Studying syntactic variation using convergent evidence from psycholinguistics and usage

Marilyn Ford
Griffith University

Joan Bresnan
Stanford University

1 Introduction

It is becoming increasingly accepted that speakers have richer knowledge of linguistic constructions than the knowledge captured by their categorical judgments of grammaticality (Bard, Robertson, and Sorace, 1996; Bender, 2005; Bresnan, 2007a,b; Bresnan, Cueni, Nikitina, and Baayen, 2007; Bresnan and Ford, In Press; Bresnan and Hay, 2008; Bresnan and Nikitina, 2009; Chater and Manning, 2006; Gries, 2003a; Manning, 2003). Speakers have reactions to linguistic expressions that are more fine-grained than can be captured by a categorical *yes – no* response or even by a system allowing responses like *good*, ?, ??, * and ** (Bard et al 1996; Bresnan, 2007a; Bresnan et al 2007; Bresnan and Ford, In Press; Featherston, 2007; Gilquin and Gries, 2009). Moreover, expressions that linguists have sometimes categorised as "ungrammatical" have been found to be accepted by people (Bresnan, 2006; Wasow and Arnold, 2005) or found to be used by speakers and to sound good compared to the examples contrived by linguists (Bresnan and Nikitina, 2009; Stefanowitsch, 2007). Many researchers today thus see the need for placing less emphasis on linguists' judgments of grammaticality and more emphasis on usage and experimental data, procedures that would be in line with Labov's (1975) call for the use of convergent evidence and a recognition of the inconsistency of intuitions about constructed examples typically used by linguists.

When judgments along a broad continuum of acceptability are acknowledged, rather than being denied, ignored, or underestimated, a potentially rich set of data about people's knowledge of language becomes available. Moreover, researchers are not forced to make distinctions that have no firm basis. Clearly, though, embracing people's fine-grained judgments about language requires new paradigms for collecting and analysing data. These methods must recognise that judgments vary and must allow for measurement of very fine judgments about expressions. Further, statistical methods to analyse these fine-grained judgments are required. It is also apparent, given the evidence that speakers accept and produce many structures categorised by some linguists as "ungrammatical", that more attention must be paid to methods of analysing the occurrence of expressions.

This chapter considers some of the methodological issues raised by a move away from a reliance on using categorical judgments to investigate linguistic knowledge. It will focus on one concrete case study, the dative alternation, using Australian and US participants. Speakers may produce a dative expression as a double object, as in *showed the woman the ticket,* with the recipient (*the woman*) preceding the theme (*the ticket*), or as a prepositional dative, as in *showed the ticket to the woman*, with the theme now preceding the recipient. Careful analysis of the *occurrence* of the two types of datives, *judgments* about datives, people's *processing* of datives, and people's choices in the *production* of datives, given a context, allows for a rich understanding of the basis for the choice of dative structure. As we shall show, these

methods open up an exciting new approach to studying syntactic variation across macro-regional varieties of language.

2 Analysing the Actual Occurrence of Expressions

2.1 Corpora

There are now many large computer-readable corpora that contain text collected from a variety of written and spoken sources, such as the Switchboard Corpus (Godfrey, Holliman, and McDaniel, 1992), the British National Corpus (Burnard, 1995), and the International Corpus of English (http://ice-corpora.net/ice/). Such corpora can provide a rich source of data for analysing the occurrence of expressions by using multivariable methods of statistical analysis; to gain evidence, for example, about whether certain syntactic structures are used, the relative frequency of their usage, and in what contexts (syntactic, semantic, or social) they are used (Gries, 2003b; Szmrecsányi, 2005; Wasow and Arnold, 2005; Jaeger, 2006; Roland, Elman, and Ferreira, 2006; Bresnan, 2007; Bresnan et al., 2007). The web is also sometimes used as a source of information about occurrence (Fellbaum, 2005; Keller and Lapata, 2003; also, see Hundt, Nesselhauf, and Biewer, 2007). It is particularly useful when the data on the structures of interest are rare (see Keller, Lapata, and Ourioupina, 2002). Thus, for example, Bresnan and Nikitina (2009) used the web to show that certain dative forms considered as "ungrammatical" actually occur, such as "she muttered him a hurried apology" and "… a kind few (three to be exact) came forward and whispered me the answer". Similarly, Stefanowitsch (2007) used the web to show that there are many instances of double object datives with the verb "donate" which would traditionally be classified as "ungrammatical". Thus, examples such as the following are quite easy to find on the internet: "If anyone would like to donate me a couple of million pounds in order for me to do that, that'd be great." Fellbaum (2005) also emphasizes the occurrence of so-called "ungrammatical" structures and argues that there is "a need to substitute or augment constructed data to avoid theoretical biases and capture the full range of rule-governed linguistic behavior" (p. 209).

As Bresnan et al (2007) have noted, however, many linguists put forward objections to the interpretation and thus relevance of "usage data" from corpora for theories of grammar. By using modern statistical theory and modeling techniques, Bresnan et al show not only that these objections are unfounded, but that the proper use of corpora can allow the development of models that solve problems previously considered too difficult, such as predicting the choice of dative structure given many possible contributing factors. The type of corpus model developed by Bresnan et al (2007) not only provides a solution to a linguistic problem, but it can in turn be used to further investigate judgments of language (Bresnan, 2007a; Bresnan and Ford, In Press). Interestingly, the nature of the corpus model is such that when it is used as a basis for further investigating judgments, it calls out for fine-grained data to be collected. The model gives quantitative values for the relative importance of different variables in influencing whether a dative is realised as a double object or a prepositional dative. Given a set of variables, the model yields the probability of a prepositional dative, and hence the probability of the alternative double object dative.

2.2 Developing the Corpus Model

While there are many variables, or one might say predictors, that influence the form of a dative, the dative alternation is characterized by just two forms. If one wants to assess the influence of each predictor in determining the dative form and to give the probability of a form given a set of predictors, then one needs to be able to control simultaneously for multiple predictors, so that the role each predictor exerts by itself can be determined. Generalized mixed effects modeling, sometimes termed multilevel or hierarchical regression, is one of the techniques most suitable for this task (Baayen, Davidson, and Bates 2008; Pinheiro and Bates, 2000; Quené and van den Bergh, 2004, 2008; Richter, 2006). The free statistical software environment R (www.r-project.org) includes various libraries that allow different types of regression modeling to be done relatively easily. In doing this modeling, the researcher is essentially attempting to capture the structure of the data. An initial model can be specified as a formula stating that the *response* (for example, double object or prepositional dative) is a function of a set of possible *predictors*. Predictors can then be eliminated from the model when they are shown to be having no effect in the model, leaving only those that have an influence. The model may include fixed effects and random effects. Participants and items are typical examples of random effects. These effects are sampled from a larger population over which the experimenter wishes to generalize. Random effects are thus not usually of linguistic interest; they are included in the formula so that the non-independence of multiple responses from the same speakers and to the same items can be modeled and controlled for. Pronominality, animacy, probability of occurrence, variety of English, and gender are examples of fixed effects. Typically, fixed effects include all or a large range of possible values of the effect (such as male/female or probability of occurrence) or levels selected by an non-random process (such as American English and Australian English).

To develop their model, Bresnan et al (2007) developed a database of 2360 instances of datives from the three-million word Switchboard corpus of American English telephone conversations (Godfrey et al 1992). Their initial model incorporated fourteen variables considered likely to influence the choice of dative form. These were the fixed effects. They also annotated each instance for the verb sense used; for example, “give in the communication sense”, “give in the transfer sense”, “pay in the transfer sense”, “pay in the abstract sense”, “cost in the prevention of possession sense”, “charge in the prevention of possession sense”, “owe in the future transfer of possession sense”, or “owe in the abstract sense.” Verb sense was a random effect; the verb senses were effectively random samples from a larger population.

By performing a series of analyses, Bresnan et al were able to show that the influence of a number of explanatory predictors remained even when differences in speakers and in verb senses were taken into account and that the predictors each contribute to the response without being reducible to any one variable such as syntactic complexity (cf. Hawkins 1994, Arnold et al 2000). Further, having developed the model on the basis of all 2360 instances from the Switchboard corpus, Bresnan et al determined how well the model generalized to unseen data by performing iterations of a training and testing sequence on different random samples of the data. They found that the model predicted choice of dative form with an average of 94% accuracy for the unseen data, compared with a possible 79% accuracy if a double object dative is

always predicted. They also tested the model with quite a different corpus, the Treebank Wall Street Journal (Marcus, Santorini, and Marcinkiewicz (1993), which has a smaller percentage of double object datives than the Switchboard corpus (62% compared with 79%). The model predicted choice of dative form in the Wall Street Journal corpus with 93% accuracy. The model could make these accurate predictions with quite different corpora on the basis of differences in the occurrence of predictors in the input data to the model – for example, there is a greater presence of longer theme arguments in written than spoken English leading to less frequent double object constructions in the Wall Street Journal corpus.

Bresnan and Ford (In Press) refit the model using 2349 instances in a corrected time-aligned version of the database derived from a resegmentation project (Deshmukh, Ganapathiraju, Gleeson, Hamaker, and Picone, 1998). They also made some improvements to the model and used newer software available in R for mixed effects modelling. The model formula resulting from the modeling of datives in the Switchboard corpus shows the effects of different predictors and how the probability of a prepositional or a double object dative is derived. A better understanding of the model can be gained by considering the formula. The parameters were estimated by the glmer algorithm of the lme4 library in R (Bates, Maechler, and Dai, 2009) . In using the glmer algorithm in R, the dependent variable is given as being a function of a list of possible predictors, as in *Response ~ Predictor 1 + Predictor 2 + ... Predictor n*. The initial glmer formula used by Bresnan and Ford (In Press) is given in (1).

(1) glmer(DativeForm ~
PronominalityRec + PronominalityTheme +
DefinitenessRec + DefinitenessTheme +
AnimacyRec + AnimacyTheme +
NumberRec + NumberTheme +
AccessibilityRec + AccessibilityTheme +
PersonRec + PersonTheme +
ConcretenessTheme +
PreviousDative +
LogLengthDifferenceBtnRecTheme +
(1|VerbSense), family = "binomial",
data = corpusdata)

It can be seen that the possible predictors of dative form were pronominality, definiteness, animacy, number, accessibility, and person of the recipient and theme, as well as concreteness of the theme, type of the nearest preceding dative, and the difference between the log length of the recipient and the log length of the theme. Predictors where the magnitude of the estimated coefficient was found to be less than the standard error were eliminated. The resulting model formula of Bresnan and Ford, with the values obtained for each predictor, is given in Figure 1.

Probability of the prepositional dative = $1 / 1 + e^{-(X\beta+u_i)}$
where

$\hat{X\beta}$ = 1.1583
−3.3718{pronominality of recipient = pronoun}
+4.2391{pronominality of theme = pronoun}
+0.5412{definiteness of recipient = indefinite}
−1.5075{definiteness of theme = indefinite}
+1.7397{animacy of recipient = inanimate}
+0.4592{number of theme = plural}
+0.5516{previous = prepositional}
−0.2237{previous = none}
+1.1819 · [log(length(recipient)) − log(length(theme))]
and $\hat{u}_i \sim N(0, 2.5246)$

Figure 1. The model formula for datives

The coefficients of the model formula show that the predictors either increase or decrease the likelihood of a prepositional dative; positive values indicate greater likelihood of the prepositional rather than the double object dative, while negative values indicate less likelihood of the prepositional dative. Each of the predictors contributes significantly to the model quality of fit, except for definiteness of the recipient, which is trending in that it just fails to reach significance[1]. The parameter $\hat{u}_i$ refers to the random effect of verb sense. Each verb sense has a positive or negative tendency to being expressed with a prepositional dative construction. Thus, for example, the model yields a random effect adjustment of –0.1314 for "give in the transfer sense", showing a slight tendency against the prepositional form, while for "sell in the transfer sense" the model yields a random effect adjustment of 1.5342, indicating a relatively strong bias to the prepositional object form. The mean of the verb sense random effects is approximately 0 and the standard deviation is 2.5246. It is quite easy to see how the model formula yields a prediction of the probability of a dative appearing in the prepositional form. Consider (2), which is an example from the Switchboard corpus, with the observed dative in italics followed by its possible alternative.

(2) *Speaker:*
I'm in college, and I'm only twenty-one but I had a speech class last semester, and there was a girl in my class who did a speech on home care of the elderly. And I was so surprised to hear how many people, you know, the older people, are like, fastened to their beds so they can't get out just because, you know, they wander the halls. And they get the wrong medicine, just because, you know, the aides or whatever just *give them the wrong medicine*/ give the wrong medicine to them.

For this example, the recipient is a pronoun, there is no previous dative and there is a recipient-theme length difference of log(1) – log(3), that is –1.0986. Using this information and the values given by the model formula, the probability of the dative being given as a prepositional dative can be calculated. Thus $X\beta$ = 1.1583 – 3.3718

[1] An effect is considered significant if the probability of it occurring by chance alone is < 0.05, that is, $p < 0.05$.

– 0.2237 + (1.1819 x –1.0986) = –3.7356. The verb sense is "give in the transfer sense", for which the random effect was –0.1314, which must be added to –3.7356, yielding –3.867. We thus have $1/(1 + e^{-(-3.867)}) = 0.0205$. Thus, there is a very low probability of a prepositional dative (0.0205), while the probability for a double object dative is very high (0.9795).

2.3 Implications of the Model

The model predicts probabilities of a dative form based on information about the distribution of the predictors in a corpus of spontaneous language production. If we think of the model as representing people's experience of occurrences of the constructions, the question arises whether people can similarly predict probabilities of occurrence of construction types based on this kind of information. Given that the model, rather than simply giving a binary preference, actually yields the probability of a particular form on a continuous scale form 0 to 1, investigations of judgment data should go beyond yes/no intuitions to finer-grained data. So, how can fine-grained judgments be measured so that the effects of predictors can be seen and how can the data be statistically analysed?

3 Intuitions about Datives

3.1 Methods for measuring people's intuitions

3.1.1 Tasks Bard et al (1996), who were interested in acceptability judgments, argued that valuable information is lost if people give judgments on a small scale, even a 5 or 6 point scale. They suggested that a task used in psychophysics, the magnitude estimation task, could be adapted to allow fine-grained measurements of acceptability. The task has been used in a number of linguistic studies by several researchers (e.g. Featherston, 2005; Hemforth, Konieczny, Seelig, and Walter, 2000; Keller and Sorace, 2003; Sprouse, 2009). Basically, in a magnitude estimation task, one stimulus (such as a sentence) is given a numerical value (either by the experimenter or the participant) and then for each subsequent stimulus the participant is asked to give a number that relates on a magnitude scale to the initial number; for example, if the initial value is 100 and a participant considers another sentence to be twice as acceptable, they would give it the value of 200.

A variant of the magnitude estimation task was developed by Featherston (2007, 2008, 2009) who wanted to allow participants to use a linear scale due to concern that when people use the magnitude estimation task for linguistic judgments they are not in fact able to use a magnitude scale. Featherston (2009) noted that the pattern of data in linguistic magnitude estimation studies does not suggest a magnitude scale and that there is a preference to give responses near zero. In Featherston's task, participants are not instructed to give responses in terms of magnitudes and they are given two reference points, 20 (associated with a bad instance) and 30 (associated with a good instance). Participants are told they can give values below 20 or above 30. The 20 and 30 points are thus meant to be like 0 and 100 in the celsius scale; hence the name thermometer judgment is given to the task.

Neither the magnitude estimation nor the thermometer judgment task seems suitable for investigating whether people's intuitions indicate they have implicit linguistic knowledge of the influence of predictors and the likelihood of one form compared with another. A method that will yield values like probabilities for the alternatives is needed. Rosenbach (2002, 2003, 2005) used a task where participants were asked to choose which alternative construction of a genitive sounded better as a continuation of a given text. This task is not suitable because it yields only binary responses. However, Bresnan (2007a) modified the task to allow responses from 0 - 100. In this task, which we will call the 100-split task, participants rate the naturalness of alternative forms as continuations of a context by distributing 100 points between the alternatives. Thus, for example, participants might give pairs of values to the alternatives like 25-75, 0-100, or 36-64. From such values, one can determine whether the participants give responses in line with the probabilities given by the model and whether people are influenced by the predictors in the same manner as the model.

3.1.2 Items It is customary in psycholinguistics, and to a large extent in linguistics, to devise artificial examples for study. Sentences are usually presented without context and are designed in sets containing instances that vary in some way that is of interest to the investigator. Typically, for psycholinguistic studies, multiple sets will be constructed in order to do statistical analyses. Thus, for example, if the interest is in the effect of animacy of a theme in prepositional and double object datives, then multiple sets of four sentences like the following might be devised: "The salesman brought the customers to the manager", "The salesman brought the brochures to the manager", "The salesman brought the manager the customers", and "The salesman brought the manager the brochures." These items are constructed to vary by two factors, animacy of the theme and dative form, each with two levels and thus being suitable for a 2x2 factorial study. Psycholinguistics is today still largely carried out with the creation of multiple sets of instances like these where the items are developed to differ in limited ways, reflecting the factors in which the investigator is interested In fact, this led Myers (2009) to develop software, Minijudge, that will create multiple sentence sets for linguists who wish to perform small-scale "experiments". When psycholinguists develop such sets they typically go on to analyse the data using analysis of variance.

Carefully constructing sets is a way of attempting to control for variables of no interest to the researcher, such as word length and word frequency, as well as variations in syntactic structure and in semantics. The aim is to vary the sentences just by the factors to be tested in the analysis of variance. However, as Roland and Jurafsky (2002, p.327) have noted, even "seemingly innocuous methodological devices" such as beginning each sentence with a proper name followed by a verb can introduce unrecognized factors which may influence results. Constructing items may also lead to experimenter bias, with experimenters unconsciously constructing items that favour their hypothesis (Forster, 2000; Baayen, 2004). Moreover, as Roland and Jurafsky note in reference to isolated sentences, " 'test-tube' sentences are not the same as 'wild' sentences" (2002, p. 327). The sentences we deal with in every day life are 'wild' and it would be advantageous to be able to study them. Fortunately, there are now sophisticated statistical techniques that allow for the use of less constrained items and that allow one to determine the independent effects of multiple possible predictors. As Baayen (2004, p. 8) notes, with more sophisticated

techniques the items studied can be "random samples, instead of the highly non-random samples of factorial studies."
Consider (3), also from the Switchboard corpus.

(3) *Speaker A:*
We just moved here from Minneapolis and to get the very nice townhouse that we're in, the property management firm that was representing a husband and wife, owners, who had never done this before, asked us for an astounding amount of information and we really didn't have the same opportunity, you know. And I guess that's when I also get upset that if you're going to do it then I want to do it too.
Speaker B:
Yeah, exactly.
Speaker A:
In terms of credit, we're also going through adoption now, and I mean after we *gave our fingerprints to the FBI*/ gave the FBI our fingerprints
Speaker B:
My God.
Speaker A:
you look at each other and say, Well, it's too late now.

This example differs in multiple ways from (2), as shown in Table 1.

Table 1. Comparison of items 2 and 3

Predictor	Item 2	Item 3
pronominality of recipient	pronoun	nonpronoun
pronominality of theme	nonpronoun	nonpronoun
definiteness of recipient	definite	definite
definiteness of theme	definite	definite
animacy of recipient	animate	animate
number of theme	singular	plural
previous	none	none
log length difference	-1.0986	0
verb sense	give–transfer	give–abstract
Probability of prepositional dative	0.0205	0.5234

Examples like (2) and (3) are very different from examples of items typically used in psycholinguistic experiments because they differ in multiple ways in regard to the variables of interest. However, with new statistical techniques, such examples can be used not only to determine whether ratings that participants give in a 100-split task are in line with the probabilities given by a corpus model, but also to determine whether people are influenced by the predictors in the same manner as the model. Thus, Bresnan (2007a) and Bresnan and Ford (In Press) were able to use 30 items from the Switchboard corpus randomly chosen from throughout the probability range as items for their psycholinguistic studies. All participants responded to all items.

3.2 Statistically analysing intuitions

For many years in psycholinguistic studies, if a researcher wished to determine whether there is an effect of particular factors, sets of items would be constructed carefully, as indicated, and two analyses of variance would be performed. For example, to determine whether there is a significant effect of animate versus inanimate themes in prepositional versus double object datives for *participants*, averaging over items would occur, with the mean values per participant per condition being obtained. To see whether there was a similar effect for *items*, averaging over participants would occur, with the mean values per item per condition being obtained. Sometimes the two results would be combined, producing a quasi-F statistic, in an attempt to determine whether results could be generalized simultaneously over participants and items (Clark, 1973). It is now recognized that mixed-effects modelling is a better alternative to these traditional analyses (Baayen, Davidson, and Bates, 2008; Quené and van den Bergh, 2008; Jaeger, 2008). Mixed-effects modeling can cope well with crossed random effects, typical in psycholinguistic studies, where all the participants (one random effect) respond to items (another random effect) in all conditions. With mixed-effects modelling with crossed random effects, researchers can simultaneously consider a large range of possible predictors, fixed or random, that could help in understanding the structure of the data. There is great flexibility and less potential for loss of power compared with the traditional methods.

Thus, just as mixed-effects modelling can be used to model corpus data, mixed-effects modelling can be used to analyse psycholinguistic data. A large range of possible predictors of responses can be considered simultaneously and each item given to participants can vary in many ways in regard to the set of predictors. As shown in (1), the glmer algorithm (with family = "binomial") was used to fit a mixed effects model to the corpus data. This was because there were only two types of "responses" to the choice of constructions in the dative alternation: double object or prepositional dative. However, with fine-grained, quantitative responses, the lmer algorithm in the lme4 package in R is suitable for mixed-effects modelling with crossed random effects. Thus, Bresnan (2007a) and Bresnan and Ford (In Press) used the lmer algorithm to analyse the data from participants responding with the 100-split task to 30 Switchboard corpus examples randomly sampled from throughout the probability range, as determined by the corpus model.

3.3 Correspondence between a corpus model and intuitions

3.3.1 Corpus probabilities Consider (2) and (3) again. According to the corpus model, for item (2) the probability of the dative being given in the prepositional form is 0.0205, while for item (3) it is 0.5234. If the corpus model captures people's knowledge of language and if the 100-split task taps knowledge of probability of production, then it would be expected that when people use the task, they would give a low value to the prepositional continuation in (2) and a higher value to the prepositional continuation in (3), representing a more even probability of occurrence. People will differ in how they distribute their values between prepositional and double object datives for the items. They will differ in their *baseline*, with their mean ratings for prepositional datives varying. They will also differ in the *range* of their ratings, with some using more of the possible ratings range than others, leading to a steeper regression line for such participants. With the lmer algorithm, these two factors can

be included in the model as random effects. It is also the case that certain verbs would have more of a bias toward the prepositional dative. Notice that in the corpus model, verb sense was used as a random effect. With 2349 items, it was quite feasible to use verb sense as a random effect even though there were 55 verb senses. With only 30 items in the ratings study, it seems more appropriate to use verb as the random effect, and in fact analyses showed that the model for the ratings data with verb as a random effect complies with assumptions of model fitting better than a model using verb sense. Using the three random effects considered and corpus probability as the fixed effect, the appropriate formula to do the modeling with lmer is the one given in (4).

(4) lmer(ParticipantRatings ~
CorpusProbs +
(1|Verb) + (1|Participant) + (0 + CorpusProbs | Participant),
data = usdata)

The lmer algorithm would now model the obtained data making adjustments for differences in verb bias towards a prepositional dative, participant differences in their baseline, and participant differences in the range of probabilities they give. Bresnan (2007a) obtained data from 19 US participants. Using the formula in (4) with the data from these participants leads to a model which gives the relationship between corpus probabilities and ratings illustrated in Figure 2.

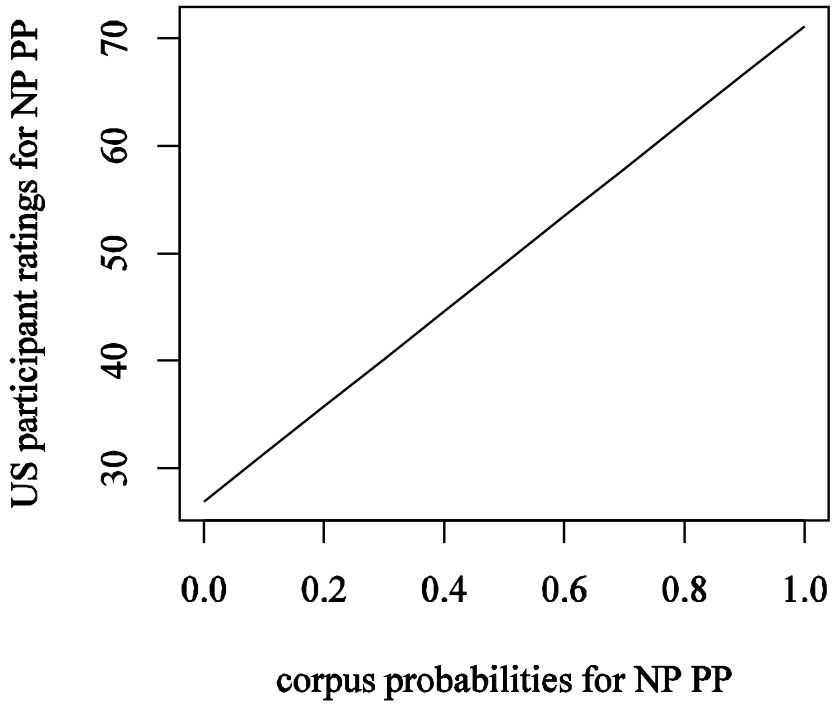

Figure 2. Model relationship between corpus probabilities and participant ratings

It can be seen that the model defines a linear relation between the corpus probabilities and participant ratings. The model estimate for the intercept is 26.910. The intercept is the value of the dependent variable (the ratings in this case) if all predictors are 0. Notice that this is the point where the line in Figure 2 would cross the y-axis. The model estimate for the fixed effect of corpus probability is 44.216. This is the model estimate of the change that the fixed effect can make to the intercept. Notice that 26.910 + 44.216 = 71.126, and that this is the point where the line in Figure 2 extends when the corpus probability is 1. The adjustments for random effects can be explained by considering the plot, together with the model adjustments for verb, participants, and the interaction of corpus probability and participant. The random effect adjustments are given in Table 2.

Table 2. Random effect adjustments of US participants, using (4)

verb		participant		corpus probability | participant	
bring	-11.937	1	10.014	1	-15.074
give	2.821	2	4.174	2	8.003
owe	-11.935	3	-3.143	3	-2.494
pay	12.716	4	7.520	4	-10.109
sell	3.678	5	-8.331	5	10.887
show	-0.391	6	-8.497	6	0.929
take	9.644	7	-1.773	7	4.045
teach	5.099	8	-3.138	8	10.961
tell	-9.695	9	-6.090	9	7.489
		10	3.674	10	8.673
		11	-1.413	11	-1.105
		12	-2.541	12	17.351
		13	8.873	13	-9.231
		14	3.429	14	-6.926
		15	-4.977	15	-1.941
		16	-2.461	16	-2.890
		17	8.169	17	-6.321
		18	-7.089	18	2.296
		19	3.601	19	-14.543

Notice that some adjustments are positive and some negative. The plot in Figure 2 shows the relationship between corpus probabilities and ratings assuming a verb and a participant with a 0 random effect adjustment. Also, for any new item with a new verb, 0 would be assumed for the random effect of verb since the random effects are assumed to be normally distributed with a mean of 0. For any new participant, 0 would assumed for the random effects of participant and corpus probability interacting with participant. Thus, if the corpus probability calculated for this new item is .30, the model would predict the following "rating" by a new participant: 26.910 + 44.216*.30 + 0 + 0 + 0 = 40.17. However, the adjustments show how the model adjusts for each known verb and participant. Notice, that the verb *bring*. for example, has an adjustment of –11.937; compared with the other verbs, it has a strong bias against the prepositional dative form. This means, effectively, that the model

recognizes that for an item with a corpus probability of .30 and the verb *bring* the actual rating for a new participant would be expected to be 26.910 + 44.216*.30 -11.937 + 0 + 0 = 28.24. In contrast, for an item of .30 probability with the verb *take*, which has a relatively strong preference for the prepositional dative form, the actual rating would be expected to be 26.910 + 44.216*.30 + 9.644 + 0 + 0 = 49.82.

Now let's consider a known participant, participant 1. The *participant* adjustment is 10.014 and the *corpus probability|participant* adjustment is –15.074. Compared to other participants, this participant has a greater preference for prepositional datives over double object datives and shows less of an increase in ratings for prepositional datives as probability increases. For an item of .30 probability with the verb *bring* the expected rating for participant 1 given the model is: 26.910 + 44.216*.30 – 11.937 + 10.014 – 15.074*.30 = 33.73. Compare this with participant 5: 26.910 + 44.216*.30 – 11.937 – 8.331 + 10.887 *.30 = 23.17.

It is apparent from Figure 2 that the relationship between corpus probabilities and the ratings that participants give to the items in the 100-split task is one where, in general, ratings increase as probabilities increase. The significance of the relationship can be found using the languageR package and the pvals.fnc function in R. The results show that there is a significant relationship between corpus probabilities and participant ratings, with $p = 0.0001$. Bresnan and Ford (In Press) also obtained data from 20 Australian participants living in Australia. They were given the same 30 items but, where necessary, place names, spelling, and atypical lexical items in the context passages were changed; for example, for (3), *Minneapolis* was changed to *Sydney*. Using the lmer algorithm with the Australian data shows that there is a significant relationship between corpus probabilities and ratings of the Australians, $p = 0.0001$.

The lmer equation in (4) uses the corpus probabilities, obtained from $1/(1 + e^{-(X\beta)})$, as a fixed effect. An alternative is to use $X\beta$, which is the log odds of a prepositional dative to a double object dative. That is, it is equivalent to log (P/(1-P)), where P is the probability of a prepositional dative. The log odds are often preferred over probabilities in regression analyses because, unlike probabilities, they are not bounded: they range from – infinity (as the probability approaches 0) to + infinity (as the probability approaches 1), with a zero on the scale meaning that two alternatives have the same likelihood of occurrence. Given that the regression analysis yields a linear function, which is inherently unbounded, it is best to use an unbounded scale. Replacing the corpus probabilities with the corpus log odds does not change the significance of the results but, with this transformation, the model will better be able to characterize the relationship of interest with a straight line. In line with Bresnan et al. (2007) and Bresnan and Ford (In Press), we will use the corpus log odds in later analyses in this paper.

The fact that people give ratings in line with the probabilities obtained from the corpus model suggests that people are sensitive to the variables that influence the choice of dative form. Just as mixed effects modelling with crossed random effects can be used with corpus probabilities as a fixed effect, so too can it be used with the different predictors of dative form.

3.3.2 Corpus predictors and variety of English If people are sensitive to the variables that influence the choice of dative form, it may be that this sensitivity is due to their experience with the actual occurrence in usage of prepositional and double object datives in the context of the presence or absence of the different predictors. An interesting possibility emerges. People growing up speaking English in different countries might have been exposed to subtly different experiences and thus might show somewhat different effects for the various predictors. To determine the effects of the different predictors for two varieties of English, mixed effects modelling can be carried out using the predictors of dative form from the corpus model as fixed effects, with variety interacting with these predictors. Random effects of verb, participant, and corpus log odds interacting with participant can be included. Predictors that are found to not be influencing the model can be eliminated. The final lmer model is given in (5).

(5) lmer(ParticipantRatings ~
Variety*LogLengthDiffRTh +
PronominalityRec + PronominalityTheme +
DefinitenessRec + DefinitenessTheme +
AnimacyRec +
NumberTheme +
PreviousDative +
(1|Verb) + (1|Participant) +
(0 + CorpusLogOdds | Participant),
data = ratingsdata)

The resulting model parameters, together with the *p*-values and associated 95% confidence limits are given in Table 3.

Table 3. Model parameters for the ratings experiment

Fixed effects:

	Estimate	95% Confidence Limits		*p*-values
		lower	upper	
(Intercept)	50.251	39.817	61.175	0.0001
variety = Aus	-1.802	-5.270	1.956	0.3176
log rec-theme diff	3.406	-0.904	7.891	0.1114
previous *to*-dative	11.032	4.625	15.767	0.0004
recipient = pronoun	-16.791	-22.096	-10.237	0.0001
theme = pronoun	14.445	5.974	23.351	0.0010
theme = indefinite	-25.800	-29.779	-20.730	0.0001
recipient = indefinite	16.304	10.242	22.033	0.0001
recipient = inanimate	21.609	15.475	28.489	0.0001
number of theme = singular	11.889	7.162	15.797	0.0001
variety = Aus:log rec-theme diff	3.224	0.383	6.168	0.0378

The 95% confidence limits give the range of values within which the true value for the whole population is likely to be, with 95% surety. The given estimates fall within the 95% confidence limits. Notice that variety and log recipient length - log theme

length are not significant ($p > .05$) and that the confidence limits for these cross zero, in that they range from a negative to a positive value. The estimates for these are thus unstable and cannot be interpreted. However, their interaction is significant. The positive value (3.224) shows that as the length of the recipient increases relative to the theme, the Australians increasingly favor the prepositional dative compared to the US participants. Thus, the Australians show a length effect that is in line with the corpus model (see Figure 1), while the US participants show no such effect. The recipient being a pronoun and the theme being indefinite lead to a preference for the double object dative (as in the corpus model). The previous dative being a prepositional dative, the theme being pronoun, and singular, and the recipient being indefinite, and inanimate, all lead to a preference for the prepositional dative, again in line with the corpus model. Given that ratings increase as corpus probabilities increase, then the predictors that contribute to the probabilities are, as expected, also related to ratings. However, the interaction between variety and length difference between the arguments shows that some predictors may be more important to some people than others. In this case, the Australians have a greater preference than the US participants for long recipients to be after a shorter theme, with the preference increasing as the recipient gets comparatively longer.

The results of analysing the occurrence of datives in a large corpus and then analysing people's intuitions about datives with the 100-split task and mixed effects modeling suggests that people are sensitive to the variables that influence choice of dative and that this sensitivity may be due to their experience with the actual occurrence in usage of alternative dative forms in the context of the presence or absence of different predictors. It is clear that obtaining people's fine-grained intuitions can capture complex, meaningful data that should not be ignored. However, intuitions that are given after a sentence has been read, do not show whether online processing has been affected. While some linguists might be particularly interested in judgments, psycholinguists are typically interested in online processing. Of course, if one can obtain converging evidence from different sources, then understanding is increased. As Ferreira (2005) and Kaplan (2009) both lament, while there was great promise of advancement due to the emergence of joint linguistic and psycholinguistic work in the 1980s, the two fields gradually separated again. With the increasing acceptance that speakers have greater knowledge of language than knowledge captured by categorical judgments of linguistic expressions, there is an increasing acceptance that there is a need to look for converging evidence (Wasow and Arnold, 2005; Arppe and Järvikivi, 2007; Bresnan, 2007b; Gilquin and Gries, 2009.). Thus, Bresnan and Ford (In Press) studied the online processing of datives, with both US and Australian participants. They were interested in reaction times to the word *to* in prepositional datives as a function of the predictors in the corpus model and variety of English. This study was inspired partly by the work of Tily, Gahl, Arnon, Snider, Kothari, and Bresnan (2009) showing that durations of the pronunciation of the word *to* in prepositional datives varied as a function of the corpus model probabilities.

4 Online Processing of Datives

4.1 Methods for investigating online processing

4.1.1 Tasks One task used often in psycholinguistics to obtain data during sentence processing is the self-paced reading task using a "moving window" display (Just,

Carpenter and Woolley 1982). In this task, lines of dashes first appear on a computer screen in place of words of a sentence. When a participant presses a space bar the first word of the sentence appears. When the space bar is pressed again, that word is replaced by dashes again and the next word appears. Thus, the participant reads the sentence at their own pace with each word being revealed in its correct position as the space bar is pressed. Sometimes more than one word is presented at a time. Typically, when the last word has been read and the space bar pressed, a simple yes-no question to test comprehension of the sentence appears. Reaction times from the appearance of a word (or segment) to the pressing of the space bar are recorded in milliseconds. To run the task, experimenters sometimes use open source software called Linger, written by Doug Rohde (Warren and Gibson, 2005; Fedorenko, Gibson, Rohde, 2006; Wagers, Lau, and Phillips, 2009). Sometimes commercial software, such as E-prime (Schneider, Eschman, and Zuccolotto 2002a,b), is used (Swets, Desmet, Clifton, and Ferreira, 2008; Pickering, McElree, Frisson, Chen, and Traxler, 2006; Hwang and Schafer, 2009).

Forster, Guerrera, and Elliot (2009) have noted that participants doing the self-paced reading task may start to respond at a constant rate which could lead to a lessening in sensitivity of the reaction times to the material being presented. They also suggested that, in this task, participants may delay some of the processing of the word until after they have pressed the button for the next word. Forster et al used a different task, the maze task, in which participants are presented with two words at a time, one of which is a continuation of the preceding sentence fragment and the other not. The participant presses a key to indicate whether the left or the right word is the correct word for a continuation. Thus, for example, a participant might receive *"The ..."* followed by *"gone dog"* followed by *"chased sink"* followed by *"our hosed"* followed by *"into. cat."*. Alternatively, they might be presented, for example, with *"The ..."* followed by *"blung dog"* followed by *"chased nene"* followed by *"our chis"* followed by *"denant. cat."*. When a key is pressed the reaction time is recorded and the next pair of words is presented. This task had also previously been used by Freedman and Forster (1985) and Nicol, Forster, and Vereš (1997).

Ford (1983), too, suggested that participants in a self-paced reading task might start to respond rhythmically at a steady pace, lessening the sensitivity of reaction times. She also proposed an alternative task. To prevent participants getting into a rhythm of pressing the space bar to get the next word, Ford suggested the Continuous Lexical Decision (CLD) task, where participants are presented with a sentence one word at a time, but where they must press a "yes" or "no" button depending on whether the "word" is a real word or a non-word. The presentation is the same as the "moving window" self-paced reading task, though Ford does not put spaces between the dashes representing words; rather there is a long continuous line of dashes representing a sequence of words. All experimental items contain only words, at least up to the final point of interest. Other, "filler" items contain non-words. Ford showed that the task is sensitive to both semantic and syntactic processing. It seems, on comparing reaction times reported by Ford (1983) and Forster et al. (2009), that reaction times in the CLD task are about half those in the maze task, which is not surprising since participants in the maze task must process two words at each point. To examine online processing of datives, Bresnan and Ford (In Press) used the CLD task and used the E-Prime software to run the experiment. A possible alternative would have been to run an eye-tracking experiment, with eye movements and gaze durations tracked

during reading. Eye-tracking studies have the advantage that participants can read in quite a natural manner, though they require specialised equipment (Carpenter and Daneman, 1981; Dopkins, Morris, and Rayner, 1992; Gordon, Hendrick, Johnson, and Lee 2006; Traxler and Frazier, 2008; Tily, Hemforth, Arnon, Shuval, Snider, and Wasow 2008; Kuperman and Piai, submitted). The CLD task did, however, prove useful in obtaining a measure of processing at the word *to* in prepositional datives.

4.1.2 Items The same items as those used in the 100-split task could be used in the CLD task. However, it is not necessary for participants to read every word making a lexical decision. An example of the initial appearance of an item is given in (5).

(5) *Speaker:*
A lot of women I know now do job sharing. And one of my supervisors, when she went on LOA to have her baby, we hooked up a terminal at her house and we could send her messages, and she kept in touch like that, and basically, just worked out of her house. I would

just ----------

The participants were instructed to read the passage and then make the lexical decision for the word at the beginning of the dashes. For each experimental item, the initial word in the continuation was always the word before the dative verb.

Given that participants are making a lexical decision, there need to be some items with non-words. Bresnan and Ford (In Press) constructed 10 items with a context passage and a continuation, where the continuation contained one or more non-words and was not a dative construction. Also, given that they were interested in reaction times to the word *to*, some experimental items included non-words after that point. Further, 6 of the 30 items from the 100-split task were used as "filler" items. These items were from the middle of the probability range and thus less interesting because they do not show an overall preference for one dative structure over another. They were presented in the double object form and contained one or more non-words. Thus there were 24 experimental items. The continuation of these items up to and including the word *to* was either the same as the item from the corpus or it was the prepositional alternative to the original double object construction. After the last word of each item, both experimental and filler, the context and continuation disappeared and a yes/no question relating to what had just been read appeared. Responses to the comprehension questions showed that participants did comprehend what they were reading and that there was no significant difference in correct responding for the US and Australian participants.

4.2 Analysing reaction times

Notice that at the stage where a participant is reading the word *to* in a prepositional dative they have no information about the recipient. Therefore, the corpus model probabilities and log odds, being based partly on predictors relating to the recipient, are not relevant. New "partial-construction log odds" were calculated by running the glmer algorithm again, but without predictors relating to recipients. The following questions can now be asked: are reaction times to the word *to* related to the corpus

partial-construction log odds? what predictors influence reaction times to the word *to*? and, do these predictors interact in some way with variety of English?

With any reaction time study, there is the chance that some very long reaction times will be due to extraneous influences, such as temporary participant distraction, and there is also the possibility of very short reaction times due to participant error, such as pressing a button twice. Thus extreme outliers should be eliminated. Bresnan and Ford (In Press) eliminated two very long reaction times (5584 and 10156 milliseconds, compared with the closest reaction time of 1496 milliseconds) and one very short reaction time (99 milliseconds, compared to the closest reaction time of 239 milliseconds). It is also common to log reaction times to reduce the effect of extreme reaction times (see Baayen, 2008). Thus, Bresnan and Ford logged reaction times.

To determine the significance of the corpus partial-construction log odds in determining subjects' reaction times to the word *to*, the lmer algorithm can be used with partial-construction log odds as a main effect. Verb and participant can be used as random effects. In the ratings analysis, corpus log odds interacting with participant was also used as a random effect to control for the fact that participants varied in how much of the rating range they used. For the reaction time study, the concern is that participants might show different effects of item order, some perhaps tiring, others perhaps speeding up as they become more confident. Thus, a random effect of item order interacting with participant can be added. There could be other extraneous variables unrelated to the partial-construction log odds that could influence reaction times and such variables can be added to see if the log odds are significant even when these extra variables are added as controls. Controls were added for item order and reaction time to the word preceding *to*. These were added as fixed effects. Item order and the reaction time to the word preceding *to* were centered, so that their central value would be 0. Centering of numerical variables in the fixed effects is often done to enhance interpretability where a zero value otherwise has no real meaning (as in a 0 millisecond or 0 log reaction time or an item order of 0). For numerical variables in the random effects, centering is required to avoid statistical artifacts (Baayen, 2008: pp 254-255). The lmer formula is given in (6).

(6) lmer(LogRTTo ~
PartialLogOdds +
LogRTBeforeToCentered +
ItemOrderCentered +
(1|Verb) + (1|Participant) +
(0 + ItemOrderCentered | Participant)+
data=RTdata)

It was found that the partial construction log odds were a significant predictor of reaction time to the word *to*, with $t = -2.14$, $p = 0.0324$.

To determine what predictors influence reaction times to the word *to* and whether any predictors interact with variety of English the formula in (6) can be modified by deleting the partial log odds and substituting predictors from the corpus model that are unrelated to the recipient, since the recipient occurs after the word *to*. Predictors that are found to have no influence of the model can be deleted. The resulting model

parameters, together with the *p*-values and associated 95% confidence limits are given in Table 4.

Table 4. Model parameters for the reaction time experiment

Fixed effects:

	Estimate	95% Confidence Limits		p-values
		lower	upper	
(Intercept)	5.9998	5.9064	6.0913	0.0001
variety = Aus	0.1098	0.0455	0.1708	0.0008
log.theme length centered	0.1164	0.0820	0.1542	0.0001
theme = indefinite	0.0337	0.0046	0.0632	0.0190
log.RT pre.to centered	0.3378	0.2905	0.3874	0.0001
itemorder centered	-0.0021	-0.0034	-0.0008	0.0026
variety = Aus**:** log.theme length centered	-0.0741	-0.1133	-0.0364	0.0002

All effects are significant. The positive estimates show the following: the Australian participants are slower than the US participants, as theme length increases reaction times to *to* increase, indefinite themes lead to increased reaction times, and as the reaction time before *to* increases reaction times to *to* increase. The length of theme and definiteness of theme effects are both in line with the corpus model. A prepositional dative is less favored where the theme is indefinite and as theme length increases. The negative estimate for item order shows that reaction times decrease as the item order increases. The negative estimate for the *variety* ***:*** *log length of theme* interaction shows that even though increased length of theme leads to an increase in reaction time, the Australians are less influenced by increases in theme length than the US participants.

Given that the Australians have slower reaction times than the US participants, it could be suggested that the *variety* ***:*** *log length of theme* effect is really a *speed* ***:*** *log length of theme* effect. That is, perhaps the Australians happen to be slower at the task than US participants and for some reason slower participants show a smaller length of theme effect. If this were the case then one would expect that if participants are divided into two groups according to whether their mean reaction time was above or below the mean of all participants and this speed factor used to replace variety, then there should be a large *speed* ***:*** *length of theme* interaction. However, it turns out that this is not the case. Table 5 shows the model parameters obtained if the reaction time data is modelled starting with the same initial model as previously considered, but using speed instead of variety.

Table 5. Model parameters for the reaction time experiment with speed not variety

Fixed effects:

	Estimate	95% Confidence Limits		p-values
		lower	upper	
(Intercept)	5.9654	5.8812	6.0503	0.0001
speed = slow	0.1987	0.1486	0.2476	0.0001
log.theme length centered	0.0798	0.0452	0.1148	0.0001
theme = indefinite	0.0336	0.0048	0.0645	0.0254
log.RT pre.to centered	0.3352	0.2859	0.3844	0.0001
itemorder centered	-0.0022	-0.0035	-0.0008	0.0034
speed = slow: log.theme length centered	-0.0012	-0.0406	0.0382	0.9648

It can be seen that there is no significant interaction between speed and length of theme. The *variety : length of theme* interaction is not due to a difference in speed between the two groups.

5 Exploring Differences Between Varieties

The ratings study and the reading study show that both the US and Australian participants are sensitive to the corpus probabilities of dative form and to the linguistic predictors underlying these probabilities. There is now quite a large body of research suggesting that when using language, people are influenced by their past experience with occurrence of constructions (Ford, Bresnan, and Kaplan, 1982; MacDonald, Pearlmutter, and Seidenberg, 1994; Jurafsky, 1996; Tabo, Juliano, and Tanehnaus (1997); Jurafsky and Martin, 2000; Bybee and Hopper, 2001, Bod, Hay, and Jannedy, 2003; Diessel, 2007) and with their wider knowledge of the preceding discourse (Van Berkum, Brown, Zwitserlood, Kooijman, and Hagoort, 2005; DeLong, Urbach, and Kutas, 2005). Apart from studying speakers from two varieties of English to generalize the finding of sensitivity to corpus probability and the underlying linguistic predictors, the inclusion of two varieties of the same language opened up the possibility of finding differences between varieties where there might be exposure to subtly different probabilities of linguistic experience. We have seen that both in the ratings and reading tasks, the US and Australians showed some subtly different effects; but can these be related to exposure to different probabilities of occurrence for some linguistic experience?

First, let's consider the differences found between the two varieties. In the ratings study, as the relative length of the recipient increased, the Australians, but not the US participants, showed an increase in preference for the prepositional dative form V NP PP(LongRecipient). Such behaviour is in accord with the end-weight principle, that is, the strong tendency to place a longer argument after a shorter argument (see Quirk, Greenbaum, Leech, and Svartvik 1972; Wasow, 1997; Wasow and Arnold 2003). It is as though, compared to the Australians, the US participants have more tolerance of the double object form V NP(LongRecipient*)* NP, even though it would go against the end-weight principle. On this finding alone, one might be tempted to suggest that the Australians show a greater end-weight effect; that is, that they, more

than the US participants, need a longer argument to come after a shorter argument. However, the results of the reading study suggest differently. In the Continuous Lexical Decision task, the US participants showed a greater end-weight effect with prepositional datives, showing less tolerance of the prepositional dative form V NP(LongTheme) PP than the Australians: as the length of the theme increased, the US participants slowed down at the word *to* in the PP more than the Australian participants. In fact, even though the US participants in general had faster reaction times than the Australians, when the theme was longer than three words the US participants were found to be slower than Australians at the word *to* in the PP (Ford and Bresnan, In Press). Thus, it seems that the Australians have more tolerance of V NP(LongTheme) PP than the US participants. Table 6 summarizes these apparent differences between the varieties shown in the ratings and reading studies.

Table 6. Comparison of Australian and US participants in tolerance to dative structures with long first arguments

	Variety	
Structure	Australian	US
double object dative: V NP(LongRecipient) NP	less tolerant	more tolerant
prepositional dative: V NP(LongTheme) PP	more tolerant	less tolerant

We see that the Australians are more tolerant of the first argument being long in the V NP PP form, while the US participants are more tolerant of the first argument being long in the V NP NP form. The difference in tolerance of the dative forms when the initial argument is long (hence going against an end-weight effect) suggests that there might be a difference in expectations for the dative alternatives for the two varieties: the expectation of NP PP might be greater for the Australian participants while the expectation of NP NP might be greater for the US participants.

There is evidence that different varieties of English do differ in rates of the alternative dative forms and that such differences are in some cases increasing: Indian English has a higher rate of prepositional dative than British English (Mukherjee and Hoffman 2006), the frequencies of double object constructions in the 19th and 20th centuries have been diverging for British and American English (Rohdenburg 2007); and the overall probability of use of prepositional datives with the verb *give* in New Zealand English has been increasing since the early 1900s (Bresnan and Hay 2008). There is a suggestion that the relative frequency of prepositional datives might be higher in Australian English than US English. Collins (1995) reported a relative frequency of 34.5% for Australian English, while Bresnan et al. (2007) reported a relative frequency if 25% for US English. However, the two datasets are not fully comparable. Collins included both *to* and *for* datives, while Bresnan et al. included only *to* datives. The question, then, is: if comparable datasets are not available for two varieties, how can the relative frequencies for varieties be found?

One possibility might be to do some type of web search using US versus Australian domains. However, it is often unclear who has produced the language. Also, there are many more US websites than Australian, so that the first 1000 Australian hits, for

example, might encompass all Australian occurrences, while the first 1000 US hits might represent a small, non-random, fraction of US occurrences. Since search engines such as Google do not give all results, bias could be introduced by a difference in the number of websites for the two countries. An alternative, used by Bresnan and Ford (In Press), is to develop a dataset by asking participants to complete sentences.

Bresnan and Ford (In Press) gave 20 US and 20 Australian participants, equally balanced for gender, all 30 items that they had used in their ratings study. The participants were given the context and the beginning of the dative, up to and including the dative verb. They were required to write a completion for the final sentence. The best feature of this methodology is that the datasets are produced with predictors being the same up to and including the verb. An analysis of the completions showed that the average level of production of datives was the same for both varieties, being 0.55 for the Australians and 0.56 for the US participants. However, 0.42 of the datives produced by the Australians were NP PP *to*-datives, while for the US participants the corresponding figure was 0.33. A logistic generalized linear model was fitted to the counts of *to*-datives and double-object datives per subject, with variety and gender and their interaction as fixed effects. Variety was significant, with $p = 0.0408$, gender was not significant, with $p = 0.2121$. The interaction of variety and gender did not quite reach significance, with $p = 0.0566$. The nearly significant interaction between variety and gender was due to Australian males being more than three times as likely to produce *to*-datives as the US males, in the same contexts, while the Australian females were more similar to the US participants.

The sentence completion data allowed a simple comparison of two varieties in the absence of comparable corpora of the varieties. It also confirmed the prediction, stemming from the interactions in the ratings and reading studies, that there is a greater preference for NP PP datives amongst Australian participants and a concomitant greater preference for NP NP datives amongst US participants. The ratings and reading studies suggest that speakers of two varieties of English are sensitive to corpus probabilities of dative form and their underlying predictors and that this sensitivity influence ratings of naturalness and reading[2].

6 Conclusion

There are limitations, of course, to the studies presented here. Thus, for example, only two varieties of English were studied and the samples from the varieties were quite small and from small regions. In the future, more participants will be asked to do the sentence completion task so that more detailed analyses can be performed. Also, more items could be used in a variety of tasks to allow further study of more predictors. However, the set of studies considered in this chapter represent a movement away from an emphasis on linguists' judgments of grammaticality and a greater emphasis on analyses of usage, fine-grained judgments given by participants under experimental conditions, and the use of diverse methods, such as ratings, reading, and production tasks to study variation. By using such diverse methods, a rich set of data about people's knowledge of language is obtained. The studies also show how one can explore both similarities and differences between groups by using

[2] See Bresnan and Ford (In Press) for further discussion of the theoretical implications of their findings.

participants speaking two varieties of English and by using modern statistical techniques. They also show that researchers can successfully use items that are much richer than the simple, constructed, items, typically used by linguists and psycholinguists in the past.

It is clear that speakers can give fine-grained judgments that are meaningful. In the 100-split task, they give ratings of naturalness of the alternative dative forms that turn out to be a function of the probabilities of occurrence and associated predictors found in corpus data. The Continuous Lexical Decision task shows that even reading processes are sensitive to probabilities of occurrence and associated linguistic predictors. The sentence completion task proved to be of great value where no comparable corpora for two varieties existed. The rich set of data obtained from diverse methods using participants from two varieties of English and analysed with modern statistical techniques shows that speakers have strong predictive capacities, using their sensitivity to spoken English corpus probabilities to rate naturalness of language, to read, and to produce sentence completions.

References

Arnold, J., Wasow, T., Losongco, A. and Ginstrom, R. 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76: 28–55.

Arppe, A. and Järvikivi, J. 2007. Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory* 3, 99-109.

Baayen, R. H. 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers* 1, 1–45.

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

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

Bard, E. G., Robertson, D., and Sorace, A. 1996. Magnitude estimation of linguistic acceptability. *Language* 72, 32-68.

Bates, D., Maechler, M., and Dai, B. 2009. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-31.

Bender, E. M. 2005. On the boundaries of linguistic competence: matched-guise experiments as evidence of knowledge of grammar. *Lingua* 115, 1579–1598.

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

Bresnan, J. 2007a. Is knowledge of syntax probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base, Series: Studies in Generative Grammar*, 75–96. Berlin and New York: Mouton de Gruyter.

Bresnan, J. 2007b. A few lessons from typology. *Linguistic Typology* 11, 297–306.

Bresnan, J, Cueni, A., Nikitina, T. and Baayen, R. H. 2007. Predicting the dative alternation. In G. Boume, I. Krämer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and Ford, M. In Press. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language.*

Bresnan, J. and Hay, J. 2008. Gradient grammar: an effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118, 245–259. (Special Issue *Animacy, Argument Structure, and Argument Encoding* edited by M. Lamers, S. Lestrade, and P. de Swart).

Bresnan, J. and Nikitina, T. 2009. The gradience of the dative alternation. In L. Uyechi and L. H. We (Eds.), *Reality Exploration and Discovery: Pattern Interacyion in Language and Life*, 161–184. CSLI.

Burnard, L. 1995. *Users guide for the British National Corpus.* British National Corpus Consortium, Oxford University Computing Service.

Bybee, J., and Hopper, P. (Eds.). (2001). *Frequency and the Emergence of Linguistic Structure.* Amsterdam: John Benjamins.

Carpenter, P.A., and Daneman, M. (1981). Lexical retrieval and error recovery in reading: A model based on eye fixations. *Journal of Verbal Learning and Verbal Behavior* 20, 137-160.

Chater, N. and Manning, C. D. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Science,* 10, 335–344.

Clark, H. H. 1973. The language-as-fixed-effect- fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.

Collins, P. 1995. The indirect object construction approach. *Linguistics* 33, 35–49.

DeLong, K. A., Urbach, T. P. and Kutas, M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117–1121.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. 1998. Resegmentation of Switchboard. In *ICSLP-1998. Proceedings of the* 5th *International Conference on Spoken Language Processing, Sydney, Australia*. On-line ICSA archive: http://www.isca-speech.org/archive/icslp 1998/i98 0685.html.

Diessel, H. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25, 108–127.

Dopkins, S., Morris, R. K., and Rayner, K. 1992. Lexical ambiguity and eye fixations in reading: A test of competing models of lexical ambiguity resolution. Journal of Memory and Language 31, 461–476.

Featherston, S. 2005. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua*, 115 1525–1550.

Featherston, S. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics*, 33, 269–318.

Featherston, S. 2008. Thermometer judgements as linguistic evidence. In C. M. Riehl and A. Rothe (Eds.), Was Ist Linguistische Evidenz? Aachen, Shaker.

Featherston, S. 2009. A scale for measuring well-formedness: Why syntax needs boiling and freezing points. In S. Featherston and S. Winkler (Eds.), *The Fruits of*

Empirical Linguistics Volume 1: Process, 47–74. Berlin and New York: Mouton de Gruyter.

Fellbaum, C. 2005. Examining the constraints on the benefactive alternation by using the World Wide Web as a corpus. In M. Reis and S. Kepser (Eds.) *Evidence in Linguistics: Empirical, Theoretical, and Computational Perspectives*, Mouton de Gruyter.

Ferreira, F. 2005. Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review,* 22, 365–380.

Fedorenko, E., Gibson, E., Rohde, R. 2006. The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language* 54, 541–553.

Ford, M. 1983. A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior* 22, 203–218.

Forster, K. I. 2000. The potential for experimenter bias effects in word recognition experiments. *Memory and Cognition* 28, 1109–1115.

Forster, K. I., Guerrera, C., and Elliot, L. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods* 41, 163–171.

Freedman, S. E. and Forster, K. I. 1985. The psychological status of overgenerated sentences. Cognition, 19, 101–131.

Gilquin, G. and Gries, S.T., 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and LinguisticTheory,* 5, 1–26.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, San Francisco, pp. 517–520.

Gordon, P. C.; Hendrick, R.; Johnson, M.; and Lee, Y. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 32, 1304-1321.

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

Gries, S.T. 2003b. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement.* New York / London: Continuum International Publishing Group.

Hawkins, J. A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hundt, M., Nesselhauf, N. and Biewer, C. (Eds.) (2006). *Corpus Linguistics and the Web*. Amsterdam and New York: Rodopi.

Hwang, H. and Schafer, A. J. 2009. Constituent length affects prosody and processing for a dative NP ambiguity in Korean. *Journal of Psycholinguistic Research* 38, 151–175.

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. PhD thesis, Stanford University, Stanford, CA.

Jaeger, T. F.. 2008 Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*. 59, 434–446.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20, 137–194.

Jurafsky, D., and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River: Prentice-Hall.

Just, M. A., Carpenter, P. A., and Woolley, J. D. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 111, 228–238.

Kaplan, R. 2009. Deep natural language processing for web-scale search. Keynote address at LFG 09 Conference, Cambridge, England.

Keller, F. and Lapata, M. 2003. Using the web to obtain frequencies for unseen bigrams *Computational Linguistics* 29, 459–484.

Keller, F., Lapata, M., and Ourioupina, O. 2002. Using the web to overcome data sparseness. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, 230-237.

Keller, F. and Sorace, A. 2003. Gradient auxiliary selection and impersonal passivization in German: an experimental investigation. *Journal of Linguistics* 39, 57-108.

Kuperman, V. and Piai, V. M. (submitted). Processing of discontinuous syntactic dependencies: Verb-particle constructions in Dutch and English.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review* 191, 676–703.

Marcus, M., Santorini, B., Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19, 313–330.

Mukherjee, J. and Hoffman, S. 2006. Describing verb-complementational profiles of New Englishes. *English World-Wide* 27, 147–173.

Nicol, J. L., Forster, K. I., and Vereš, C. 1997. Subject-verb agreement processes in comprehension. *Journal of Memory and Language* 36, 569–587.

Pickering, M. J., McElree, B., Frisson, S., Chen, L. and Traxler, M. J. 2006. Underspecification and aspectual coercion, *Discourse Processes* 42, 131–155

Pinheiro, J. C. and Bates, D. M. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Quené, H. and van den Bergh, H. 2004. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication* 43, 103–121.

Quené, H. and van den Bergh, H. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59, 413–425.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. 1972 *A Grammar of Contemporary English*. London: Longman.

Richter, T. 2006. What is wrong with ANOVA and multiple regression? analysing sentence reading times with hierarchical linear models. *Discourse Processes* 41, 221–250.

Rohdenburg, G. 2007. Grammatical divergence between British and American English in the 19th and Early 20th centuries. Paper presented at the Third Late Modern English Conference, the University of Leiden.

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

Roland, D. and Jurafsky, D. 2002. Verb sense and verb subcategorization probabilities. In P. Merlo and S. Stevenson (Eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, 325– 345. Amsterdam, John Benjamins.

Rosenbach, A., 2002. *Genitive Variation in English: Concdeptual Factors in Synchronic and Diachronic Studies*. Berlin and New York: Mouton de Gruyter.

Rosenbach, A., 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenburg, B. Mondorf, (Eds.), *Determinants of Grammatical Variation in English,* 379–411. Berlin: Mouton de Gruyter.

Rosenbach, A., 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81, 613–644.

Schneider, W., Eschman, A. and Zuccolotto, A. 2002a. *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., Eschman, A. and Zuccolotto, A. 2002b. *E-Prime Users Guide*. Pittsburgh: Psychology Software Tools Inc.

Sprouse, J. 2009. Evidence for an equalization response strategy. *Linguistic Inquiry*, 40, 329–341.

Stefanowitsch, A. 2007. Linguistics beyond grammaticality. *Corpus Linguistics and Linguistic Theory* 3, 57–71.

Swets, B., Desmet, T., Clifton, C. and Ferreira, F. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition* 36, 201-216.

Szmrecsányi, B. (2005). Language users as creatures of habit: a corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1, 113-49.

Tabor, W., Juliano, C., and Tanehnaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* 12, 211–271.

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

Tily, H., Hemforth, B., Arnon, I., Shuval, N., Snider, N., and Wasow, T. 2008. Eye movements reflect comprehenders' knowledge of syntactic structure probability. Paper presented at the 14th Annual Conference on Architectures and Mechanisms for Language Processing, Cambridge, UK.

Traxler, M. J. and Frazier. L. 2008. The role of pragmatic principles in resolving attachment ambiguities: Evidence from eye movements. *Memory and Cognition* 36, 314-328.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. 2005. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 31, 443–467.

Venables, W. N. and Ripley, B. D. 2002, *Modern Applied Statistics with S*, Springer-Verlag, 4th ed.

Wagers, M. W., Lau, E. F. and Phillips, C. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61, 206–237.

Wasow, T. 1997. End-weight from the speaker's perspective. *Journal of Psycholinguistic Research* 26, 347–361.

Wasow, T. and Arnold, J. 2003. Post-verbal constituent ordering in English. In G. Rohdenburg and B. Mondorf (Eds) *Determinants of Grammatical Variation in English,* 119–154. Berlin: Mouton de Gruyter.

Wasow, T. and Arnold, J. 2005. Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496.