

# UC Berkeley

## UC Berkeley PhonLab Annual Report

### Title

Syntactic Probabilities Affect Pronunciation Variation in Spontaneous Speech

### Permalink

<https://escholarship.org/uc/item/22h474d2>

### Journal

UC Berkeley PhonLab Annual Report, 4(4)

### ISSN

2768-5047

### Authors

Tily, Hal  
Gahl, Susanne  
Arnon, Inbal  
[et al.](#)

### Publication Date

2008

### DOI

10.5070/P722h474d2

(Manuscript under review. Please check with Hal Tily ([hjt@stanford.edu](mailto:hjt@stanford.edu)) or Susanne Gahl ([gahl@berkeley.edu](mailto:gahl@berkeley.edu)) before citing)

**Syntactic probabilities affect pronunciation variation in spontaneous speech**

Hal Tily\*, Susanne Gahl+, Inbal Arnon\*, Neal Snider\*, Anubha Kothari\*, and Joan Bresnan\*

**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 (e.g. 2000), Bod, Hay et al. (2003), Bybee and Hopper (2001), and Gahl and Yu (2006). 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. Furthermore, all stress the influence of the individual's exposure to language on production and comprehension.

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, Cueni, Nikitina and Baayen (Bresnan et al. 2007) examined what drives speakers' choice of syntactic realization patterns in the so-

---

\* Stanford University, +University of California at Berkeley. This work was supported by NSF grant BCS-9818077. We thank Dan Jurafsky, Tom Wasow, and the audience at the 2007 AMLaP conference for useful comments and suggestions. Email addresses for correspondence: [hjt@stanford.edu](mailto:hjt@stanford.edu) or [gahl@berkeley.edu](mailto:gahl@berkeley.edu).

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). Any single one of these generalizations, whether categorical or probabilistic, might cover a great number of cases – but exceptions always remain. Indeed, corpus analysis shows the choice between the two constructions to be far more flexible than first appears to intuition (Bresnan 2008; Bresnan & Nikitina 2007; Fellbaum 2005). 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 (Bell et al. in press; Lieberman 1963), morphemes (Pluymaekers et al. 2005), and syntactic structures (Gahl & Garnsey 2004; Gahl & Garnsey 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 idiosyncratic

properties of single words such as orthographic regularity, and *probabilistic* 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 (ref Bell et al, Gahl, 2008). Other things being equal, the production of low-probability linguistic units 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 (Bell et al. in press; Jurafsky et al. 2001). However, if grammars are indeed probabilistic, one should expect to see similar pronunciation effects of more abstract syntactic probabilities, as pointed out in Gahl & Garnsey (2004). In fact, 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 (Garnsey et al. 1997; Trueswell et al. 1993).

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

In Gahl & 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 & Garnsey, 2006, for discussion, and Gahl, Garnsey, Fischer & Matzen (2006) for a similar effect in a different pair of constructions).

While the observations in Gahl & 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 (e.g.) Bresnan et al. (2007), Szmrecsanyi & Hinrichs (to appear), and Wasow (2002).

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. The current examines 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 & Garnsey, 2004; Gahl et al., 2006) were based on read speech. 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 & Garnsey 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, perhaps in an effort to make the sentences easier to comprehend for an imaginary listener. 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. An analysis of spontaneous speech alleviates this problem, 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).

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. 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 & Garnsey, 2006), but they did so indirectly; the dative alternation provides a much more direct means of teasing apart probability of constructions and of real-world denotata.

### **Background: The dative alternation**

In Bresnan, Cueni, Nikitina & Baayen (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), 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 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 preceding 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. 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 a coin flip between the two, or even -- in a few cases -- the *less* likely outcome?

## 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  
b. if they can test the teachers, that **gives them the full right to test the kids**

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 (Bell et al. 2003; Clark & Wasow 1998).

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 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. in press; Bell et al. 2003; Jurafsky et al. 2001), 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 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. in press) obtained from the Web 1T ngram corpus (Brants & Franz 2006)

We removed cases where there was a disfluency 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 well 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.

## 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 milliseconds. 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) and *excluding* the word *to* itself, 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. Together with the other controls described above, these variables were entered into a linear regression model.

Regression inputs were standardized by 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 & Singhal 1978): forward bigram probability ( $p=.43$ ), speech rate of the surrounding region ( $p=.27$ ) and verb bias ( $p=.95$ ).

	Value	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

Table 1: Final model for *to* duration in the PP outcome

The three factors shown in Table 1 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.

We now turn to our second variable of interest: disfluency. We coded sentences for whether they contain 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 PP and DO 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 PP and DO outcomes were included, and the variable was not the absolute probability of a PP, but the probability of the actual outcome chosen (i.e. one minus the probability of the PP in the DO 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.

	Coef	S.E.	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

Table 2: Final model for disfluency in the dative VP

The probability of the outcome (PP vs DO) is a significant predictor of disfluency: more probable PPs and more probable DOs 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 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 datapoints closer to zero 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.

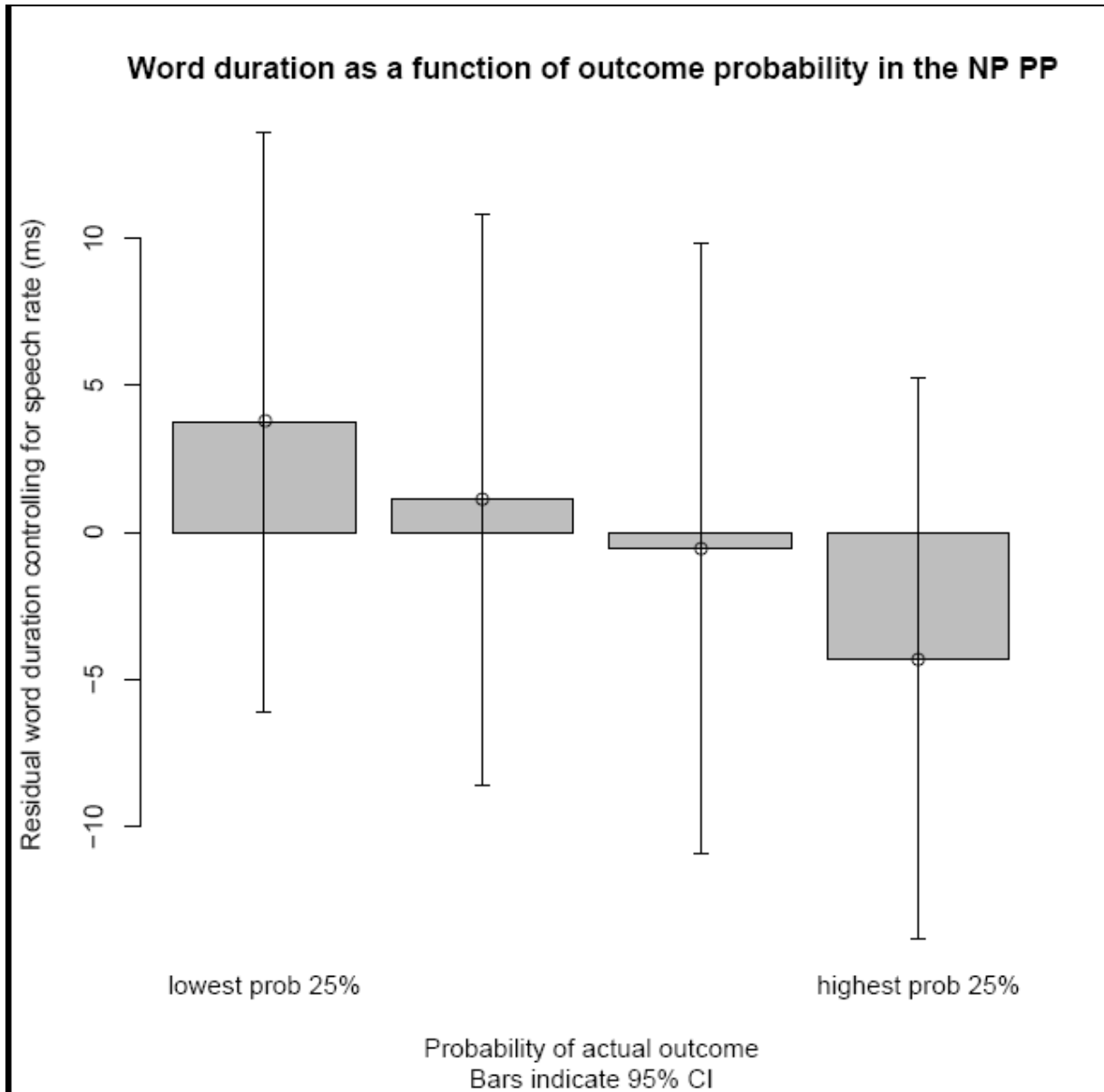


Figure 1

For disfluency, the effect is a little more pronounced: 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 theoretical 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 "to" model.

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.

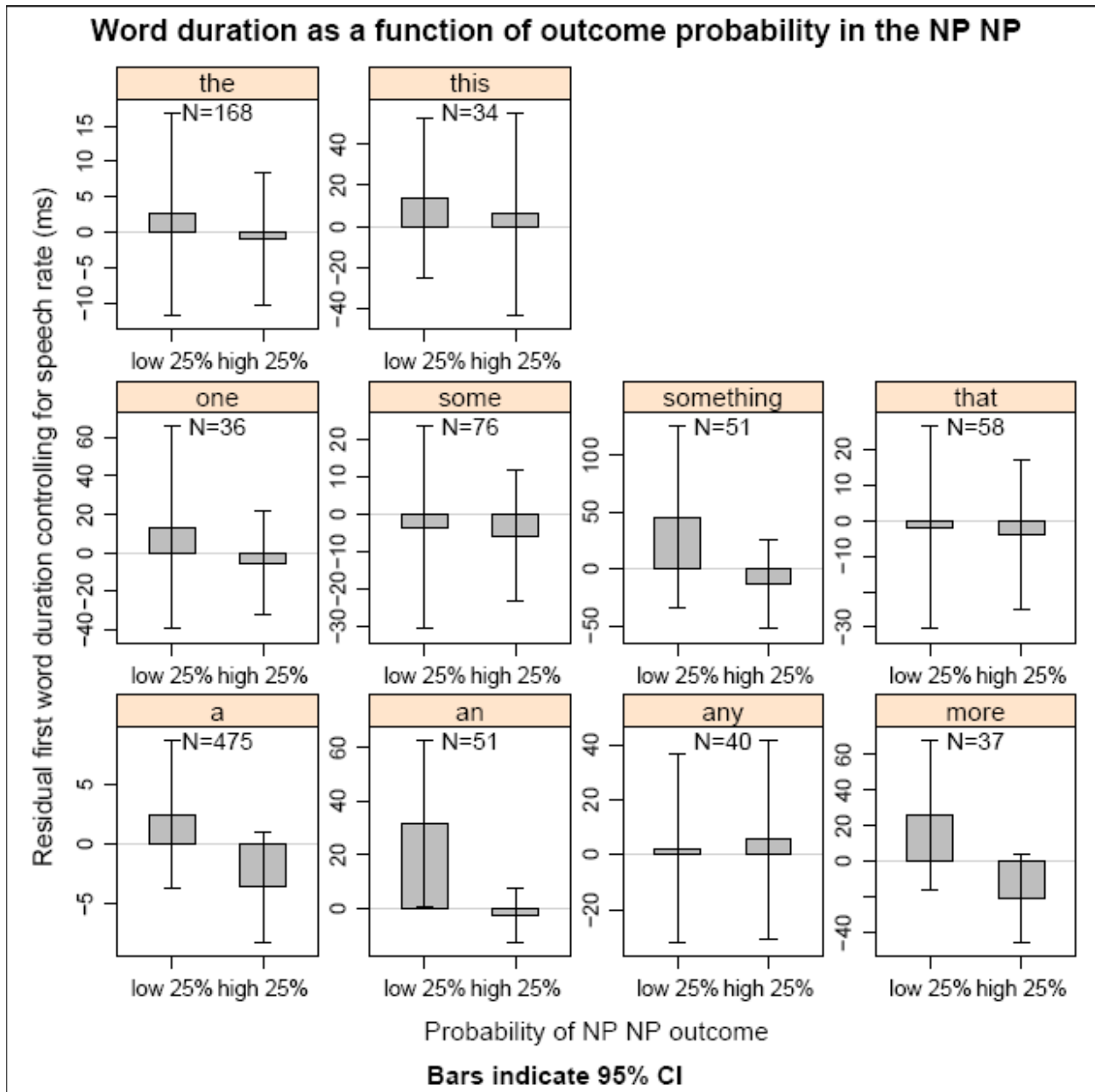


Figure 2

## 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 our 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 fact that 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 appeared 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 may be the result of the entire collection of factors working in concert.

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, and here we show that meaning-equivalent alternants differ in pronunciation, as a function of the syntactic probability.

### **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, David W., Gail McKoon & Roger Ratcliff. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 22.714-35.
- Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to Statistics using R* Cambridge: Cambridge University Press.
- Barlow, Michael & Suzanne Kemmer (eds) 2000. *Usage-based models of language*. Chicago: CSLI.
- Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. in press. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*.
- Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society*



- of America 113.1001-24.
- Bod, Rens, Jennifer Hay & Stephanie Jannedy (eds) 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.
- Brants, Thorsten & Alex Franz. 2006. Web 1T 5-gram. Philadelphia, PA: LDC Data Consortium.
- Bresnan, Joan. 2008. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in search of its evidential base*, ed. by S. Featherston & W. Sternefeld, 75-96. Berlin and New York: Mouton de Gruyter.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. *Cognitive Foundations of Interpretation*, ed. by G. Bourne, I. Kraemer & J. Zwarts, 69-94. Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, Joan & Tatiana Nikitina. 2007. The gradience of the dative alternation. *Reality exploration and discovery: Pattern interaction in language and life*, ed. by L. Uyechi & L.H. Wee. Stanford: CSLI.
- Bybee, Joan & Paul Hopper (eds) 2001. *Frequency and the emergence of linguistic structure* (Typological studies in language, vol. 45. Amsterdam, Netherlands: John Benjamins Publishing Company. (2001).
- Clark, Herbert H. & Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37.201-42.
- Deshmukh, Neeraj, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker & Joseph Picone. 1998. Resegmentation of Switchboard. Paper presented to the International Conference on Spoken Language Processing, Sydney, Australia, 1998.
- Erteschik-Shir, Nomi. 1979. Discourse constraints on dative movement. *Discourse and syntax*, ed. by T. Givon, 441-67. New York: Academic Press.
- Fellbaum, Christiane. 2005. Examining the constraints on the benefactive alternation by using the World Wide Web as a corpus. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, ed. by M. Reis & S. Kepsner, 209-40. Berlin and New York: Mouton de Gruyter.
- Gahl, Susanne & Susan M. Garnsey. 2004. Knowledge Of Grammar, Knowledge Of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language* 80.748-75.
- . 2006. Syntactic probabilities affect pronunciation variation. *Language* 82.405-10.
- Gahl, Susanne, Susan M. Garnsey, Cynthia Fisher & Laura Matzen. 2006. "That sounds unlikely": Syntactic probabilities affect pronunciation. 28th Annual Conference of the Cognitive Science Society. CD-ROM.
- Gahl, Susanne & Alan C. L. Yu (eds) 2006. *Special Issue on Exemplar-based Models in Linguistics* (The Linguistic Review 23 (3).
- Garnsey, Susan M., Neal J. Pearlmutter, Elizabeth Myers & Melanie A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of

- temporarily ambiguous sentences. *Journal of Memory & Language* 37.58-93.
- Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27.2865-73.
- Godfrey, J, E. Holliman & J McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. Paper presented to the International Conference on Acoustics, Speech, and Signal Processing, 1992.
- Green, Georgia. 1971. Some implications of an interaction among constraints. *Papers from the seventh regional meeting*, 85-100. Chicago: Chicago Linguistic Society.
- . 1974. *Semantics and syntactic regularity* Bloomington: Indiana University Press.
- Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg & Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. 65.203-57.
- Hawkins, Jack. 1994. *A performance theory of order and constituency* Cambridge: Cambridge University Press.
- Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. [References]. Bybee, Joan (Ed); Hopper, Paul. (2001). *Frequency and the emergence of linguistic structure*. *Typological studies in language*, vol. 45., 229-54. Amsterdam, Netherlands: John Benjamins Publishing Company.
- Lawless, J. & K. Singhal. 1978. Efficient screening on nonnormal regression models. *Biometrics* 34.318-27.
- Lieberman, Philip. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6.172-87.
- Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118.2561-69.
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. Vienna.
- Schafer, Amy J., Shari R. Speer & Paul Warren. 2005. Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. *Approaches to studying world-situated language use*, ed. by J.C. Trueswell & M.K. Tanenhaus, 209-25. Cambridge, MA: MIT Press.
- Trueswell, John C., Michael K. Tanenhaus & Christopher Kello. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 19.528-53.