & = & f((W^1, \ldots, W^d), V) \end{array} \right. \tag{4}$$

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

$$\left\{ \begin{array}{lcl} W^j & = & F^j(U^j), \quad j = 1, \ldots, d \\ Y_0^1 & = & f((0, W^2, \ldots, W^d), V) \\ Y_1^1 & = & f((1, W^2, \ldots, W^d), V) \\ Y & = & Y^1_{W^1} \end{array} \right. , \tag{5}$$

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, (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 Ψ^1 which maps the set $\mathcal{M}$ to $[-1, 1]$ and is given at any $P \in \mathcal{M}$ by

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

because $Q(P)(W) = P(Y = 1|W)$ (see (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 (5), (4) also yields the following system

$$\left\{ \begin{array}{lcl} W^j & = & F^j(U^j), \quad j = 1, \ldots, d \\ Y_k^2 & = & f((W^1, k, W^3, \ldots, W^d), V), \quad k = 0, \ldots, K \\ Y & = & Y^2_{W^2} \end{array} \right. , \tag{7}$$

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

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

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.

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 (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

$$\left\{ \begin{array}{lcl} W^j & = & F^j(U^j), \quad j = 1, \ldots, d \\ Y^3_w & = & f((W^1, W^2, w, W^4, \ldots, W^d), V), \quad \text{all } w \in \mathcal{W}^3 \\ Y & = & Y^3_{W^3} \end{array} \right. , \tag{9}$$

provided that the second equation always makes sense. Among other things, system (9) induces a model for $\mathbb{P}^3_0$, 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 $\mathbb{P}^1_0$, $\mathbb{P}^2_0$, and some parameters of the latter which are interpretable as importance measures. In the present situation, though, we cannot yet introduce our parameter of $\mathbb{P}^3_0$ 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 $w \mapsto E_{\mathbb{P}^3_0}\{Y^3_w\}$ 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 $\mathbb{P}^3_0$ as a measure of the importance of W^3 on Y:

$$\underset{\theta \in \Theta}{\arg\max} \sum_{w \in \mathcal{W}^3} h(w) \Lambda \left(E_{\mathbb{P}^3_0}\{Y^3_w\}, f_\theta(w) \right) = \underset{\theta \in \Theta}{\arg\max} \sum_{w \in \mathcal{W}^3} h(w) \Lambda \left(\mathbb{P}^3_0(Y^3_w = 1), f_\theta(w) \right), \tag{10}$$

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 (1997) first introduced marginal structural models in causal analysis. Robins et al. (2000) discusses their use in epidemiology. More recently, Rosenblum et al. (2009) 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 $E_{\mathbb{P}^1_0}\{Y^1_1\} - E_{\mathbb{P}^1_0}\{Y^1_0\}$ on one hand and $(E_{\mathbb{P}^2_0}\{Y^2_k\} - E_{\mathbb{P}^2_0}\{Y^2_0\} : k = 1, \ldots, K)$ on the other hand, (10) has no closed-form explicit expression in terms of $\mathbb{P}^3_0$ 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 $w \mapsto E_{\mathbb{P}^3_0}\{Y^3_w\}$ than every other $f_{\theta'}$, where the gap between two functions f, f' mapping $\mathcal{W}^3$ to $[0,1]$ is measured by

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

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 (10) involves an arg max and not an arg min. In particular, if $w \mapsto E_{\mathbb{P}^3_0}\{Y^3_w\}$ coincides with f_θ for some $\theta \in \Theta$ and if the weight function h only takes positive values then (10) equals θ. This is very unlikely. If, on the contrary, no f_θ equals $w \mapsto E_{\mathbb{P}^3_0}\{Y^3_w\}$ then (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 5 to hammer home this important point.

Because the set $\mathcal{F}$ 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 $\text{expit}(q) = 1/(1+e^{-q})$ and $\text{logit}(p) = \log(p/(1-p))$, respectively. In this article, we consider the set

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

and assume that (10) uniquely defines a single element of Θ for this specific choice of $\mathcal{F}$, an assumption that cannot be tested on data. Thus, the parameter (10) should be understood as the best second-order polynomial approximation to $w \mapsto \text{logit}\left(E_{\mathbb{P}_0^3}\{Y_w^3\}\right)$ 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 (10) in the sense that *(a)* under appropriate assumptions it is equal to (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 $P \in \mathcal{M}$,

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

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 (10) under suitable, untestable assumptions.

4 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 4.1, followed by targeted minimum loss explanation, see Section 4.2.

4.1 Machine learning prediction

We consider first the inference of $Q(P_0)$ as defined in (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 Schapire, 1990; Wolpert, 1992; Breiman, 1996a,b; Hoeting et al., 1999, to cite only a few seminal works, with an emphasis on methods using the cross-validation principle). Specifically, we choose to rely on the super-learning methodology (van der Laan et al., 2007; Polley et al., 2011).

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 $\{T(\nu) \subset \{1,\ldots,n\} : 1 \leq \nu \leq V\}$ such that $\cup_{\nu=1}^V T(\nu) = \{1,\ldots,n\}$ (their union covers $\{1,\ldots,n\}$) and for each $1 \leq \nu_1 \neq \nu_2 \leq V$, $T(\nu_1) \cap T(\nu_2) = \emptyset$ (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

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

(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 (13), the subset of data used to "train" $\hat{Q}_\alpha$, represented by $P_{n,T(\nu)}$, and the subset used to evaluate its performances, represented by $P_{n,T(\nu)^c}$, 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.

4.2 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 (6), (8) and (12). We take the route of TMLE, a paradigm of inference based on semiparametrics and estimating functions (see van der Laan and Robins, 2003; van der Vaart, 1998, Chapter 25, for recent and comprehensive introductions). Introduced by van der Laan and Rubin (2006), TMLE has been studied and applied in a variety of contexts since then (we refer to van der Laan and Rose, 2011, for an overview). An accessible introduction to TMLE is given in (Chambaz et al., 2014, Sections 12, 13 and 14).

It is apparent in (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_n^{init}—that could be, for instance, the super-learner $\hat{Q}_{\alpha_n}(P_n)$ whose construction we described in

Section 4.1. Here, the superscript "init" indicates that we think of Q_n^{init} as an initial estimator of $Q(P_0)$ built for the sake of predicting, not explaining.

Taking a closer look at (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_n^{init}. Consider (6): if we substitute Q_n^{init} 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 law. 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 (6) yields an initial estimator of $\Psi^1(P_0)$, say $\psi_n^{1,\text{init}}$, writing as

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

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

$$\begin{aligned}\psi_{k,n}^{2,\text{init}} &= E_{P_n}\left\{Q_n^{\text{init}}(W^1,k,W^3,\ldots,W^d) - Q_n^{\text{init}}(W^1,0,W^3,\ldots,W^d)\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Q_n^{\text{init}}(W_i^1,k,W_i^3,\ldots,W_i^d) - Q_n^{\text{init}}(W_i^1,0,W_i^3,\ldots,W_i^d)\right]\end{aligned}$$

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

$$\psi_n^{3,\text{init}} = \underset{\theta\in\Theta}{\arg\max}\sum_{w\in\mathcal{W}^3} h(w)\Lambda\left(E_{P_n}\left\{Q_n^{\text{init}}(W^1,W^2,w,W^4,\ldots,W^d)\right\}, f_\theta(w)\right) \tag{15}$$

$$= \underset{\theta\in\Theta}{\arg\max}\sum_{w\in\mathcal{W}^3} h(w)\Lambda\left(\frac{1}{n}\sum_{i=1}^{n} Q_n^{\text{init}}(W_i^1,W_i^2,w,W_i^4,\ldots,W_i^d), f_\theta(w)\right). \tag{16}$$

Interestingly, the optimization problem (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

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

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

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

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

$$= \underset{\theta\in\Theta}{\arg\max} \sum_{w\in\mathcal{W}^3}\sum_{i=1}^{n} h(w)\Lambda\left(Q_n^{3,\text{targ}}(W_i^1, W_i^2, w, W_i^4, \ldots, W_i^d), f_\theta(w)\right).$$

The optimization problem (19) can be solved easily just like (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 (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.

5 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 3.2.2 along the lines presented in Section 4. 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 3.1, the forthcoming results are obtained with consideration for speaker-related dependency, see Section A.3.4.

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 w_0. 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 (xvii) and (xviii);
- 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 (xv);
- 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 (xvi);
- a 3.95% increase when definiteness of recipient switches from definite to indefinite.

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 3.2.2.

Consider for instance example (xv): under the constraint "set the animacy of recipient to inanimate", the speaker selects either (xv a) or (xv b); under the constraint "set the animacy of recipient to animate", she selects either (xv c) or (xv 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 (xv d) slightly more natural than (xv c), (xv a) is undoubtedly more natural than (xv 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.

(xv) 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 (xvi), one can conceive that the preference of (xvi d) over (xvi c) is slightly stronger than that of (xvi b) over (xvi 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.

(xvi) 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 (xvi) 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. This 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 (20) so that the above coincide with our estimate of the effect of definiteness of theme based on (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:

(xvii) 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 (xvii) holds when the theme is indefinite. If the theme is definite, then the preference for PD is not so marked anymore:

(xviii) 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 (xvii) are as a rule more frequent in the data set than those such as (xviii). 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.

Integer valued contextual information variables. Just like Ψ^1 and Ψ^2 differ from Ψ^3 (only Ψ^3 involves a working model), Table 2 is different in nature from Table 1. Instead of commenting on the values in Table 2, 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 (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 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 (xiv).

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 (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.

6 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.

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 it 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.

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 gratefully acknowledge that this research was partially supported by the French National Center for Scientific Research (CNRS) through the interdisciplinary

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

Table 1: Estimated effects of the categorical information variables. 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).

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

Table 2: Estimated effects of the integer valued information variables. 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).

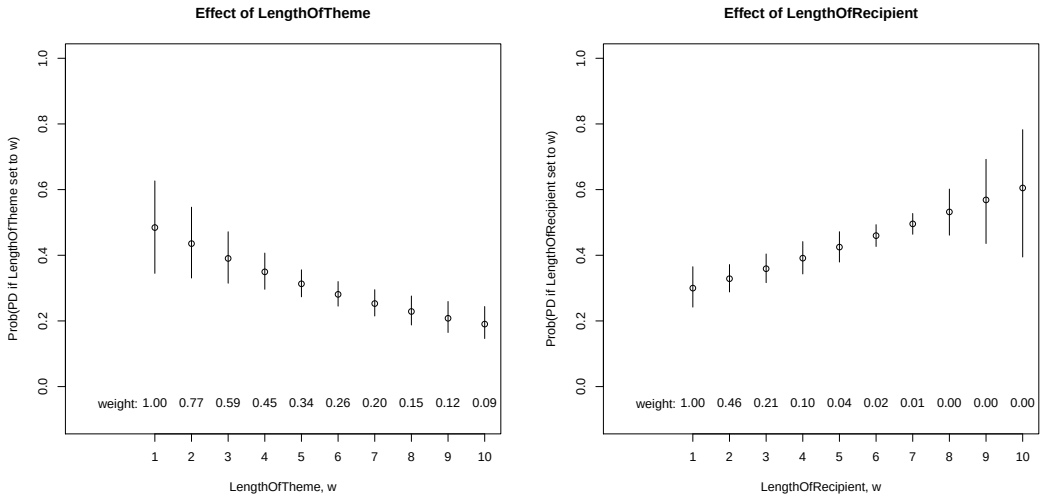

Figure 1: Representing the effects of the integer valued information variables.

A Appendix

A.1 A lemma

We claimed in Section 3.2.2 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 (4) can be extended to (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 P_0 whereas the second assumption, dubbed the "randomization assumption", cannot. Then $\Psi^1(P_0) = E_{\mathbb{P}_0^1}\{Y_1^1\} - E_{\mathbb{P}_0^1}\{Y_0^1\}$.*

Proof. The conditional independence of U^1 and V given $(U^2, \ldots, U^d)$ implies the conditional independence of W^1 and (Y_0^1, Y_1^1) given W^{-1} under $\mathbb{P}_0^1$. This justifies the second equality below:

$$E_{P_0}\{P_0(Y = 1|W^1 = 1, W^{-1})\} = E_{\mathbb{P}_0^1}\{\mathbb{P}_0^1(Y_1^1 = 1|W^1 = 1, W^{-1})\} = E_{\mathbb{P}_0^1}\{\mathbb{P}_0^1(Y_1^1 = 1|W^{-1})\}.$$

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 $\mathbb{P}_0^1(Y_1^1 = 1|W^{-1}) = E_{\mathbb{P}_0^1}(Y_1^1|W^{-1})$ imply the equality $E_{P_0}\{P_0(Y = 1|W^1 = 1, W^{-1})\} = E_{\mathbb{P}_0^1}\{Y_1^1\}$. By symmetry, we also have $E_{P_0}\{P_0(Y = 1|W^1 = 0, W^{-1})\} = E_{\mathbb{P}_0^1}\{Y_0^1\}$. 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

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

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 4.1. This is made easy thanks to the `SuperLearner` package (Polley and van der Laan, 2011) 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. (2014)), random forests (package `randomForest` by Liaw and Wiener (2002)), multivariate adaptive polynomial spline regression (`polymars` function from the package `polspline` by Kooperberg (2013)), and the NDL predicting methodology (`ndlClassify` function from the package `ndl` by Antti Arppe et al. (2014)).

Incidentally, the minimizer α_n of the cross-validated risk (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 van der Laan and Rose, 2011, Chapter 5, for instance) that $\nabla\Psi^1$ is characterized by

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

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

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

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

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

(Rosenblum and van der Laan, 2010; Rosenblum, 2011). Note that there is actually one single non-zero term in the second sum of the RHS of (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 multinational 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

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

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

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

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

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

We refer the interested reader to (van der Laan and Rose, 2011, Chapter 5) and (Rosenblum, 2011) 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:

$$\begin{aligned}
\varepsilon^1_n &= \arg\max_{\varepsilon\in\mathbb{R}} \sum_{i=1}^n \left[Y_i \log(Q^{1,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{1,\text{init}}_{n,\varepsilon}(W_i))\right], \\
\varepsilon^2_n &= \arg\max_{\varepsilon\in\mathbb{R}^K} \sum_{i=1}^n \left[Y_i \log(Q^{2,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{2,\text{init}}_{n,\varepsilon}(W_i))\right], \\
\varepsilon^3_n &= \arg\max_{\varepsilon\in\mathbb{R}^3} \sum_{i=1}^n \left[Y_i \log(Q^{3,\text{init}}_{n,\varepsilon}(W_i)) + (1-Y_i)\log(1-Q^{3,\text{init}}_{n,\varepsilon}(W_i))\right].
\end{aligned}$$

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 (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 (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, $w \in \mathcal{W}^3$ 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 (Rosenblum, 2011).

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 (21), (22), (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 (2014). This, however, is beyond the scope of this article.

References

Antti Arppe, Peter Hendrix, Petar Milin, R. Harald Baayen, and Cyrus Shaoul. *ndl: Naive Discriminative Learning*, 2014. URL `http://CRAN.R-project.org/package=ndl`. R package version 0.2.16.

Joseph Aoun and Yen-hui Li. Scope and constituency. *Linguistic Inquiry*, 1989.

Jennifer E. Arnold, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1), 2000.

R. Harald Baayen. *languageR: data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*, 2009. URL `http://CRAN.R-project.org/package=languageR`.

R. Harald Baayen. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada*, 11(2):295–328, 2011.

Mark C. Baker. Review of S. Pinker, *Learnability and Cognition: The Acquisition of Argument Structure*. *Language*, 68:402–413, 1992.

Mark C. Baker. Thematic roles and syntactic structure. In Liliane Haegeman, editor, *Elements of Grammar. Handbook of generative syntax*, pages 73–137. Kluwer, Dordrecht, 1997.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996a.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996b.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Joan Bresnan. *Lexical-Functional Syntax*. Blackwell, Oxford, 2001.

Joan Bresnan and Tatiana Nikitina. The gradience of the dative alternation. In Linda Uyechi and Lian-Hee Wee, editors, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, pages 161–184. CSLI, Stanford, 2009.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.

Marina K. Burt. *From deep to surface structure*. Harper Row, New York, 1971.

Antoine Chambaz, Isabelle Drouet, and Jean-Christophe Thalabard. Causality, a trialogue. *Journal of Causal Inference*, pages 1–41, 2014. Ahead of print.

Noam Chomsky. *Syntactic structures*. Mouton, The Hague, 1957.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1962.

Walter Daelemans and Antal van den Bosch. *Memory-based language processing*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, 2009.

David Danks. Equilibria of the rescorla-wagner model. *Journal of Mathematical Psychology*, 47: 109–121, 2003.

Kristin Davidse. Ditransitivity and possession. In Ruqaiya Hasan, Carmel Cloran, and David G. Butt, editors, *Functional Descriptions: Theory in Practice*, pages 85–144. John Benjamins, Amsterdam, 1996.

Kristin Davidse. Agnates, verb classes and the meaning of construals: the case of ditransivity in english. *Leuvense Bijdragen*, 87:281–313, 1998.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7. doi: 10.1007/978-1-4612-0711-5. URL `http://dx.doi.org/10.1007/978-1-4612-0711-5`.

D.R. Dowty. Thematic proto-roles and argument selection. *Language*, 67:547–619, 1991.

Nick Ellis. Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1):1–24, 2006.

Nick Ellis and Fernando Ferreira-Junior. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7:187–220, 2009.

Joseph E. Emons. Evidence that indirect-object movement is a structure-preserving rule. *Foundations of Language*, 1972.

Charles Fillmore. *Indirect object constructions in English and the ordering of transformations*. Mouton, The Hague, 1965.

Adele E. Goldberg. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago, 1995.

Georgia Green. *Semantics and syntactic regularity*. Indiana University Press, Bloomington, 1974.

Stefan Thomas Gries. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, 1:1–27, 2003.

Stefan Thomas Gries. Frequency tables: tests, effect sizes, and explorations. In Dylan Glynn and Justyna Robinson, editors, *Polysemy and synonymy: Corpus methods and applications in Cognitive Linguistics*. John Benjamins, 2014.

Stefan Thomas Gries and Anatol Stefanowitsch. Extending collostructional analysis – a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1):97–129, 2004.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. The learnability and acquisition of the dative alternation in english. *Language*, 65(2):pp. 203–257, 1989.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12, 1943.

Ian Hacking. *The taming of chance*, volume 17. Cambridge University Press, Cambridge, 1990.

Barbara Corey Hall. *Subject and object in English*. PhD thesis, Massachusetts Institute of Technology, 1975.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial. *Statist. Sci.*, 14(4):382–417, 1999. ISSN 0883-4237. doi: 10.1214/ss/1009212519. URL `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.

Ray S. Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge, MA, 1983.

Ray S. Jackendoff. *Semantic Structures*. MIT Press, Cambridge, MA, 1990.

Charles Kooperberg. *polspline: Polynomial spline routines*, 2013. URL `http://CRAN.R-project.org/package=polspline`. R package version 1.1.8.

Manfred Krifka. Semantic and pragmatic conditions for the dative alternation. *Korean Journal of English Language and Linguistics*, 2004.

Beth Levin. *English verb classes and alternations: a preliminary investigation*. The University of Chicago Press, Chicago and London, 1993.

Beth Levin and Malka Rappaport Hovav. *Argument Realization*. Cambridge University Press, Cambridge, 2005.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Richard T. Oehrle. *The grammatical status of the English dative alternation*. PhD thesis, Massachusetts Institute of Technology, 1976.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge University Press, Cambridge, 2000.

Stephen Pinker. *Learnability and cognition: The acquisition of argument structure*. MIT Press, Cambridge, 1989.

Eric Polley and Mark J. van der Laan. *SuperLearner*, 2011. URL `http://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-4.

Eric C. Polley, Sherri Rose, and Mark J. van der Laan. Super learning. In Mark J. van der Laan and Sherri Rose, editors, *Targeted learning*, Springer Series in Statistics, chapter 3, pages 43–66. Springer, New-York, 2011.

Malka Rappaport Hovav and Beth Levin. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, 44(1):129–167, 2008.

James M. Robins. Marginal structural models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pages 1–10, 1997.

James M. Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Michael Rosenblum. Marginal structural models. In Mark J. van der Laan and Sherri Rose, editors, *Targeted learning*, Springer Series in Statistics, chapter 9, pages 145–160. Springer, New-York, 2011.

Michael Rosenblum and Mark J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2), 2010. Article 19.

Michael Rosenblum, Steven G. Deeks, Mark J. van der Laan, and David R. Bangsberg. The risk of virologic failure decreases with duration of hiv suppression, at greater than 50% adherence to antiretroviral therapy. *PLoS ONE*, 2009. doi: 10.1371/journal.pone.0007196.

Robert E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Royal Skousen, Deryle Lonsdale, and Dilworth B. Parkinson, editors. *Analogical Modeling: An exemplar-based approach to language*. John Benjamins, 2002.

Kieran Margaret Snyder. *The Relationship between Form and Function in Ditransitive Constructions*. PhD thesis, University of Pennsylvania, PA, 2003.

Margaret J. Speas. *Phrase Structure in Natural Language*. Kluwer, Dordrecht, 1990.

Dirk Speelman. Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In Dylan Glynn and Justyna A. Robinson, editors, *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, pages 487—-533. John Benjamins, 2014.

Daphne Theijssen, Louis ten Bosch, Lou Boves, Bert Cranen, and Hans van Halteren. Choosing alternatives: using bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*, 9(2):227–262, 2013.

Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. URL `http://CRAN.R-project.org/package=rpart`. R package version 4.1-8.

Mark J. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *International Journal of Biostatistics*, 10(1):29–57, 2014.

Mark J. van der Laan and James M. Robins. *Unified methods for censored longitudinal data and causality*. Springer-Verlag, New York, 2003. ISBN 0-387-95556-9.

Mark J. van der Laan and Sherri Rose. *Targeted learning*. Springer, New York, 2011. ISBN 978-1-4419-9781-4. doi: 10.1007/978-1-4419-9782-1.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *Int. J. Biostat.*, 2:Art. 11, 40, 2006. ISSN 1557-4679. doi: 10.2202/1557-4679.1043.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Article 25, 2007.

Aad W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, Berlin, 1995.

Allan R. Wagner and Robert A. Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black and William F. Prokasy, editors, *Classical Conditioning ii*, pages 64–99. Appleton-Century-Crofts, New York, 1972.

Thomas Wasow. Remarks on grammatical weight. *Language Variation and Change*, 9:81–105, 2008.

Robert S. Williams. A statistical analysis of english double object alternation. *Issues in Applied Linguistics*, 5:37–58, 1994.

David H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.