The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets

TYLER KENDALL, JOAN BRESNAN, and GERARD VAN HERK*

Abstract

Recent research has shown the dative alternation in English to be a productive arena for examining the relationship between group-level variation and the internalization of individuals' grammars. Experimental methods (e.g., Bresnan and Ford 2010) and the analysis of large published corpora (e.g., Bresnan et al. 2007) have revealed subtle cross-dialect differences for this variable. The current paper seeks to improve our understanding of this feature and its bearings on experience-based models of grammar by examining African American English (AAE) data from sociolinguistic interviews and from historical letters written by semi-literate ex-slaves. We also consider some methodological problems of conducting corpus-like analyses on non-standard varieties.

Keywords: *syntactic variation, dative alternation, African American English, sociolinguistics*

1. Introduction

The dative alternation is the variable choice between a double NP object structure and an NP PP object structure that occurs with some common verbs in English, such as *give*, as exemplified in (1) from Bresnan and Hay (2008).

(1) a. *Who gave that wonderful watch to you?* prepositional (*to*-)dative
b. *Who gave you that wonderful watch?* double object construction

The alternation has been found to be a useful window into variable syntactic processes and, increasingly in recent years, has been the object of corpus-based study (e.g., Gries 2003, 2005; Bresnan, Cueni, Nikitina, and Baayen 2007). In particular, the alternation has recently been used to examine experience-based models of grammar, which consider individuals' grammars not as invariant or as idealized-to-invariant but as probabilistic, influenced by usage and

Corpus Linguistics and Linguistic Theory 7–2 (2011), 229–244 1613-7027/11/0007–0229
DOI 10.1515/CLLT.2011.011

experience and variable across both individuals and groups. Examining usage data, Bresnan et al. (2007) show that a probabilistic model achieves around 94% accuracy predicting the alternation on unseen data in the Switchboard and Wall Street Journal corpora, based on aspects of the objects like discourse accessibility, pronominality, and argument length (in lexical units) and that meaning alone cannot predict the alternation. The exploration of experience-based models of grammar, which posit inherent variability in each individual's grammar from statistical learning or the storage of linguistic exemplars in memory (e.g., Bybee 2001; Jurafsky 2003), has some similarities to many central inquiries in sociolinguistics (e.g., Weinreich, Labov, and Herzog 1968) and it is at this nexus that we consider the current paper.

In fact, recently researchers have used the alternation to examine differences between varieties of English. For instance, Bresnan and Hay (2008) found that the statistical model of Bresnan et al. (2007) extended well to data from the ONZE project's corpus of New Zealand English (Gordon, Maclagan, and Hay 2007), but that there were subtle differences between the two varieties. At a probabilistic level, New Zealand speakers were found to be more sensitive to animacy, with the U.S. English data less likely to have animate recipients in the double object construction than the New Zealand data (Bresnan and Hay 2008). Bresnan and Ford (2010) examined the alternation in experimental data to compare American and Australian subjects' knowledge of probabilistic grammatical choices and found that subtle differences between American and Australian English varieties are apparent in speakers' psycholinguistic judgments and word recognition times during reading. Mukherjee and Hoffmann (2006), following up on a study by Olavarría de Ersson and Shaw (2003), compared ICE-GB and ICE-India and demonstrated that the *to*-dative form – the prepositional dative – is more common in Indian English than in British English. To quote Mukherjee and Hoffmann (2006: 149), "verb complementation has so far been underestimated as an area of the language system in which regional differentiation figures prominently."

These studies have looked for, and found, variability in the dative alternation across different macro-regional varieties of English (such as Indian English versus British English and American English versus New Zealand English). From this perspective, the dative alternation appears to be a *sociolinguistic variable*, in that its outcome shows correlations with nonlinguistic aspects of its realization (cf. Labov 1972b; Wolfram 1993). In this short paper, we examine whether these sorts of subtle grammatical differences are also found within regionally embedded, but socially distinct, varieties of English, in particular considering African American English in comparison to "standard" American English. That is, we ask, is the dative alternation a sociolinguistic variable in the dimension of ethnicity in the U.S. or is it (probabilistically speaking) stable within the larger umbrella of American English? Several recent papers have

detailed the dative alternation at length (cf. Bresnan et al. 2007; Bresnan and Ford 2010) and we limit our general overview in this paper, referring interested readers to those sources for fuller discussions of the alternation and its statistical analysis. While a number of English verbs take two objects and participate in the dative alternation, we follow Bresnan and Hay (2008) in looking specifically at *give* in this paper. It is by far the most common alternating verb, accounting for 51% of the Bresnan et al. (2007) data.

2. African American English and sociolinguistic variation

African American English (AAE, sometimes called African American Vernacular English, AAVE) has long been a central object of study in North American sociolinguistics (e.g., Wolfram 1969; Labov 1972a; Fasold 1972; Rickford 1999; Poplack and Tagliamonte 2001; Wolfram and Thomas 2002). In fact, it has inspired more than five times as many sociolinguistic publications as any other ethnic or regional dialect (Schneider 1996: 3). These studies have resulted in our knowing quite a lot about AAE and about many of the sociolinguistic features that differentiate it from white varieties of American English at both the qualitative (e.g., the use of invariant *be*, copula absence, etc.) and quantitative level (e.g., higher rates of common features in English varieties like consonant cluster reduction and velar nasal fronting)[1]. Studies of AAE have for the most part focused on features that are uniquely identified or stereotypically associated with the variety. We know of no studies, for instance, that have looked at the dative alternation in AAE and it does not appear that there are salient patterns of the alternation that listeners associate with AAE.[2]

Labov (1972b) established a three-tiered conception of sociolinguistic variables, where a variable can be considered as a sociolinguistic *indicator*, *marker*, or *stereotype*. *Indicators*, the most subtle type of variables, vary with social attributes of speakers but are *not* socially marked or interpreted. They are not manipulated by speakers or commented on by hearers but they do show patterns that correlate with social stratification and/or ethnicity and so on. Labov (1972b: 314), for instance, provides the merger of the vowels in *hock* and *hawk* as an example of a sociolinguistic indicator. The degree to which these vowels are merged varies across groups and individuals, but is often below the level of speakers' conscious awareness and outside of speakers' active control. Sociolinguistic *markers* are features that vary stylistically as well as socially and carry observable meaning, like the production of *-in'* for *-ing* which numerous studies (as early as Fischer 1958) have shown to correlate with social features like class and ethnicity, as well as the formality of a speech event. *Stereotypes* are the most marked type of variables. They are readily commented on by hearers and often become actively manipulated (or avoided due to stigmatization).

Multiple negation or *ain't* in certain speech communities are typical examples of sociolinguistic stereotypes.

While we consider the dative alternation to be a sociolinguistic variable, it seems clear to us that it is not a sociolinguistic marker or stereotype – and markers and stereotypes have been the focus of most studies of AAE. Examining finer-grained variation, such as the dative alternation, in these sorts of regionally embedded language varieties should provide richer knowledge about both the scope of sociolinguistic variation and, more theoretically, the influence of experience on speakers' grammars. Sociolinguists have long shown that social orientations and affiliations have linguistic consequences (cf. Eckert 2000), but an understanding of the relationship between these social and sociolinguistic patterns has not yet been fully integrated into theories of grammar. Put differently, if patterns of the dative alternation correlate with finer levels of sociolinguistic differentiation than the macro-level regional varieties that have so far been studied, it would give us some evidence of the scope of influence on individuals' probabilistic grammars, the degree to which experience is localized and the degree to which it is a function of larger social interactions outside of the variety with which one most closely associates. Importantly, ethnic varieties like African American English are always embedded within a larger macro-regional matrix and examining differences or similarities in variable structures like the dative alternation *within* these macro-regional varieties (instead of just *between*) would impact our understanding of experience-based grammatical models.

While AAE has been and continues to be so extensively studied by sociolinguists, examining its syntactic features in a thorough, quantitative way has remained difficult due to the large amount of transcribed data needed for systematic analysis, as well as the relatively small size of most sociolinguistic studies and the fragmented nature of sociolinguistic data collections (cf. Kendall 2008). Traditionally – and actually with very few exceptions – the field recordings that arise in the course of sociolinguistic studies of AAE have remained closed resources, available only to the original research group. This is perhaps changing, and our current project, we hope, represents some steps in a positive direction; as we point out again shortly, a large amount of the data we examine here comes from very generous colleagues.

We also must note that individual sociolinguistic collections of AAE data are typically small – on the order of perhaps 20 to 30 one hour long interviews. If we estimate about 10,000 words per hour of interview talk, that means a collection may contain somewhere between 200,000 and 300,000 total words, but even then not all of that talk is by the persons of interest. There are often white, standard-speaking interviewers and sometimes as little as only half or two-thirds of the recording supplies actually relevant talk. This is a "problem" that rarely surfaces when looking at macro-regional language varieties, since it is

more often the case in studies of macro-regional or standard varieties that all participants in a conversation are talkers of the relevant variety. Finally, rarely are all or even many of the recordings in a sociolinguistic study transcribed, or transcribed in a standardized way, so much of the data from these field projects are not readily available for corpus-based inquiries. To put this in perspective, Bresnan et al.'s (2007) analysis obtained about 7.87 tokens of the alternation per 10,000 words of corpus for the Switchboard Corpus for all alternating verbs. When interested in rare variables, such as syntactic features, individual sets of recordings are often simply insufficient to generate enough data.

3. Compiling and analyzing our Corpus

We have compiled the data for this project from a number of different sociolinguistic sources. These fall into two primary categories, contemporary sociolinguistic interview recordings and historical letters from antebellum ex-slaves. About half of the spoken interview data – comprising about 165,000 words – come from the Sociolinguistic Archive and Analysis Project (SLAAP;[3] Kendall 2007, 2008). SLAAP is a growing online archive of sociolinguistic recordings (featuring digitized audio from over 1,600 sociolinguistic interviews, a small but growing collection of time-aligned orthographic transcripts, and web-based analytic software). The other half of the spoken data – about 160,000 words – come from transcripts of sociolinguistic interviews that have been generously shared with us by colleagues (whom we thank in our acknowledgments). It is difficult to describe the exact size of the dataset with any definiteness, due to the differing natures of the transcripts, the fact that not all of the talk in the transcripts is relevant data (i.e. there is much talk by speakers of non-AAE varieties), and so on. In the end, our complete collection of spoken language transcripts pares down to about 250,000 words of African American English talk.[4]

As a second source of data, we examine historical written letters by African American ex-slaves. Our historical written letters data come from the Ottawa Repository of Early African American Correspondence (OREAAC; Van Herk and Poplack 2003), which supplied about 140,000 words from "427 letters written between 1834 and 1866 by African American immigrants to Liberia" (Van Herk and Poplack 2003: 233). In previous research (e.g., Van Herk and Poplack 2003; Van Herk and Walker 2005), these letters have been shown to be useful windows into the past and to be representative of the linguistic features of their semi-literate authors.

Our current work only considers data that come from these sociolinguistic sources, as they are collected using methods specifically designed to elicit vernacular language and avoid some mediating problems that arise when studying

ethnic language varieties in other settings (e.g., literature, media). For instance, we have not made use of other possible sources, such as African American literature or other materials. While papers such as Mukherjee and Hoffmann (2006), and other work by researchers like Hoffmann (2007), have shown the relative ease with which one can generate large amounts of corpus data using the Internet, we did not feel that this was a reasonable route to go for our project. First, the determination of ethnic or racial identity on the Internet is not a straightforward issue, though we acknowledge that there are some online sources that could productively be mined for data. For example, we considered using transcripts from the Tavis Smiley Show, an interview and news program on public television hosted by a well-known African American and often featuring African American guests. But this too would be complicated by the fact that not all African Americans speak AAE – it is not as simple as determining a speaker's ethnicity to determine whether or not he or she speaks an ethnic dialect or the degree to which that person has features of the ethnic dialect. In the end we decided to limit our data for this inquiry to materials that come from previously existing sociolinguistic research, where we can make use of that previous research to ensure we examine data that accurately represents African American English. Future work will need to ask whether our database is usefully improved by extending our data collection to include other sorts of data sources.

Since our data come from many sources in a variety of formats, the first step in preparing the tokens of *give* was converting the data to comparable plain text files for parsing. The materials from SLAAP (again, about 1/3 of the total data) were extracted from SLAAP's time-aligned relational database (see Kendall 2007, 2008). SLAAP has an "export" feature and these transcripts were simply exported to plain text files through the SLAAP software. The other spoken language transcripts were in formats including Praat TextGrids, Transcriber transcripts, and Word documents. The Transcriber and Praat files were converted to plain text using tools available online,[5] and the Word documents were converted to plain text using Word and then cleaned up slightly in Emacs, an open-source text editing program. For the spoken language transcript data, we then wrote a Perl script that used a manually generated spreadsheet of speaker identifiers to determine which speakers were appropriate for data extraction (i.e. which of the speakers in the transcripts were African American English speakers) and extracted all lines of text that contained words matching the regular expression pattern "\bg[ai]v\w*" for those speakers. As the OREAAC transcripts retain the highly non-standard spellings of the originals, tokens of *give* and all its variant spellings were extracted by a manual search through the materials. All extracted tokens were then reviewed by hand to remove the (numerous) tokens outside the variable context, such as non-double object instances of *give* (e.g., "he gave it all away"), idiomatic or

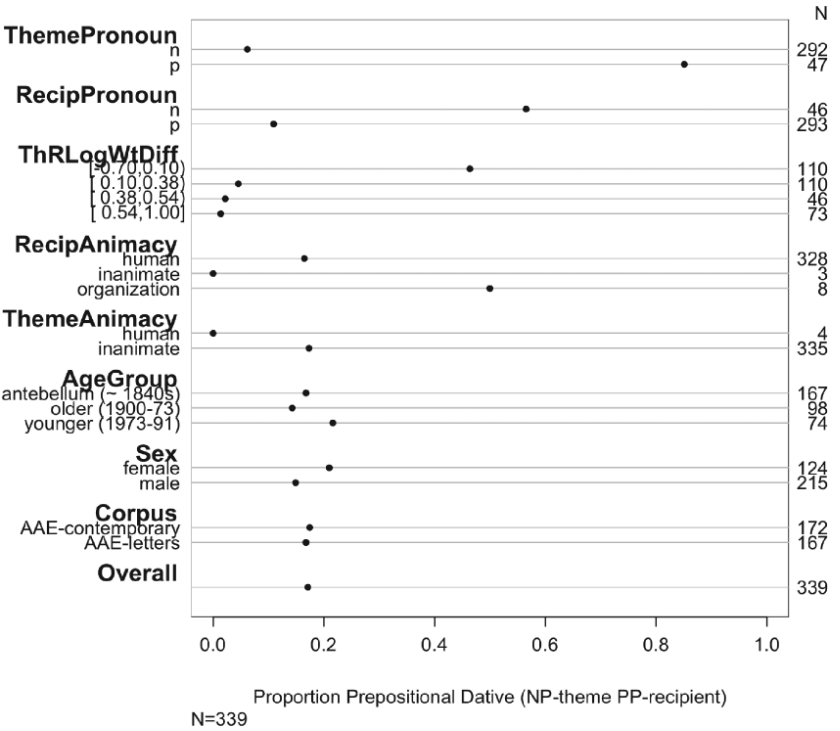

Figure 1. *Summary of African American English* give *data*

formulaic expressions (such as "give my love/respects to . . .") and so forth. Altogether we obtained 339 relevant tokens of *give*.

We coded the data (by hand) for a number of predictors based on Bresnan and Hay's (2008) work. These factors are shown in Figure 1, along with their corresponding proportion of prepositional datives. (The plot is made using the Design library for R, Harrell 2009; see also R Development Core Team 2009.) In all cases, dots to the right represent higher proportions of prepositional datives, while dots to the left represent higher proportions of double NP object structures. The predictors are ordered and discussed roughly according to importance as determined by the models outlined later in this paper. We explain each of the predictor factors in turn.

We originally coded the pronominal status of the recipient and theme into 4 factors following Cueni (2004) although ultimately we collapsed these into binary variables with simply pronoun, "p", versus non-pronouns and indefinite pronouns ("someone", "anyone"), "n". The pronominality of the theme appears to favor the prepositional dative most strongly; 85% of the data with

pronominal themes (ThemePronoun = "p") appear in the NP PP form. Meanwhile, we see that when the recipient is not a pronoun (RecipPronoun = "n") about 56% of the outcomes are prepositional datives as opposed to about 11% when the recipient is a pronoun.

We measured the log of the lexical lengths of the recipient (LogRecipWt) and the theme (LogThemeWt) and then took the difference of these values (theme minus recipient; ThRLogWtDiff) to obtain a positive or negative continuous variable. This is shown in Figure 1 grouped into four bins.[6] We note that this predictor – the log weight of the theme minus the log weight of the recipient – is much further to the right (46% favoring NP PP) when the recipients are heavier than the themes (i.e. when we have negative or near negative values). That is, the data relatively more favor the prepositional dative structure when the recipient is longer than the theme.

We also coded for the animacy of the recipient and theme – that is, whether each is "human" or "inanimate", or also for recipient "organization". We seem to see some possible effects for animacy, but most importantly we notice that these data are massively skewed towards "human" recipients and "inanimate" themes (only 11 recipients are not "human" and 4 themes are not "inanimate").

Each token was coded for the speaker's year of birth (YOB), though for the historical letters component all years of birth were designated 1840. We then collapsed these, as shown in Figure 1, into categories of "antebellum", "older contemporary", "younger contemporary". There is some indication here that the younger AAE speakers have higher rates of the prepositional forms.[7] We also coded speaker/writer sex, which shows slight differences with females realizing 21% NP PP and males realizing 15% NP PP.[8]

The final coding category, Corpus, characterizes whether a token comes from the OREAAC historical letters data (AAE-letters) or whether it comes from the more contemporary sociolinguistic interview data (AAE-contemporary). As is apparent in Figure 1, there is barely any difference between the two in terms of the overall percentage of prepositional forms, and overall there is a low rate of prepositional forms in the AAE data – only 17% (58 of 339).

4. Modeling the dative alternation in AAE and "standard" American English

The AAE data were examined through logistic regression (using the Design library in R, Harrell 2009) in order to determine the relative importance of each of the factors discussed above. For sake of space, we only summarize the outcome of a logistic regression model for these data here, which reflects the impression given by Figure 1 above: the pronominality of the theme and recipient are highly significant (ThemePronoun = "p": log-odds 5.96, $p < 0.0001$;

RecipPronoun = "p": log-odds −3.96, $p < 0.0001$), the difference between the log of the theme and recipient lexical lengths is marginally significant (ThRLogWtDiff: log-odds −1.94; $p = 0.075$), and the extra-linguistic factors (AgeGroup or YOB, Sex, Corpus) are not significant. Due to the very few non-human recipients and non-inanimate themes in the data, the animacy of theme and recipient were excluded from the modeling. Rather than dwell on these basic findings, we turn now to ask the question of most interest: how do these AAE data relate to the macro-regional "standard" American English?

To examine this, we extracted the tokens of *give* from Bresnan et al.'s (2007) data to obtain 1,263 tokens representing spoken "standard" American English from the Switchboard Corpus (Godfrey et al. 1992) and 403 tokens representing written language from the Treebank Wall Street Journal Corpus (Marcus et al. 1993). We then combined these datasets and modeled them, again using logistic regression. Although this was not shown in Figure 1, all of our data were also coded for individual speaker/writer. Similar to Bresnan et al. (2007)'s results, however, we did not find an effect of individual speaker. This was tested by including the individual speakers/writers as a random intercept in a mixed-effect model (cf. Baayen 2008), but the model indicated that speakers/writers had zero variance. Thus, individual speaker/writers are not included in the model as random effects. Also, since all of our data come from the single verb *give*, we do not use a random effect item for verb. In addition to the predictors discussed above, we added the predictor factor Variety to test whether there is a significant overall effect based on language variety, "standard" American English versus AAE. The predictor Corpus, which for the AAE data only had two levels, AAE-contemporary and AAE-letters, now has four levels, to account for the Switchboard and Wall Street Journal corpora. Finally, a predictor, Modality, was included to differentiate the two spoken datasets (AAE-contemporary and Switchboard) from the two written datasets (AAE-letters and Wall Street Journal).

Our best model – obtained through the analysis and comparison of possible models and the use of model criticism (cf. Baayen 2008; Bresnan et al. 2007; Bresnan and Ford 2010) – is presented in Tables 1 and 2. The model statistics include $C = 0.960$, Somers' $D_{xy} = 0.919$, Nagelkerke $R^2 = 0.690$, all of which indicate a quite tight-fitting model. Bootstrap validation obtains less than 1% optimism, indicating that the model is not over-fitting the data (Harrell 2001).

Figure 2 displays the significant effects in the model, including the significant interaction between Modality, whether a token is from spoken or written language, and ThRLogWtDiff, the difference between the theme and recipient log lengths. Most importantly, we note that one model accounts for the data extremely well (recall the high model C and D_{xy} statistics) and that specific Corpus (which one of the 4 particular datasets the data come from) does not arise as significant, nor does Variety. In fact, Modality does not surface as a

Table 1. *Logistic regression model for all* give *data*

Factor	Log-odds	*p*
Intercept	−0.4798	0.0037
Recipient = Pronoun (RecipPronoun = "p")	−3.1300	<0.0001
Theme = Pronoun (ThemePronoun = "p")	4.8766	<0.0001
Theme–Recipient Log Weight Difference (ThRLogWtDiff)	−0.9969	<0.0001
Modality = Written	0.1607	non-sig
ThRLogWtDiff * Modality = Written	−1.3106	0.0001

Table 2. *Wald Statistics for logistic regression model for all* give *data*

Factor	X^2	d.f.	*p*
Recipient Pronominality (RecipPronoun)	110.06	1	<0.0001
Theme Pronominality (ThemePronoun)	237.67	1	<0.0001
Theme–Recipient Log Weight Difference (ThRLogWtDiff) (Factor+Higher Order Factors)	104.00	2	<0.0001
All Interactions	16.19	1	0.0001
Modality (Factor+Higher Order Factors)	16.20	2	0.0003
All Interactions	16.19	1	0.0001
ThRLogWtDiff * Modality (Factor+Higher Order Factors)	16.19	1	0.0001
Total	366.42	5	<0.0001

significant main effect, but only in its interaction with the argument weights. As we see in the plot of Figure 2 and model results tables, the written data are more sensitive to "end weight", the later placement of the longer complement.

Since our AAE data are one-fifth the size of the "standard" English dataset, it is fair to ask whether this model is primarily the result of the Bresnan et al. data overpowering our smaller dataset. To ensure this is not the case, we tested this by using a model trained on just the Bresnan et al. (2007) data to predict the alternation in our current AAE dataset. That model generalized well to our AAE data (with a concordance statistic C of 0.97 and a Somers' D_{xy} of 0.94) indicating that the patterns in the AAE data are, in fact, extremely similar to the patterns found by Bresnan et al. Various further tests also support this. For instance, a model built on just the AAE data also accurately predicts the Bresnan et al. data quite well ($C = 0.95$, $D_{xy} = 0.90$) despite it being a much smaller dataset. It seems clear that the dative alternation in AAE is probabilistically equivalent to the alternation in "standard" English, at least as represented by our samples.

So, then, are there any differences between our African American English data and the Switchboard and Wall Street Journal data? In fact, it does appear that there are, though the differences that are identifiable are primarily inputs to the models and not differences within the models themselves. As an example

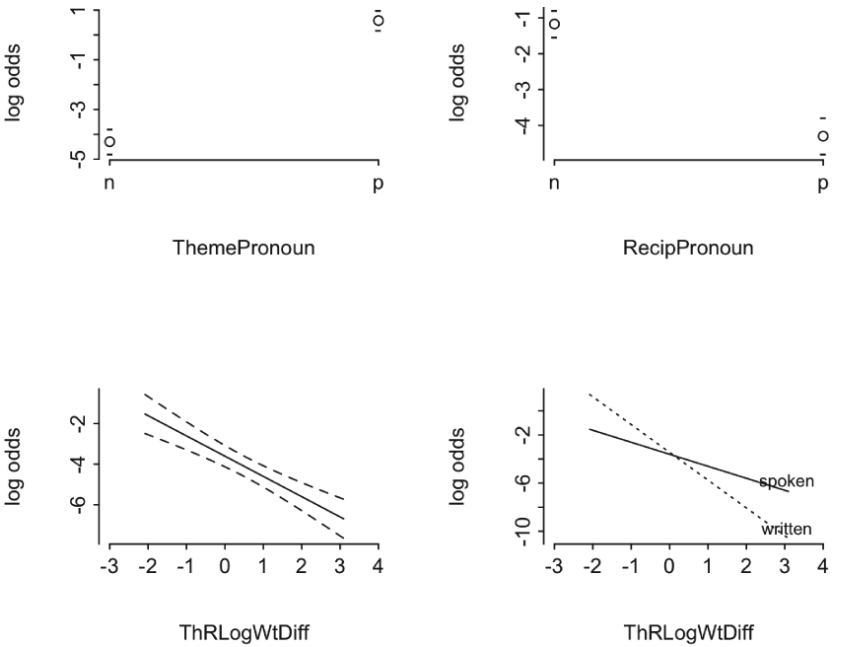

Figure 2. *Logistic regression model effects*

of this, the mosaic plot in Figure 3 represents the proportions contributed by the source corpora data to the total compiled dataset as horizontally divided areas, and the conditional probability of a recipient pronoun in each source by the vertical proportions of the areas. The figure shows that there are more pronominal recipients and more pronominal themes in the AAE data than in the non-AAE data. Only 7% of the recipients in the spoken AAE data (AAE-C in the Figure) are noun phrases compared to 16% in the Switchboard Corpus (SWBD) and only 20% in the AAE letters (AAE-L) compared to 77% in the Wall Street Journal (WSJ).

However, interpreting the cause of these differences in our data is hindered by an important confound in our AAE data. Our AAE spoken data come from in-person conversational interviews, while the Switchboard data come from telephone conversations between strangers. Our AAE written data come from personal letters written by semi-literate authors, while the non-AAE written data come from the professional journalistic writing of the Wall Street Journal. Research on historical pragmatics and genres (e.g., Biber and Finegan 1989) leads us to expect that face-to-face conversation (as in the AAE spoken data) would have higher pronoun use than telephone conversations (the Switchboard data) and even colloquial letters written by semi-literate ex-slaves. More

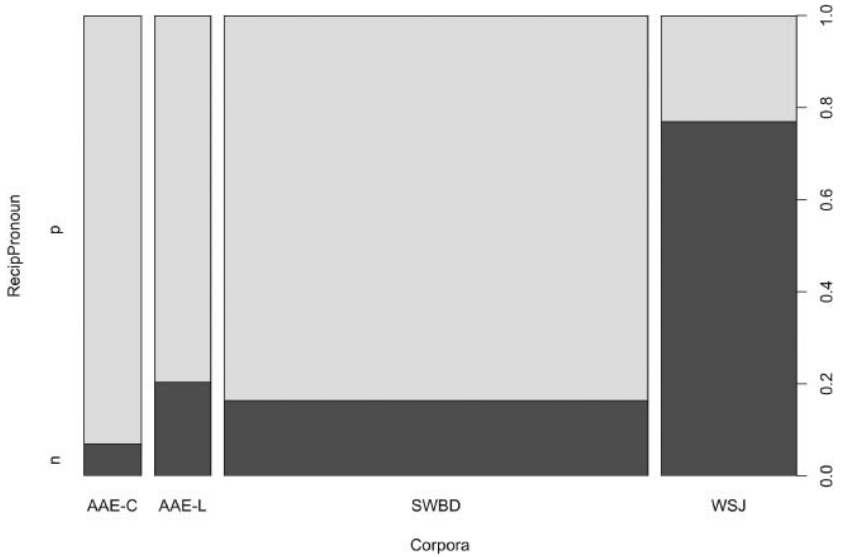

Figure 3. *Recipient pronominality across the four datasets*

full nouns are needed to express reference in 19th century written correspondence than in talk, regardless of how colloquial or vernacular the letters are. It appears likely, then, that the differences in the model inputs have more to do with differences of genre than with differences in language variety.

5. Conclusions

While other recent work has demonstrated that varieties of English can exhibit discernible probabilistic differences in the patterns of the dative alternation, we do not see here any evidence that the alternation in African American English is substantially different than it is in the macro-regional standard of "standard" American English. Based on previous sociolinguistic work, such as the continuum from sociolinguistic indicator to stereotype discussed earlier (again Labov 1972b), perhaps we should not be surprised by this finding. For one reason or another, it appears that the dative alternation has not reached a level where its patterns are socially salient. It further appears that it is not a sociolinguistic indicator, at least in the dimension of ethnicity in the U.S. After all, not all linguistic variables will be *socio*linguistic variables in all cases and language variation and change at the macro-regional level is subject to different forces than language variation and change at, for instance, the level of ethnicity.

Nonetheless, variability in the dative alternation exists and pervades the English language and, as previous studies like Mukherjee and Hoffmann (2006), Bresnan and Hay (2008), Grimm and Bresnan (2009), and Bresnan and Ford (2010) have indicated, there exist subtle differences between macro-regional varieties. These previous findings indicate that the patterns in the dative alternation are not simply the outcome of processing or language-internal constraints but are impacted at least to some measurable degree by extra-linguistic factors.[9] Our current finding, that the dative alternation in AAE is not significantly different from that in "standard" American English, reinforces the view that speakers acquire (and likely continually refine) their grammars based at least in part on their linguistic experiences. Macro-regional varieties may drift apart in terms of small probabilistic differences due to the separation of (and therefore lack of direct communication between) the majority of speakers, but, without even a subconscious social impetus, embedded varieties may pick up their exemplars and preferences passively, through normal daily contact. Should the dative alternation ever acquire a social significance, we might then expect to see differentiation within regions – in varieties like African American English – as speakers actively (though still possibly subconsciously) select as models variants with which they associate in social space.

We end by noting that this short paper represents only a first attempt at understanding the dative alternation in African American English and its variability *within* co-existing but socially distinct English varieties. We hope this work inspires further research on the alternation. The data examined here could also usefully be compared with other post-colonial varieties of English and with current and historical pidgin and creole varieties of English. In addition to increasing our knowledge of gradient grammar and the role of social information and social diversity in experience-based models of grammar, research such as this could shed important light into the origins and development of African American English.

Bionotes

Tyler Kendall is an assistant professor of linguistics at the University of Oregon. His research focuses on language variation and change, primarily through the examination of North American English varieties. He is also the architect and project coordinator for the Sociolinguistic Archive and Analysis Project (SLAAP; http://ncslaap.lib.ncsu.edu/) and the Online Speech/Corpora Archive and Analysis Resource (OSCAAR; http://oscaar.ling.northwestern.edu/).
Email: tsk@uoregon.edu

Joan Bresnan is Sadie Dernham Patek Professor Emerita in Humanities at Stanford University and Senior Researcher at the Center for the Study of

Language and Information, where she directs research in the Spoken Syntax Lab on the dynamics of probabilistic grammar and the development of syntactic alternations. Email: bresnan@stanford.edu

Gerard Van Herk is the Canada Research Chair in Regional Language and Oral Text at Memorial University of Newfoundland. His interests include language variation and change, African American English, Newfoundland English, and the integration of research and teaching.
Email: gvanherk@mun.ca

Notes

* We are extremely grateful to the following researchers for generously sharing data with us for this project: Valerie Fridland, Kirk Hazen, Christine Mallinson, Shana Poplack, John Rickford, Natalie Schilling, Erik Thomas, and Walt Wolfram. Without their help this study would not have been possible. We also thank the audience members at the 2009 AACL conference and the two anonymous reviewers for very helpful comments on this project and paper.

1. Since AAE has been discussed at such length in the literature, in this short paper we limit our general treatment of the variety to its use as a case study for the investigation of the dative alternation. Readers are referred to the cited sources (e.g., Labov 1972a; Rickford 1999; Poplack and Tagliamonte 2001; Green 2002; Wolfram and Thomas 2002) for further background information about AAE.
2. Green's (2002) linguistic overview of AAE, for instance, makes no mention of the alternation.
3. Online at <http://ncslaap.lib.ncsu.edu/>
4. 250,000 words may seem like a meager amount of data after our discussion above about the size of many sociolinguistic data collections – and it certainly is in comparison to the size of most published corpora – but we remind the reader that most sociolinguistic recording collections, especially of AAE, are not transcribed or available publicly. A part of our project was motivated by a desire to determine just how much transcribed data is available on AAE if several groups of scholars pool their resources. As we have already commented, we are extremely grateful to the generosity of our colleagues for sharing data with us for this purpose. Clearly, work in the future will need to compile – and hopefully make publicly available – larger sets of AAE data for large-scale quantitative research.
5. Online at <http://ncslaap.lib.ncsu.edu/tools/>
6. The Design library (Harrell 2009) automatically determines the bins for continuous variables like ThRLogWtDiff. The four groups displayed for such variables are the quantiles .05, .25, .75, and .95 – that is, the means of the lowest and highest 5% of the data and of the lowest and highest quartiles of the data. While the bins are not equally sized or populated, they display the general patterns in the data in a way that is sufficient for the present purposes.
7. Neither the age groupings nor raw year of birth (as a continuous predictor) surface as significant in any of the statistical models of the data (discussed in Section 4). In passing, we note that if future analysis finds this age-effect to be significant, it runs contrary to some previous findings, which indicate an increasing tendency toward the double object (NP NP) construction over time. For example, Grimm and Bresnan (2009) found that datives in LOB/FLOB and Brown/Frown show a change toward double object constructions over a thirty-year period from the 60s to the 90s in both U.K. and U.S. English. Wolk, Ehret, Bresnan, and Szmrecsanyi (2010) find a historical change in the same direction in Late Modern English datives collected from the Archer corpus.

8. In fact, males have a slightly higher ThRLogWtDiff than females, which likely accounts for some of this difference. Sex is not found to be significant in the statistical models (see Section 4).
9. While we have not discussed in this paper work on the genitive alternation in English (cf. Rosenbach 2003; Hinrichs and Szmrecsányi 2007), we note that such research has also found probabilistic differences and different sensitivities in this somewhat related alternation across macro-regional varieties of English, such as between British English and American English. See Bresnan and Ford (2010) for a discussion relating the dative alternation and genitive alternation.

References

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487–517.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, & R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer, and J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Royal Netherlands Academy of Science: Amsterdam.

Cueni, Anna. 2004. Coding notes. Stanford, CA: Stanford University, ms.

Eckert, Penelope. 2000. *Linguistic variation as social practice*. Oxford & Malden, MA: Blackwell Publishing.

Fasold, Ralph. 1972. *Tense marking in Black English: a linguistic and social analysis*. Arlington, VA: Center for Applied Linguistics.

Godfrey, J., Holliman, E., & McDaniel, J. 1992. Telephone speech corpus for research and development. *Proceedings of ICASSP-92*. 517–520.

Green, Lisa J. 2002. *African American English: a linguistic introduction*. Cambridge: Cambridge University Press.

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

Gries, Stefan Th. 2005. Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.

Grimm, Scott & Joan Bresnan. 2009. Spatiotemporal variation in the dative alternation: a study of four corpora of British and American English. Paper presented at the Third International Conference on Grammar and Corpora. Mannheim, Germany.

Gordon, Elizabeth, Margaret Maclagan, & Jennifer Hay. 2007. The ONZE Corpus. In J. Beal, K. Corrigan, and H. Moisl (eds.), *Creating and Digitizing Language Corpora: Volume 2, Diachronic Databases*, 82–104. Houndmills, Basingstoke: Palgrave Macmillan.

Harrell, Frank Jr. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.

Harrell, Frank Jr. 2009. Design: Design Package. R package version 2.3-0. [URL: http://biostat.mc.vanderbilt.edu/s/Design]

Hinrichs, Lars, & Benedikt Szmrecsányi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3). 437–474.

Hoffmann, Sebastian. 2007. Processing Internet-derived text: Creating a corpus of usenet messages. *Literary and Linguistic Computing* 22(2). 151–165.

Jurafsky, Dan. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jen Hay, & Stephanie Jannedy. (eds.), *Probabilistic linguistics*, 39–95. Cambridge, MA: MIT Press.

Kendall, Tyler. 2007. The Sociolinguistic Archive and Analysis Project: Empowering the sociolinguistic archive. *Penn Working Papers in Linguistics* 13(2). 15–26.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2(2). 332–351.

Labov, William. 1972a. *Language in the inner city: studies in the Black English Vernacular*. Philadelphia, PA: University of Pennsylvania.

Labov, William. 1972b. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania.

Marcus, Mitchell, Beatrice Santorini, & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Mukherjee, Joybrato & Sebastian Hoffmann. 2006. Describing verb-complementation profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27(2). 147–173.

Poplack, Shana & Sali Tagliamonte. 2001. *African American English in the diaspora*. Oxford/Malden, MA: Blackwell Publishing.

R Development Core Team. 2009. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. [URL: http://www.R-project.org/]

Rickford, John R. 1999. *African American Vernacular English: features, evolution, and educational Implications*. Oxford/Malden, MA: Blackwell Publishing.

Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–411. Berlin: Mouton de Gruyter.

Schneider, Edgar. 1996. *Focus on the USA*. Philadelphia: John Benjamins.

Van Herk, Gerard & Shana Poplack. 2003. Rewriting the past: Bare verbs in the Ottawa Repository of Early African American Correspondence. *Journal of Pidgin and Creole Languages* 18(2). 231–266.

Van Herk, Gerard & James Walker. 2005. *S* marks the spot? Regional variation and early African American correspondence. *Language Variation and Change* 17(2). 113–131.

Wolfram, Walt. 1969. *A sociolinguistic description of Detroit Negro speech*. Washington, D.C.: Center for Applied Linguistics.

Wolfram, Walt. 1993. Identifying and interpreting variables. In D. Preston (ed.), *American dialect research*. 193–221. Amsterdam & Philadelphia: John Benjamins.

Wolfram, Walt & Erik R. Thomas. 2002. *The development of African American English*. Oxford/Malden, MA: Blackwell Publishing.

Wolk, Christoph, Ekaterina Ehret, Joan Bresnan, & Benedikt Szmrecsanyi. 2010. Dative and genitive variability in Late Modern English. Paper presented at the workshop Probabilistic Syntax: Phonetics, Diachrony, and Synchrony, Freiburg Institute for Advanced Studies, Freiburg, Germany.