# Syntactic probabilities affect pronunciation variation in spontaneous speech

HARRY TILY, SUSANNE GAHL, INBAL ARNON, NEAL SNIDER, ANUBHA KOTHARI, AND JOAN BRESNAN*

*Stanford University*
*University of California, Berkeley*
*University of Rochester*

*Abstract*

*Speakers frequently have a choice among multiple ways of expressing one and the same thought. When choosing between syntactic constructions for expressing a given meaning, speakers are sensitive to probabilistic tendencies for syntactic, semantic or contextual properties of an utterance to favor one construction or another. Taken together, such tendencies may align to make one construction overwhelmingly more probable, marginally more probable, or no more probable than another. Here, we present evidence that acoustic features of spontaneous speech reflect these probabilities: when speakers choose a less probable construction, they are more likely to be disfluent, and their fluent words are likely to have a relatively longer duration. Conversely, words in more probable constructions are shorter and spoken more fluently. Our findings suggest that the differing probabilities of a syntactic construction in context are not epiphenomenal, but reflect a part of a speakers' knowledge of their language.*

*Keywords*
*pronunciation variation, gradience, disfluency, ditransitive, word duration, speech production, syntactic alternation*

* Correspondence address: Harry Tily, Linguistics, Margaret Jacks Hall, Stanford University, CA 94305, USA. E-mails: hjt@stanford.edu; gahl@berkeley.edu. This material is based in part upon work supported by the National Science Foundation under grant number IIS-0624345 to Stanford University for the research project "The Dynamics of Probabilistic Grammar" (PI Joan Bresnan). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Dan Jurafsky, Tom Wasow, and the audience at the 2007 AMLaP conference for useful comments and suggestions.

*Language and Cognition* 1–2 (2009), 147–165 1866–9808/09/0001–0147
DOI 10.1515/LANGCOG.2009.008

## 1. Introduction

Empirical methods have become ubiquitous in all subfields of Linguistics. For example, the 2003 meeting of the Linguistic Society of America featured a symposium on "Probability theory and Linguistics", but only a single regular session on psycholinguistics and none on corpus linguistics. By contrast, the 2008 meeting had several sessions devoted to psycholinguistics and corpus linguistics, and, moreover, featured corpus-based and experimental psycholinguistic research in practically every session, on topics ranging from syntactic theory to morphology to lexical semantics. This methodological change has gone hand in hand with the emergence of new theoretical approaches. Most major models of grammar until recently cast linguistic structure as discrete, static, and categorical. Recent years, however, have seen the emergence of more and more models that conceive of structure as gradient, malleable, and probabilistic (see for example the papers in Barlow and Kemmer (2000); Bod et al. (2003); Bybee and Hopper (2001); and Gahl and Yu (2006)). In these models, knowledge of language includes not just knowledge of syntactic, morphological, and phonological categories, but also knowledge of the frequency and probability of use of these categories in speakers' experience. Families of frameworks such as "probabilistic linguistics", "usage-based" and "exemplar-based" models all recognize gradient activation of linguistic units and probabilistic and gradient effects of linguistic form and meaning. The linguistic units in question include structures at all levels of linguistic representation and varying degrees of abstraction (see e.g. Borensztajn et al. 2009; Pierrehumbert 2001, 2002; Bybee 2002, 2006; Johnson 1997). Taken together, these proposals constitute a major departure from a research tradition that imposed rigid boundaries between competence and performance, sought to minimize redundancy in lexicon and grammar, and assumed linguistic representations to be categorical and discrete.

The development of these models has been possible in part thanks to rich, large-scale corpora of naturalistic usage data and the availability of statistical techniques for analyzing complex interactions of multiple factors. These tools have made it possible to build sophisticated models of the many factors affecting how speakers encode meaning in linguistic form. For example, Bresnan et al. (2007) examined what drives speakers' choice of syntactic realization patterns in the so-called dative alternation. A given scenario can be expressed with either of two syntactic patterns, either NP NP or NP PP, exemplified in (1a) and (1b), respectively:

(1) a. They sent us two of our coach tickets (NP NP)
    b. They sent two of our coach tickets to us (NP PP)

Attempts to account for speakers' choice between the dative alternants have tended to invoke semantic differences between the forms (Green 1974; Gropen et al. 1989), or constraints on the pronominality (Green 1971), information structure (Erteschik-Shir 1979) or length of the two arguments involved (Hawkins 1994). Each one of these generalizations covers many cases—but each is subject to exceptions. Indeed, corpus analysis shows the choice between the two constructions to be far more flexible than first appears to intuition (Fellbaum 2005; Bresnan and Nikitina 2007; Bresnan 2008). Analyzing a large corpus of such "dative" sentences, Bresnan et al. (2007) showed that a multitude of such factors, taken together, jointly predict speakers' syntactic choice between NP NP or NP PP alternants at very high accuracy. No analysis considering just one factor at a time, be it semantic, phonological, or pragmatic, does justice to the facts about the dative alternation. Grammatical models seeking to describe syntactic realization patterns with any degree of accuracy must therefore take into account many factors at once.

Speakers' syntactic choices can be accurately modeled using statistical models incorporating interacting constraints that jointly estimate the outcome probability. Moreover, Bresnan (2008) found that acceptability judgments reflect these factors, as well. However, the off-line judgment task does not show whether the language production process is sensitive to similar constraints as it unfolds: the models may achieve mere "descriptive adequacy". What constraints are speakers in fact sensitive to? One means of investigating that question draws on observations about pronunciation. Different tokens of one and the same word or phrase typically sound slightly different. This variation may be random to some degree; to some extent, however, it reflects planning processes during language production: A large body of evidence suggests that the duration of words and pauses provides a sensitive diagnostic revealing speakers' sensitivity to probabilities at various levels of linguistic structure, such as the frequency and contextual predictability of words (Lieberman 1963; Bell et al. 2009), morphemes (Pluymaekers et al. 2005; Kuperman et al. 2007), and syntactic structures (Gahl and Garnsey 2004, 2006; Gahl et al. 2006).

Just as with research on syntactic alternations, research on pronunciation variation reveals speakers' sensitivity to many probabilistic factors at once. This point is firmly established in the study of word durations, which simultaneously reflect static properties of single words such as orthographic regularity, and dynamically-changing properties related to the speaker's experience with that word: for example, its frequency, and its likelihood of appearing in the context of the words before and after it (Gahl 2008; Bell et al. 2009). Other things being equal, the production of low-probability linguistic units—that is, low-frequency words and words which are unlikely in a given context—tends to involve lengthening of words and pauses. By contrast, the pronunciation of high-probability linguistic units is characterized by phonetic reduction and durational shortening.

Research on probabilistic pronunciation variation has often focused on "string probability" measures such as n-grams, or transitional probabilities, i.e. the probability of a word conditioned on the word(s) that precede or follow it (Jurafsky et al. 2001; Bell et al. 2009). However, if grammars are indeed probabilistic, one should expect to see similar pronunciation effects of more abstract syntactic probabilities, as pointed out in Gahl and Garnsey (2004). Jaeger et al. (2006) found that construction probability had no effect on relativiser duration and limited effect on complementiser duration, but their measures were also conditioned on adjacent words only, leaving open the possibility that better estimates might yet reveal an effect. In our previous research, we have shown that syntactic probabilities can affect pronunciation. That research was based on the so-called subcategorization bias of a verb, or "verb bias". Verb bias refers to the probability with which a given verb appears with each of the subcategorization frames it is compatible with, such as the sentential complement (SC) and double object (DO) frames shown in (2). Effects of verb bias, i.e. a syntactic property, on sentence comprehension are well established (Trueswell et al. 1993; Garnsey et al. 1997).

(2) a. We confirmed the date was correct (SC)
    b. We confirmed the date (DO)

In Gahl and Garnsey (2004), we examined pronunciation variation in these types of sentence, and showed that, among other things, the acoustic-phonetic realization of the clause boundary following "confirmed" in the SC-variant was in part a function of the probability of encountering an SC following that verb. SCs after verbs that are highly likely to take direct objects ("DO-bias verbs") are realized differently from SCs following verbs that are likely to take SCs ("SC-bias verbs"), independently of the specific words appearing in those structures. Importantly, this difference was not due to the real-life probability of scenarios described by sentences with high and low syntactic probability (cf. Gahl and Garnsey 2006, for discussion, and Gahl et al. 2006) for a similar effect in a different pair of constructions).

While the observations in Gahl and Garnsey (2004) suggest that pronunciation variation reflects probabilities associated with syntactic structure, it is clear that the probability measure used there is overly simple. To look only at a verb's subcategorization bias, estimated from corpus counts of various subcategorization frames in corpora, is to throw away the mass of rich information available in sentences which speakers' choices may be sensitive to. Subcategorization biases exist in tandem with (and in some part, result from) a host of local and discourse-level factors, as can be seen in the rich and detailed analyses in Bresnan et al. (2007), Szmrecsanyi and Hinrichs (2008), and Wasow (2002), among others.

The goal of the current study is to bring the tool of pronunciation variation to bear on understanding the richness of speakers' probabilistic knowledge of language. We examine pronunciation variation in the dative alternation. If pronunciation variation is a sufficiently sensitive reflection of the multiple probabilistic cues predicting the choice between syntactic structures, then it can help show whether the human language production system does indeed rely on the full range of available cues.

The current study also allows us to address serious questions left open by previous research. Previous studies of syntactic probabilities (Gahl and Garnsey 2004; Gahl et al. 2006) elicited speech from participants by asking them to read sentences. That fact constitutes a limitation: For one thing, the prosody of read speech differs from that of spontaneous speech (Schafer et al. 2005). An even more serious problem is that the observed effect may have resulted from comprehension difficulty, rather than directly reflecting the workings of the language production system. Sentences with local ambiguities often induce "garden-paths", i.e. incorrect parses that temporarily throw the comprehension system off-track. Gahl and Garnsey (2004) excluded tokens from the analysis that showed self-correction or marked overemphasis (*"we confirmed, no wait, oh now I get it, . . . we conFIRMED the date was correct"*). Still, the possibility cannot be ruled out that the subjects in those studies initially misunderstood some of the sentences they were asked to read and then decided to emphasize low-probability prosodic phrasings. In fact, to keep subjects from feeling self-conscious knowing their speech would be analyzed, they were falsely given the impression that the researchers needed the recordings for a future comprehension experiment. Perhaps, then, speakers were attempting to make the sentences easy to comprehend for an imaginary listener. An analysis of spontaneous speech alleviates the problems caused by possible garden-path effects experienced by the speakers, if it is assumed that talkers are unlikely to induce garden-path effects in themselves by their own speech. That assumption appears plausible, given that talkers do not generally appear to be aware of local ambiguities in their own speech here (Allbritton et al. 1996). In addition, though this is an active area of research, it appears that speakers do not consistently provide cues to listeners that would maximize ease of comprehension (Ferreira and Dell 2000).

The dative alternation provides a particularly useful tool for an investigation of syntactic probabilities in that the two alternants (*They sent us two tickets* ~ *They sent two tickets to us*) denote identical real-life scenarios (semantic differences between the alternants notwithstanding, cf. Green 1974; Gropen et al. 1989). If the phonetic realization of dative sentences indeed reflects probability of construction choice, then it does not simply reflect probability of real-world scenarios. Speakers' choices of dative alternants are subject to a range of probabilistic constraints at least some of which are based on linguistic facts alone, not on real-world denotata. Therefore, differences in planning or processing difficulty between the two alternants must be due to speakers' store of linguistic experiences, not to differences in the frequency of events in the world. Our earlier studies controlled for real-life probability of denoted scenarios (cf. the discussion in Gahl and Garnsey 2006), but they did so indirectly; the dative alternation provides a direct means of teasing apart probability of constructions and of real-world denotata.

## 2. Background: The dative alternation

In Bresnan et al. (2007), we used multivariate statistical analysis to investigate the many factors that have been claimed to influence speakers' choice between the dative alternants. As mentioned above, previous accounts explain the choice in terms of a single variable. Surprisingly perhaps, all of these accounts work fairly well despite the different constraints they invoke. This is because the properties that have shown to be relevant tend to pattern together: For instance, pronominal themes tend to favour the NP PP construction and pronominal recipients the NP NP construction; but pronouns also tend to be short, definite, concrete, and given. Using a logistic regression model, however, Bresnan et al. (2007) were able to include many such correlated factors and test whether speakers' choices were influenced by each independently, controlling for the others.

Bresnan et al.'s analysis used data from the Switchboard corpus of spoken American English, which consists of recorded telephone conversations between strangers (Godfrey et al. 1992). Bresnan et al. hand-annotated each sentence containing one of the two dative alternants (NP NP or NP PP; in a total of 2360 sentences), tracking a host of syntactic and semantic variables that might have influenced the syntactic choice. All of the variables were previously claimed to be relevant to the alternation in the theoretical or experimental literature. All in all, fourteen variables were chosen and annotated in the data: the semantic class of the verb (coding the type of relationship held between the recipient and

Table 1. *Factors found by Bresnan et al. (2007) to favor the NP NP or NP PP constructions*

| NP NP more likely | NP PP more likely |
|---|---|
| given recipient or nongiven theme | given theme or nongiven recipient |
| pronominal recipient or nonpronominal theme | pronominal theme or nonpronominal recipient |
| animate recipient or inanimate theme | animate theme or inanimate recipient |
| definite recipient or indefinite theme | definite theme or indefinite recipient |
| short recipient or long theme | short theme or long recipient |
| singular theme | plural theme |

theme); the givenness, pronominality, definiteness, animacy, person and number of the recipient; the givenness, pronominality, definiteness, number and concreteness of the theme; the (log) difference in the number of words of the recipient and theme; structural parallelism (whether there had been instances of the same syntactic pattern in the dialogue). A logistic regression model was then estimated which could predict the speaker's choice between NP NP and NP PP as a function of these variables. Except for number and person of recipient, and concreteness of theme, all of the factors were found to have an effect on the choice of NP NP or NP PP: the nature of some of these effects is illustrated in Table 1. On previously unseen data, the model correctly predicted in 94% of cases whether the NP NP or NP PP would be used.

The outcome variable in a logistic regression model is a continuous number ranging between 0 and 1. This number can be interpreted as the probability with which the model "expects" (or "predicts") the NP PP construction—or equivalently, 1 minus the probability of the NP NP. For example, when all the cues converge to make the outcome very certain, the output will be close to 1 or 0; in cases where the cues are more equivocal, the output will be closer to .5. We can consider this output as a measure of the probability of the construction choice, given the cues: for each NP NP or NP PP, was the speakers' choice of that construction inevitable? Or was the choice more of coin flip between the two, or even—in a few cases—the *less* likely outcome?

## 3. Methods

The Bresnan et al. data and model give us a set of tokens of NP NP and NP PP sentences, along with an estimate of the probability of the alternant that was chosen: In some cases, the choice of the alternant that the speaker in fact chose received strong support from the various factors in the model. Other cases are assigned a lower probability by the model. For example, the two sentences below had predicted probabilities of 0.01 and 0.99, respectively:

(3) a. Yeah. I haven't given much thought to it. I'm kind of busy raising my kids ($p = 0.01$)
    b. if they can test the teachers, that gives them the full right to test the kids ($p = 0.99$)

With these probabilities in hand, we examined the effect of syntactic probability on the phonetic realization of dative sentences. We examined two aspects of phonetic realization: word duration and the presence of disfluencies. Word durations and the presence of disfluencies are two well established measures for fluctuations in processing speed and processing difficulty (Fox Tree and Clark 1997; Shriberg 2001; Clark and Fox Tree 2002; Bell et al. 2003; Arnold et al. 2000).

To study word durations, we focused on the preposition *to* in the NP PP alternants, using durations extracted from the time-aligned transcript of the Switchboard corpus (Deshmukh et al. 1998). Our choice of the word *to* as our target was motivated largely by concerns about effect size: Previous studies of probabilistic pronunciation variation led us to expect that the size of any effect of duration reduction would be quite small (Bell et al. 2003; Pluymaekers et al. 2005; Kuperman et al. 2007; Gahl 2008; Bell et al. 2009), so it is important to minimize other effects that are not in the model, such as the length or frequency of other words in the dative constructions. Examining many instances of the same word is a way to control for word-specific information; hence we use the duration of this word in all of the NP PP outcomes as our dependent variable.

Our models also included the following other variables as controls:

- *Rate of speech*, measured in syllables per second, for the intonational phrase surrounding the word *to* (excluding the duration of *to* itself). Following Bell et al. (2003), we define the intonational phrase as the longest region containing the word of interest that contains no sentence boundaries or pauses of 500ms or more.
- *Segmental context*, specifically the presence of a preceding and following vowel, as this environment may favor flapping and other contextually-induced articulatory changes.
- Other measures of contextual probability:
  - *Verb bias*, i.e. the probability of NP NP or NP PP conditioned only on the verb,
  - *Forward and backward bigrams*, i.e. the probability of the word *to* given the immediately preceding or following word (Bell et al. 2009) obtained from the Web 1T ngram corpus (Brants and Franz 2006)

We removed cases with disfluencies immediately preceding or following *to*. We consider the following to be disfluencies: a pause of 500ms or more; repetition of a word; a filled pause (*"uh", "um"*); or a repair or restart (*"give thi- that to them"*).

We then built a multiple linear regression model to test the effects of these variables. A linear regression model relates a set of predictor variables to an outcome variable, by considering the influence of all independent variables simultaneously. The model determines a coefficient for each independent variable which shows how strongly it correlates with the outcome variable when all other variables in the model are controlled. The outcome variable in our case was the duration of the word *to*. The critical predictor variable of interest was syntactic probability, i.e. the probability assigned a sentence in the Bresnan et al. model. The coefficient for probability showed the average difference, in milliseconds, of the word *to* in high versus low probability instances of the construction, after controlling for all other factors in the model. If this difference is significantly different from zero, i.e. if it is large relative to the difference that would be expected due to random variation in the data, the influence of syntactic probability on duration is considered to be statistically significant.

A second outcome variable of interest was the presence of disfluencies in the dative sentences. A second regression model was constructed, this time predicting the presence of disfluencies preceding or following the verb or within either of its two arguments (the recipient or the theme) in the NP NP and NP PP sentences. As this outcome variable is categorical, we used logistic regression. Like linear regression, logistic regression relates a set of predictor variables to an outcome variable. Unlike in the case of linear regression, the outcome variable in a logistic regression model is a probability estimate, namely the probability of observing particular values of a categorical variable, here, the probability that the utterance contains a disfluency.

The only predictor variables in this model were verb bias, speech rate, and the probability of the NP NP or NP PP variant, from the Bresnan et al. database. Note that the other predictor variables in the model of *to*-duration, such as the bigram probability measures, vary for each word in a sentence. It would be possible to estimate the values of these variables for every word in the sentences and to combine those measures with the construction outcome probability to predict disfluency at each point in the sentence. We are currently exploring this and other variants of the disfluency model.

Data preparation and statistical analysis was carried out using the statistical package R (R Development Core Team 2008) and in particular the Design (Harrell 2007) and languageR (Baayen 2008) packages.

Table 2. *Final model for* to *duration in the PP outcome*

| | $\beta$ | Std. Error | $t$ | $p$ |
|---|---|---|---|---|
| Intercept | 0.17557 | 0.01220 | 14.397 | 0.000000 |
| Outcome probability | −0.34147 | 0.13782 | −2.478 | 0.013603 |
| Backward bigram | −6.92303 | 2.12904 | −3.252 | 0.001235 |
| Previous vowel | 0.02486 | 0.01038 | 2.396 | 0.017001 |

## 4. Results

We first turn to the model of the duration of the word *to* at the start of the PP. Our dependent measure was the duration of this word in seconds. We removed datapoints with disfluencies adjacent to the word of interest, or with durations more than 2.5 standard deviations from the mean (8.4% of the data). 446 cases remained. A speech rate control variable was calculated by taking the duration of the intonational phrase containing the word *to* (i.e. the maximum period containing no pause of 500ms or more and no sentence boundaries). We excluded the word *to* itself from the region over which speech rate was calculated, to avoid collinearity with the dependent variable. The number of syllables in the region was divided by this duration, to determine the speaking rate, measured as syllables per second. The independent variable of interest, the probability of the actual outcome spoken, was calculated using the Bresnan et al. model refitted to the Recchia (2007) spoken dative dataset. Together with the other controls described above, these variables were entered into a linear regression model. Regression inputs were standardized by subtracting the mean and dividing by two standard deviations, as recommended in Gelman (2008).

Although some of the predictor variables might be expected to co-vary, in fact collinearity turned out to be unproblematic. All VIFs were less than 1.2, meaning that the predictors were almost orthogonal. Because the number of datapoints from each speaker varied greatly and because speech rate accounted for much of the inter-speaker variability, we did not use any random or fixed effect for speaker.

The following controls were not significant, and were removed from the model during model comparison by fast backwards elimination of factors (Lawless and Singhal 1978): Forward bigram probability ($p = .43$), Speech rate of the surrounding region ($p = .27$) and Verb bias ($p = .95$).

The three factors shown in Table 2 were determined (by likelihood ratio tests) to improve model quality (at $p < .05$). Importantly, the probability of the PP outcome is a statistically significant predictor of the duration of *to*, with higher probability outcomes resulting in shorter pronunciations.

Table 3. *Final model for disfluency in the dative VP*

| | $\beta$ | Std. Error | Wald Z | $p$ |
|---|---|---|---|---|
| Intercept | 0.6782 | 0.27773 | 2.44 | 0.0146 |
| speech rate | −0.8168 | 0.09997 | −8.17 | 0.0000 |
| outcome probability | −0.2020 | 0.09403 | −2.15 | 0.0317 |

We now turn to our second variable of interest: disfluency. Arnold et al. (2000) found disfluency to correlate with word-order choice in dative sentences, although the restricted referential domain and wide variety of constructions produced in their experiment make it hard to infer specific predictions for the current study from their findings. We coded sentences for whether they contained a disfluency in the intonational phrase surrounding the "dative" verb. Utterances were identified as disfluent if the longest stretch of pause-free speech surrounding the verb contained repetitions, filled pauses, repairs or restarts. Both NP PP and NP NP outcomes were included. We removed sentences with speech rate 2.5 standard deviations from the mean (0.43% of the data). This left 2061 cases, of which 594 contained a disfluency in the verb region. Again, our independent variable of interest was calculated using the Bresnan et al. model. This time, because both NP PP and NP NP outcomes were included, and the variable was not the absolute probability of a NP PP, but the probability of the actual outcome chosen (i.e. one minus the probability of the NP PP in the NP NP case). Collinearity between predictors was found not to pose a problem: all VIFs were less than 1.3.

Verb bias proved non-significant by likelihood ratio tests during model comparison ($p = .26$), and so was removed from the model.

The probability of the outcome (NP PP vs NP NP) is a significant predictor of disfluency: more probable NP PPs and more probable NP NPs are less likely to contain disfluencies. Additionally, sentences that are spoken more quickly are less likely to contain disfluencies.

The size of the effect of probability on duration is small. For the *to*-model, the predicted difference between the least and most probable outcome in the actual data is just over 20ms, but since the data is so heavily skewed towards likely outcomes, most datapoints are predicted to have much more similar durations. The difference between an utterance at the 25th percentile (the probability value which is greater than the least probable 25% of the data) and the 75th percentile, for instance, is predicted to be 15ms. Figure 1 shows the distribution of durations for each utterances falling in each quartile. It is not entirely surprising that the effect on duration should be so small: the word *to* is very short (mean duration of 129ms). Although standardizing the regression inputs

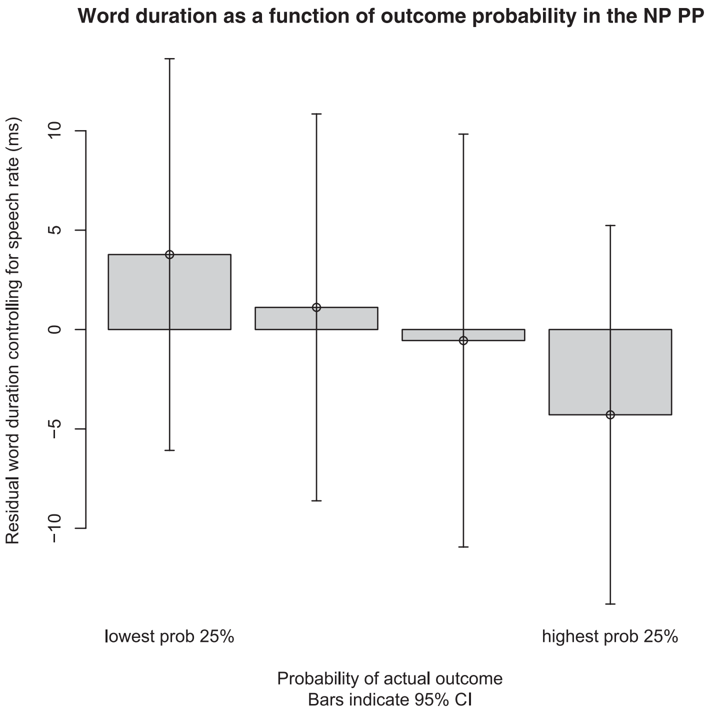


Figure 1. *Duration of tokens of* to *as a function of outcome probability. Bars indicate one standard deviation*

does make coefficients more comparable (see Gelman 2008), the probability measures used here have quite skewed distributions: In particular, the bigram probabilities are much less evenly distributed than the outcome probabilities, with roughly two thirds of the probabilities smaller than 0.05. This skewed distribution exaggerates the standardized effect size of the bigram relative to the outcome probability. As a result, we cannot directly compare the bigram and outcome probability effect sizes. Even so, it is safe to say that the bigram probability has a greater effect on duration than the outcome probability.

The size of the effect of probabilities on disfluency was slightly stronger: The probability of a disfluency in an average speed utterance jumps from .27 among the highest probability outcomes to .40 among the lowest probability outcomes.

To explore the effect of syntactic probability further, we additionally examined its effect on the duration of other words besides *to.* Recall that we chose the word *to* in the PP for methodological, rather than theoreti-

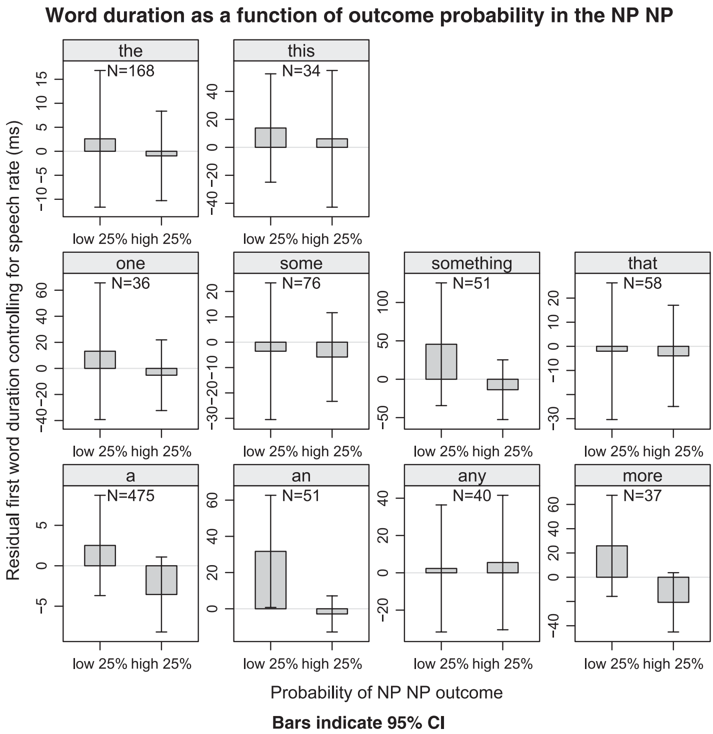


Figure 2. *Average durations of words in initial position in the second argument of the NP NP construction. Bars indicate 95% confidence intervals*

cal reasons: the within-item analyses allowed us to minimize noise, as well as to avoid prosodic and structural confounds. Even more importantly, we needed a word that was sufficiently frequent in our database to allow this kind of statistical analysis. To supplement our analyses, we in addition investigated the words which appear as the first word of the second argument in the NP NP outcome. We extracted all words that appeared in this position at least 30 times in the database, and used the entire Switchboard corpus to determine the average duration for each of these words overall, to control to some extent for differences between words. We do not report the resulting regression models here, except to note that a duration effect on the aggregated data is significant and similar to the model of *to*-duration. Figure 2 shows the durations for each of these words in low and high probability NP NP outcomes. It is clear that almost all the words show a similar effect: shorter duration when the actual outcome NP NP is more likely than the alternative. This suggests that the effect is not limited to the word *to* and that it shows up in both the NP NP and NP PP constructions.

## 5. Discussion

The goal of this study was to explore ways in which the probabilistic constraints on syntactic choice might be reflected in speakers' pronunciation of dative sentences. An additional goal was to ascertain whether this effect existed in spontaneous speech, or whether it was limited to the tightly-constrained artificial stimulus material used in previous studies.

Our crucial finding is that the probability of speakers' choice between alternants is indeed reflected in pronunciation, in spontaneous speech. While our previous findings on syntactic probabilities and pronunciation variation in read speech might have arisen from garden-path effects, i.e. a comprehension-based effect, the current results suggest that syntactic probabilities affect language production. Several caveats are in order: First, the observed effect on the duration of *to* was very small, and the unexplained variability substantial. A related caveat concerns the fact that the corpus data are heavily skewed towards likely syntactic choices: low-probability outcomes are rare by their nature—a persistent problem facing corpus-based research.

The small effect sizes and the sparseness of low-probability data raise the question whether the observed effect was spurious. However, we found the same probability estimate to be a significant predictor of disfluency in both constructions. Moreover, the effect consistently seemed to appear on other words in the NP NP construction. The pervasiveness of these related patterns increase our confidence in their stability and generalizability.

It may seem surprising that verb bias, a measure that had revealed itself as a significant predictor of probabilistic pronunciation variation in previous research, did not emerge as a significant predictor in the current data. On closer consideration, this fact is to be expected: verb bias is a crude measure of the probability with which a speaker will choose each construction. The detailed analysis in Bresnan et al. of the factors affecting the dative alternation reveals that verb bias is overridden in many cases by the host of other factors shown to play a role. Naturally, a crude measure only reveals large effects—or small effects as long as other factors are tightly controlled, as was the case in the scripted stimuli in our earlier work.

Our data do not enable us to say which of the many factors influencing the choice of syntactic alternant carried the effect, or indeed whether any single factor carried it. Our insistence that the dative choice is conditioned on a multitude of factors might invite the objection that we only included one summary measure in our models of phonetic variation, viz. the probability of the outcome conditioned on all of those factors. However, a model including all factors as predictors of pronunciation variation would be problematic, as it would unduly reflect phonetic properties of particular words that tend to occur in one level of certain factors, rather than the properties of those words that influence the syntactic outcome. Hence, such a model would not have shed light on the role of syntactic probabilities. Furthermore, the relatively small amount of data and the large number of factors would have left us in danger of overfitting the model to the specific data in our corpus; and the collinearity between factors wouldn't have permitted us to see the importance of individual factors with certainty.

The most promising way to tease apart the role of individual factors probably lies in experimental research, for factorial manipulation of individual factors. In this way, corpus studies and experimental research can be mutually supportive. But again, it is possible that no single factor or small set of factors would emerge as significant even then: the overall pattern result from the entire collection of factors working in concert.

Our results add to the growing body of evidence that the acoustic realization of words reflects higher-level linguistic information (Clark and Wasow 1998; Gahl and Garnsey 2004). Taken together, these findings argue for a model of language production in which high-level linguistic representations of syntax and meaning are not strictly isolated from low-level processes such as articulation and speech rate control, but where information can flow between levels of representation. Computational accounts that are consistent with the descriptive generalizations do exist. For instance, *Uniform Information Density* (Jaeger 2006; Levy and Jaeger 2007; based on Genzel and Charniak 2002; Aylett and Turk 2004) posits that speakers tend to make the rate at which information is conveyed over the speech stream roughly constant, and therefore more predictable words (which carry little information) should be produced to take up a shorter duration than less predictable words. This would be an efficient strategy for communication over the speech channel, in that it makes utterances shorter without reducing the words that the hearer would have the most difficulty reconstructing. While psycholinguistic models of the language production system underlying these effects that could accommodate these findings are not yet available, we believe that current work on exemplar-based and usage-based models may yield a useful formalization of the relevant processing units, thanks to its ability to represent linguistic units at arbitrary levels of abstraction and probabilistic tendencies between them.

Finally, our results add further evidence to the view that probabilistic effects in language production are not due to probability of real-world scenarios: There are multiple ways to express a given meaning. What we have shown here is that meaning-equivalent alternants differ in pronunciation, as a function of the syntactic probability.

## 6. General conclusion

Language production requires integrating many types of information. The view of the mind that underlies this research is that language production system is an adaptive system that comes to process those structures most efficiently that it has processed most often in prior experience. But what aspects of prior language experience does the language production system keep track of? The present work supports the view that many factors jointly shape speakers' probabilistic knowledge of language. We have arrived at this view based on corpus evidence, experimentation, and statistical modeling. It is thanks to this methodological grounding that our theoretical models can explore the consequences of abandoning the simplifying assumptions of grammar as categorical and deterministic.

## References

Allbritton, D. W., G. McKoon & R. Ratcliff. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 22(3). 714–735.

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

Aylett, M. & A. Turk. 2004. The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

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

Barlow, M. & S. Kemmer (eds.). 2000. *Usage-based models of language*. Chicago: CSLI.

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

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory & D. Gildea. 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.

Bod, R., J. Hay & S. Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Borensztajn, G., W. Zuidema & R. Bod. 2009. Children's grammars grow more abstract with age—Evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science* 1. 175–188.

Brants, T. & A. Franz. 2006. *Web 1T 5-gram*. Philadelphia, PA: LDC Data Consortium.

Bresnan, J. 2008. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (eds.)*, Roots: Linguistics in search of its evidential base*, 75–96. Berlin & New York: Mouton de Gruyter.

Bresnan, J., A. Cueni, T. Nikitina & R. H. Baayen. 2007. Predicting the dative alternation. In G. Bourne, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. & T. Nikitina. 2007. The gradience of the dative alternation. In L. H. Wee & L. Uyechi (eds.), *Reality exploration and discovery: Pattern interaction in language and life*. Stanford: CSLI.

Bybee, J. & P. Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure* (Typological studies in language 45). Amsterdam: John Benjamins.

Bybee, J. 2002. Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition* 24(2). 215–222.

Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 529–551.

Clark, H. H. & T. Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37(3). 201–242.

Clark, H. H. & J. E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1). 73–111.

Deshmukh, N., A. Ganapathiraju, A. Gleeson, J. Hamaker & J. Picone. 1998. Resegmentation of Switchboard. International Conference on Spoken Language Processing, Sydney, Australia, Australian Speech Science and Technology Association.

Erteschik-Shir, N. 1979. Discourse constraints on dative movement. In T. Givon (ed.), *Discourse and syntax*, 441–467. New York: Academic Press.

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.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 209–240. Berlin & New York: Mouton de Gruyter.

Ferreira, V. S. & G. S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40(4). 296–340.

Fox Tree, J. E. & H. H. Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition* 62(2). 151–167.

Gahl, S. 2008. "Time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language* 84(3). 474–496.

Gahl, S. & S. M. Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80(4). 748–775.

Gahl, S. & S. M. Garnsey. 2006. Syntactic probabilities affect pronunciation variation. *Language* 82(2). 405–410.

Gahl, S., S. M. Garnse, C. Fisher & L. Matzen. 2006. "That sounds unlikely": Syntactic probabilities affect pronunciation. 28th Annual Conference of the Cognitive Science Society, CD-ROM.

Gahl, S. & A. C. L. Yu. (eds.). 2006. Special issue on Exemplar-based Models in Linguistics. *The Linguistic Review* 23(3).

Garnsey, S. M., N. J. Pearlmutter, E. Myers & M. A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory & Language* 37(1). 58–93.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27. 2865–2873.

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

Godfrey, J., E. Holliman & J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. International Conference on Acoustics, Speech and Signal Processing.

Green, G. 1971. Some implications of an interaction among constraints. In *Papers from the seventh regional meeting*, 85–100. Chicago: Chicago Linguistic Society.

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

Gropen, J., S. Pinker M. Hollander, R. Goldberg & R. Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65(2). 203–257.

Harrell, F. E. 2007. *Design: Design package. R Package version 2.1-1* [Computer software]. http://CRAN.R-project.org/package=Design

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

Jaeger, T. F. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Stanford University, CA: Unpublished PhD dissertation.

Jaeger, T. F., N. Snider, L. Staum & D. Jurafsky. 2006. (In)dependence of lexical and syntactic production: That-reduction and omission in spontaneous speech. Poster presented at the 19th annual CUNY conference on Human Sentence Processing, New York, NY.

Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In K. Johnson & Mullennix (eds.), *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Jurafsky, D., A. Bell, M. Gregory & W. D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production [References]. In Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 229–254. Amsterdam: John Benjamins.

Kuperman, V. & M. Pluymaekers, M. Ernestus & R. H. Baayen. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America* 121(4). 2261–2271.

Lawless, J. & K. Singhal. 1978. Efficient screening on nonnormal regression models. *Biometrics* 34. 318–327.

Levy, R. & T. F. Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt & T. Hoffman (eds.), *Advances in neural information processing systems 19*, 849–856. Cambridge, MA: MIT Press.

Lieberman, P. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6. 172–187.

Pierrehumbert, J. B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. [References]. In Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 137–157. Amsterdam: John Benjamins.

Pierrehumbert, J. B. 2002. Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.), *Laboratory phonology VII*, 101–140. Berlin: Mouton de Gruyter.

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

R Development Core Team. 2008. R: A language and environment for statistical computing. Vienna.

Recchia, G. 2007. *STRATA: Search Tools for Richly Annotated and Time-Aligned Linguistic Data*. Stanford University, CA: Unpublished undergraduate honors thesis.

Schafer, A. J., S. R. Speer & P. Warren. 2005. Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. In J. C. Trueswell & M. K. Tanenhaus (eds.), *Approaches to studying world-situated language use,* 209–225. Cambridge, MA: MIT Press.

Shriberg, E. 2001. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(1). 153–169.

Szmrecsanyi, B. & L. Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta & M. Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*. Amsterdam: John Benjamins.

Trueswell, J. C., M. K. Tanenhaus & C. Kello. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory & Cognition* 19(3). 528–553.

Wasow, T. 2002. *Postverbal behavior*. Stanford, CA: CSLI Publications.