Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The effects of construction probability on word durations during spontaneous incremental sentence production

Victor Kuperman [a,*], Joan Bresnan [b]

[a] Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 626, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M2
[b] Department of Linguistics, Jordan Hall 420-022, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 22 June 2011
revision received 30 March 2012
Available online xxxx

Keywords:
Syntactic alternation
Speech production
Statistical modeling
Information theory
Multimodel inference

ABSTRACT

In a series of seven studies, this paper examines acoustic characteristics of the spontaneous speech production of the English dative alternation (*gave the book to the boy/ the boy the book*) as a function of the probability of the choice between alternating constructions. Probabilistic effects on the acoustic duration were observed in the acoustic signal at the choice point (the first word that commits the speaker to one of the alternatives), before the choice point, but not after the choice point. These findings speak in favor of the simultaneous operation of production mechanisms consistent with both information-smoothing theories and availability-based models of speech production: they are incompatible with a number of competing theoretical accounts. Finally, we outline the statistical modeling procedure of multimodel inference suitable for addressing our multiple working hypotheses and the ultimate question of the explanatory role of probability.

Introduction

Grammars code best what speakers do most—that is, grammars provide the most economical expressions for the speech functions that speakers utilize most often (Du Bois, 1985, pp. 362–363). This core idea has been mathematically developed and empirically investigated at least as early as Zipf (1929, 1935): higher-probability linguistic units are more likely to be easier to pronounce, lexically shorter and phonetically more reduced. The data that support the idea reflect linguistic changes on very different time scales. On the one hand, in both writing and spontaneous speech speakers tend to use reduced (covert or cliticized) rather than overt or full word forms in more probable syntactic contexts (e.g., Krug, 1998; Bybee & Scheibman, 1999; Roland, Elman, & Ferreira, 2006; Frank & Jaeger, 2008; Jaeger, 2010). For example, the use of *don't* is reduced compared to *do not*, and *I think you lost* is reduced compared to *I think that you lost*. These reduced forms, which linguists commonly regard as lexically stored grammatical alternatives or allomorphs, often arise from diachronic changes in the usage of speaker populations over historical time through processes of sound change and grammaticalization (Bybee, 2001; Bybee & Hopper, 2001). On the other hand, a strikingly similar relation to probability appears in the gradient phonetic properties of the acoustic speech signal on the timescale of milliseconds: numerous studies have shown a positive correlation between the contextual probability of a phoneme, syllable, morpheme, or word in spontaneous speech and its reduction in acoustic salience, typically duration or intensity (e.g., Aylett & Turk, 2004, 2006; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Bell, Jurafsky, Fosler-Lussier, Girand, & Gildea, 2003; Fidelholtz, 1975; Gahl & Garnsey, 2004, 2006; Jurafsky, Bell, Fosler-Lussier, Girand, & Raymon, 1998; Jurafsky, Bell, Gregory, & Raymond, 2001; Pluymaekers, Ernestus, & Baayen, 2005; Tily et al., 2009; Van Son & Pols, 2003; Van Son & Van Santen, 2005; Whalen, 1991). Unlike cases of syntactic variation and allomorphy, these phonetic reductions constitute variation on *continuous* scales. Yet the findings of recent work suggest that they

[*] Corresponding author.
E-mail address: vickup@mcmaster.ca (V. Kuperman).

0749-596X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2012.04.003

may be conditioned by the same kinds of high-level construction probabilities as syntactic variation (Gahl & Garnsey, 2004, 2006; Tily et al., 2009). How do these very different kinds of 'reduction'—the alternative discrete grammaticalized constructions, and the continuous acoustic compressions and expansions during speech production—follow from the same probabilistic theory?

Several theories postulate direct linkages between continuous phonetic variation and higher-level grammatical probabilities. One of the most recent and thorough of the phonetic studies of word reduction (Bell et al., 2009) adopts and modifies the widely accepted standard staged model of production (building on Bock & Levelt, 1994 and much subsequent work). The modified model assumes that lexical access is slowed or speeded by frequency-driven activation, and proposes "a mechanism of fluent speech that helps coordinate lexical access and/or phonological encoding and the execution of the articulatory plan" (Bell et al., 2009: 106-7). However, in their analyses of evidence for the proposed model, the authors admit that "predictability from neighboring words is not distinguished from predictability from syntactic constructions" (Bell et al. 2009: 107), a point that has been underscored elsewhere in the literature (Gahl & Garnsey, 2004, 2006) and shown to matter empirically (Levy & Jaeger, 2007; Tily et al., 2009). So we cannot tell from this and similar studies whether word duration variation is in fact sensitive to syntactic construction probabilities. The latter are potentially long-distance effects not necessarily spanned in adjacent words that influence the coordination of lexical access with articulation.

Another kind of theory predicts a direct linkage between continuous phonetic variation and higher-level construction frequencies: (multilevel) exemplar theory (cf. Bybee, 2002, 2007, 2010; Gahl and Yu, 2006; Walsh, Moebius, Wade, & Schuetze, 2010). Theories of this kind vary in the role attributed to mechanisms such as repetition and automatization of articulatory routines (proposed in Bybee's work) or to the threshold activation of dual constituent and unitary exemplars (proposed by Walsh et al., 2010). But they share the proposal that syntactic constructions are stored in memory together with fine-grained phonetic information; more frequently encountered exemplars of a construction can gradually shape the stored representations, leading to sound changes (including but not limited to reduction) and grammaticalization. To mention just one of many examples, New Zealand English has been undergoing a sound change in its vowels, notably raising æ and centralizing ɪ, so that *black widow* pronounced by a New Zealander sounds to speakers of American English much like "bleck wuddow". Hay and Bresnan (2006) studied this centralization in the words *hand* and *give* in a spoken corpus of New Zealand English. They found that the vowel in *give* is more likely to be centralized when it occurs in the context of more frequent dative construction types such as abstract uses of *give* (*give me a hand, give her a chance*), compared to concrete transfer uses (*give us presents, give us a plate full of food*). Similarly, the vowel in *hand* is more likely to be raised ([æ̝] or [ϵ̝]) when the word designates the limb, which is the more frequent use. They relate these findings to multilevel exemplar models that store phonetically detailed instances of constructions. It must be noted, however, that multilevel syntactic representation and variation are just beginning to be explored in this general conceptual framework, and the storage of productive combinatorial syntactic exemplars faces theoretical and practical computational challenges (cf. Daelemans & van den Bosch, 2005; Bod, 2006; Walsh et al., 2010).

Yet another class of theories that predict phonetic effects of construction probability derives from the idea of information-smoothing (e.g., Aylett & Turk, 2004, 2006; Jurafsky et al., 2001; Van Son & Pols, 2003; Van Son & Van Santen, 2005). These theories propose that the information content of a unit, defined as the binary logarithm of the inverse of the probability of the unit (Shannon, 1948), is smoothed over the acoustic signal in order to optimize communication: more informative units (which are less predictable) are expanded in the signal and less informative units (which are more predictable) are compressed. The core ideas have been applied to texts (Genzel & Charniak, 2002), to eye-tracking data (Keller, 2004), and to syntactic variation in speech under the rubric of "Uniform Information Density" (UID) (Jaeger, 2006; Levy & Jaeger, 2007; Jaeger, 2010).

Notably, UID applies these information-theoretic concepts from phonetic reduction research to "all levels of linguistic representation" (Jaeger, 2010: 24). As a "computational theory" that characterizes functional relations at an abstract level (Marr, 1982), it does not postulate specific mechanisms of production, although it makes specific empirical predictions about the locus of production effects. And although the works supporting this theory apply it to all levels of linguistic representation, the inquiry into construction probabilities has mainly focused on the choice between lexically stored discrete alternative word forms or allomorphs such as clitics and null complementizers (e.g., Gomez Gallo, Jaeger, & Smyth, 2008; Gries, 2003; Jaeger, 2010; Frank & Jaeger, 2008; Roland et al., 2006), rather than gradient phonetic reduction and continuous acoustic variables.

One study of phonetic effects of high-level construction probabilities is Gahl and Garnsey (2004, 2006). In an experimental task in which participants read out loud from written texts, Gahl and Garnsey measured phonetic reduction in contexts of differing probabilities of complement types (direct object or sentential complement) given a verb that can take alternative complement types (such as *He believed the rumor yesterday, He believed the rumor was false*). They found that in sentences in which the verbs are biased toward either one or the other type of complement, speakers' pronunciations of the verb and its arguments are shorter when the sentence matches the verb bias than when it does not. However, among several caveats to this work noted in subsequent work with one of the co-authors, the experimental task differs from spontaneous speech in combining both comprehension and production, and hence "the observed effect may have resulted from comprehension difficulty, rather than directly reflecting the workings of the language production system" (Tily et al., 2009, p. 151).

In the subsequent study of gradient acoustic reduction by Tily et al. (2009), word duration is reported to be sensitive to construction probability in spontaneous speech production: for example, the spoken preposition *to* in dative constructions like *he brought the pony to my children* or *give a backpack to me* is reported to vary in duration as a function of the probability of this construction compared to its alternative paraphrase as a double object construction (respectively *he brought my children the pony, give me a backpack*), even after adjusting for low-level transitional probabilites between words. By the time that speakers begin articulating the words of the immediately postverbal phrase—*the pony* or *my children* after *brought*—they have committed to overtly expressing the construction type they have chosen (prepositional or double-object) in describing the event; therefore the reported variation in duration of the preposition *to* occurs relatively far downstream in the flow of spontaneous speech from the point of choice between the alternative constructions, where the information derived from their relative probabilities is available to be smoothed in the acoustic signal (see below). We were not able to replicate the findings (see Study 3 below).

The state of the art in previous research has demonstrated that low-level continuous phonetic variation is sensitive to high-level construction probabilities. In the present paper we revisit some of the prior findings and investigate (i) whether the effects are localized in a way consistent with theoretical predictions and (ii) whether the higher-level probabilities themselves make an independent contribution or merely serve as a summary measure of the individual factors that influence construction outcome. This investigation is based on a corpus sample of spontaneously spoken sentences. In order to distinguish among theoretical predictions of the loci of effects, we extend our dataset beyond the scope of previous studies to multiple data points in the incremental unfolding of spontaneously spoken sentences. Finally, we use powerful but elegant statistical modeling methods to address the issue of whether probability has a role in itself or serves merely as a summary statistic for individual accessibility factors.

Theoretical predictions

Information smoothing—the idea that speakers accommodate the amount of information in the acoustic signal by modulating properties of this signal—has been advocated as a pervasive operational mechanism of speech production in a number of conceptually related proposals, such as the Smooth Signal Redundancy hypothesis (Aylett & Turk, 2004, 2006), the Probabilistic Reduction hypothesis (Jurafsky et al., 2001), or research on speech efficiency (Pluymaekers et al., 2005; Van Son & Pols, 2003; Van Son & Van Santen, 2005). We focus here on one version of the theory, Uniform Information Density, described in the Introduction.

Uniform Information Density (UID): Uniform Information Density theory applies the theory of information smoothing to higher-level syntactic probabilities. But where do these syntactic probabilities come from? The answer is that they are inherent in syntactic *variation*: wherever variation exists, it provides a choice between alternatives; the relative frequencies of the alternants in a sample of language provide the most basic estimate of their probability of occurrence in the language, all else being equal. While phonetic variation is present in every segment during speech, the large size of higher-level syntactic structures during the incremental production of sentences limits the loci of syntactic variation in the speech stream. According to the UID, these loci are where the speaker smooths syntactic information over the signal during incremental word-by-word production to avoid peaks and troughs in information density (Levy & Jaeger, 2007; Jaeger, 2010). For example, variation in postverbal complement structures occurs in the immediate postverbal position, where the speaker commits to articulating one of the alternants. The 'choice point' is operationalized here as the part of the acoustic stream where the speaker's commitment to articulating one of the variants is manifest, *from the speaker's point of view*.[1]

Smoothing of the information carried by a syntactic choice consists of distributing the information over the signal as uniformly as possible. At a choice point, therefore, the speaker will tend to lengthen (or shorten) the relative acoustic duration of the word that commits her to articulating a less (or more) probable syntactic variant. More probable speech units encode less information and are easier to reconstruct from context, so reducing the amount of signal associated with these units is less likely to jeopardize successful communication. Conversely, a less probable unit by virtue of encoding more information needs to be more salient—typically, stretched over production time—to ensure that speech perception is not overly effortful for the hearer and the unit's transmission is successful.

While information smoothing at the lowest phonetic levels can flow continuously, smoothing of *syntactic* information on this theory is expected to occur word by word at the choice points during incremental production where the speaker controls the articulation of syntactic variants. For this reason, UID predicts syntactic reduction effects in production only at such 'choice points' (Jaeger, 2010, p. 26, Fig. 1a & b; p. 28). As Frank and Jaeger (2008, p. 940) emphasize, "Since UID makes predictions about speakers' choices, we are only interested in cases where speakers actually have a choice between two different realizations of the target . . .." This point is further illustrated by an explicit theoretical model of top-down stochastic incremental sentence production (Levy & Jaeger, 2007, Eq. (2) & Fig. 2).

As noted above, although UID has been claimed to apply to all levels of linguistic representation, its applications to syntactic information have focused primarily on the choice between lexically stored discrete alternative word forms or allomorphs such as clitics and null complementizers,

[1] Unlike the comprehender, the speaker is assumed to know what message she is formulating, and hence lexical variation between homonyms, for example, need not affect the speaker's commitment to the syntactic formulation of her message. Of course, the speaker could conceivably delay her own commitment when the first word of both alternative constructions is the same, as is postverbal *the* in *give the dog the bone, give the bone to the dog*. But by the time *dog* or *bone* is articulated in this example, construction choice has been made and is manifest for the speaker regardless of any uncertainty for the comprehender.

rather than continuous acoustic measures such as duration. The specific loci of temporal duration effects it predicts provide a strong test of the theory.

Syntactic reduction in the sense of Levy and Jaeger (2007) can only take place (by definition) where there is variation in the choice of full and reduced word forms—hence at "choice points" as defined above. But (syntactically conditioned) *phonetic* reduction is not limited in this way; it can apply more broadly with all types of syntactic variation, including variations of word order rather than word forms. For these cases as well, we assume that syntactic information will be smoothed across the speech signal at the point where the speaker's commitment to articulating one of the syntactic variants is manifest, from the speaker's point of view—the choice point again. Articulatory theory (e.g., Browman & Goldstein, 1992) tells us that words can be compressed or expanded during speech production by dynamically modifying the spatiotemporal structure of their stored articulatory routines. Since these must be executed sequentially word by word during incremental production, and since higher-level syntax has a relatively small effect on phonetic reduction (Bell et al., 2009; Tily et al., 2009), the speaker gains the most communicative leverage in smoothing syntactic information just where the word choice manifests the syntactic choice. Starting earlier before the choice point, where an abstract syntactic construction choice may have been covertly made, or later, where the construction choice is already made, are communicatively inefficient.

The dative alternation in spontaneous American English provides a convenient illustration. This alternation is a syntactic choice in which the alternatives (ditransitive, or NP NP *I'll give Tom the book* and prepositional, or NP PP *I'll give the book to Tom*) differ in their word order but not in their core meaning (for discussion see Bresnan, Cueni, Nikitina, & Baayen, 2007; Bresnan & Nikitina, 2009; Fellbaum, 2005; Bresnan & Ford, 2010). The choice between alternatives depends on multiple and often conflicting properties of the verb *give*, the recipient *Tom* and the theme *the book* (Arnold, Losongco, Wasow, & Ginstrom, 2000; Bock & Irwin, 1980; Bock, Loebell, & Morey, 1992; Collins, 1995; Gries, 2005; Hawkins, 1994; McDonald, Bock, & Kelly, 1993; Lapata, 1999; Prat-Sala & Branigan, 2000; Snyder, 2003; Thompson, 1990, 1995; Wasow, 2002). For instance, the probability of a ditransitive construction is increased when the first phrase following the verb is a pronoun, is definite, is mentioned in prior discourse, is animate, or is short. Incorporating these and other variables such as the previous occurrence of a parallel structure (Weiner & Labov, 1983; Bock, 1986; Gries, 2005; Pickering, Branigan, & McLean, 2002), a probability model of the dative alternation predicts the choice of construction for dative verbs in spoken English with very high accuracy (Bresnan et al., 2007).

The evidence thus suggests that the differences in alternative constructions are preferences, not categorical regularities. To give an example, the probability of the prepositional dative *afford some time to the military or helping elderly people* is estimated at 89%, and that of its counterpart *afford the military or helping elderly people some time* at 11%. For the dative alternation, the choice point—or the point in articulation when the speaker manifests her commitment to one of the available variants during incremental speech production (Ferreira, 1996)—is typically the first word of the proximate object (e.g., *some* in *afford some time to the military or helping elderly people*). On UID, the word that serves as the point of choice between the dative alternants for the speaker is where a lower or higher construction probability of the selected alternant would yield a more or less salient realization of the word. Importantly, UID predicts no effect of the dative construction probability *past* the choice point. As soon as the speaker begins articulating the first words of the recipient *some time* or the theme *the military or helping elderly people*, the order of upcoming objects is solidified and fully determined as either a theme, or a preposition plus a recipient, respectively. During word-by-word incremental production, then, the probability of the spoken alternant becomes exactly 1 after the choice point, and that of the rejected alternant becomes exactly 0. Hence no variability is expected in the articulation of the remainder of the selected construction that can be attributed to its dative construction probability.

These predictions of UID contrast with those of certain other theories, under the simplest assumptions. For example, the exemplar theories discussed in the Introduction could treat the *entire* syntactic construction under production as a potential locus for the effect of construction probability during incremental production (although further development and elaboration of the syntactic exemplars assumed in these theories could result in different predictions). Thus, one could expect that all, or at least multiple, words across the preferred construction *afford some time to the military or helping elderly people* would be realized acoustically in the way that reflects a higher probability of this construction as compared to its alternant. (This expectation would generally hold true even if one allows for the possibility that construction probability affects the realization of different words within the multi-word dative to a different degree.)

For example, in explaining the reasoning that led to their predictions concerning duration, Gahl and Garnsey (2004, p. 754) write (emphasis added):

> Previous research has shown that high-frequency and high-probability words tend to be short. *By extension, we hypothesized that words and phrases instantiating high-probability syntactic structures would also be short.* Sentential complements—and hence, clause boundaries—have a higher probability following SC-bias verbs than following DO-bias verbs, and direct objects have a higher probability following DO-bias verbs than following SC-bias verbs. *We reasoned that, in bias-matching contexts, the lengthening typically observed near clause boundaries and phrase-finally might be offset by phonetic reduction found in high-probability items generally.* As a result, we hypothesized that the lengthening near prosodic boundaries would be observed to a greater extent in bias-violating contexts than in bias-matching ones.

The prediction that "words and phrases instantiating high-probability syntactic structures would also be short" and thereby offset prosodic boundary lengthening effects, we will refer to as *broad-scope*, in contrast to the specific prediction of UID at the choice point, as in Table 1A and B.

Table 1
The timeline of effects of syntactic probability on the ease of incremental production at different syntactic positions in the dative construction (verb, proximate object, distant object), as predicted by the Uniform Information Density theory and theories that make broad-scope predictions of construction-probability effects.

	Verb	Proximate Object	(Preposition +) Distant Object
	Before the choice point	*At the choice point*	*After the choice point*
A. UID	None	Probability of the spoken alternant	None
B. Broad scope	Probability of the spoken alternant	Probability of the spoken alternant	Probability of the spoken alternant

Gahl and Garnsey (2004, p. 769) observe that exemplar-based models most readily accommodate their findings of phonetic effects of high-level construction probabilities. Their broad-scope predictions of an extended temporal locus of the probabilistic effects of syntactic construction are compatible with their own findings described in the Introduction, which go beyond the word-to-word transitions at the point of the postverbal clause or phrase boundary to the duration of the entire postverbal noun phrase (Gahl & Garnsey, 2004, p. 754 & 763). And they are compatible with the subsequent report that words past the choice point—the preposition *to* in the prepositional dative and the first word of the distant object in the ditransitive dative—were realized longer in the less probable dative constructions (Tily et al., 2009). The predictions are also in line with the finding of Wagner Cook, Jaeger, and Tanenhaus (2009) that the entire utterance containing a dative construction elicits more gesturing and disfluencies (also confirmed in Tily et al. (2009)) if the construction is of a relatively low probability.

The contrasting UID and broad-scope predictions for the case of the dative alternation are summarized in Table 1A and B, respectively. In the present work we tested these contrasting predictions of the loci of construction probability effects by investigating word durations at multiple syntactic positions in the incremental acoustic production of English datives.

Method

Materials

We performed a time-locked comparison of theoretical predictions using the set of 2328 sentences with datives extracted from the Switchboard corpus of spontaneous spoken US English (Godfrey, Holliman, & McDaniel, 1992; Greenberg, Hollenback, & Ellis, 1996) and annotated by Bresnan et al., (2007), Bresnan and Nikitina (2009), and Recchia (2007), described most fully by Bresnan and Ford (2010). The Switchboard corpus is a collection of 240 h of spontaneous dialogs recorded as telephone conversations between pairs of speakers. A list of topics was suggested to speakers, though they were not required to maintain that topic throughout the conversation. The speakers were not familiar with each other. The choice of spontaneous speech in the present and previous studies of Bresnan and colleagues provides relative ecological validity in comparison to speech production conditioned by experimental tasks. This study builds on Bresnan et al.'s (2007) statistical model of the construction probability of the dative alternation as a function of multiple factors, and on Tily et al.'s (2009) investigation of the effects of construction probability on the acoustic duration and the rate of disfluencies in the production of preposition *to* in the NP PP datives and in the production of the distant object (recipient) of the NP PP dative. Our use of the same data source that Bresnan et al. (2007) and Tily et al. (2009) based their observations on ensures direct comparability between the present and the earlier studies.

Dependent variables

The time-aligned transcript of the Switchboard corpus produced by Deshmukh, Ganapathiraju, Gleeson, Hamaker, and Picone (1998) provides an estimate of the acoustic duration for each word in each sentence based upon an improved resegmentation of the corpus. In a series of five studies we consider as the dependent variable the acoustic duration of the following (words within) syntactic constituents:

1. The first (recipient) object in the NP NP dative (e.g., *John wrote* **him** *a letter*).
2. The first (theme) object in the NP PP dative (e.g., *She gave* **a book** *to them*).
3. The preposition *to* in the NP PP dative (e.g., *She gave a book* **to** *them*).
4. The second (theme) object in the NP NP dative (e.g., *John wrote him* **a letter**).
5. The second (recipient) object in the NP PP dative (e.g., *She gave a book to* **them**).

The distribution of the raw durational data in each dataset is skewed (see Studies 1–5 below for descriptive statistics). We conducted Box–Cox tests to establish what power transformation would be optimal in rendering the distribution closer to normality (Box & Cox, 1964; for a detailed discussion in psycholinguistic literature, see Kliegl, Masson, & Richter, 2010). The Box–Cox power transformation is defined as $y(\lambda) = \frac{y^{\lambda}-1}{\lambda}$, if $\lambda \neq 0$, and $y(\lambda) = \log(y)$, if $\lambda = 0$, where λ is the power coefficient. The method implemented as the function `boxcox` in the library `MASS` in `R` (Venables & Ripley, 2002) estimates the optimal value for λ for the given linear regression model. An optimal value of λ close to -1 indicates the reciprocal transformation of the skewed variable, an optimal value of 0 indicates the logarithmic transformation, while an optimal value close to 1 suggests that no transformation is warranted (Kliegl et al., 2010). We estimated the optimal values of λ for each of the five data sets by providing linear regression models (fitted using the `lm` function in `R`) with untransformed acoustic

Please cite this article in press as: Kuperman, V., & Bresnan, J. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* (2012), http://dx.doi.org/10.1016/j.jml.2012.04.003

durations as dependent variables and multiple predictors (see below for the detailed descriptions) as the input for the function `boxcox`. The respective optimal values of λ for the given studies enumerated above were: 0.3, 0.3, 0.2, 0.0, 0.0. As all values of λ were close to zero, we log-transformed acoustic durations in all datasets to avoid skewness, attenuate the disproportional influence of outliers and approximate the normal distribution of production data.

Critical predictors

Measures of syntactic probability: Our critical predictor is the probability of each instance of the alternative dative constructions used by the speakers, estimated by the statistical model of Bresnan et al. (2007) on the basis of multiple structural, semantic and pragmatic characteristics of dative verbs and their objects (see the list below in this section). To give an example, the probability of an NP PP alternative is 0.89 for the prepositional dative *afford some time to the military or helping elderly people* and only 0.008 for the double object dative *allot each of us enough time to go out*. Since the model's probability in Bresnan et al. is defined with respect to the NP PP alternative, we either used this probability directly in utterances with the NP PP order of constituents ($p = 0.89$ in the first example), or its inverse in the utterances with the NP NP realization of dative ($1 - p = 1 - 0.008 = 0.992$ in the second example). The resulting probability measure (labeled as *Prob*) ranges from 0 to 1 and its distribution is highly skewed such that high-probability outcomes are (unsurprisingly) very common. The mean value of construction probability is extremely high for ditransitive or NP NP datives (mean = 0.953, SD = 0.127; min = 0.008, 1st quartile = 0.972, 2nd quartile = 0.995, 3rd quartile = 0.998, max = 0.999), and slightly lower for prepositional or NP PP datives (mean = 0.830, SD = 0.251; min = 0.005, 1st quartile = 0.757, 2nd quartile = 0.937, 3rd quartile = 0.996, max = 0.999).

Components of syntactic probability: We aimed to assess whether variables that reliably predicted the binary syntactic choice in dative alternation (Bresnan et al., 2007) would have an effect on the acoustic signal associated with the production of alternants. The variables were dubbed here as "components of probability". In sum, the following components were considered: length, definiteness, givenness, person, animacy, pronominality and number of the recipient (coded as *length.rec* gauged in orthographic words; *def.rec* with values "def" and "indef"; *given.rec* with values "given" and "ngiven"; *person.rec* with a value "local" when referring to interlocutors in the dialog *I* or *you* and with a value "non-local" for other referents; *animacy.rec* with values "a" and "ina"; *pron.rec* with values "pron" and "npron" for pronominal and lexical nouns; and *number.rec* with values "singular" and "plural"). A similar set of the theme properties was considered: *length.theme, def.theme, given.theme, person.theme, animacy.theme, pron.theme, number.theme* with values as defined above. We also took into account syntactic parallelism in the dialog (whether or not a dative construction of the same type was present in the same dialogue), and semantic class of the verb.

Control variables

Frequency-based controls: It is a robust finding in the production literature that the frequency of occurence of word N co-determines its acoustic realization, such that more frequent words are pronounced with less acoustic salience, for instance, shorter (e.g., Fidelholtz, 1975; Rhodes, 1996; Zipf, 1929). Also probabilities of word N given the preceding word and, separately, the following word are known to co-determine the acoustic duration of word N: the higher the probability (frequency) of a backward or forward bigram of word N, the more acoustically reduced word N is (cf. Bell et al., 2003, 2009; Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Jurafsky et al., 2001). We obtained target word frequency as well as backward and forward bigram probabilities from the Web 1T ngram corpus (Brants & Franz, 2006). If the bigram from the Switchboard dataset was not registered in the Brants and Franz corpus, we assigned the frequency of 1 to the bigram; all unigrams were attested in the corpus. Following up on findings of Jaeger and Post (2010), we also included frequency of the upcoming word and the probability of the upcoming word given preceding word in the list of predictors. We log-transformed (base e) all frequency-based measures to remove skewness from their distributions.

Acoustic controls: Speech rate was defined as the number of phonemes realized per second. To avoid a circular use of the dependent variable in the calculation of speech rate (*SpeechRate*), we obtained this measure by subtracting the duration of the target word from the total duration of the utterance and subtracting the word's number of phonemes from the total count of phonemes in the utterance, and dividing the latter by the former. We also took into account the length of the target word (in phonemes) labeled. Since the immediate phonological environment may alter acoustic characteristics of production, we coded whether or not the target word ended in a vowel and whether or not the following word began with a vowel. Furthermore, we coded the position of the stressed syllable in the word, the presence of the consonant cluster at the end of the word, and the average diphone frequency for the word. All phonological information was extracted from the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) or, in the case this source lacked the necessary information, added manually.

Other controls: We coded the ordinal word position of the verb in the sentence, as it may codetermine the amount of syntactic and semantic information available at the position in which the realization of the dative construction begins. Finally, the total number of words in the utterance was computed, as the overall complexity of planning and producing a larger utterance may have consequences for the production effort at each word (Quené, 2008; Yuan, Liberman, & Cieri, 2006).

Prosodic factors might influence acoustic durations as well. There is evidence (Selkirk, 2003) that prosodic phrasing is mainly driven by syntactic structure, also specifically in English datives (Anttila, Adams, & Speriosu, 2010). Heads (including functional heads) form a prosodic phrase with their lexical complements. In this sense, all the syntactic positions that we examine are controlled for their positions within prosodic units. Verbs as heads were in the beginning of respective phonological phrases. First words of the proximate object, as well as prepositions, were located within phonological phrases. The one-word distant objects (accounting for over 90% of respective data sets) were at the end of those prosodic phrases. Thus, in each study the position within the phonological phrase was held constant across all data points. As phonological words are often tied to the boundaries of syntactic constituents, our measures of lengths of syntactic objects afford a certain level of control over the position of a lexical word in the phonological word, and of the size of the prosodic unit.

Data trimming

Lengths of recipient and of theme showed skewed Zipfian distributions and varied widely between the dative alternatives: the average recipient length was 1.104 words (SD = 0.436) in NP NP datives and 1.850 words (SD = 1.144) in NP PP datives, while the average length of the theme was 3.742 words (SD = 2.861) in NP NP datives and 1.591 words (SD = 1.144) in NP PP datives. We set off the maximum length as 15 words for the theme (original range 1–46) and 6 words for the recipient (original range 1–16). The resulting data set accounts for over 90% of the probability mass for each variable and excluded disproportionately long syntactic objects (see Roland, 2009). To remove skewness even further, we log-transformed lengths of the theme and recipient. We also standardized log lengths (by subtracting the mean log length from the given log length and dividing the difference by one standard deviation).

From the original dataset of 2328 utterances, we excluded utterances with coding errors as well as those utterances in which disfluencies (pauses, fillers, or repetitions) immediately preceded the verb or occurred at any position within or up to two words after the dative construction: the total of 311 data points (13%) were removed. Finally, in each of the 5 datasets, we additionally excluded data points where any numerical covariate (speech rate, backward and forward bigram and word frequency) is further than 3 standard deviations away from the log mean. Depending on the data set, this step of the trimming procedure removed 5–7% of the data points. Finally, in each data set we removed words with acoustic durations that were 3 or more standard deviations away from the mean duration. This step shaved off additional 0–2% of the total of data points across data sets. The sizes of resulting data sets for respective dependent variables are reported in Table 2.

Most of our studies involve both a large number of predictors (construction probability and its components, as well as multiple controls) and a relatively small number of observations (with the minimum $N = 235$). As models with a low observation/predictor ratio are susceptible of overfitting the data, Harrell (2001, p. 61) recommends that the ratio is kept no lower than at 10 or 20 observations per coefficient in the statistical model. To avoid overfitting, we lowered the number of predictors in our models using the following methods. First, we did not enter in the model those control variables that showed a weak Spearman's pairwise rank correlation with the respective dependent variable ($|\rho| < 0.1$). This step consistently filtered out variables with low predictivity of acoustic duration (e.g., utterance length, position of stress, quality of the segment preceding or following the target word, and such). In this we followed the statistical literature (e.g., Freedman, 1983; Lukacs, Burnham, & Anderson, 2010) for the demonstration that explanatory variables with no relation to the response may result in spurious effects, especially when the sample size is small relative to the number of models fitted and the stepwise model selection procedure is applied. Second, we applied principal component analysis in the regression model (Harrell, 2001) to highly collinear frequency-based controls and phonological word length, and only used as predictors those principal components that accounted for at least 5% of the variance in the original collinear space (1–3 principal components across models). Third, binary variables with less than 5% of data points representing one of their levels were excluded from consideration as predictors in respective models.

Statistical modeling

We use linear mixed-effects regression models with crossed random effects, implemented in package `lme4` of the statistical software `R` (R Development Core Team, 2007), that allow for the simultaneous consideration of multiple covariates, while keeping under statistical control the between-speaker and between-item variance (cf. Baayen, 2008; Baayen, Davidson, & Bates, 2008; Bates &

Table 2
Summary of the effects of construction probability and its components on acoustic durations of words across the dative construction. Sample sizes N are reported after the trimming procedures were applied. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$), and the t-test based p-values are reported for those predictors that reach significance at the 0.05 level. ES stands for effect size, a model-estimated contrast in acoustic durations of words showing the maximum and the minumum values of the predictor.

NP NP (double)	NP PP (prepositional)
At the choice point	
Study 1, recipient (first object): $N = 1317$	Study 2, theme (first object): $N = 235$
Probability ($\bar{\beta} = -0.453$; $\bar{\sigma} = 0.145$; $p = .003$; ES = −109 ms)	Probability ($\bar{\beta} = -0.431$; $\bar{\sigma} = 0.186$; $p = .028$; ES = −61 ms)
Length.theme ($\bar{\beta} = 0.051$; $\bar{\sigma} = 0.014$; $p = .001$; ES = 58 ms)	
After the choice point	
NA	Study 3, preposition *to*: $N = 422$
Study 4, theme (second object): $N = 889$	Study 5, recipient (second object): $N = 286$
–	Length.theme ($\bar{\beta} = 0.085$; $\bar{\sigma} = 0.040$; $p = .042$; ES = 41 ms)

Sarkar, 2007; Pinheiro & Bates, 2000). All models include speaker and (except for the model for the duration of *to* where word is held constant) word as random effects: both random intercepts significantly ($p < 0.001$) improve the performance of the models as established by likelihood ratio tests. We report the outputs of these models below. In a separate set of analyses, we also included in each model the random by-speaker slopes or contrasts for all predictors in that model. In this way we effectively model all sources of random variation between speakers and target words with respect to our predictors. The observed patterns converged perfectly in terms of the polarity and statistical significance (at the 0.05 level) of effects with the results reported here.

Because of the large number of predictors and controls required to answer the crucial theoretical questions, the problem of model selection loomed very large in the present work. The widely accepted practice of selecting a single best-performing model as an approximation of the information about full reality present in the data fails to account for *model uncertainty*—the existence of acceptable alternative models of the same data, each telling a different story. We therefore decided to employ an information-theoretic method of multimodel inference developed by Burnham and Anderson (2002, 2004). This method estimates the strength of evidence for each model in the set of possible models, ranging from one having zero predictors (the intercept-only model) to the model with the full list of predictors. A set of *n* predictors generates a model set consisting of 2^n models. The strength of evidence for the model is defined as the amount of information lost when that model is used to approximate full reality or truth, or equivalently, the distance between the model and full reality (Burnham & Anderson, 2004): the required metric is provided by the information-theoretic measure of the Kullback–Leibler (KL) distance (Kullback & Leibler, 1951).

Multimodel inference crucially takes into account the distribution of the strengths of evidence over the model set and factors this distribution in when estimating inferential statistics for and the relative importance of predictors. Basing estimates on the *entire* model set rather than any particular model or subset safeguards researchers from founding their interpretations on estimated model parameters that are potentially specific to either the model selection algorithm or to a single model. It also obviates the need for multiple comparisons between models and/or variables, which are inherent in some model selection algorithms and as such may require corrections for nominal significance levels. The motivation and mathematical foundations of this approach, as well as the criticism of the methods of single best model selection are exhaustively presented in Burnham and Anderson (2002, 2004) and Lukacs et al. (2010). In the Appendix, we outline in detail the multimodel inference procedure (Burnham & Anderson, 2002, 2004) that we used for statistical data analyses in all studies. We also discuss there how the multimodel inference technique handles collinearity, a crucial issue for our analyses that simultaneously consider such correlated variables as construction probability and its components. In what follows, we report the outcomes of the multimodel inference procedure for all of the studies that we conducted.

Results and discussion

Below we present statistical analyses for the five studies that probe a range of timepoints at and past the choice point during the incremental production of alternating dative constructions. Our main emphasis is on the time-course of effects that the critical predictors (construction probability and its component predictors) register in the acoustic signal, as diagnostics for a range of proposed speech production theories. The effects of critical predictors that reached significance at the 0.05 level under our multimodel inference procedure (see Appendix) are summarized across studies in Table 2, while the detailed modeling outputs for particular studies are reported in Tables 3–7. For each predictor, we report the model-averaged estimates of its regression coefficient ($\bar{\beta}$) and standard error ($\bar{\sigma}$), as well as the boundaries of the 95% confidence interval and *p*-values based on *t*-tests, with the number of models fitted to the dependent variable minus one as degrees of freedom. We also report for each critical predictor the cumulative probability or the relative importance it accrues in the model set of the respective study.

At the choice point

Studies 1 and 2 target the point in the production of datives where speakers realize their choice between alternative placements of constituents. On UID account, it is at

Table 3
Outcome of the multimodel inference procedure for the duration of the initial word of the first (recipient) object in the NP NP dative, $N = 1317$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's *p*-value, and its cumulative probability. Marked in bold are predictors with *p*-values below 0.05.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	*p*	Cumul.Prob
Intercept	5.708	0.170	5.374	6.041	<.001	NA
SpeechRate	−0.029	0.004	−0.036	−0.021	<.001	NA
VerbPos	0.001	0.001	−0.001	0.003	.40	NA
PC1	0.001	0.003	−0.005	0.006	.37	NA
PC2	−0.030	0.009	−0.047	−0.012	.002	NA
given.theme = ngiven	−0.000	0.047	−0.091	0.091	.40	0.20
def.theme = indef	−0.017	0.032	−0.081	0.046	.34	0.29
pron.theme = pron	−0.092	0.049	−0.189	0.004	.07	0.74
length.of.theme	**0.051**	**0.014**	**0.024**	**0.077**	**.001**	**1**
Prob	**−0.453**	**0.145**	**−0.738**	**−0.168**	**.003**	**1**

Table 4
Outcome of the multimodel inference procedure for the duration of the initial word of the first (theme) object in the NP PP dative, $N = 235$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's p-value, and its cumulative probability. Marked in bold are predictors with p-values below 0.05.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	p	Cumul.Prob
Intercept	5.311	0.386	4.554	6.067	<.001	NA
SpeechRate	−0.043	0.011	−0.064	−0.023	<.001	NA
PC1	0.020	0.009	0.003	0.038	.034	NA
PC2	−0.005	0.036	−0.077	0.066	.39	NA
PC3	0.113	0.130	−0.143	0.368	.38	NA
given.rec = ngiven	−0.141	0.075	−0.287	0.005	0.07	0.10
given.theme = ngiven	−0.043	0.210	−0.454	0.368	0.39	0.10
pron.theme = pron	0.381	0.214	−0.038	0.800	0.08	0.46
Prob	**−0.431**	**0.186**	**−0.796**	**−0.066**	**0.028**	**0.64**

Table 5
Outcome of the multimodel inference procedure for the duration of the preposition *to* in the NP PP dative, $N = 422$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's p-value, and its cumulative probability.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	p	Cumul.Prob
Intercept	4.898	0.271	4.368	5.429		NA
SpeechRate	−0.028	0.013	−0.053	−0.003	.039	NA
VerbPos	0.003	0.003	−0.004	0.010	.242	NA
PC1	−0.016	0.023	−0.061	0.028	.312	NA
PC2	0.001	0.040	−0.079	0.080	.398	NA
given.rec = ngiven	−0.135	0.106	−0.343	0.073	.177	0.24
given.theme = ngiven	0.171	0.124	−0.073	0.415	.154	0.19
def.rec = indef	0.045	0.105	−0.161	0.250	.364	0.15
def.theme = indef	0.036	0.132	−0.223	0.296	.384	0.19
pron.rec = pron	−0.102	0.116	−0.330	0.126	.271	0.24
pron.theme = pron	0.030	0.174	−0.312	0.371	.393	0.19
length.rec	0.016	0.048	−0.079	0.111	.377	0.14
length.theme	−0.002	0.049	−0.098	0.093	.398	0.11
Prob	−0.193	0.171	−0.528	0.142	.211	0.32

Table 6
Outcome of the multimodel inference procedure for the duration of the second (theme) object in the NP NP dative, $N = 889$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's p-value, and its cumulative probability.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	p	Cumul.Prob
Intercept	5.598	3.144	−0.565	11.760	<.001	NA
SpeechRate	−0.022	0.007	−0.035	−0.008	.003	NA
VerbPos	−0.013	0.199	−0.404	0.377	.398	NA
PC1	−0.008	0.118	−0.238	0.222	.340	NA
PC2	0.008	0.161	−0.308	0.323	.400	NA
given.rec = ngiven	0.021	0.074	−0.124	0.165	.383	0.13
given.theme = ngiven	−0.055	0.070	−0.192	0.082	.293	0.20
def.theme = indef	0.099	0.198	−0.289	0.487	.352	0.20
pron.rec = pron	−0.063	0.067	−0.195	0.069	.256	0.18
length.rec	0.013	0.018	−0.022	0.048	.307	0.18
length.theme	−0.006	0.019	−0.043	0.031	.379	0.12
Prob	−0.196	0.176	−0.541	0.149	.214	0.26

this point that speakers accommodate in the acoustic signal the information density associated with the choice, such that a less probable (more informative) alternative would show an increased amount of signal (slower or more disfluent speech, or a higher prevalence of gestures accompanying speech, e.g., Jaeger, 2010; Wagner Cook et al., 2009).

Study 1: Initial word of the first (recipient) object in the NP NP dative

For our purpose of comparing acoustic durations, the variability in the length and the syntactic complexity of recipients in the NP NP datives is staggering. To circumvent this problem, we zoomed in on the initial word of the first

Table 7
Outcome of the multimodel inference procedure for the duration of the second (recipient) object in the NP PP dative, $N = 286$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's p-value (based on the t-test with the number of models in the set minus one, as degrees of freedom), and its cumulative probability. Marked in bold are predictors with p-values below 0.05.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	p	Cumul.Prob
Intercept	4.763	0.237	4.299	5.227	<.001	NA
SpeechRate	−0.008	0.010	−0.027	0.012	.289	NA
VerbPos	−0.004	0.003	−0.009	0.001	.164	NA
PC1	0.161	0.065	0.033	0.289	.019	NA
PC2	−0.026	0.032	−0.090	0.038	.286	NA
given.rec = ngiven	−0.096	0.124	−0.340	0.148	.295	0.23
given.theme = ngiven	0.058	0.086	−0.110	0.226	.318	0.19
pron.rec = pron	−0.101	0.229	−0.549	0.347	.362	0.15
pron.theme = pron	0.014	0.100	−0.182	0.211	.394	0.14
length.rec	−0.028	0.066	−0.157	0.102	.364	0.13
length.theme	**0.085**	**0.040**	**0.007**	**0.163**	**.042**	**0.72**
Prob	0.173	0.141	−0.103	0.450	.187	0.33

(recipient) object in the NP NP dative. We selected monosyllabic function words, with at least 10 occurrences in this syntactic position. Seven pronouns satisfied these selection criteria, *it, her, him, me, us, them* and *you*, which left us with 1317 data points after the standard trimming procedure. The acoustic duration of target words ranged from 40 to 480 ms (mean = 138; SD = 68). Virtually all selected recipients (1312 out of 1317) consisted of a single pronoun (e.g., the full recipient being *her* rather than *her coat*), so we excluded the length, pronominality and definiteness of recipients from the list of predictors. Our choice of recipients also ensured that no pair of the objects in any construction began with the same word, as in *gave the boy the book*. (If this were the case, then the point of choice between alternating constructions might shift to the second word of the recipient, rather than the initial word that we zoom in on.) Two principal components (PC1 and PC2) account for 89% of variance in the frequency-based independent measures and phonological word length, and were entered into the model along with speech rate as fixed predictors. A set of five critical predictors formed the model space for the multimodel inference; see Table 3. No other variable passed the pre-screening requirements.

Construction probability is a significant predictor of the acoustic duration of the initial word of recipient $[\bar{\beta} = -0.453;\ \bar{\sigma} = 0.145;\ p = 0.003]$. Initial words of recipients in the NP NP constructions that have a higher probability of this constituent order are realized faster: the contrast in the recipient's acoustic duration between the most and least probable construction is estimated at 109 ms. We note that the effect is mostly driven by the 25% (330 datapoints) of cases with the lowest probability (range: 0.008–0.980), as the datapoints from the 26th to the 100th percentiles have the probability of the NP NP realization at the ceiling (range: 0.981–0.999). Even if confined to the subset of the data showing a substantial variability in syntactic probability, the effect is noteworthy. It supports the claims of UID (Levy & Jaeger, 2007; Jaeger, 2010) that information carried by the syntactic structure is accommodated in the acoustic signal at the choice point, in such a way that the realization of a less probable (more informative) choice comes with an inflated acoustic duration of the speech unit (here, the word), revealing the speaker's commitment to the spoken alternant, and not to the rejected one. The observed probabilistic effect is also compatible with those probabilistic approaches that would argue for a broader scope of acoustic reduction effects.

We also observed a strong effect of the (standardized log) length of theme on the acoustic duration of the recipient $[\bar{\beta} = 0.051;\ \bar{\sigma} = 0.014;\ p = 0.001]$. Recipients have a longer realization if followed by longer themes: the durational contrast between constructions with the longest and the shortest themes is estimated at 58 ms. Both the length of theme and construction probability have the highest possible cumulative probability of 1 and thus are strongly supported as important predictors of the acoustic duration in this data set. Construction probability and the length of theme are correlated: in this data set, Pearson's $r = 0.4$. As we demonstrate in the Appendix, given the relatively weak magnitude of both effects on acoustic duration and our choice of multimodel inference as the analytical technique, this fact did not pose a problem for the accuracy of model estimates or the statistical inferences in this and other studies.

Study 2: Word in the initial position in the first object (theme) in the NP PP dative

In order to assess the effects of probability and its components on the first object (theme) in the NP PP construction and to avoid the excessive variability in the length of the theme, we selected a subset of monosyllabic function words with at least 10 occurrences in the syntactic position under consideration. The subset includes six pronouns and determiners (*a, her, it, that, the, them*), yielding the total of 235 datapoints. Acoustic durations of target words range from 40 to 340 ms (mean = 109; SD = 53). Three principal components (PC1, PC2 and PC3) account for 94% of variance in the frequency-based independent measures and phonological word length, and these were entered into the model along with speech rate as fixed predictors. A set of eight critical predictors form the model space for the multimodel inference, see Table 4: no other variable passed the pre-screening requirements.

Construction probability exerts a statistically significant influence on the acoustic duration of the initial word of the theme object [$\bar{\beta} = -0.431$; $\bar{\sigma} = 0.186$; $p = 0.027$]. Again, themes that are more likely to occur in the theme + recipient (NP PP) order are pronounced shorter: the durational contrast between the most and the least probable constructions is estimated at 61 ms. Similarly to Study 1, however, most of the effect is driven by the constructions in the lower half of the probability range (range: 0.008–0.9510), as the construction probability in the upper half is virtually at the ceiling (probability range: 0.9512–0.9999). Construction probability has a moderate weight of support given the data and the candidate models: its cumulative probability is 0.64. This implies that one or more well-performing models in the model set does not include construction probability as a predictor. In fact the best-performing model does not include any of the critical predictors and is composed of the intercept, PC1 and SpeechRate. Yet construction probability shows the greatest weight of support given the data and the candidate models, as compared to other critical predictors (ranging from 0.10 to 0.46). It is also the only one to pass the inferential significance threshold of 0.05.

In sum, Studies 1 and 2 both reveal the negative correlation of construction probability with the acoustic duration of the word at the point where the choice between dative alternants is made. The more difficult (longer) production of the less probable alternant is also in line with the increased rate of disfluencies and gestures that production of datives revealed in Wagner Cook et al. (2009) and (for disfluencies) in Tily et al. (2009). Likewise, our findings are compatible with Gahl and Garnsey's (2004) report that direct objects of the verbs were pronounced shorter if they were constituents in the syntactic construction of the type biased by the verb. Our results confirm the predictions of UID that identifies the choice point as the locus for the speaker's (rational) behavior aimed at maximizing the efficiency of speech communication. Study 1 additionally suggests the simultaneous operation of (i) whatever mechanisms produce the effect of construction probability and (ii) the mechanisms responsible for the concurrent effect of the complexity of the upcoming theme.

After the choice point

As shown in the 'Theoretical predictions' discussion, UID makes distinctive predictions for role of construction probability (or information density) in the post-choice syntactic positions. Studies 3–5 probe words in syntactic positions past the choice point for dative alternation, and test whether or not their production is affected by the construction probability.

Study 3: Preposition to in the NP PP construction

This study follows up on Tily et al.'s (2009) examination of the acoustic duration of the preposition in the NP PP construction as a function of construction probability. We expand on their study by additionally examining the effects of components of construction probability. After trimming, our data set consisted of 422 data points: this number differs from the 446 cases in Tily et al.'s data set, as we applied an additional trimming criterion of removing constructions with overly long recipients (>6 words) and themes (>15 words). Acoustic durations of instances of *to* range from 20 to 400 ms (mean = 108; SD = 73). Since this study focuses on a single word, most predictors that tap into lexical variability are immaterial here. We entered into our models speech rate and two principal components that account for over 90% of variance in relevant frequency-based measures as fixed predictors, i.e., those present in each model. A set of nine critical predictors form the model space for multimodel inference, see Table 5: no other variable, including those coding the phonological environment, passed the pre-screening requirements.

Neither construction probability nor its components reach statistical significance as predictors of the acoustic duration of the preposition. The observed null effect of construction probability [$p = 0.167$] runs counter to the earlier finding of Tily et al. (2009) obtained from the same data set that is used here. In what follows, we review the discrepancies in the data trimming and analysis to identify our failure to replicate an earlier result.

Tily et al. (2009) used (a) more liberal criteria for data trimming than the present work (see above), (b) a slightly different set of other predictors (e.g., no principal components for frequency-based measures) than in the present study; (c) linear regression models rather than the linear mixed-effects model with a random effects structure, and (d) raw acoustic durations as the dependent variables, and not the log-transformed durations as in the present work. Discrepancies (a), (b) and (c) proved to be inconsequential. When fitted to raw acoustic duration of preposition in our data set of 422 data points, both linear regression models and linear mixed-effects regression models showed significant effects of construction probability that were just above the 0.05 threshold of significance. The marginally significant effects held true whether the Tily et al.'s set of predictors is used or the one that we identified as relevant [e.g., Tily et al.'s set of predictors, 422 data points, linear model: $\hat{\beta} = -0.026$; $SE = 0.014$; $p = 0.059$; mixed-effects model: $\hat{\beta} = -0.026$; $SE = 0.014$; $p = 0.059$, where p-values were obtained using the `pvals.fnc` function with 10,000 simulations]. Our reanalysis with the same array of predictors as in Tily et al. estimated the contrast in the acoustic duration of *to* between the most and the least probable NP PP construction at 25 ms, which is virtually identical to the 20 ms-constrast in Tily et al. (2009).

The important discrepancy between the present work and Tily et al. ensues from our use of log-transformed acoustic durations and not the raw ones as in Tily et al. The logarithmic transformation of duration renders the effect of construction probability not significant ($p > 0.2$). We also fitted a model that had both acoustic duration and probability log-transformed: the effect of log probability is not reliable either ($p > 0.3$). We remind the reader that the logarithmic transformation is indicated by the Box–Cox test as a power transformation that would attenuate skewness in the distribution of acoustic durations and would render the distribution closer to normality, as required for the improved accuracy of parameters in regression models (see 'Data trimming' above). The quantile–quantile plot (not shown) demonstrates that the distribution of log-transformed durations is indeed closer to normal than the distribution of raw durations.

We conclude that the effect of construction probability on the acoustic duration of the preposition reported in Tily et al. (2009) may have been due to the disproportionate influence of extreme values in the skewed distribution of raw durations, and is not replicated because the influence of those values was attenuated in our study with the help of the logarithmic transformation.

Study 4: Second object (theme) in the NP NP dative

This study considered the second (theme) object in the NP NP constructions (*the book* in *I gave Tom the book.*). Since themes vary widely in their length as the second object in the NP NP dative, we opted for considering only the initial word of the theme. The set was restricted to 12 monosyllabic function words (pronouns and determiners) with a minumum of 10 occurrences: *a, an, any, more, one, some, that, the, this, three, two*, and *your* yielding a total of 889 observations. Acoustic durations range from 20 to 640 ms (mean = 119; SD = 111). Two principal components (PC1 and PC2) account for 91% of variance in the frequency-based independent measures and phonological word length, and were entered into the model along with speech rate, and verb position in the utterance. A set of seven critical predictors formed the model space for the multimodel inference, see Table 6: no other variable passed the pre-screening requirements.

Construction probability showed an effect in the expected negative direction, but failed to reach statistical significance in the multimodel inference procedure [$\bar{\beta} = -0.2048$; $\bar{\sigma} = 0.1752$; $p = 0.201$]. This runs counter to the findings in Tily et al. (2009): for a possible source of discrepancies with the present work see our discussion in Study 3. No other critical predictor reached statistical significance either.

Study 5: Second object (recipient) in the NP PP dative

In this study we considered the first words of the recipient object in the NP PP construction (*Tom* in *I gave the book to Tom.*). We selected a subset of 10 pronouns and determiners that had at least 10 occurrences in the dataset: *a, her, him, it, me, my, the, them, us, you*: 286 observations in total. Acoustic durations range from 20 to 420 ms (mean = 137; SD = 87). One principal component (PC1) accounts for 92% of variance in the frequency-based independent measures and phonological word length, and was entered into the model along with speech rate, and a set of seven critical predictors, see Table 7: no other variable passed the pre-screening requirements.

Construction probability did not register a significant effect on the acoustic duration of the recipient's first word [$p = 0.187$]. The multimodel analysis reveals, however, a reliable effect of the length of theme [$\bar{\beta} = 0.085$; $\bar{\sigma} = 0.040$; $p = 0.042$]. This variable also surfaced as the most likely candidate, out of critical variables, for being included in the best-approximating model (its relative importance is 0.72). If the recipient is preceded by a longer theme, its acoustic duration is longer: the estimated durational contrast between the longest (15 words) and the shortest (1 word) theme is 41 ms. Importantly, since themes precede recipients in NP PP constructions, the effect in question is the lagging effect of the already-produced linguistic material on the current production. Lagging effects are well established in comprehension but have been only recently described in production studies of syntactic alternation (see the effect of lemma frequency of the matrix verb on complementizer *that*-mentioning in Roland et al., 2006; Jaeger, 2010). One common explanation for such effects is the notion of "spillover". Speakers are assumed to have a limited capacity of cognitive resources recruited in speech production. If this capacity overflows for speech unit N, its processing difficulty may spill over to the following unit $N + 1$ and make its production more effortful. Thus, the relatively long theme of the NP PP dative may lead to processing overload during the production of the theme and a deficient planning of the upcoming recipient. The deficit in planning and/or the processing overload accumulated during the theme production may spill over to the production of the recipient and make it more effortful (longer).

To summarize Studies 3–5, we have not confirmed the effects of construction probability on the acoustic duration of any word past the point at which the speaker commits to articulating a dative alternant (i.e., the first object of the dative verb). Taken together with Studies 1 and 2, our data have shown that the probability of dative construction choice is reliably related to word durations in incremental spontaneous speech. Moreover, the effects are narrowly localized at the postverbal 'choice point' in the way theoretically predicted by UID, in contrast to other theories which are currently compatible with a broad constructional scope for syntactic probabilistic effects on phonetic variation.

The role of probability: further predictions

While our results shed light on the time-course of probabilistic effects, questions remain about the explanatory power of the probability of syntactic choice between semantically equivalent alternants. The probability of the syntactic outcome is a complex function of multiple factors, mostly the ones indexing relative *accessibility* (see below) of syntactic constituents whose order is under choice. Logistic regression models of the probabilities of syntactic alternations reach high accuracy in estimating those probabilities as a mathematical function of the weighted sum of accessibility factors (Bresnan et al., 2007; Gries, 2003; Hinrichs & Szmrecsanyi, 2007; Jaeger, 2010; Roland et al., 2008): e.g., 94% in Bresnan et al.'s (2007) study on English datives; 84% in Gries (2003) study of English particle placement; and 86% (out of the maximum of 90%) in Roland et al.'s (2006) study of the English direct object/sentential complementizer ambiguity. (The lower bounds of accuracy for these three studies were 78%, 52% and 73% respectively.)

Could it be these components of the probability models—the accessibility factors—that actually carry the behavioral effects seen in the data? If so, construction probability would be a mere mediator variable masking the direct causal links between accessibility of syntactic constituents and the speech production behavior.

Alternatively, perhaps the accessibility of representations of alternating syntactic constituents can only affect production behavior to the extent that accessibility affects the probability (information content) of those syntactic constituents (cf. Levy (2008) for a similar argument in comprehension behavior). No direct causal link between accessibility factors and the ease of production is expected under this account.

Finally, might both construction probability and one or more accessibility factors make independent contributions to the observed variability in behavior? This finding would be in line with the accruing evidence that speech production can be simultaneously affected both by the probability of the produced speech unit and the effects of accessibility of upcoming units: for demonstrations at the morpheme and word level see Kuperman, Pluymaekers, Ernestus, and Baayen (2007) and Jaeger and Post (2010), at the syntactic level see Gahl and Garnsey (2004) as well as the "continuous articulatory planning" proposal of Pluymaekers et al. (2005).

In order to address these questions we must first consider the theoretical predictions of accessibility-based theories for the locus of word duration effects in the production of alternative dative constructions. In this section we draw out these predictions and present two additional word duration studies that we performed on data points *before* the choice points in the dative constructions to further test the predictions. In the General discussion the implications of all seven studies for the three questions above are then considered together.

Accessibility: Experimental and corpus-based work has established a significant influence of the conceptual and informational accessibility of arguments on the choice between alternating constructions. For instance, the choice between an overt and a covert complementizer (*this is the friend (that) I told you about*) has been argued to depend on the accessibility (availability) of the subject (pronoun *I*) under planning: a less available constituent is more likely to elicit the overt complementizer *that* (cf. Jaeger, 2010; Roland et al., 2006; Torres Cacoullos & Walker, 2009). Similarly, the choice of active or passive constructions (*the girl hit the ball, the ball was hit by the girl*) is a function of such indices of accessibility as animacy of syntactic constituents (Bock et al., 1992) and the prior mention of either one of constituents or the active/passive construction in recent discourse (Bock, 1982, 1986; Bock & Griffin, 2000; Bock & Irwin, 1980; Bock & Loebell, 1990; Prat-Sala & Branigan, 2000). In dative constructions too, the probabilities of syntactic alternants are to a large extent determined by accessibility of syntactic constituents in these alternants (Bresnan et al., 2007; Gries, 2003; Roland, Dick, & Elman, 2007). As previously mentioned, a more accessible (shorter, animate, pronominal, discourse-given, etc.) constituent is more likely to occur *before* a less accessible one (Bresnan et al., 2007).

To account for the wealth of accessibility effects, Ferreira and Dell, 2000, p. 289 propose the Principle of Immediate Mention: driven by the pressure of efficiently producing fluent speech, speakers tend to choose syntactic structures that "permit quickly selected lemmas to be mentioned as soon as possible" and buy the time to retrieve the less available material. Depending on the circumstances of speech production, the spoken structure may then be the one that is generally less accessible and dispreferred: in this case an inflated duration of the preceding spoken unit is expected. This principle accounts for both types of phenomena mentioned above: the preferred early placement of accessible constituents in passive, dative and other alternations, as well as the lower likelihood of overt complementizers.

Importantly, in the situation of syntactic choice, availability effects are expected to take place when syntactic constituents (such as theme and recipient objects in the dative construction) are assessed for their relative accessibility and their alignment is planned for subsequent production. In other words, the Principle of Immediate Mention predicts the effects of constituent accessibility to *precede* the production of either syntactic constituent and to occur prior to the choice point, for example at the dative verb (*give* in *give Tom the book*). See Table 8C for the summary of predictions. (Table 8A and B includes predictions of the approaches outlined above in Table 1A and B.) This early temporal locus makes the availability-based account of speech production readily distinguishable from information-theoretical approaches which place their respective loci further downstream. As discussed in 'Theoretical predictions' above, the UID predicts the probability of the choice to affect production only at the choice point, while the broad-scope probabilistic predictions are for potential effects of probability on all words and phrases that instantiate the constructions of interest, which would include their head verbs (see Gahl & Garnsey, 2004, p. 754; Hay & Bresnan, 2006).

The extensive literature on accessibility effects on the phonological, morphological and lexical levels suggests that availability of upcoming syntactic arguments will influence production even in the absence of syntactic choice, that is past the choice point. We may expect then accessibility of the distant object (the theme in the NP NP dative or the recipient in the NP PP dative) to affect production of the proximate object (the recipient and the theme, respectively). Specifically, less accessible upcoming objects are predicted to correlate with an increased difficulty (in our study, an inflated acoustic duration) of current production (see however Wagner Cook et al., 2009).

Competition: The competition theory of speech production builds on availability; it proposes the mechanism of competition between alternants as a determinant of the production effort at the choice point (Haskell & MacDonald, 2003; Race & MacDonald, 2003; Solomon & Pearlmutter, 2004), see Table 8D for the summary of predictions. The claim is that the alternant chosen for production competes with the rejected alternant, and their competition is the stronger, the more active the rejected alternant is. Activation of alternants is argued to be a function of their accessibility. Thus, the maximum of competition is

Table 8
The timeline of effects of constituent availability and the probability of syntactic choice on the ease of incremental production at different syntactic positions in the dative construction (verb, proximate object, distant object), as predicted by the UID, theories that make broad-scope predictions of construction-probability effects, the availability-based theory, and the competition theory of speech production.

	Verb	Proximate object	(Preposition +) distant object
	Before the choice point	*At the choice point*	*After the choice point*
A. UID	None	Probability of the spoken alternant	None
B. Broad-scope	Probability of the spoken alternant	Probability of the spoken alternant	Probability of the spoken alternant
C. Availability	Accessibility of the proximate object, or accessibility of both objects	Accessibility of distant object (difficult production if less accessible)	None
D. Competition	None	Accessibility of distant object (difficult production if supports the alternative)	Accessibility of proximate object (difficult production if supports the alternative)

expected if a generally dispreferred alternant is realized in speech while its highly active generally preferred counterpart is rejected. Cast in information-theoretical terms, the competition account predicts more effortful production if a less probable (more informative) alternant is chosen, and a more probable one rejected. The competition account also places the temporal locus of the effect at the choice point. Thus in the narrow sense of the time-course of probabilistic effects, the competition account is indistinguishable from UID with respect to its predictions before and at the choice point (for a similar conclusion see Wagner Cook et al., 2009), and we hypothesize that the conclusions one could draw regarding the validity of UID given the present data would equally hold for the competition account.

The competition account does differ from the UID in predicting effects after the choice point because the rejected alternant may interfere with the spoken one, and the interference is the stronger, the more active the rejected alternant is. The competition might continue even after the alternant is chosen and is well under production, if the rejected alternant maintains partial activation. A longer theme is known to favor the recipient-theme (NP NP) order of the constituents (Bresnan et al., 2007), which on competition account adds to its activation. Thus, an increased length of theme is expected to translate into a more effortful production of the recipient in NP PP datives, as was indeed the case in Study 5, although spillover remains a possible alternative explanation. Distinguishing these accounts will require further research.

Method

For these additional studies of verb duration before the choice point the materials, dependent variables, critical predictors, control variables, data trimming, and statistical modeling were the same as in the preceding studies. The Box–Cox optimal values of λ for transforming the raw durational data were respectively 0.0, 0.2—again close to zero. We therefore log-transformed the acoustic durations of the verbs as we did the other word durations in the previous studies. The sizes of the data sets resulting from our trimming procedure are reported in Table 9.

Results and discussion

As specified in Table 8, only the availability-based account of speech production predicts effects of either accessibility or probability (as an accessibility index) before the choice point and none after the choice point. Neither the competition-based account nor the UID predicts the early simultaneous access to objects, as neither implements a production mechanism that would necessitate the speaker's consulting the syntactic objects before either of the objects is articulated. Studies 6 and 7 test these hypotheses against acoustic characteristics of the ditransitive verb, which lies in the syntactic position that precedes the choice point between alternative orders of dative arguments.

Study 6: Duration of the verb followed by the NP NP dative

The data set of verbs followed by the NP NP alternant (**give** *Tom the book*) includes 1490 data points. Verb duration ranges from 80 to 620 ms (mean = 222, sd = 79). Three principal components account for 92% of the variance in the frequency-based independent measures and phonological word length, and were entered into the model along with speech rate as fixed predictors (the ones that were present in every model). There is also a set of 8 critical

Table 9
Summary of the effects of construction probability and its components on acoustic durations of ditransitive verbs. Sample sizes *N* are reported after the trimming procedures were applied. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$), and *t*-test based *p*-values are reported for those predictors that reach significance at the 0.05 level. ES stands for effect size, a model-estimated contrast in acoustic durations of words showing the maximum and the minumum values of the predictor.

NP NP (double)	NP PP (prepositional)
Before the choice point	*Before the choice point*
Study 6, verb: $N = 1490$	Study 7, verb: $N = 367$
recipient = NotGiven ($\bar{\beta} = 0.075$; $\bar{\sigma} = 0.032$; $p = .026$; ES = 21 ms)	Probability ($\bar{\beta} = -0.184$; $\bar{\sigma} = 0.061$; $p = .004$; ES = −64 ms)
Length.theme ($\bar{\beta} = 0.016$; $\bar{\sigma} = 0.007$; $p = .029$; ES = 18 ms)	

Table 10
Outcome of the multimodel inference procedure for the duration of verb followed by the NP NP dative, $N = 1490$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's *p*-value, and its cumulative probability. Marked in bold are predictors with *p*-values below 0.05.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	*p*	Cumul.Prob
Intercept	5.686	0.062	5.564	5.808	<.001	NA
SpeechRate	−0.014	0.002	−0.017	−0.010	<.001	NA
PC1	0.043	0.005	0.033	0.054	<.001	NA
PC2	−0.043	0.006	−0.055	−0.030	<.001	NA
PC3	−0.134	0.018	−0.170	−0.098	<.001	NA
given.rec = ngiven	**0.075**	**0.032**	**0.013**	**0.137**	**.026**	**1**
given.theme = ngiven	0.002	0.021	−0.040	0.044	1	0.18
pron.rec = pron	0.046	0.041	−0.033	0.126	.213	0.39
pron.theme = pron	0.023	0.024	−0.025	0.071	.219	0.30
length.rec	0.005	0.011	−0.016	0.027	.360	0.26
length.theme	**0.016**	**0.007**	**0.003**	**0.029**	**.029**	**0.98**
Prob	−0.042	0.060	−0.159	0.076	.31	0.26

predictors which forms the model space over which multimodel inference is estimated, see Table 10. No other variable passed the pre-screening requirements.

Construction probability does not reach significance as a predictor of the verb's acoustic duration [$\bar{\beta} = -0.042$, $\bar{\sigma} = 0.06$, $p = 0.26$]. Yet several indices of accessibility of constituents of dative show effects in the expected direction: lower accessibility correlated with longer production durations. First, the dative verb had a longer acoustic realization by approximately 21 ms, if the recipient of the verb is not-given and thus less accessible [$\bar{\beta} = 0.075,^{-}$ $\sigma = 0.032$, $p = 0.026$]. Second, the verb is pronounced longer if it is followed by a longer theme [$\bar{\beta} = 0.016$, $\bar{\sigma} = 0.007$, $p = 0.029$]: the contrast in verb duration between the longest and shorter theme is estimated at 18 ms. Again, heavier constitutents are typically construed as less accessible, and thus are predicted to incur higher costs of planning. Cumulative probabilities provided by the multimodel inference procedure indicate that both the givenness of recipient and the length of the theme had a very strong support from the model space (1 and 0.98, respectively) and had higher likelihoods to be found in the best-approximating model than other candidate variables.

Study 7: Duration of the verb followed by the NP PP dative

The dataset includes 361 verbs followed by the NP PP dative (**give** *the book to Tom*). Verb duration ranges from 90 to 600 ms (mean = 232; SD = 81). Three principal components account for 93% of the variance in the frequency-based independent measures and phonological word length, and they were entered into the model along with speech rate as fixed predictors (those present in each model). A set of eight critical predictors forms the model space for multimodel inference, see Table 11: no other variable passed the pre-screening requirements.

Multimodel inference reveals a sizable effect of construction probability as a critical predictor [$\bar{\beta} = -0.184$, $\bar{\sigma} = 0.061$, $p = 0.004$]. If the NP PP ordering of constituents is more probable, the verbs are acoustically shorter, suggesting that the planning of dative constructions at the verb is facilitated by the relatively probable ordering of upcoming constituents. The contrast in durations of the verbs followed by the most and the least probable NP PP constructions is estimated at 64 ms, though given the prevalence of high-probability constructions in the data set (mean probability = 0.82), the effect is in fact more subtle: for example, the estimated contrast between the 25%

Table 11
Outcome of the multimodel inference procedure for the duration of verb followed by the NP PP dative, $N = 361$. Model-averaged estimates of regression coefficients ($\bar{\beta}$) and standard errors ($\bar{\sigma}$) are reported, as well as the lower and upper boundaries of the 95% confidence intervals (LoCI and HiCI), the predictor's *p*-value, and its cumulative probability. Marked in bold are predictors with *p*-values below 0.05.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	*p*	Cumul.Prob
Intercept	5.959	0.104	5.755	6.163	<.001	NA
SpeechRate	−0.017	0.004	−0.025	−0.009	<.001	NA
VerbPos	−0.001	0.001	−0.003	0.001	.24	NA
PC1	−0.047	0.008	−0.063	−0.031	<.001	NA
PC2	−0.015	0.009	−0.033	0.003	.39	NA
PC3	0.093	0.022	0.050	0.137	<.001	NA
given.rec = ngiven	−0.020	0.034	−0.088	0.047	.33	0.42
given.theme = ngiven	0.017	0.038	−0.057	0.091	.36	0.33
def.theme = indef	−0.018	0.038	−0.092	0.057	.35	0.34
pron.rec = pron	−0.032	0.040	−0.110	0.046	.29	0.55
pron.theme = pron	−0.004	0.019	−0.041	0.034	.39	0.20
length.rec	0.003	0.010	−0.016	0.023	.38	0.26
length.theme	−0.001	0.007	−0.015	0.014	.39	0.16
Prob	**−0.184**	**0.061**	**−0.304**	**−0.064**	**.004**	**1**

and 75% percentiles is only 15 ms. Cumulative probability of the construction probability predictor (0.92) clearly identifies it as an excellent candidate for the best-approximating model given the data and the candidate models. No other critical predictor reached significance.

Albeit subtle, the effect of construction probability on the verb shows that availability of higher-level syntactic information can affect words even before the choice point. This falls in line with Gahl and Garnsey's (2004) report of early construction probability effects on preceding constituents in read speech, as well as the effects of word-level distributional information on preceding words in Jaeger and Post (2010). We interpret this early probabilistic effect as an availability effect consistent with the Principle of Immediate Mention (Ferreira & Dell, 2000).

General discussion

The present work examines how syntactic construction probabilities influence word durations during the course of spontaneous sentence production, asking two questions: (i) whether the effects are localized in incremental sentence production in a way consistent with theoretical predictions and (ii) whether the higher-level probabilities themselves make an independent contribution rather than merely serve as a summary measure of the individual factors that influence construction outcome. Our seven studies provide affirmative answers to these questions.

Predicted loci of construction probability effects

We observe reliable effects of the probability of syntactic choice on acoustic durations in three out of seven syntactic positions that we investigated; see Tables 2 and 9. These are the verb preceding the realization of the NP PP dative (Study 7; position before the choice point), the initial word of the first (recipient) object in the NP NP dative (Study 1, the choice point position), and the initial word of the first (theme) object in the NP PP dative (Study 2, the choice point position). Throughout the studies, a lower probability of the selected alternant correlated with its longer realization in speech. No position past the choice point (Studies 3–5) revealed a statistically significant (at the 0.05 level) effect of construction probability, nor did the verb preceding the NP NP dative alternant (Study 6).

Consider first the data obtained at the choice points (Studies 1 and 2) and the syntactic positions downstream in incremental dative production (Studies 3–5). Predicted effects of construction probability occur at the choice points in both alternants, and not in subsequent production after the choice points. These findings lend strong support to UID, with its emphasis on the choice point as the temporal locus of probabilistic effects in the situation of (syntactic or other) choice. The post-choice null effects in Studies 3–5 are not consistent with theories that consider words and phrases in the entire spoken unit as the potential scope of the influence of that unit's probability. These null effects might be due to relatively small sizes of our data sets (422, 889 and 286 data points in Studies 3–5, respectively) and insufficient statistical power. We note however that probabilistic effects are strong enough to be estimated as statistically reliable by models with a similar number of predictors fitted to even smaller datasets (235 and 367 data points in Studies 2 and 7, respectively). A more plausible interpretation is that construction probability has no role to play in incremental speech production after the speaker has chosen to align the semantic role of recipient or theme with the first object of the ditransitive verb and thus the other role with the second object. In other words, by the time the second object or (where applicable) the preposition is realized, the probability of the construction being ditransitive or prepositional is exactly 1 or exactly 0. This lack of constructional variability does not give grounds to expect dative-construction probability to drive differences in production behavior after the choice point, true to fact. We note that our argument runs counter to earlier information-theoretical investigations of the dative alternation (Tily et al., 2009; see also Wagner Cook et al., 2009), which reported evidence supporting a broader scope of probabilistic effects: see Study 3 for a detailed discussion. We conclude that, as far as the production of alternating multi-word constructions is concerned, the position of UID appears justified: at the choice points (but not after) speakers are affected by the relative probability (amount of information or information density) of continuations that are compatible with the intended meaning (Jaeger, 2010).

The reliable effect of construction probability on the verb followed by the NP PP dative in Study 7 suggests that it is not only the production, but also the *planning* of the upcoming syntactic choice, that can be made more effortful if the choice is made in favor of a less probable (more informative) alternant. While unexpected under the information-smoothing account that we consider here (UID), the effect is compatible with the availability-based account of speech production and particularly the Principle of Immediate Mention. (It is also compatible with the broad-scope predictions, but these are incompatible with our previous findings in Studies 3–5.) When planning a syntactic choice, the principle predicts that a more accessible (more probable) alternant will be mentioned first. If a less accessible alternant is planned for production, the planning time will increase leading to an inflated production time of the syntactic constituent preceding the realization of the alternant, namely, the dative verb, true to fact. Our finding that planning can be affected by the probability of upcoming choice corroborates an earlier report based on the read-speech data set. In Gahl and Garnsey (2004), verbs followed by a higher-probability subcategorization (either a direct object or a sentential complement) showed a higher rate of the word-final *t*/*d* deletion as well as a shorter acoustic duration. At the lexical level, the availability of the upcoming word (defined as its frequency and probability in the context of the preceding bigram) has also been shown to influence the acoustic duration of word N (Jaeger & Post, 2010). At the morphological level, the entropy of the morphological family of the upcoming morpheme influenced the acoustic duration of the interfix in Dutch compounds (e.g., *-s-* in *herdershond* "shepherd's dog", Kuperman et al., 2007).

While there is support in the literature for the availability-related interpretation of probabilistic effects, it is only partially confirmed here: unlike the prepositional constructions, the verb followed by the ditransitive construction (Study 6) was not affected by the construction probability. One explanation for the null effect in Study 1 and the reliable effect in Study 2 is such that the variability in construction probability is extremely small in the NP NP datives, with only 25% of the data occupying the probability range from 0.008 to 0.972. The range is more uniformly populated in prepositional datives, with 50% of the sentences having a construction probability below 0.937 (for detailed distributional information see *Critical predictors*). It is possible then that the at-ceiling probabilities in the ditransitive datives, unlike the prepositional datives, did not register an effect on acoustic durations of verbs that would be sufficiently strong to reach the level of statistical significance. The conditions under which probability-driven accessibility of upcoming constituents affects planning requires further research. Importantly, however, there is a theoretical framework accommodating such effects, the "involvement-in-planning" account advocated in Pluymaekers et al. (2005). Pluymaekers et al. claim first, that articulatory planning is not based on units (segments, syllables, or words) but rather is continuous, spanning, in our case, the boundaries of syntactic arguments (verbs and object noun phrases). Second, speech production of a word is simultaneously influenced by pressures of the current articulation and also affected by the demands of planning the upcoming production. Our data support this argument by showing several cases of simultaneous effects of planned *and* produced syntactic constituents. These include the simultaneous effects of the length of upcoming theme and the probability of concurrent syntactic choice observed at the initial word of the recipient in the ditransitive construction (Study 1), as well as effects of upcoming constituents (length of theme, givenness of recipient) and the concurrent effects of word and bigram frequency and speech rate observed in the production of the verb followed by the ditransitive construction (Study 6). Thus, the availability-based account of Pluymaekers et al. (2005), which originally tackled word-based distributional properties such as word frequency and predictability, gains support from our study of information transmission in multi-word syntactic phrases.

To sum up, the present body of evidence reveals that a full account of speech production as it unfolds in time requires simultaneous, complementary operation of the mechanisms underlying theories that incorporate construction probability—either information-theoretic or implemented as activation (represented respectively by the UID and competition theories)—and construction availability (as represented by the Principle of Immediate Mention)[2].

The role of construction probability

As Tables 2 and 9 show, two variables are pervasive in predicting the speaker's behavior. One is the probability of the syntactic alternant, which correlates negatively with the word durations in Studies 2, 3, and 7. The other predictor is the length of the theme, which correlates positively with word durations in Studies 1, 5, and 6. We also observed in Study 6 a shorter acoustic duration of the verb followed by the recipients of the ditransitive dative that were given (mentioned in prior discourse) as compared to non-given ones.

The distribution of the effects over the time-course of production of both ditransitive and prepositional datives does not support the view that accessibility effects work only through probability. Specifically, this view is disconfirmed by the fact that (a) accessibility indices of both the proximate recipient (givenness) and the distant theme (length) affect the production of the ditransitive verb (Study 6) in the absence of the effect of construction probability and (b) the effects of construction probability and the length of the upcoming theme are detected simultaneously as predictors of the recipient's acoustic duration in Study 3.

Nor does the evidence support the view that probability is a summary measure of effects carried by the underlying accessibility factors used to compute it. The effects of construction probability on the acoustic duration of the verb in Study 7 and that of the proximate theme object in Study 2 appear in the absence of reliable effects stemming from components of probability. The observed patterns are most compatible with the view that accessibility of syntactic constituents affects production behavior both directly and indirectly, that is, by co-determining the probability of one or the other alignment of the constituents.

In conclusion, the present set of studies reports empirical evidence confirming the linkage between high-level probabilities of syntactic constructions and continuous variation at the phonetic level of speech production. It also points to a narrow temporal locus over which the probabilistic effects of construction choice operate. Over and above the substantive implications of these findings for current theories of speech production already discussed, we believe that our strategy of analyzing word-by-word effects as spontaneous sentence production unfolds has value for future research.

Acknowledgments

This material is based upon work supported by the Rubicon grant of the Netherlands Organization for Scientific Research (NWO) to the first author and by the National Science Foundation under Grants No. BCS-1025602 and IIS-0624345 to Stanford University (PI the second author). We thank Florian Jaeger for extensive discussions of earlier drafts of this paper, three anonymous reviewers for insightful comments, and Gabriel Recchia for his help with data collection. This work has also benefited from presentation and discussion at the workshop

[2] We assume here the well-known interpretation of connectionist model activation as the *probability* of a spike in action potential in neural networks. An accessibility-based connectionist implementation of construction competition therefore converts the accessibility factors that influence construction choice into a nonlinear function of the probabilities of these factors.

'The Development of Syntactic Alternations' (Stanford, 2010) and the workshop 'Probabilistic syntax' (Freiburg, 2010), supported by a grant from the Freiburg Institute of Advanced Studies for the project "Predicting Syntax in Space and Time" (to the second author with Benedikt Szmrecsányi).

Appendix

Multimodel inference

The multiple working hypotheses that we pursue in this paper require statistical inference for the effects of numerous, potentially collinear, independent variables (Chamberlin [1890] 1995). Typically, such inferences are made on the basis of a single model identified as the best among candidate models. The criteria for selecting the single best model out of the model space are often defined as a maximum goodness-of-fit value: the amount of explained variance, the values of the Akaike or Bayesian Information Criterion, the residual sum of squares or others. Alternatively, the single best model is identified as a constellation of variables whose goodness-of-fit to the data would be detrimentally affected if (i) any of the variables it contains is removed or (ii) any variable is added to the set of variables that it contains, or a combination of (i) and (ii). Available best-model selection algorithms vary in whether they exhaustively search through the entire model set (the power set of 2^n models that represent all possible combinations of n predictors) or proceed along a certain path through the model set where each step weeds out models that underperform in terms of the selection criterion. Whatever algorithm is chosen, its end-result is one model (and one constellation of variables) picked out of the model set and selected as the basis for statistical inferences regarding the variables represented in or left out of this model.

As argued eloquently by, among others, Burnham and Anderson (2004, p. 261), the standard practice of the single best model selection is open to criticism on several counts (for similar criticism in the psycholinguistic literature, see Keuleers & Daelemans, 2007):

> For a model selection context, we assume that there are data and a set of models and that statistical inference is to be model based. Classically, it is assumed that there is a single correct (or even true) or, at least, best model, and that model suffices as the sole model for making inferences from the data. Although the identity (and parameter values) of that model is unknown, it seems to be assumed that it can be estimated—in fact, well estimated. Therefore, classical inference often involves a data-based search, over the model set, for (i.e., selection of) that single correct model (but with estimated parameters). Then inference is based on the fitted selected model as if it were the only model considered. Model selection uncertainty is ignored. This is considered justified because, after all, the single best model has been found. However, many selection methods used (e.g., classical stepwise selection) are not even based on an explicit criterion of what is a best model.

For reasons given in our Methods (Statistical Modeling) section, in this work we depart from the practice of single best model selection and employ instead the method of multimodel inference, as developed by Burnham and Anderson (2002, 2004). This method estimates the strength of evidence for each model in the model set as the amount of information lost when that model is used to approximate full reality or truth, or equivalently, the distance between the model and full reality (Burnham & Anderson, 2004): the required metric is provided by an information-theoretic measure, the Kullback–Leibler (KL) distance (Kullback & Leibler, 1951). As shown in Akaike (1973, 1974), for the case when model parameters have to be estimated rather than are known, the expected value of the model's KL distance is inversely proportional to the model's Akaike Information Criterion (AIC). AIC values are routinely reported in the outputs of regression models and are defined as a function of the maximized log-likelihood of the model given the data ($\log L$) and the number of the model's estimable parameters (K): $\text{AIC} = -2\log L + 2K$. The difference Δ between the AIC values of any two models A and B is the index of how much information is lost if model A is used to approximate the data rather than B. The difference from the model with the minimum value of AIC_{min} is construed as the strength of evidence in favor of the given model, given the data and the n variables that generate the model set.

The motivation and mathematical foundations of this multimodel inference approach, as well as the criticism of the methods of single best model selection are exhaustively presented in Burnham and Anderson (2002, 2004) and Lukacs et al. (2010), see Anderson (2008) and Conroy (2006) for a worked example of multimodel inference. In what follows, we confine ourselves to outlining the steps of our multimodel inference procedure closely following Burnham and Anderson's work, as well as Conroy (2006).

Multimodel inference procedure

The model set for n critical variables consists of 2^n models fitted to a dependent variable. This power set of models contains all possible combinations of critical variables, including as extremes a model with none of these variables and a model with all n variables as predictors. Each critical variable is thus present in 50% of the models. A set of control variables can be entered into each model in the respective model set: these are dubbed "fixed" variables, and appear, together with the intercept, in all models in the power set. Variables of interest are identified on theoretical grounds and—in our case—with data sparsity in mind (see section "Data trimming" above). Multimodel inference applies to essentially any kind of regression model. Our models were linear mixed-effects models with speaker and word as random effects (except for Study 3 of preposition "to" where speaker was the only random effect), and a set of control variables that were held constant across models. (All data patterns were also confirmed in the analyses where random by-speaker slopes or contrasts were additionally defined for all fixed effects in the model.)

For each of the models in the model set, the value of the Akaike Information Criterion (AIC) is computed. Burnham and Anderson (2004) warn that AIC estimates may be biased in small samples where the ratio of the number of observations to the largest number of the model's parameters is below 40: this situation is pervasive in our studies. For such cases, a value corrected for sample size (AICc) is recommended. It is calculated as follows (Burnham & Anderson, 2002, 2002):

$$\mathrm{AICc} = -2\ln\ L + 2K + \frac{2K(K+1)}{m-K-1}, \tag{1}$$

where m is the number of observations.

AIC or AICc values for models are not independently interpretable and are only meaningful for between-model comparison. The difference Δ was computed between the model with the minimum AICc value and each model in the respective model set:

$$\Delta_i = AICc_i - AICc_{min}, \tag{2}$$

where i is the model index ranging from 1 to R, the number of models in the model set. The AIC values decrease as the model's goodness-of-fit increases, so the minimum AIC or AICc value corresponds to the best-performing model. Δ_{min} for the best-performing model is zero, and values of Δ_i are positive for all other models. The value of Δ_i is then the information-theoretic distance of model i from the best-approximating model in the set: the smaller it is, the stronger the evidence that model i is the best-approximating model.

The likelihood of model i being the best-approximating model given the data is estimated as: $L_i = e^{-\Delta_i/2}$. The likelihood of the best-performing model is 1. The ratio of the likelihoods of two models is called the *evidence ratio* and indicates the relative quality of the models given the data. The larger the evidence ratio between models A and B, the better model A approximates the data, relative to B. It is useful to norm the relative likelihood of each model by dividing it by the sum of likelihoods of the full model set:

$$w_i = \frac{e^{-\Delta_i/2}}{\sum_{r=1}^{R} e^{-\Delta_r/2}}, \tag{3}$$

where R is the total number of models in the model set. The resulting quantity w_i is dubbed the Akaike weight and is construed as the *weight of evidence* that model i is the best performing model given the data and the candidate models. The normalization step in (3) ensures that Akaike weights w add up to 1 over the model set, and can be treated as probabilities. Model A that has the Akaike weight w of 0.4 is two times more likely to be the best model than model B with the Akaike weight of 0.2: this *evidence ratio* does not depend on any models in the model set besides A and B. The model with the minimum AICc$_{min}$ value is associated with the greatest Akaike weight, yet the absolute value of its weight w depends on how well other models in the set performed, and it will be lower if, say, the difference between the best and the second best model is small. Akaike weights are thus indices of model selection uncertainty: the greater the weight of the model, the more certain it is to be the best model among the candidates.

To illustrate the procedure outlined so far, we present calculations for the set of mixed-effects models formed by three critical variables fitted to the log acoustic duration of the verb followed by the NP NP construction: the variables are construction probability (*Prob*) and standardized log lengths of recipient (*length.recipient*) and theme (*length.theme*). All models additionally contain random effects for verb and speaker, as well as the constant term for intercept. Table 12 below reports for each of the 8 (=2^3) models its ordinal number in the original model set, regression coefficents for the intercept (present in every model), and regression coefficients for the three critical variables (confined to the models in which they appear). Additionally, Table 12 shows for each model the following: k, the number of estimable parameters in the respective model, AIC, AICc, the difference between the model's AICc and the minimum AICc$_{min}$ in the model set (Δ), as well as the model's Akaike weight w. Models are sorted in the decreasing order of their Akaike weights. The best-performing model has length of recipient and length of theme as critical predictors, the lowest AICc (-15.86), and the highest Akaike weight (w = 0.58) in the set. Values in the column Δ demonstrate how distant candidate models are from the best-performing model or, equivalently, how much information is lost if the candidate model is used to approximate the data instead of the best-performing model. The rule of the thumb is that models with $\Delta \leqslant 2$ show a substantial strength of evidence (Burnham & Anderson, 2004), and it is customary not to consider candidate models with $\Delta > 4$, as there is little evidence in their favor (Barton, 2011). Akaike weights can be used to quantify the evidence ratio: the best model in Table 12 (line 1, original ordinal number 7) is 1.8 times more likely to be the best-performing model than the second best model (line 2, original ordinal number 8): 0.58/0.33 = 1.8.

Table 12
Summary of the model set generated by the multimodel inference procedure. Each model reports the estimated regression coefficients for construction probability and standardized log lengths of theme and recipient, as well as the number of estimable coefficients in the model (k), values of AIC and the corrected AICc, the difference between the model's AICc and the minimum AICc in the model set (Δ), as well as the model's Akaike weight w. Models are sorted in the decreasing order of Akaike weights.

Model #	Original model #	Intercept	Prob	Length.rec	Length.theme	k	AIC	AICc	Δ	w
1	7	5.572		0.029	0.016	6	−15.92	−15.86	0.00	0.58
2	8	5.622	−.055	0.026	0.018	7	−14.82	−14.75	1.11	0.33
3	3	5.568		0.029		5	−11.47	−11.42	4.44	0.06
4	5	5.575	−.008	0.029		6	−9.49	−9.43	6.43	0.02
5	6	5.714	−.153		0.021	6	−1.96	−1.90	13.96	0.00
6	4	5.574			0.021	5	4.59	4.63	20.49	0.00
7	2	5.668	−.107			5	6.56	6.60	22.47	0.00
8	1	5.570				4	9.05	9.08	24.94	0.00

The next step of the multimodel inference procedure evaluates the relative importance of variables of interest. For each variable, Akaike weights are summed up for the models that contain that variable: this sum is the cumulative probability p_{cuml} of the variable given the data and the set of variables. The cumulative probabilities reveal the relative importance of variables as predictors of the observed data, and allow for principled variable ranking and selection. Importantly, this method of variable selection is defined over the full model space and is not conditional on any specific model. Thus, the cumulative probabilities of variables in Table 12 computed over the full model set are as follows:

$$p_{cuml}(\text{Intercept}) = 1 \text{ (present in all models)};$$
$$p_{cuml}(\text{prob}) = 0.33 + 0.02 + 0.0 + 0.0 = 0.35;$$
$$p_{cuml}(\text{length.recipient}) = 0.58 + 0.33 + 0.06 + 0.02 = 0.99;$$
$$p_{cuml}(\text{length.theme}) = 0.58 + 0.33 + 0.00 + 0.00 = 0.91.$$

Thus, the length of the recipient is the most important critical predictor of the acoustic duration in this example and is closely followed by length of theme, while construction probability is the least important of critical predictors. Notably, these estimates of the relative importance of predictors are not dependent on any single model, but rather are computed over the entire model set, which makes these estimates more reliable especially in situations "when the second or third best model is nearly as well supported as the best model or when all models have nearly equal support" (Burnham & Anderson, 2004, p. 274).

The multimodel selection method has the further advantage of offering the *unconditional* estimates for the model parameters (regression coefficients) that are not dependent on any given model, but rather are estimated over the entire model set. One possibility for computing these would be to take the average value for each of regression coefficients across the model set. This approach, however, would ignore the hard-won knowledge of model selection uncertainty, i.e., how likely models are to be the best-approximating model given the candidates and the data. Alternatively, the model-averaging method of Burnham and Anderson suggests that parameter estimates $\hat{\beta}_{p_i}$ for predictor p in each individual model i be multiplied by the model's Akaike weight w_i, i.e., the probabilistic index of the strength of evidence in favor of the model. The products are then summed up to obtain the unconditional estimate of the predictor's regression coefficient $\bar{\beta}_p$. The literature on model-averaging discusses two options of computing $\bar{\beta}_p$. One is to compute the weighted average of the predictor's regression coefficients based on all models in the model set while assigning the value of zero to the regression coefficient in the models that do not include the respective predictor. The other one is to base the computation on those models only that include the respective predictor. (While the intercept and "fixed" control variables are defined in all models, critical variables that form the power set are present in one half of the model set each.) We chose the second option as it provides more accurate estimates for the unconditional model-averaged regression coefficients and standard errors (see below).

We illustrate the procedure by computing the unconditional mean for construction probability in our example model set; see Table 12. Regression coefficients for construction probability vary across models by a factor of 15, depending on what other variables enter those models. This variability makes a strong case for the need of an accurate estimator for the regression coefficient, one that would not be dependent on any single model, and yet would take into account the weight of evidence for the models in the set. The Akaike weights w_j for the models in the subset j that contains construction probability as a predictor (lines 2, 4, 5 and 7 in Table 12) are 0.33, 0.02, 0.00, and 0.00: their sum is 0.35. Since Akaike weights are construed as probabilities, the weights are re-parameterized by dividing each weight by their sum, such that they add up to one. The new weights w' yield: 0.33/0.35 = 0.943; 0.02/0.35 = 0.057; 0; and 0. The unconditional, model-averaged regression coefficient of construction probability, given the candidate models and the data, is: $\bar{\beta}_p = \sum_j \hat{\beta}_p * w'_j \quad = -0.055 * 0.943 - 0.008 * 0.057 - 0.153 * 0.00 - 0.107 * 0.00 = -0.052$, where p is predictor and j the subset of the model set that contains this predictor. This weighted average is heavily influenced by the regression coefficient in the model which has the greatest weight of evidence in Table 12 (line 2), while other models (lines 4, 5, and 7) make little to no contribution to the weighted average.

Model-averaged regression coefficients come with an estimate of their reliability, or the unconditional standard error ($\bar{\sigma}$). Unconditional standard errors account for two sources of variance. One is the variance gauged by an individual model and reported as the estimated standard error of the regression coefficient for the predictor ($\hat{\sigma}$): this variance is obviously conditional on the invididual model. The other source is the model selection variance, the measure of how different the individual model's estimate of a regression coefficient ($\hat{\beta}$) is from the model-averaged estimate ($\bar{\beta}$). By adding this second source of variance, the model-averaging method overcomes the drawback of typical inferential procedures that use estimates of sampling variance that are conditional on a given model and ignore the uncertainty as to whether this model is the best-approximating model for the data out of the model set. Eq. (4) defines the unconditional model-averaged standard error $\bar{\sigma}_p$ for predictor p as follows:

$$\bar{\sigma}_p = \sum_j \sqrt{\hat{\sigma}_p^2 + (\bar{\beta}_p - \hat{\beta}_p)^2}. \tag{4}$$

Table 13 reports the results of model averaging for the model set: the unconditional estimates of regression coefficients $\bar{\beta}$ and standard errors $\bar{\sigma}$, as well as the lower and upper boundaries of the 95% confidence interval, computed as $\bar{\beta} \pm 1.96 * \bar{\sigma}$, where 1.96 is a (roughly) estimated t-value for the 95% confidence interval and samples with over 20 observations.

Inferential statistics in Table 13 demonstrate that only lengths of recipient and theme, but not construction probability, reach the significance level of 0.05 (their 95% confidence intervals do not straddle the zero). In the body of

Please cite this article in press as: Kuperman, V., & Bresnan, J. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* (2012), http://dx.doi.org/10.1016/j.jml.2012.04.003

Table 13
Summary of the model-averaging procedure. For each predictor we report the unconditional estimates of its regression coefficient ($\bar{\beta}$) and standard error ($\bar{\sigma}$), the boundaries of the 95% confidence interval, the *t*-test based probability and the cumulative probability.

Predictor	$\bar{\beta}$	$\bar{\sigma}$	LoCI	HiCI	*P*	Cumul.Prob
Intercept	5.5884	0.0472	5.496	5.6808	<0.0001	NA
length.of.theme	0.0164	0.0063	0.004	0.0288	0.0142	0.91
length.of.recipient	0.0282	0.0065	0.0155	0.041	<0.0001	1.00
Prob	−0.0522	0.0589	−0.1677	0.0634	0.2696	0.36

this article we report the model-averaged estimates of regression coefficients for the seven studies that form the empirical core of this work. While we calculated and reported the estimates using our own code, our calculations are nearly identical to ones that can be obtained with the help of the function `model.avg` in the recently developed library `MuMIn` in `R` (Barton, 2011), with the method of model-averaging set in `model.avg` to "NA" (only those models that contain a predictor contribute their estimated coefficients and standard errors to the model-averaging computations for a predictor.). The minor discrepancies in model-averaged estimates between our script and Barton (2011) are due to rounding errors and to the use in Barton (2011) of the cut-off of $\Delta \leqslant 4$ as a criterion for including models into the model-averaging computation: For the sake of completeness, we did not set any such cut-offs here.

It is important to realize that, while offering increased accuracy, multimodel inference is subject to the same methodological caveats that apply to any modeling practice. Thus Burnham and Anderson (2002, 2004) and subsequent literature (e.g., Smith, Koper, Francis, & Fahrig, 2009; Freckleton, 2011) warn against mechanistic initial selection and data dredging of variables that are not substantiated by either theoretical considerations or prior empirical research. Based on extensive research into dative alternation, we provide the grounds for the selection and exclusion of our variables in the Method section. Likewise, both predictor and outcome variables with skewed distributions may need to undergo transformations that would render their distribution closer to meeting the assumptions of the modeling method (in this case, normality). Above we present the tests that identify log-transformation as the most appropriate for the distribution observed in acoustic durations. Similar tests led to the (log) transformation of all continuous predictors that had skewed distributions, i.e., object lengths. As part of model criticism, our inspection of residuals in three best-performing models in each of the seven model sets did not reveal any systematic patterns either. Another required step is an a priori examination of potential non-linear functional relationships between predictor variables and outcomes (Baayen, 2008; Harrell, 2001): there were no indications of non-linearity in our data sets.

Finally, high collinearity or non-independence of predictors is prevalent in psycholinguistic studies, and the present one is no exception. There are correlations between individual components of construction probability (e.g., pronominal objects of dative verbs tend to be short, definite, given and tend to precede their alternants). There are also obvious correlations between those components and the actual estimate of construction probability that the components contribute to Bresnan et al. (2007). The harmful consequence of having two or more collinear predictors in the model is the inflation of estimated standard errors and the concomitant inaccuracy of inferences for those predictors. Thus the relationship between the method of statistical modeling and model selection and collinearity is crucial. In what follows we demonstrate that non-independence of our predictors did not affect the accuracy of models in any appreciable way.

Model averaging and collinearity

An important benefit of using model-averaging is that it alleviates (though does not completely resolve) the issue of collinearity, as compared to single best-model approaches. First, collinearity does not affect the estimation of strength of evidence or Akaike weights or the values of relative importance that multimodel inference provides. This is because AIC values, which form the basis for the computation of strengths of evidence, are only dependent on the performance of the model as a whole given the data, and not on how the variance is shared between (collinear or non-collinear) predictors. This is in contrast to several other algorithms of model selection for which collinearity is demonstrably harmful (Smith et al., 2009). So collinearity is not an issue if the interpretation of the effects is based on the relative importance of predictors, rather than on their inferential statistics.

Second, the model-averaged estimate of the unconditional standard error for a predictor relies on standard errors from *all* models in the model set that contain that predictor. Suppose predictors A and B are highly collinear. For predictor A, only one half of the models that contain A will also contain B; for another half of models no inflation of standard error is expected. As a result, the inflated standard errors will be factored into the unconditional standard error along with uninflated standard errors, and the influence of the former is expected to be weaker as compared to the single best model with collinear predictors A and B. Freckleton (2011) lends support to this intuition in a series of simulations that vary the strength of correlation between collinear predictors and the strength of the correlation between each of those predictors and the dependent variable. Freckleton demonstrates that model averaging techniques achieve stronger reduction of the variance inflation that collinearity causes, as compared to the ordinary least squares methods commonly used for regression analysis. Thus, collinearity does not lead to biased estimates of the regression slopes for predictors in the cases where the collinearity between predictors is weak to moderate (Pearson's $r \leqslant 0.5$) and when the effects of predictors on the dependent variable are weak to moderate too (Pearson's $r \leqslant 0.5$). In our data set, the pair of variables that keeps re-emerging as pervasive predictors of acoustic duration contains construction probability and length of theme. Across the seven data sets, the strength of correlation between these predictors ranges from 0.25 to 0.40. The absolute value of correlations of either predictor with the acoustic duration did not exceed $r = 0.2$ either. Given the relatively weak collinearity and weak effects of construction probability and length of theme, collinearity is not harfmul for the accuracy of estimating the variance for their slopes and their p-values. We additionally confirmed Freckleton's results by fitting to each dataset a mixed-effects model with the length of theme as the only critical predictor, and a separate mixed-effects model with construction probability as the only critical predictor: all random effects and fixed predictors were kept in both models as controls. The estimates obtained for length of theme in the absence of construction probability were identical in terms of their polarity and their significance level (at 0.05) with the estimates and p-values estimated by the model averaging technique. Same was true for the estimates of construction probability. We conclude, on the basis of simulations of Freckleton (2011) and our additional analyses, that collinearity was not an issue for the accuracy of our inferential statistics.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (Vol. 1, pp. 267–281).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Anderson, D. (2008). *Model based inference in the life sciences: A primer on evidence*. Springer Verlag.

Anttila, A., Adams, M., & Speriosu, M. (2010). The role of prosody in the English dative alternation. *Language and Cognitive Processes, 25*(7-9), 946–981.

Arnold, J., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language, 76*(1), 28–55.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*, 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllabic nuclei. *Journal of the Acoustic Society of America, 119*, 3048–3058.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM). Linguistic data consortium*. Philadelphia, PA: University of Pennsylvania.

Barton, K. (2011). The MuMIn library. <http://lib.stat.cmu.edu/R/CRAN/>.

Bates, D. M., & Sarkar, D. (2007). The lme4 library. <http://lib.stat.cmu.edu/R/CRAN/>.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language, 60*(1), 92–111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical society of America, 113*(2), 1001–1024.

Bock, J. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*, 1–47.

Bock, J. (1986). Syntactic persistence in language production. *Cognitive psychology, 18*, 355–387.

Bock, K., & Griffin, Z. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129*, 177–192.

Bock, J., & Irwin, D. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior, 19*, 467–484.

Bock, J. K., & Levelt, W. J. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego: Academic Press.

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition, 35*, 1–39.

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological review, 99*, 150.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review, 23*, 291–320.

Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211–252.

Brants, T., & Franz, A. (2006). Web 1t 5-gram version 1.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language, 86*(1), 186–213.

Bresnan, J., & Nikitina, T. (2009). The gradience of the dative alternation. In L. Uyechi & L. H. Wee (Eds.), *Reality exploration and discovery: Pattern interaction in language and life* (pp. 161–184). CSLI Publications.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. KrSmer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Koninklijke Nederlandse Akademie van Wetenschapen.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155–180.

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer Verlag.

Burnham, K., & Anderson, D. (2004). Multimodel inference. *Sociological Methods & Research, 33*(2), 261.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition, 24*(2), 215–221.

Bybee, J. (2007). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711–733.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamin.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in english. *Linguistics, 37*(4), 575–596.

Chamberlin, T. (1995). The method of multiple working hypotheses. *Journal of Geology, 103*(3), 349–354.

Collins, P. (1995). The indirect object construction in English: An informational approach. *Linguistics, 33*(1), 35–49.

Conroy, M. (2006). Methods for evaluating the influence of biotic and abiotic factors on biological populations. *Paper presented at the workshop on statistical sampling and estimation for fisheries and wildlife, University of Georgia, Athens, May 2006*.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing (studies in natural language processing)*. Cambridge: Cambridge University Press.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998). Resegmentation of switchboard. In *ICSLP-1998. Proceedings of the 5th international conference on spoken language processing, Sydney, Australia November 30–December 4, 1998*.

Du Bois, J. (1985). Competing motivations. In J. Haiman (Ed.), *Iconicity in syntax* (pp. 343–365). Amsterdam: Benjamin.

Fellbaum, C. (2005). Examining the constraints on the benefactive alternation by using the world wide web as a corpus. In M. Reis & S. Kepser (Eds.), *Evidence in linguistics: Empirical, theoretical, and computational perspectives* (pp. 209–240). Berlin and New York: Mouton de Gruyter.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language, 35*(5), 724–755.

Ferreira, V., & Dell, G. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*, 296–340.

Fidelholtz, J. (1975). Word frequency and vowel reduction in English. In *Chicago Linguistic Society* (Vol. 11, pp. 200–213).

Frank, A., & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the Cognitive Science Society*, Washington, DC (pp. 939–944).

Freckleton, R. (2011). Dealing with collinearity in behavioural and ecological data: Model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, 1–11.

Freedman, D. (1983). A note on screening regression equations. *The American Statistician, 37*(2), 152–155.

Gahl, S., & Garnsey, S. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language, 80*(4), 748–774.

Gahl, S., & Garnsey, S. (2006). Knowledge of grammar includes knowledge of syntactic probabilities. *Language, 82*, 405–410.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA*, July 2002 (pp. 199–206).

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *IEEE international conference on acoustics, speech, and signal processing, 1992. ICASSP-92* (Vol. 1, pp. 517–520). IEEE.

Gomez Gallo, C., Jaeger, T., & Smyth, R. (2008). Incremental syntactic planning across clauses. In *Proceedings of the 30th annual meeting of the Cognitive Science Society* (pp. 845–850).

Greenberg, S., Hollenback, J., & Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *International conference on spoken language processing* (pp. S32–S35).

Gregory, M., Raymond, W., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *CLS, 35*, 151–166.

Gries, S. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum International Publishing Group Ltd.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research, 34*, 365–399.

Harrell, F. (2001). *Regression modeling strategies*. Berlin: Springer.

Haskell, T., & MacDonald, M. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language, 48*, 760–778.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hay, J., & Bresnan, J. (2006). Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review, 23*, 321–349.

Hinrichs, L., & Szmrecsanyi, B. (2007). Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics, 11*, 437–474.

Jaeger, R. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. PhD thesis. Stanford University.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

Jaeger, T., & Post, M. (2010). Word production in spontaneous speech: Availability and communicative efficiency. *Paper presented at the CUNY-2010 conference on human sentence processing, New York*, NY, March 2010.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: Benjamin.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymon, W. (1998). Reduction of English function words in switchboard. In *Proceedings of ICSLP-98* (pp. 3111–3114).

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the conference on empirical methods in natural language processing, Barcelona* (pp. 317–324).

Keuleers, E., & Daelemans, W. (2007). Memory-based learning models of inflectional morphology: A methodological case study. *Lingue e linguaggio, 6*(2), 151–174.

Kliegl, R., Masson, M., & Richter, E. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655–681.

Krug, M. (1998). String frequency. *Journal of English Linguistics, 26*(4), 286.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America, 121*, 2261–2271.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th meeting of the North American chapter of the Association for Computational Linguistics*, College Park, Maryland.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. *Proceedings of the twentieth annual conference on neural information processing systems* (pp. 29–37). Vancouver: NIPS.

Lukacs, P., Burnham, K., & Anderson, D. (2010). Model selection bias and freedmans paradox. *Annals of the Institute of Statistical Mathematics, 62*(1), 117–125.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman and Co.

McDonald, J. L., Bock, J. K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology, 25*(2), 188–230.

Pickering, M. J., Branigan, H. P., & McLean, J. F. (2002). Constituent structure is formulated in one stage. *Journal of Memory and Language, 46*(3), 586–605.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS. Statistics and computing*. New York: Springer.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America, 118*(4), :2561–2569.

Prat-Sala, M., & Branigan, H. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language, 42*, 168–182.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America, 123*, 1104–1113.

Race, D., & MacDonald, M. (2003). The use of that in the production and comprehension of object relative clauses. In 26th annual meeting of the Cognitive Science Society (pp. 946–951).

R Development Core Team (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Recchia, G. (2007). *STRATA: Search tools for richly annotated and time-aligned linguistic data*. Stanford University symbolic systems program honors thesis.

Rhodes, R. (1996). English reduced vowels and the nature of natural processes. *Natural phonology: The state of the art*, 239–259.

Roland, D. (2009). Relative clauses remodeled: The problem with mixed-effect models. In *Poster presented at the CUNY 2009 conference on human sentence processing*.

Roland, D., Dick, F., & Elman, J. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57*(3), 348–379.

Roland, D., Elman, J., & Ferreira, V. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition, 98*, 245–272.

Selkirk, E. (2003). The prosodic structure of function words. In J. McCarthy (Ed.), *Optimality theory in phonology: A reader* (pp. 464–482). Malden, MA: Blackwell.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.

Smith, A., Koper, N., Francis, C., & Fahrig, L. (2009). Confronting collinearity: Comparing methods for disentangling the effects of habitat loss and fragmentation. *Landscape Ecology, 24*(10), 1271–1285.

Snyder, K. (2003). *The relationship between form and function in ditransitive constructions*. PhD thesis. University of Pennsylvania.

Solomon, E., & Pearlmutter, N. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology, 49*, 1–46.

Thompson, S. A. (1990). Information flow and dative shift in English discourse. In J. A. Edmondson, C. Feagin, & P. Mühlhausler (Eds.), *Development and diversity: Language variation across time and space* (pp. 239–253). Dallas: Summer Institue of Linguistics and University of Texas at Arlington.

Thompson, S. (1995). The iconicity of dative shift in English: Considerations from information flow in discourse. In M. E. Landsberg (Ed.), *Syntactic iconicity and linguistic freezes* (pp. 155–175). Berlin: Mouton de Gruyter.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition, 1*(2), 147–165.

Torres Cacoullos, R., & Walker, J. (2009). On the persistence of grammar in discourse formulas: a variationist study of that. *Linguistics, 47*, 1–43.

Van Son, R., & Pols, L. (2003). Information structure and efficiency in speech production. In *Proceedings of EUROSPEECH 2003, Geneva, Switzerland* (pp. 769–772).

Van Son, R., & Van Santen, J. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication, 47*, 100–123.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S-Plus* (4th ed.). New York: Springer.

Wagner Cook, S., Jaeger, T., & Tanenhaus, M. (2009). Producing less preferred structures: More gestures, less fluency. In *The 31st annual meeting of the Cognitive Science Society (CogSci09)* (pp. 62–67).

Walsh, M., Moebius, B., Wade, T., & Schuetze, H. (2010). Multilevel exemplar theory. *Cognitive Science*, 537–582.

Wasow, T. (2002). *Postverbal behavior*. Stanford: CSLI.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics, 19*, 29–58.

Whalen, D. (1991). Infrequent words are longer in duration than frequent words. *The Journal of the Acoustical Society of America, 90*, 2311.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *INTERSPEECH-2006, Pittsburg, PA* (pp. 1–4).

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology, 15*, 1–95.

Zipf, G. K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.

Please cite this article in press as: Kuperman, V., & Bresnan, J. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* (2012), http://dx.doi.org/10.1016/j.jml.2012.04.003