Here's you a thesis:

An evaluation of methodology for research in micro-comparative syntax

Rachel Martinez Regan
Advisor: Jim Wood

Submitted to the faculty of the Department of Linguistics in partial fulfillment of the requirements for the degree of Bachelor of Arts

Yale University
May 2018

Abstract

Recent micro-comparative syntax research has centered around electronically-distributed written surveys that allow for language data to be collected from many speakers across a large area and rely on participants independently rating the acceptability of written sentences. While the use of acceptability judgments and intuitions more generally as an empirical research method have often come under fire from linguists who prefer to study corpora, many researchers have shown them to be robust, reliable, and replicable. However, it is true that people do not always give judgments that correspond to their actual language use. I observed this effect first-hand in a previous project, in which one of my survey participants, a friend, consistently rated test sentences containing punctual whenever as unacceptable when I knew that he produced this feature in casual speech.

This discovery led me to wonder whether surveys with audio rather than written sentences might evoke a conversational setting where participants would give more accurate judgments. I designed an experiment to compare acceptability judgments of dative presentative constructions using either written and audio sentences in electronic surveys, closely modeled after those used by the Yale Grammatical Diversity Project. I also compared acceptability judgments for sentences read in Southern and standard American English accents to examine whether accent might also have an effect, as dative presentatives are mainly accepted in the American South.

Overall, the results of this study show minimum acceptance rates of 20% and maximum rejection rates under 60% for dative presentative constructions regardless of whether these sentences appear in a written, mainstream audio, or Southern audio format, suggesting that all of these formats are conducive to examining this feature. We also see that grammatical and ungrammatical control sentences are appropriately accepted or rejected in the vast majority of cases after excluding participants who fail control measures, suggesting that these measures and established protocol for exclusion do function adequately for audio surveys. However, differences in rating distributions of certain sentences do indicate that modality and/or accent have the potential to impact the judgment task.

Frequent abbreviations

AJ: acceptability judgment
HYT: "have yet to"
K-W test: Kruskal-Wallis rank sum test
microsyntax: micro-comparative syntax
MTurk: Amazon Mechanical Turk
NZE: New Zealand English
PD: personal dative
SDP: Southern dative presentative
YGDP: Yale Grammatical Diversity Project

Acknowledgments

I am very grateful to have received funding for this project from the Yale Grammatical Diversity Project and through a Mellon independent research award from Yale's Davenport College. Thank you to Professor Jim Wood for guiding this project at all stages, to Alexa Little for her comments on this thesis in its late stages, and to Professor Jason Shaw and Fabian Schellhaas at Yale StatLab for consultation on experimental design and statistical analysis techniques respectively. Thank you also to John Baxter for bringing my attention to discrepancies between acceptability judgments and language use, and for serving as the speaker who recorded all the audio sentences for this experiment. Thank you to Charlotte Killiam for holding my hand through the stats analysis and senior year more generally. Lastly, thank you to my fellow linguistics seniors and Raffaella for the snacks and the support.

Contents

1 Introduction

1.1 What is micro-comparative syntax?

1.2 Current state of research in micro-comparative syntax

Syntax, and generative linguistics more broadly, centers around one main question: how is language represented structurally in the minds of individuals? In other words, the goal is to describe what linguists call a "mental grammar"; what rules for language use does a persons mental grammar contain, and how do those rules work? Comparative syntax asks a sub-part of this question, at the language level: what does it mean to speak the same language as someone else? More technically, what are the systematic structural differences in the mental grammars of speakers of say, Russian and Arabic? In more recent decades, researchers in comparative syntax have become interested in honing in on more minute differences in language structure by studying closely related national/established languages like Spanish and Italian (both Romance languages descended from Latin), regional varieties like Piedmontese and Sicilian (both dialects of Italian, spoken in different regions of Italy), and even surveying large populations of speakers of a certain language to examine regional variety where distinct "dialects" might not be so firmly established. Comparative syntax at this finer-grained level, examining regional variety within a single language, is referred to a micro-comparative syntax, and asks the sub-sub-question: what variation do we observe within a language? Or: what are the systematic differences in the similar mental grammars of speakers who share the same language, but who may not share the same dialect? This still serves to inform the larger central question of generative linguistics, which is to describe the language faculty of individuals.

1.2.1 Elicitation versus written electronic surveys

Two main research methods in micro-comparative syntax are elicitation and written surveys. These two methods are very distinct from each other, in form and purpose. Below I give a basic overview of each and their respective benefits and drawbacks.

The goal of elicitation, long-form interviews which are then transcribed and compiled into a corpus, is to elicit casual speech from interviewees which will hopefully contain the features that the researcher is interested in studying. While the documentation of actual language use is very valuable, elicitation is time-intensive and often limit researchers to interviewing people within a certain region, though large-scale elicitation projects have been undertaken (Labov et al. 2006; Wolfram & Schilling-Estes 2005). Another limitation of this method is that while these researchers strive to create a comfortable, casual setting where the speaker will not police their speech, some interviewees might still find a recorded interview with a linguist to be an unnatural setting for a casual conversation and might police themselves somewhat. Finally, there is no guarantee that certain constructions will happen to crop up in conversation, even if they are grammatical for an interviewee. With written surveys, on the other hand, researchers have much more control; participants are asked to judge the grammaticality or acceptability of a set of sentences according to a pre-set scale, and the sentences at hand are specifically crafted by researchers to elucidate their particular research questions about certain language features.

Recent micro-comparative syntax research has centered around written surveys distributed electronically via crowdsourcing platforms like Amazon Mechanical Turk (MTurk). The use of written surveys is much less time-intensive than elicitation, and when they are dis-

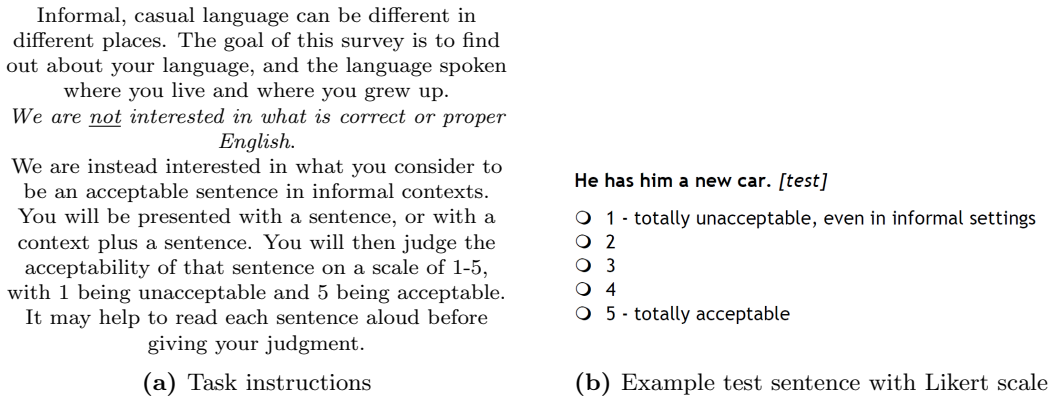
tributed electronically, written surveys allow researchers to collect judgment data from many speakers across a large area. These freedoms additionally enable researchers to more closely examine how other demographic information besides location (e.g. age, race/ethnicity, gender) is related to language use. However, the linguistic concept of acceptability can be difficult to grasp, especially for speakers of non-standard dialects who were likely told that the way they spoke and/or wrote was "ungrammatical" or "bad English". It is possible that a speaker will explicitly say that a certain feature is unacceptable to them, even if they are observed to use that feature in their speech; Labov (1996) documents several instances of what he calls a "mismatch of intuition and behavior", where Philadelphia speakers gave introspective reactions indicating they did not accept the feature positive *anymore*, when in fact they were observed to use it freely in their own speech. One woman reportedly told her interviewer, "I've never heard the expression," when she had used it twice in the same sentence earlier during the interview: "Anymore, I hate to go in town anymore," (Labov 1996). Still, techniques for obtaining judgments have improved since this study of positive anymore; this mismatch effect was observed in response to an explicit, open response interview question about the acceptability of a non-standard feature, whereas current written surveys are conducted as controlled, counterbalanced experiments using sentence judgments on a pre-set scale where the participant is not made aware of the particular features being studied. Wayne Cowart's book *Experimental Syntax* has been a particularly foundational text in establishing objective methods for the use of acceptability judgments (AJs) in experimental syntax (Cowart 1997). While some researchers continue to criticize reliance on AJs and point out their supposed deficiencies (Edelman & Christiansen 2003; Gibson & Fedorenko 2010), Cowart and others have firmly established that AJs are reliable and robust measures (Cowart 1997; Sprouse 2011; Sprouse & Almeida 2017); one study by these last authors exhaustively tests nearly 500 data points from a popular syntax textbook and finds a minimum replication rate of 98% (Sprouse & Almeida 2012). It is also of note that even those who criticize how AJs from few people have been used to defend syntactic theories do still advocate for large-scale electronic surveying on MTurk (Gibson et al. 2011).

1.2.2 Written survey composition and methodology

The Yale Grammatical Diversity Project, henceforth referred to as the YGDP, is one preeminent research group making use of surveys distributed via MTurk. A recent article published in Linguistics Vanguard written by faculty members of the YGDP outlines the overall goal and strategies of the project as looking for "theoretically significant linguistic correlations" in the domain of North American English, using large-scale sentence judgment surveys distributed via MTurk, maps of these judgments and statistical tests taking geography, judgments, and other social variables into account (Zanuttini et al. 2018). The term "large-scale" means that these surveys test multiple phenomena across multiple varieties. While there have been prominent large-scale studies in the phonological, lexical, or typological domains (Wolfram & Schilling-Estes 2005; Labov et al. 2006; Kortmann & Lunkenheimer 2013), the large-scale research done by the YGDP is in contrast to most other microsyntax studies, which most often zero in on particular constructions and/or particular varieties (Montgomery & Hall 2004; Green 2002; Feagin 1979).

Zanuttini et al. (2018) provide a detailed description of the YGDP survey methodology in their supplementary materials. YGDP surveys contain roughly 45 sentences: 15 test sentences, 15 filler sentences, and 15 control sentences, following recommendations by Cowart (1997). Test sentences contain the primary phenomenon or phenomena under investigation

in varied morpho-syntactic configurations, filler sentences often include lesser-studied phenomena that may be the subject of future work, and control sentences are either universally grammatical or ungrammatical to all North American English speakers. Control sentences are designed to measure participants understanding of the task of providing acceptability judgments according to the instructions, with some grammatical control sentences containing semantically anomalous or implausible content, or acceptable structures that are targeted by prescriptive rules, and some ungrammatical sentences which remain interpretable but violate morphosyntactic rules of English that do not display regional variation. Participants are excluded from analysis on the basis of failing controls if they: (A) judged one or more grammatical sentences as 1 or 2 and had an average grammatical judgment under 4, or (B) judged one or more ungrammatical sentences as 4 or 5 and had an average ungrammatical judgment over 2. All sentences in these surveys are provided in written form and, after being provided with the following instructional statement about how to judge acceptability of sentences, participants are asked to give their judgments on a 5-point Likert scale, with 1 corresponding to a judgment of "totally unacceptable" and 5 corresponding to a judgment of "totally acceptable".



(a) Task instructions

(b) Example test sentence with Likert scale

Figure 1: YGDP survey methodology

These surveys then rely on participants independently rating the acceptability of written sentences. Zanuttini et al. (2018) address the concern that participants may not receive enough instruction on what linguistic acceptability is and how to give an accurate judgment, defending the use of control sentences to eliminate responses from participants who interpret the task incorrectly, and noting that their results do corroborate prior work on American English dialects which used more traditional methodologies.

Ultimately, judgments from YGDP research are consolidated into maps showing geographic coverage of particular sentences, with "hot spot" and "cold spot" regions identified using the Gi* statistic in ArcGIS, a software used for mapping and geospatial analysis. The map below from Wood et al. 2015 shows the primary childhood residence of a participant and their sentence judgment, with acceptable responses in yellow and unacceptable responses in black, and identifies a "hot spot" in the Southeastern United States; here the geospatial clustering of values is significantly higher than expected if the values were distributed randomly across space. Wood (Submitted) discusses "hot spot" analysis in more detail.

The judgment patterns found in YGDP survey results are also used to inform syntactic

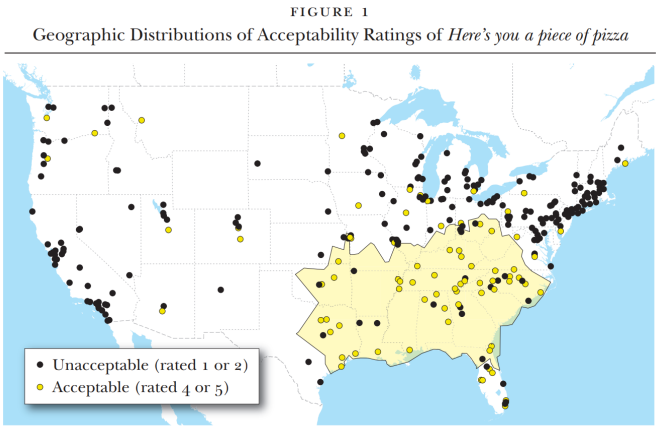


Figure 2:
An example of YGDP map results (Wood et al. 2015)

analyses of phenomena. In a recent study of the *have yet to* (HYT) construction, Wood & Tyler 2018 rely on survey results to resolve a dispute in the literature about performance of this construction under negation tests. Harves & Myler 2014 and Bybel & Johnson 2014 provide opposite grammaticality judgments for the following sentences, the first of which exhibits sentential negation while the second exhibits no sentential negation. Harves & Myler deem the first of these to be ungrammatical and the second to be grammatical, whereas Bybel & Johnson have the opposite judgment and accordingly come to opposite conclusions about the underlying structure of this construction (Wood & Tyler 2018).

(1) Negation tests applied to HYT construction
   a. John has yet to attend Mary's lecture, and neither has Jim.
   b. John has yet to attend Mary's lecture, and so has Jim.

In studying results from a large-scale YGDP survey including HYT sentences, Wood & Tyler (2018) find that some speakers are capable of treating *have* in this construction as either a main verb or an auxiliary verb, while others are restricted to treating *have* as a main verb. Wood & Tyler arrive at this conclusion based on results from testing sentences like the following pair of questions.

(2) I have yet to visit my grandmother.
   a. *main verb have* — Have you yet to visit your grandmother?
   b. *auxiliary verb have* — Do you have yet to visit your grandmother?

An asymmetry is observed in the judgments for the above two questions; speakers who accept the auxiliary verb sentence also overwhelmingly accept the main verb sentence, but not vice versa. Essentially, YGDP survey data on HYT sentences show that there is genuine speaker variation in this construction, previously assumed to exhibit unified behavior across all speakers of English who accept it. Wood & Tyler also show that there is natural variation in the how negation interacts with this construction, which explains the opposing judgments made by Harves & Myler 2014 and Bybel & Johnson 2014. This is but one example of how examining the nature of variation using electronically-distributed surveys is valuable to theoretical syntax.

1.3 My methodological experiment

As I have just discussed, recent micro-comparative syntax research by the YGDP has centered around written surveys distributed via MTurk. These large-scale surveys allow for language data to be collected from many speakers across a large area, and the regional variation we see in survey results has important implications for empirical and theoretical claims about syntactic constructions. While these surveys rely on participants to independently rate the acceptability of written sentences, those judgments have been shown to be robust, reliable, and replicable (Sprouse 2011).

However, corpora-or-bust linguists do have a point that people do not always give judgments that correspond to their actual language use. I observed this effect first-hand in a previous project, in which one of my survey participants, a friend, consistently rated test sentences containing punctual whenever as unacceptable when I knew that he produced this feature in casual speech. As my final project for a Yale linguistics course in Spring 2016, I conducted a survey on the dialectal feature punctual whenever, which I was inspired to research since one of my close friends from the American South uses this feature. Because of stigma, he had consciously moved to a more standard English grammar and accent since moving to the Northeast, but punctual whenever still remained a feature of his speech. However, my written survey did not accurately test his acceptability of this feature; he rated some of my test sentences as unacceptable even when they had been based on sentences that he had in fact uttered. This discovery led me to wonder whether participants might give more accurate judgments in a survey that uses audio rather than written sentences. While it is true that in the age of email, blogs, texts, Facebook, Twitter, etc. seeing casual and non-standard speech in a written format is becoming more and more commonplace, written English is often standardized, and participants may feel more familiar with non-standard constructions in a spoken context. Zanuttini et al. 2018 address this issue in the following excerpt, and acknowledge that audio presentation of sentences is an area of interest for this field.

> We acknowledge that there are valid arguments against presenting sentences in written form. One type of argument is that because the phenomena we are testing tend to appear in colloquial and/or stigmatized speech and tend not to appear in print, participants may be thrown off by a register or medium clash. We respond to [this] argument by pointing to our survey instructions and our results: our instructions explicitly ask participants to focus on what would be acceptable in informal situations, and our results show that even nonstandard features are often rated highly by speakers who grew up in the areas where they are known to occur. In the age of texting and Twitter, we expect that speakers are more accustomed to seeing nonstandard phenomena in print than ever before. Furthermore, the alternative of presenting audio recordings of our sentences instead of or in addition to the written presentation raises its own set of issues relating to the recorded speakers accent and how that may influence participants perception and judgment of the sentences acceptability. While we do defend our decision to present written sentences, we acknowledge that valuable future studies may use other media, including perhaps audio presentation. (Zanuttini et al. 2018)

Though audio sentences have been previously used in sentence judgment tasks, with Sprouse 2011 even linking to an HTML template for an MTurk auditory AJ task, to my

knowledge there has never been a comprehensive comparison of AJs obtained using written versus audio formats. Previous work surrounding AJ methodology has largely focused on experimenting with different scales for the the judgment task, i.e. comparing AJ results using 1-5 and 1-7 Likert, magnitude estimation and yes-no scales (Bader & Häussler 2010; Weskott & Fanselow 2011).

In the interest of exploring the feasibility of using audio sentences in a microsyntax survey, I designed an experiment to compare acceptability judgments of dative presentative constructions using either written or audio sentences in electronic surveys closely modeled after those used by the Yale Grammatical Diversity Project. It is not obvious what accent would lend itself best to this endeavor, seeing as test features are often associated with a regional accent; therefore I also compare acceptability judgments for sentences read in mainstream English accent versus sentences read in a Southern US English accent appropriate for dative presentatives, the test feature in this experiment.

2 Implementing audio in an acceptability judgment experiment

One might wonder whether participants in audio conditions might base their acceptability judgments entirely on prosodic effects how natural the speaker sounded in her/his pronunciation rather than the content of sentences, or whether judgments might be affected by issues of speaker bias. Certainly with regard to ungrammatical sentences in an auditory context, prosody is one signal to a participant that something is "off" about a sentence. But prosody could also serve as a clue for detecting ungrammaticality in written surveys since the task instructions encourage participants to read sentences aloud to themselves, and those who do so must recognize how difficult it is for them to produce ungrammatical sentences and how unnatural they sound.

In a context like this experiment where participants are asked to judge sentences based on their content alone, their judgments may be influenced by prosody but that does not mean that prosody is solely responsible for the judgment patterns to be found. Additionally, people are affected by and do process syntactic structure of sentences in tasks even when they are explicitly told to ignore what is being said and focus solely on how it is being said (Walker 2008). In her masters thesis titled *Phonetic Detail and Acceptability Judgements*, Abby Walker shows that socially-meaningful phonetic detail and morpho-syntactic constructions alter judgments of a speaker's age and social class. In Walkers experiment, five female speakers of New Zealand English (NZE) with theatrical training were coached into producing sentences with a natural sounding phrase final /t/ which exhibited glottalization followed by a release. Sentences without phrase final /t/ were obtained by manipulating the audio recordings and cutting the /t/ from the end of the sentence. The NZE constructions recorded for the experiment, showing social variation in their distribution, were nonstandard preterite forms (e.g. *come* in "George come over last night", *done* in "George done the dishes late last night") and *have-got* to denote possession ("that's all she's got"). Grammatical and ungrammatical sentences were also tested. Participants were told that they were listening to actresses reading lines, and to focus on their voices as opposed to what was actually being said. After hearing a sentence one time, participants were then asked to judge the speakers age and social class.

Importantly, Walker also made use of the same speakers to obtain recordings with the different pronunciations; the women reading the sentences "were told to make them sound

as natural and conversational as possible" while also being "coached into consistently producing /t/ with a release phrase finally" (**?**). Walker's thesis does not make any mention of specifically recruiting these women to record sentences because they used these non-standard features or spoke with a phrase-final /t/ naturally, yet clear results were still obtained: the manipulation of phrase final /t/ and the different NZE constructions had a significant effect on both age and social class ratings. For sentences with absent phrase final /t/, the speaker was rated as younger and of a lower social class compared to the same speakers sentences with the conservative variant. Walker also observed a similar effect for sentence type, such that speakers were rated as older and of a higher social class for the grammatical and HAVE-GOT sentences, compared to when they read the ungrammatical, COME and DONE sentences. The ungrammatical sentences generally received the youngest age ratings, while the preterite COME/DONE constructions received the lowest class ratings. These correlations were all observed even though these recordings were carefully coached and/or electronically manipulated, and also despite the fact that participants had been explicitly told to only base their judgments on the voice of the speaker rather than what they said.

I bring this up because this is true in my experiment also while the speaker who recorded sentences for this experiment is from the American South, Mississippi to be exact, where SDPs are accepted, his recordings required some coaching as he does not natively produce this feature. Additionally, while he is familiar with both a mainstream and Southern accent and could effectively codeswitch between them for the purposes of these recordings, this took a conscious effort on his part. However, as Walker's results indicate, clear results can still be obtained from audio sentences that are carefully choreographed. Having one speaker record all sentences for this experiment was certainly preferable to having separate speakers record the mainstream and Southern audio sentences, as this would have introduced a whole host of potential speaker bias issues related to different age/gender/race of the speakers that could affect results; Walker herself talks about facing these issues in a previous experiment where the two male speakers who recorded sentences differed substantially in age, and humans are famously capable of racially profiling over the phone after just one word, "hello" (Purnell et al. 1999).

3 Experimental design

3.1 Examining acceptability of a known feature in a restricted geographic area

In order for this experiment to serve as a methodological analysis, it needed to resemble previous research in this area in its design. This experiment is composed of electronic surveys that follow the same basic structure and format as previous microsyntax surveys done by the YGDP. My surveys shared the same instructions, ratio of test to control sentences, and distribution method as previous YGDP surveys. The only difference was the use of audio sentences rather than written sentences in the judgment task.

While it was important that the surveys in this experiment be essentially identical in structure to previous YGDP work, the goal of this experiment was quite different, and the experimental design reflects this. The YGDP has used their surveys to examine the distribution of acceptability of understudied syntactic features in the United States, and whether that distribution is affected by geography, age, race/ethnicity, education, etc. However, since

this experiment is an examination of microsyntax research methods, it was prudent that test sentences in my surveys contain a syntactic feature which (A) had already been studied in a similar capacity, (B) was firmly geographically defined. This way I could restrict survey participation to a geographic area where there would be reasonable rates of acceptance of the test feature, and examine variation in acceptance rates depending on survey condition.

3.1.1 Dative presentatives and previous research on this construction

Southern dative presentatives (SDPs) satisfy both of the requirements discussed above: (A) the YGDP has done previous research on SDPs, and (B) they have determined that acceptability of this feature is roughly confined to the American South. Importantly, this region is also associated with a non-standard accent, which lends itself to my comparison between audio conditions using different accents. Before I discuss the use of dative presentative test sentences in my survey experiment, here is a brief overview of this construction.

A presentative is a construction that a speaker uses to bring or "present" some entity to the attention of their listener. Dative presentatives contain a pronoun or noun phrase with dative case; this word or phrase functions as the beneficiary, the recipient of whatever entity is being presented. Most attested examples of dative presentatives involve a 2nd person dative, you, but 1st or 3rd person datives can be used in this construction as well. All the below examples were widely accepted by speakers in the Southeast United States in previous surveys conducted by the YGDP (Wood et al. 2015, Submitted).

(3) Dative presentatives with different dative beneficiaries

a. Here's me a good pair of jeans. *1st person singular*

b. Here's us a gas station - pull over! *1st person plural*

c. Here's you a piece of pizza. *2nd person singular*

d. Here's him a nice cup of coffee. *3rd person singular*

While the above examples of the construction all contain *here*, this word can be replaced by *where* or *there*, as seen in the sentences below from Wood et al. (2015); Wood (2005):

(4) DPs with *where* (Wood et al. 2015)

a. Where's me a screwdriver?

b. Where's us a place to eat around here?

c. Where's you a quiet place to study?

(5) DPs with *there* (Wood 2005)

a. Have you ever tried bull riding? You should do it once and put it in your show. There's you an idea.

b. Now there's me a new Easter Dress or Maybe not...

c. There's me some fantasy points.

And while the verb in dative presentative sentences is most often *'s*, a contracted form of the verb to be, it is possible for this verb to appear uncontracted in the form *are*:

(6) DPs with *are*

a. Where are me some little elves? (Wood 2005)

b. Here are some statistics to examine. (Rockwood Tennessee Police 2016)

SDP constructions appeared only sporadically in the literature (Dudley 1946; **?**) and received no special attention until an in-depth investigation of this feature in Wood et al. 2015, which identified SDPs as an understudied construction related to but unique from personal datives (PDs), whose syntactic, semantic, and pragmatic properties had previously been studied at length (Wolfram & Schilling-Estes 2005; Webelhuth & Dannenberg 2006; Horn 2008). Wood et al. (2015) go on to systematically explore connections and distinctions between SDPs and personal datives; noting, for instance, that the dative pronoun in both constructions must be immediately adjacent to the verbal form, and cannot be stressed, modified, or coordinated, but also that personal dative pronouns must be coreferential with the subject of the sentence whereas SDP pronouns are no coreferential with any overt argument. SDPs can also be rephrased using a benefactive prepositional phrase, whereas no such rephrasing is possible for PDs:

(7) Southern dative presentative with rephrasing

a.	Here's you a piece of pizza.	*SDP construction*
b.	Here's a piece of pizza for you.	*grammatical rephrasing*

(8) Personal dative with attempted rephrasing

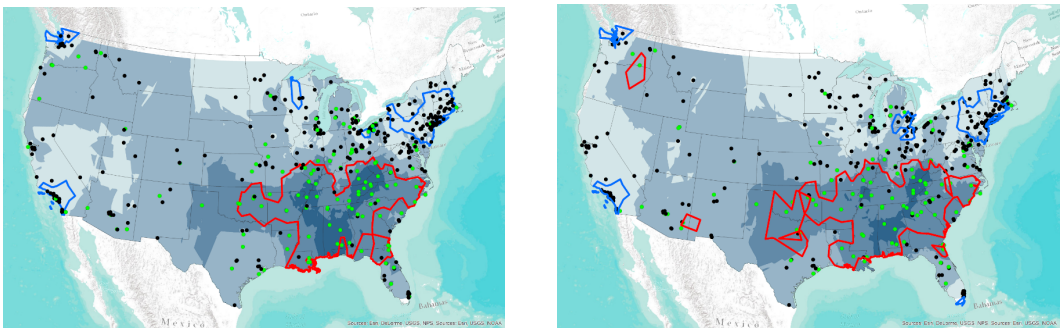
a.	I need me a screwdriver.	*PD construction*
b.	*I need a screwdriver for me.	*ungrammatical rephrasing*

And finally, SDPs cannot occur in embedded, negated or yes/no question environments whereas PDs are not restricted in this way.

The authors next mention closer corollaries to SDPs found in other languages: French *voici/voila*, Italian *ecco*, and Hebrew *hinne*, to name a few. But the centerpiece of this paper is of course the use of a written acceptability judgment survey distributed online via MTurk, which enables the testing of acceptability of SDPs in a range of contexts. SDP sentences in this survey contained one of two locative elements, *here* or *where*, and one of three pronouns *me, you,* or *us*. The goal of this survey was primarily to assess the productivity of this feature, and secondarily to look more closely at geographic distributions of acceptability of particular sentences and sentence types. Survey results from this work firmly establish SDPs as a geographically restricted feature of American English, with acceptances clustering in the South and Appalachian mountain region, with rejections being found only rarely in this region but commonplace outside this region. Continued research on SDPs by these authors reveals a hierarchy of acceptance of SDPs with regards to the locative element, choice of pronoun, and copula form present in the sentence; the less marked of each of these categories being here or there, the 2nd person singular pronoun, and contracted singular copula *s* (Wood et al. Submitted). Essentially, the least marked SDP construction is of the form *"Here's you a [Noun Phrase]"* or *"There's you a [Noun Phrase]"*.

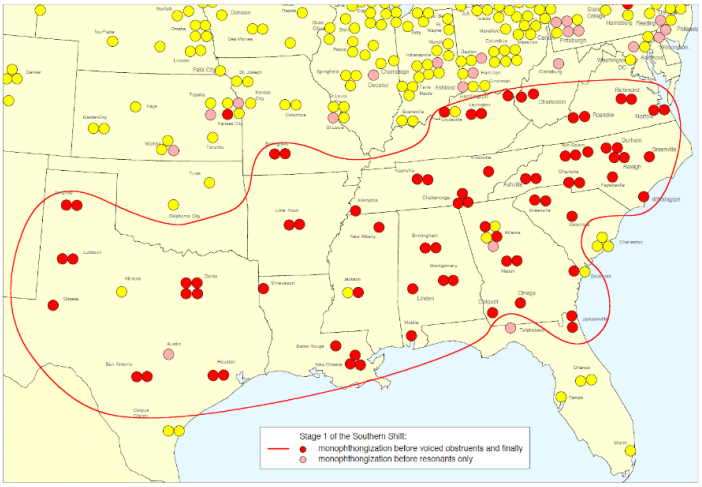
3.1.2 Restricting survey region to seven states in the American South

I decided to survey only MTurk workers who were current residents of one of seven Southern U.S. states: Alabama, Arkansas, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee. These states were chosen based on two factors: previously attested regions of 1) acceptability of dative presentatives, and 2) monophthongization of phonemic /aj/, which is a prominent feature of the US Southern accent. It is certainly true that not all Southern accents are the same monophthongization of phonemic /aj/ was merely chosen as the



(a) "Here's you some money" **(b)** "There's you a piece of pizza"

Figure 3: YGDP acceptability maps for SDP constructions (Wood et al. Submitted)



Map 11.3. Stage 1 of the Southern Shift: glide deletion of /ay/

Figure 4:
Dialect map of monophthongization of phonemic /aj/ (Labov et al. 2006)

substantive feature to quantitatively distinguish the mainstream and Southern accents used in this recording, since Labov et al. 2006 describes this feature as stage 1 of the Southern Shift.

In these YGDP acceptability maps, dots represent the primary childhood residences of participants, with green dots indicating those who judged the sentence in question to be relatively acceptable, 4 or 5 on a scale of 15, and the black dots indicating those who judged the sentence as relatively unacceptable, 1 or 2 on a scale of 15. The red and blue borders surrounding particular areas on the map indicate statistically significant hot and cold spots, where average acceptability values are statistically higher or lower than would be expected were these values randomly distributed across the map. Lastly, the shading of these maps refers to an interpolation analysis, which, at every point in the map, takes the 12 nearest points and uses an inverse-distance weighted algorithm to determine what that point might be expected to be (Wood et al. Submitted). The maps then display distinct colors for ranges corresponding to the calculated judgment: 12, 23, 34, and 45. As we can see from these maps, interpolation and hot/cold spot indications often overlap darkest shaded regions in

the South (interpolation of 4-5) are contained within red hot spot border, and the lightest shaded regions in the Northeast and along the West Coast (interpolation of 1-2) also contain blue cold spots.

The area outlined in red on the dialect map from Labov et al. 2006 encapsulates the region where monophthongization of phonemic /aj/ appears most widely red dots indicate people surveyed who monophthongize /aj/ in all phonetic contexts, pink dots indicate those who monophthtongize /aj/ only before resonants, and yellow dots indicate those who did not monophthongize at all.

Many more states in the American South beyond the seven chosen display acceptability of dative presentatives, and/or monophthongization of phonemic /aj/; Florida, Illinois, Kentucky, Louisiana, Maryland, Oklahoma, Texas, Virginia, and West Virginia were also in contention. However, I decided to limit the testing region of this experiment to only those states which appeared mostly encapsulated by both hot spots on YGDP maps and the red outlined region on the map of Southern glide deletion from Labov et al. 2006. This way, participation would be restricted to only those who were currently living in a place where they were likely to have encountered dative presentatives and also were familiar with a Southern monophthongization of /aj/ in all phonetic contexts.

3.2 Designing test and control sentences

As discussed above, Wood et al. (Submitted) find the most widely accepted or "least marked" of dative presentative constructions to be: *"Heres you a [Noun Phrase]"*. I chose to use only this basic format for all test sentences in this experiment, rather than manipulating the morpho-syntactic structure of the dative presentative construction throughout the test sentences, since the purpose of this experiment is not to make any empirical or theoretical claims about dative presentatives and/or geographic acceptability patterns of this feature, but instead just to explore the feasibility of using audio sentences.

To ensure that the distinction in accent between the two audio conditions would be salient, half of all sentences contained a lexical item with an /aj/ diphthong, which was monophthongized in the Southern audio condition but not in the mainstream audio condition. Since monophthongization is affected by the consonant following the /aj/ diphthong, sentences were designed so that /aj/ appears in a variety of phonological contexts: before stops (*dried, bike*), fricatives (*drive, nice, prize*), nasals (*limes, timer*), a laminal (*pile*) and a glide (*eyes*). Additionally, to double-check that there would be no difference in judgments solely based on the presence or absence of /aj/ items, test sentences were designed in pairs in the format "Heres you a [adjective][noun]", where the adjective would either contain /aj/ or not. In the following example, the adjective *dried* contains /aj/ diphthong while *fresh* does not.

(9) Test sentence pair

a.	Here's you some **dried** fruit.	*/aj/ present*
b.	Here's you some fresh fruit.	*/aj/ absent*

In order to avoid testing paired sentences in the same survey, each condition was tested using a Type A and Type B survey, with one sentence from a pair being assigned to Type A surveys and the other being assigned to Type B surveys. Each survey then contained a total of 8 test sentences, 4 with /aj/ lexical items and 4 without, and 16 control sentences, 8 with /aj/ lexical items and 8 without, that were the same across both types (see Appendix B for all sentences). The figure below shows a visual representation of my experimental design.

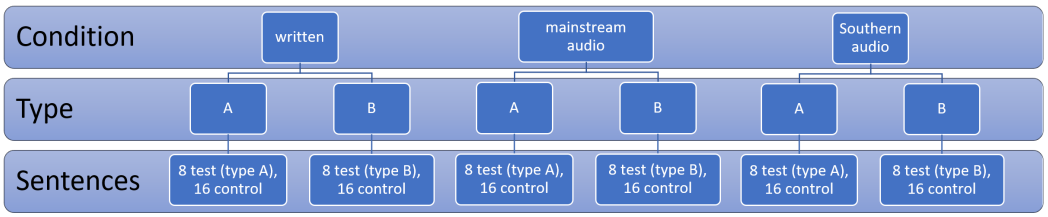


Figure 5: Experimental design flowchart

The 1:2 ratio of test to filler sentences follows suggestions from Cowart 1997 and precedent set by YGDP. While the majority of my filler sentences were grammatical and ungrammatical control sentences, I did also include two additional filler sentences containing *even* for my colleague Katie Martin's thesis research Martin (2018).

In my analysis, I was able to ascertain that the presence or absence of an /aj/ item did not affect judgments using a Wilcoxon rank sum test, a non-parametric statistical hypothesis test. I mention this here because this was not the main focus of this experiment and I would prefer to focus on differences detected between rating distributions of the separate conditions in the analysis section.

3.3 Recording audio sentences

To obtain sentence audio recordings, I enlisted the assistance of just one speaker who could codeswitch between Southern and mainstream accents, rather than one speaker to record Southern audio sentences and another speaker to record mainstream audio sentences. This was done in order to minimize speaker bias effects as discussed in a previous section - in essence, using audio sentences from one speaker eliminates the possibility that differences in ratings in the mainstream audio condition versus the Southern audio condition could be attributed to phonetic properties other than accent (e.g. gender, age, pitch, etc.) that might differ between two speakers.

All audio was recorded by a speaker from Hattiesburg, Mississippi, who has lived in the Northeast United States for 5 years and can produce both mainstream and Southern accents naturally. He was told to emphasize /aj/ monophthongization for the Southern audio recordings, and he was given clarification for how to pronounce ungrammatical sentences. By nature, ungrammatical sentences are not ones that humans naturally produce and vocalize, and it follows that the ungrammatical control sentences for this survey are awkward to say aloud. In an attempt to make the ungrammatical control audio recordings as natural-sounding as possible, I designed these sentences to have a close grammatical corollary, and instructed my speaker to model his intonation after this corollary.

(10) Ungrammatical control sentences and their grammatical corollaries

a.	*They decided would need limes.	*ungrammatical control sentence*
b.	They decided they would need limes	*grammatical corollary*
c.	*He seems that is a dishonest person.	*ungrammatical control sentence*
d.	He seems to be a dishonest person.	*grammatical corollary*
e.	*She your present put over there.	*ungrammatical control sentence*
f.	She put your present over there.	*grammatical corollary*

4 Methods: Survey format and distrbution

This experiment was composed of three pairs of surveys numbered 1 through 3, each with versions A and B, for a total of 6 surveys. In Survey 1, sentences were presented in written form. In Survey 2, sentences were presented in audio form, and were spoken in a standard English accent. In Survey 3, sentences were presented in audio form, and were spoken in a Southern US accent. Versions A and B of each survey type contained the same grammatical and ungrammatical control sentences but slightly different test sentences. This experiment included 8 pairs of test sentences (16 total) - each pair matched exactly except for one word - with A versions containing one of the pair and B versions containing the other of the pair. (see subsection 3.2 for a more detailed description, and Appendix B for a list of all sentences). The six surveys (1A, 1B, 2A, 2B, 3A, 3B) for this experiment were created using Qualtrics, a web-based survey tool for which Yale has an institutional license, and all shared the same design.

All surveys were uploaded to MTurk and made available as Human Intelligence Tasks (HITs) for MTurk workers from selected Southern U.S. states to complete. MTurk workers who elected to participate were first presented with a consent form, and were asked to confirm that they would provide sentence judgments according to their standards of informal speech, and refrain from completing multiple of these six related surveys. Participants were then presented with more detailed instructions about how to give sentence judgments; these instructions explained the 1-5 acceptability scale and showed an example sentence in the appropriate written or audio format as pertained to that particular survey (see Appendix A for a reproduction of survey instructional materials provided to participants). Next, participants rated each of twenty-five sentences for that particular survey (see Appendix B for a list of all sentences), which were presented in random order with each sentence appearing on its own page. I felt that page breaks were necessary to minimize differences between written and audio surveys - without page breaks, the written surveys would allow people to look ahead at upcoming sentences whereas the audio surveys would not. After completing the sentence judgment task, participants were asked to provide some demographic information: age, gender, race/ethnicity, education, current city & number of years lived there, hometown & number of years lived there, and parent/guardian's hometowns & number of years they lived there. Once demographic information was filled out, participants were asked to input their MTurk worker ID, and a debrief on the purpose of the experiment and a comments box was provided to those participants who might be interested. Lastly, Qualtrics generated a random code for participants to copy into MTurk to verify that they completed the survey.

I accepted 80 responses per survey, and had five days to accept or reject these responses on MTurk. All 480 responses were ready for review about 3 to 4 hours after all the surveys were first published. I paid MTurk workers $1 for completing one survey, and paid all who completed the surveys in good faith, i.e. took an appropriate amount of time to complete the task and entered a correct completion code. Responses from participants who 1) entered an incorrect completion code, 2) completed multiple surveys, or 3) spent less than half the average time to complete the task were rejected. I rejected the work of 17 participants who fit one or more of these criteria, which amounted to about 3% of the original survey responses. I then republished a small amount of response openings corresponding to the work I rejected.

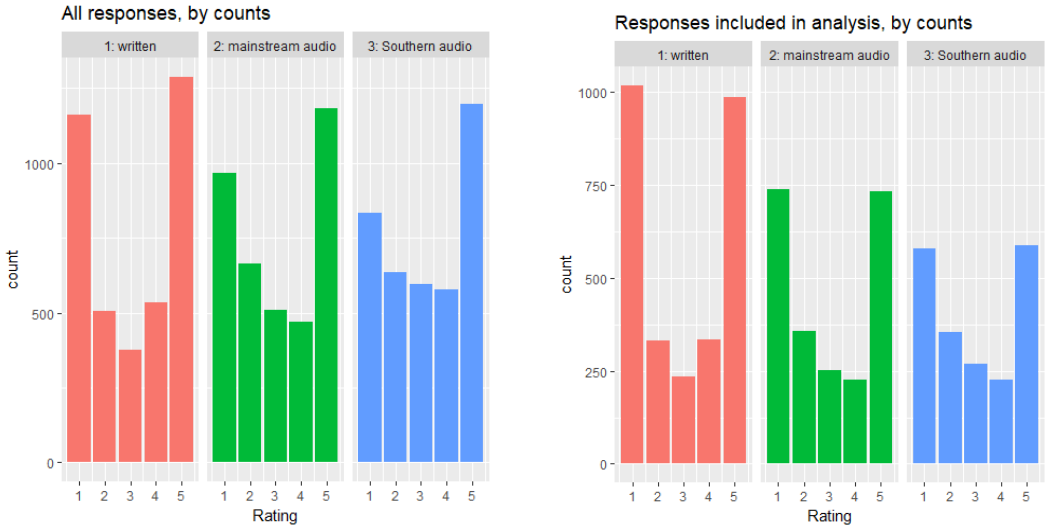
Following YGDP methodology, I excluded 17 participants from analysis who had primary childhood residences outside the United States; however, unlike YGDP research, this experiment did not involve any mapping or other geographic analysis, so I did not exclude

participants who had lived in their primary childhood residence for under 7 years. I also excluded participants on the basis of their control sentence judgments, again according to YGDP protocol. Participants failed grammatical controls if they rated one or more grammatical sentences as 1 or 2 and had an average grammatical judgment under 4, and failed ungrammatical controls if they rated one or more ungrammatical sentences as 4 or 5 and had an average ungrammatical judgment over 2.

5 Pre-Analysis Discussion

5.1 Unexpected differences in participant exclusion

One outcome of this experiment I was not expecting were the large differences between conditions in exclusion rates of participants who failed controls. The figure below shows this discrepancy visually, with many more participants in the two audio conditions failing to pass controls as compared to participants in the written condition.



(a) Before **(b)** After

Figure 6: Results by counts before and after excluding participants who failed controls

We see the most extreme difference between the written and Southern audio conditions, where the number of excluded participants is double in the Southern condition. It is also interesting to consider the reasons why participants were excluded from analysis. Taking a closer look, we see that the number of participants who failed both ungrammatical and grammatical controls varies widely. We see a much larger percentage of doubly disqualified participants in the Southern condition (20%) as compared to the written (5%); while the vast majority of excluded participants in all conditions are excluded because they failed ungrammatical controls, this higher rate of doubly disqualified participants in the Southern condition in turn makes the proportion of people who failed grammatical controls significantly higher in the Southern condition than in the written condition.

Figure 7: Number of participants excluded for failing controls

Condition	Total participants	# excluded	% excluded
Written	159	40	25.2%
Mainstream audio	155	62	40%
Southern audio	160	76	47.5%

Figure 8: Reasons for failing controls

Condition	% failed grammatical	% failed ungrammatical	% failed both
Written	5/40 = 12.5%	36/40 = 90%	1/40 = 2.5%
Mainstream audio	8/62 = 12.9%	59/62 = 95.2%	5/62 = 8.1%
Southern audio	16/76 = 21.1%	73/76 = 96.1%	13/76 = 17.1%

Essentially, participants who failed ungrammatical controls (most of the excluded participants) were much more likely to have also failed the grammatical controls in the Southern condition as compared to the other two conditions. According to an N-1 Chi-squared test for multiple proportions, this "double disqualification" rate is statistically significantly higher in the Southern condition as compared to the written condition (p-value <.0001). In the mainstream condition, we see more participants fail the controls as compared to the written condition but fewer as compared to the Southern condition, and mainstream participants fail grammatical and ungrammatical controls at roughly proportional rates to written participants - there is not significantly higher double disqualification rate in this condition that throws these proportions off.

This interesting discrepancy by condition in rates of disqualification on the basis of control sentences seems to indicate that participants in the Southern condition had a harder time completing the task accurately. It isnt immediately clear why there were so many more double disqualifications in this condition as compared to the other audio condition; this is something to keep in mind for future research.

5.2 Misperceptions of ungrammatical sentences in an auditory context

One likely contributing factor to higher rates in the audio conditions of disqualification of participants who fail controls is misperception of ungrammatical sentences. As mentioned previously, ungrammatical sentences are not ones that humans naturally produce and vocalize. I kept this in mind when designing the ungrammatical control sentences for this experiment, choosing a close grammatical corollary for the ungrammatical controls and instructed my speaker to model his intonation after the corollary; for example, the speaker was instructed to make his pronounciation of ungrammatical control sentence **She your present put over there* as close as possible to grammatical corollary *She put your present over there.* However, I did not consider how strange it would be for survey participants to encounter an ungrammatical sentence in an auditory context. Again, these are not sentences that one ever encounters in this format, and it is possible that some participants in the audio conditions of this experiment misheard ungrammatical control sentences in a desperate attempt to parse or salvage them. For ungrammatical control sentences U1026 and U1027, we see a sharp difference in rating distribution between written and one or both audio conditions

that can be explained by this mishearing effect.

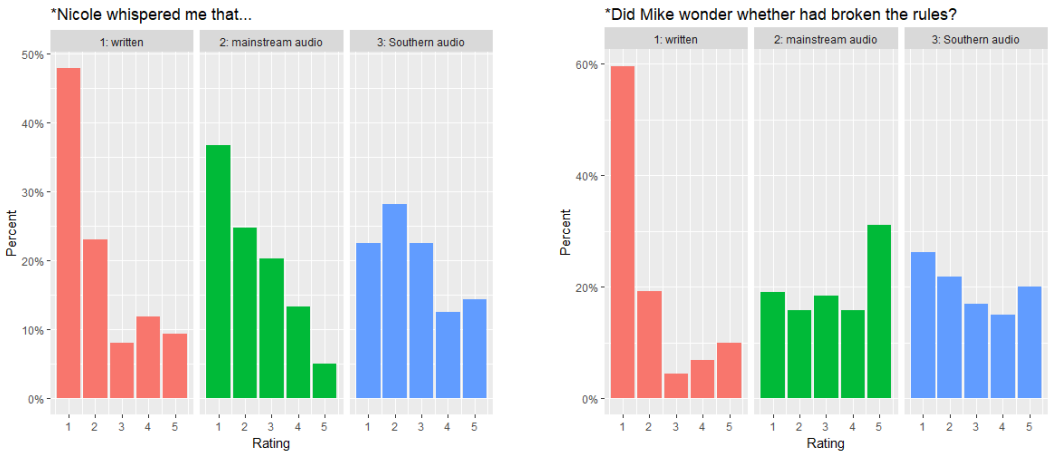
(11) Potential mishearings of ungrammatical control sentences U1026 and U1027

a. *Nicole whispered me that we should drive away from here. *U1026*

b. Nicole whispered **to** me that we should drive away from here.

a. *Did Mike wonder whether had broken the rules? *U1027*

b. Did Mike wonder whether **he'd** broken the rules?



(a) Rating distributions of U1026

(b) Rating distributions of U1027

Figure 9: Ungrammatical control sentences with results that display potential mishearing effects

As indicated in the example above, U1026 could have been misheard as grammatical by hallucinating *"to"*, which can appear very phonologically reduced after a past tense verb ending in "-ed", and U1027 could have been misheard as grammatical due people mistaking *had* for *he'd*. I did make a last minute alteration to one ungrammatical control anticipating that mishearing might be an issue, changing U1028 from *That man **liked** to you* to *That man* likes *to you*, after mishearing the original recording as grammatical myself: *That man **lied** to you*. Still, I was unable to anticipate these other two ungrammatical sentences that were susceptible to mishearings, and this goes to show that using ungrammatical sentences as controls can be difficult in an audio context.

6 Analysis

6.1 The Kruskal-Wallis test

I chose to use the K-W test, a non-parametric analog to the standard analysis of variance (ANOVA), to analyze my results. The K-W test is appropriate for ordinal data, makes no assumption about normal distribution, and compares entire distributions against each other, returning a significant result if it is likely that these distributions were obtained from different populations, i.e. if these distributions vary significantly from one another.

Experimental data on a Likert scale of 1-5, like the responses to this survey, have been established to be ordinal rather than interval data meaning that while a rating of 2 is greater than a rating of 1, it is not necessarily true that the distance between ratings of 1 and 2 is the same distance between ratings of 2 and 3. Thus, Likert scale data does not technically meet the assumptions for an ANOVA, which tests for significant differences in mean between three or more independent survey groups and requires interval data. However, as Wood & Tyler (2018) point out in this excerpt, many linguists continue to use parametric statistical tests on Likert scale data, and these tests seem to perform perfectly fine in most cases.

> ...there is much evidence showing that surveys using Likert scales do not actually yield interval datathey yield ordinal data (citations omitted). We could take this to mean that since parametric statistical tests such as t-tests and ANOVAs assume interval data, Likert data should not be submitted to such tests. However, it is overwhelmingly common to treat Likert data as if it were interval data, and submit them to parametric statistical tests when those tests are informative. These tests are sufficiently robust that violating the interval assumption is unlikely to lead to erroneous conclusions (citations omitted). Though this may remain controversial ...it is likely that few if any substantive conclusions have been in error simply because parametric statistics were used on inherently ordinal Likert data (Wood & Tyler 2018).

Regardless of this debate, using an ANOVA would be inappropriate for the purposes of this methodological experiment; it collapses distinguishing features of distributions that might look very differently from one another while still sharing similar means. Consider the following scenario. Say that we have sentence judgment data from two related sentences, Sentence A and Sentence B. Sentence A received an equal number of 1 and 5 ratings; in other words, participants had only extreme judgments of this sentence, either they completely accepted it or completely rejected it. On the other hand, sentence B received only ratings of 3; no participants were able to give a firm judgment of this sentence one way or the other. In this scenario, both Sentence A and B would have a mean of 3, and an ANOVA would not turn up a significant result when comparing these, even though their rating distributions could not be more different.

In the context of this experiment, now imagine that the same two distributions were observed, but Sentence B was actually the same as Sentence A except for the fact that one was an audio recording and the other was written sentence. Having these completely opposite distributions for the same sentence tested in two different conditions would be an extremely noteworthy result, regardless of which condition elicited which responses why would the same sentence elicit only ratings of 3 in one format, when people clearly have strong and polar judgments about this sentence when it is tested in the other format? A scenario this extreme is unlikely, to be sure, but it is entirely possible that a sentence could receive stronger judgments in one condition as compared to another; this issue has not been looked into in this depth previously. A result like this would be incredibly relevant to the methodological research questions at hand, as it would indicates that either people are not rating the same sentence due to mishearing of an audio sentence or misreading of a written sentence, or that the same sentence is being judged differently depending on what format it appears in. Both of these would have a bearing on the feasibility of using audio sentences in acceptability judgment surveys.

6.2 Overview of findings

I used the Kruskal-Wallis test to identify significant differences in rating distribution by condition at many levels. First, I applied the test to the entire dataset, comparing the overall rating distributions in each condition (and doing subsequent pairwise comparisons); next, I applied the test to each sentence type individually, comparing 1) test sentence rating distributions, 2) grammatical control sentence rating distributions, and 3) ungrammatical control sentence rating distributions by condition (and doing subsequent pairwise comparisons); and finally, I applied the test to individual sentences to compare rating distributions by condition (and doing subsequent pairwise comparisons). I used an alpha-level of 0.05 throughout and used the Bonferroni correction for pairwise comparisons. I will elaborate on the K-W significance findings and percent differences in acceptance and rejection between survey condition at each of these levels in the following sections, but first, here are two tables summarizing the significant results.

Figure 10: Summary of large Kruskal-Wallis test results

Is there a sig. diff. in rating distributions between...	**Answer**	**p-value**
All ratings from all conditions	no	0.927
All ratings from written and mainstream audio	no	0.728
All ratings from written and Southern audio	no	0.878
All ratings from mainstream and Southern audio	no	0.77
Test ratings from all conditions	**yes**	**<.0001**
Test ratings from written and mainstream audio	**yes**	**<.0001**
Test ratings from written and Southern audio	**yes**	**<.0001**
Test ratings from nainstream and Southern audio	**yes**	**<.0001**
Grammatical control ratings from all conditions	**yes**	**<.001**
Grammatical control ratings from written and mainstream audio	no	0.089
Grammatical control ratings from written and Southern audio	**yes**	**<.01**
Grammatical control ratings from mainstream and Southern audio	**yes**	**<.0001**
Ungrammatical control ratings from all conditions	**yes**	**<.0001**
Ungrammatical control ratings from written and mainstream audio	**yes**	**<.0001**
Ungrammatical control ratings from written and Southern audio	**yes**	**<.0001**
Ungrammatical control ratings from mainstream and Southern audio	no	0.434

Figure 11: Individual sentences with significant Kruskal-Wallis results

Sentence type	**Sig. across all conditions**	**Sig. between 2 conditions**
Test	P1005	P1005, P1012
Grammatical control	G1019	G1019
Ungrammatical control	U1025, U1026, U1027, U1030	U1026, U1027, U1030

6.3 Applying K-W test to entire dataset

At the level of the entire dataset including all sentence types, the rating distributions of the three conditions were not statistically significantly different (p-value = 0.87). Pairwise

comparisons of the entire dataset also fail to yield significant differences in rating distribution between any two survey conditions, though from the graph and tables above we do see higher rates of middling judgments 2-4 and lower rates of extreme judgments 1 and 5 in the audio conditions as compared to the written condition. This is seen most clearly when comparing the written and Southern audio conditions; nearly 70% of all judgments given in the written condition were 1s and 5s, as compared to about 56% in the Southern audio condition. As an aside, there could be something to this trend of milder judgments in audio conditions as compared to the written condition that could potentially be detected by imposing more effective controls or getting more participants. Still, these differences are not significant in this experiment; additionally, when collapsing judgments into the larger combinatory labels of "accept" and "reject", as we see in the tables in Figure 13, all conditions have roughly the same percentage of responses in each respective category, with only a 2% difference for total rejections and 5% difference for total acceptances.

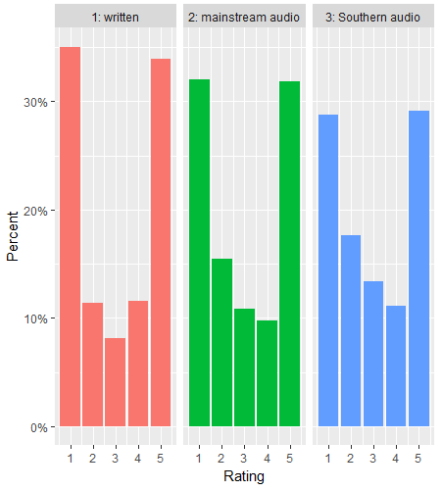


Figure 12: Rating distributions of all conditions including all sentence types

6.4 Applying K-W test to sentence types and individual sentences

While we do not see a significant difference in rating distribution between conditions when considering ratings from all sentence types, looking at sentence types individually does result in significant differences in rating distribution between conditions. This is true for all three sentence types: test sentences, grammatical controls, and ungrammatical controls. However, significant differences are not always found when comparing two conditions at a time; there was no significant difference found in grammatical control rating distribution between written and audio sentences, and likewise for the ungrammatical control rating distributions of the mainstream and Southern audio conditions. In the following subsections, I examine these results more closely and look at individual sentences within these types that contribute to these results.

6.4.1 Ungrammatical control sentences

Examining only ungrammatical control sentences, we do find significant differences in rating distribution between all three conditions (p-value $<.0001$). Pairwise comparisons reveal

Figure 13: Combinatory acceptance/rejection rates in all conditions

written judgment	N	%		combined N	combined %
1	1017	35.02%	REJECT	1349	46.45%
2	332	11.43%			
3	235	8.09%			
4	335	11.54%	ACCEPT	1320	45.46%
5	985	33.92%			
TOTAL	2904				

mainstream audio judgment	N	%		combined N	combined %
1	738	32.03%	REJECT	1095	47.52%
2	357	15.49%			
3	251	10.89%			
4	225	9.77%	ACCEPT	958	41.58%
5	733	31.81%			
TOTAL	2304				

Southern audio judgment	N	%		combined N	combined %
1	579	28.72%	REJECT	934	46.33%
2	355	17.61%			
3	270	13.39%			
4	225	11.16%	ACCEPT	812	40.28%
5	587	29.12%			
TOTAL	2016				

that the distributions of ungrammatical control sentences from the two audio conditions are not significantly different from each other (p-value = 0.43), but they are both significantly different from that of the written condition (both p-values <.0001). This result is expected, given previous discussion of the two ungrammatical sentences in particular that audio participants had a difficult time correctly identifying. However, the overall rejection rate for ungrammatical sentences across all conditions is within 6% of the same percentage, even though those two problem sentences were included in the analysis. This seems to indicate that the established procedure for excluding participants based on their performance on ungrammatical controls works perfectly well for audio conditions.

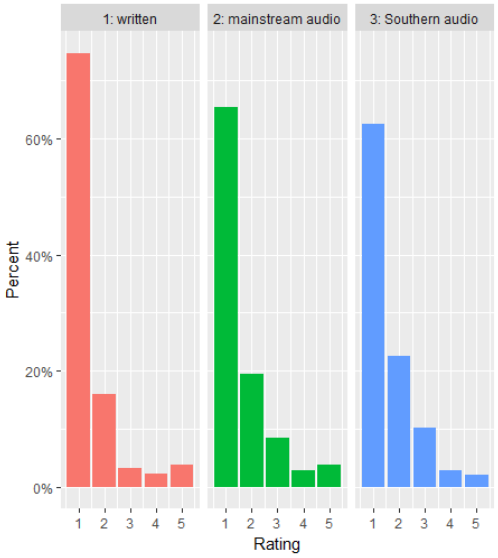


Figure 14: Rating distributions of ungrammatical control sentences in all conditions

There were four ungrammatical sentences which had significantly different distributions based on condition, U1025, U1026, U1027 and U1030.

(12) Ungrammatical sentences with significantly different rating distributions by condition

a.	*They decided would need limes.	*U1025*
b.	*Nicole whispered me that we should drive away from here.	*U1026*
c.	*Did Mike wonder whether had broken the rules?	*U1027*
d.	*He seems that is a dishonest person.	*U1030*

Sentences U1026 and U1027, previously identified as having had potential issues with mishearing, show more dramatic differences in rating distribution between conditions than sentences U1025 and U1030, for which there is no obvious grammatical mishearing I can think of. Another distinction between these two pairs of sentences is that for U1025 and U1030, the audio condition rating distributions are not both significantly different from the written rating distribution, as is true for U1026 and U1027; U1025 has a significantly different rating distribution between the written and Southern audio conditions, and U1030 has a significantly different rating distribution between the written and mainstream audio conditions. However, the fact that there is a significantly different rating distribution for U1025 and U1030 between the written condition and one of the two audio conditions still seem to demonstrate the fact that rejecting ungrammatical control sentences comes less naturally to participants in audio conditions (see Appendix B for graphs).

Figure 15: Combinatory acceptance/rejection rates for ungrammatical sentences

written judgment	N	%		combined N	combined %
1	723	74.69%	REJECT	877	90.6%
2	154	15.91%			
3	32	3.31%			
4	22	2.27%	ACCEPT	59	6.09%
5	37	3.82%			
TOTAL	968				

mainstream audio judgment	N	%		combined N	combined %
1	502	65.36%	REJECT	651	84.76%
2	149	19.4%			
3	65	8.46%			
4	22	2.86%	ACCEPT	52	6.77%
5	30	3.91%			
TOTAL	768				

Southern audio judgment	N	%		combined N	combined %
1	420	62.5%	REJECT	571	84.97%
2	151	22.47%			
3	68	10.12%			
4	19	2.83%	ACCEPT	33	4.91%
5	14	2.08%			
TOTAL	672				

6.4.2 Grammatical control sentences

Examining only grammatical control sentences, we again find significant differences in rating distribution (p-value <.001). Pairwise comparisons reveal the Southern audio condition has a grammatical control rating distribution that is significantly different from that of both the written condition (p-value <.0001), and the mainstream audio condition (p-value = 0.006), but that the written and mainstream audio condition do not have grammatical control rating distributions that are significantly different from each other (p-value = 0.08).

From the graph of rating distributions of all grammatical control sentences below, we see that these differences do not appear to be that drastic, and I have omitted tables showing the combinatory acceptance/rejection rates for grammatical sentences in the interest of space. In fact, it is possible that one sentence is largely responsible for the significantly different rating distribution in the Southern condition as compared to the other conditions.

The Kruskal-Wallis test identifies grammatical control sentence G1019 as having a significantly different rating distribution by condition, and pairwise comparisons (and our eyes) reveal that the Southern audio condition has a starkly different distribution from the other two conditions, with nearly 30% of Southern audio participants (who passed the controls!) rejecting this sentence as opposed to nearly 0% of participants in other conditions. This sentence also is likely partially responsible for making the Southern audio condition distribution for grammatical control sentence ratings significant from those of the other condition. Upon revisiting the Southern audio recording for this sentence, it seems possible that participants

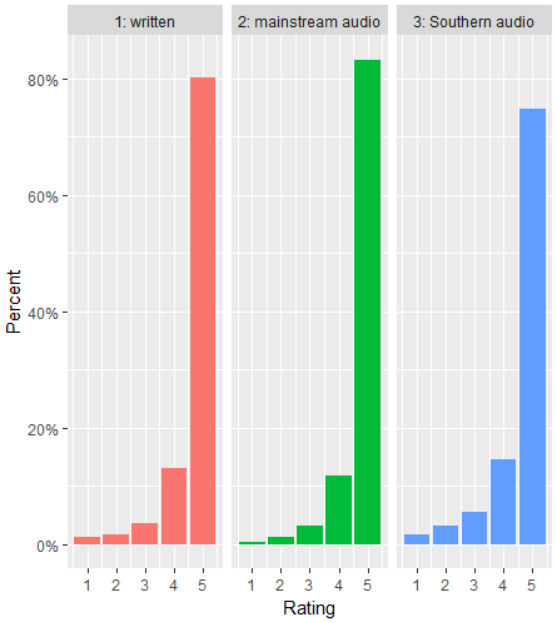


Figure 16: Rating distributions of grammatical control sentences in all conditions

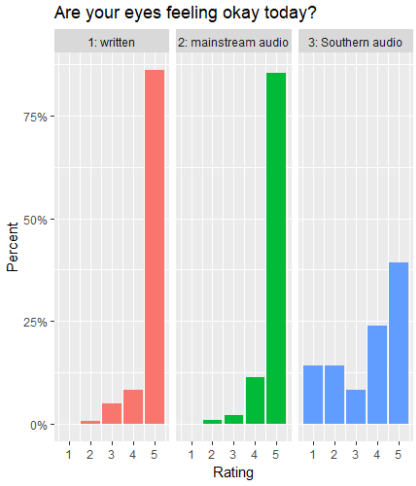


Figure 17: Rating distributions of grammatical control G1019

did not have a problem with the sentence as it was intended to be heard but they misheard it as *Are your ice/ass[?] feeling okay today?*

6.4.3 Test sentences

Understandably, it is in examining only test sentences that we find the most variation in rating distribution (p-value <.0001), since these sentences are not designed to elicit an acceptable or unacceptable response, but instead are intended to measure variable acceptability of a test feature. In this case, pairwise comparisons reveal that rating distributions from the all conditions are significantly different from each other: written vs. mainstream audio, written vs Southern audio, and Southern audio vs mainstream audio rating distributions are all significantly different from each other (all p-values <.0001). Visually, we see that ratings for test sentences in the written condition have a spike in the 1s, and roughly even rates for ratings of 2-5, while the ratings for the audio condition display a cascading pattern, with percentages decreasing as judgment values increase. We do see that the Southern condition has the highest percentage of ratings concentrated in the 2s column as compared to the mainstream condition which has the highest percentage of ratings concentrated in the 1s column, as also seen in the written condition.

Interestingly, this result suggests that fewer people were willing to fully reject a Southern construction when it appeared in a Southern accent as opposed to a mainstream accent or in written form. By the numbers, we see that using the appropriate accent for the test feature did positively impact judgments, resulting in 6% fewer rejection ratings and 4% more acceptance ratings for test sentences in the Southern audio condition as opposed to the mainstream audio condition.

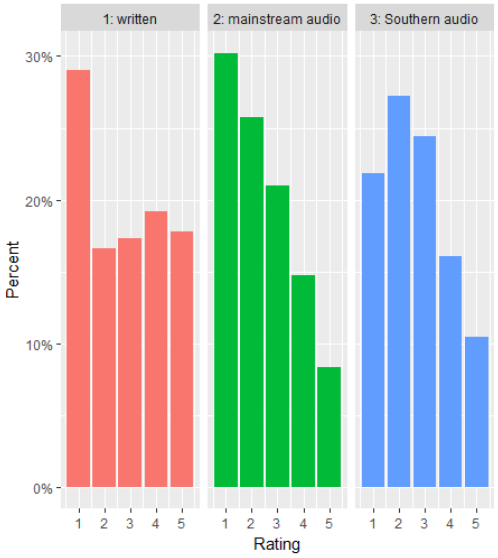


Figure 18: Rating distributions of test sentences in all conditions

Still, with regards to overall acceptance and rejection rates, the written condition has a lower rejection rate for the test feature than the Southern condition. The written condition has highest acceptance and lowest rejection, the mainstream audio condition has the lowest acceptance and highest rejection, with the Southern condition falls somewhere in between in both categories. However, it is important to note that all conditions exhibit rejection rates of less than 60% and acceptance rates of more than 20% - even the mainstream audio condition does produce a considerable acceptance rate for dative presentative constructions despite the fact that its rate was lowest out of the three conditions. In other words, all three of these survey conditions do obtain a range of judgments that display roughly similar overall results.

This is not true at the individual sentence level. Two of eight test sentence pairs are identified as having significantly different rating distributions by condition, and a closer examination of these results indicate that researchers could draw substantively different conclusions from these sentences depending on the format they were tested in. As a reminder, test sentences were designed in pairs to ascertain that the presence of an /aj/ lexical item was not solely responsible for judgment differences. A Wilcoxon rank sum test found no significant differences based on presence or absence of /aj/ in these pairs, so I will talk about each pair as a single entity though the sentences within pairs were rated separately.

(13) Test sentences with significantly different rating distributions by condition

a. There's you a white/blue jacket. *P1005/P1006*

b. Here's you a nice/hot coffee. *P10011/P1012*

Test sentence pair P1005/P1006 has a significantly different distribution based on condition, and pairwise comparisons reveal that this difference lies between the written and audio conditions, but not between the audio conditions themselves. In contrast with previous distribution differences indicated as significant by the K-W test, we still see large differences in both overall acceptance and rejection rates according to condition.

Figure 19: Combinatory acceptance/rejection rates for test sentences

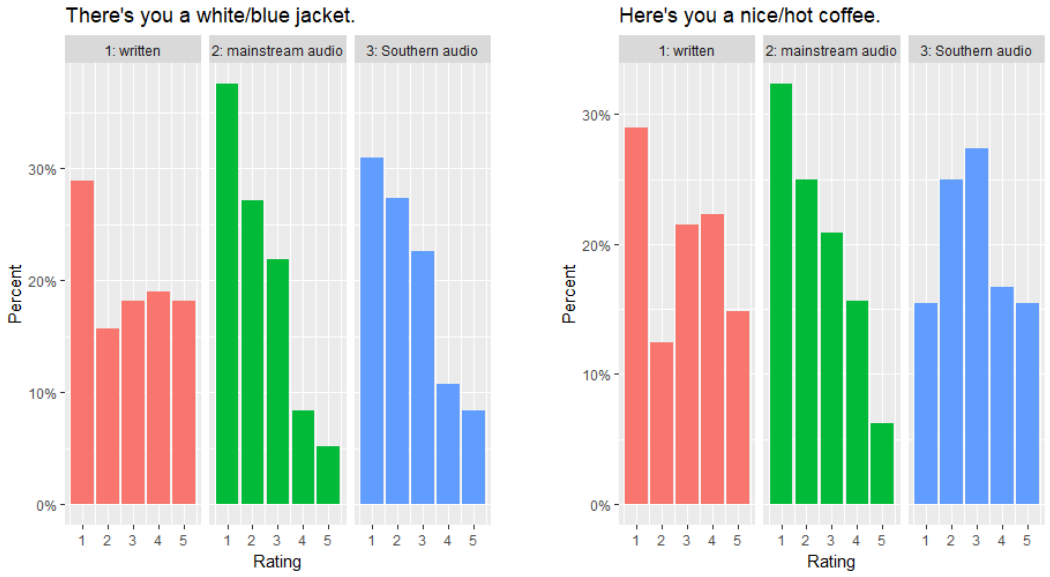
written judgment	N	%		combined N	combined %
1	281	29.03%	REJECT	442	45.66%
2	161	16.63%			
3	168	17.36%			
4	186	19.21%	ACCEPT	59	36.98%
5	172	17.77%			
TOTAL	968				

mainstream audio judgment	N	%		combined N	combined %
1	232	30.21%	REJECT	430	55.99%
2	198	25.78%			
3	161	20.96%			
4	113	14.71%	ACCEPT	177	23.04%
5	64	8.33%			
TOTAL	768				

Southern audio judgment	N	%		combined N	combined %
1	147	21.88%	REJECT	330	49.11%
2	183	27.23%			
3	164	24.4%			
4	108	16.07%	ACCEPT	178	26.49%
5	70	10.42%			
TOTAL	672				

P1005/1006 received a much lower percentage of rejection ratings in the written condition with 45%, as compared to nearly 60% in the Southern audio condition and over 67% in the mainstream audio condition (see Appendix C for complete table). The written condition also received a much higher percentage of acceptance ratings, with 38%, over double the percentage of acceptance ratings this sentence received in the mainstream condition and nearly double the acceptance percentage from the Southern condition as well. This result suggests that researchers quite possibly would have made substantively different claims about the acceptability of P1005/P1006 depending on condition: whether they tested it in a written or an audio format.

The second test sentence with a significant result, P1011/P1012, is equally interesting to consider for similar reasons. Unlike the previous test sentence, this sentence does not have significantly different rating distributions when taking all conditions into consideration, but a pairwise comparison between the two audio conditions does yield a significant result (p-value $<.01$). Like with the previous test sentence with a significant result, this significant result does manifest in differences between overall acceptance and rejection rates by condition. There is no significant difference in rating distribution between the written and Southern audio conditions, and only relatively small differences in overall acceptance and rejection rates between these two conditions; however, the mainstream audio condition received a quite noticeably larger percentage of overall rejection judgments and smaller percentage of overall acceptance judgments than these two conditions. The mainstream audio condition



(a) Rating distributions of P1005/P1006

(b) Rating distributions of P1011/P1012

Figure 20: Test sentence pairs with significantly different distributions between conditions

received about 20% more rejection judgments and 10% fewer acceptance judgments than did the written or Southern audio conditions (see Appendix C for complete table). Thus, P1011/P1012 is a clear case of accent making a substantive difference in judgment of the same sentence. Like the previous test sentence, researchers quite possibly would have made substantively different claims about the acceptability of this sentence depending on the survey condition; but in this case, what gives rise to a discrepancy in overall acceptance and rejection rates of a sentence is not a choice between written and audio format, but the choice of what accent to record sentences in if audio is to be used.

7 Conclusion and suggestions for future research

This exploratory look at the feasibility of using audio sentences for microsyntax research has been a productive one. Looking at the big picture, we see minimum acceptance rates of 20% and maximum rejection rates under 60% for dative presentative constructions regardless of whether these sentences appear in a written, mainstream audio, or Southern audio format, suggesting that all of these formats are conducive to examining this feature. We also see that grammatical and ungrammatical control sentences are appropriately accepted or rejected in the vast majority of cases after excluding participants who fail control measures, suggesting that these measures and established protocol for exclusion do function adequately for audio surveys.

While the overall rating distributions including all sentences are not significantly different based on condition, taking a closer look at the different sentence types and at individual sentences does yield some significant results. At the individual sentence level, 7 of 24 sentences have significant different rating distributions between conditions: 1 of 8 grammatical

sentences, 4 of 8 ungrammatical sentences, and 2 of 8 test sentences. Most of these significantly different distributions do not manifest in large differences in overall acceptance or rejection rates, but results from certain sentences do indicate that modality and/or accent have the potential to impact the judgment task. Ungrammatical control sentences seem to be difficult for audio participants to rate correctly and are subject to potential mishearings, and two test sentences have different enough rating distributions between conditions that researchers quite possibly would have made substantively different claims about the acceptability of these sentence depending on modality (P1005) or accent of speaker (P1012).

Improvements can certainly be made on the audio sentence methodology used in this experiment. Close attention should be paid to all sentences for potential mishearings that would affect judgments, particularly for ungrammatical controls which participants are unused to encountering in a spoken format. An experiment where participants are asked to give judgments on audio sentences but also to type what they heard would be very helpful in evaluating this issue. Along similar lines, it would be helpful to ask participants what strategies they employ when making judgments on written sentences. The YGDP task instructions, which were also used as the task instructions for this experiment, include a suggestion that participants read sentences aloud to help them give judgments. It would be interesting to know how many participants do so, and whether that could be part of the reason why for the test sentences in this experiment, the written (potentially spoken aloud by Southern participant) and Southern audio conditions have more similar patterns of overall acceptance and rejection as compared to the mainstream audio condition. It is also important to mention that the test sentences used in this experiment were extremely curated in order to create matching pairs of sentences with and without /aj/ adjectives and ensure that these items were not solely responsible for accent effects between the audio conditions. While we still see a minimum acceptance rate of 20% for the dative presentative constructions that appeared in these surveys, it would be good to replicate this study using more natural examples of the test feature.

Overall, this experiment indicates that the use of audio sentences in microsyntax sentence judgment tasks should be explored further. Audio sentences could lend themselves well to the study of constructions which do not have obvious spellings, for instance, the *should have* construction mentioned in Wood et al. 2015 could be stylized as *should have*, *should of*, or *shoulda*. These last two spellings represent the pronunciation of the feature more accurately but are instantly recognized as non-standard written forms; audio sentences might thus be better equipped to test acceptability of this feature in a less obvious way. I would also be interested to see this study replicated where there is no restriction on region and/or the feature at hand is not associated with a particular accent.

8 Appendices

8.1 Appendix A: Explanatory survey materials

8.1.1 Introductory statement and participation agreement

SENTENCE JUDGMENTS SURVEY

THIS IS 1 OF 6 RELATED SURVEYS. PLEASE TAKE ONLY 1 OF THESE SURVEYS. YOU WILL NOT BE REIMBURSED IF YOU TAKE MORE THAN 1 SURVEY.

Human Subjects Committee Consent
Purpose: We are conducting a research study to examine variation in American English.
Procedure: Participation in this study will involve judging the acceptability of a series of sentences in your own informal speech. We anticipate that your involvement will require 10 minutes or less.
Risks and Benefits: There are no known risks associated with participation in this study.
Confidentiality: All of your responses will be anonymous. While basic demographic information will be collected, no information that can identify you personally will be collected in this survey.
Voluntary participation: Participation in this study is completely voluntary. You are free to decline to participate, to end participation at any time for any reason.
Questions: There will be a comment box at the end of the survey if you have any questions about this study, or you may contact us at rachel.regan@yale.edu.

Agreement to Participate: By clicking this box, I certify that I have read the above information, have had the opportunity to have any questions about this study answered and agree to participate in this study. I understand I will not be reimbursed if I have already taken a related survey.

8.1.2 Sentence judgment instructions provided to participants

Sentence judgment instructions for this experiment were replicated from YGDP survey instructions, with a small modification depending on written or audio survey condition indicated in brackets:

Informal, casual language can be different in different places. The goal of this survey is to find out about your language, and the language spoken where you live and where you grew up.
We are <u>not</u> interested in what is correct or proper English.
We are instead interested in what you consider to be an acceptable sentence in informal contexts. You will be presented with a [written or audio recording of a] sentence. You will then judge the acceptability of that sentence on a scale of 1-5, with 1 being unacceptable and 5 being acceptable. It may help to read each sentence aloud before giving your judgment.

8.1.3 Debrief provided to participants

Participants were provided with a debrief of the experiment if they wished to know more.

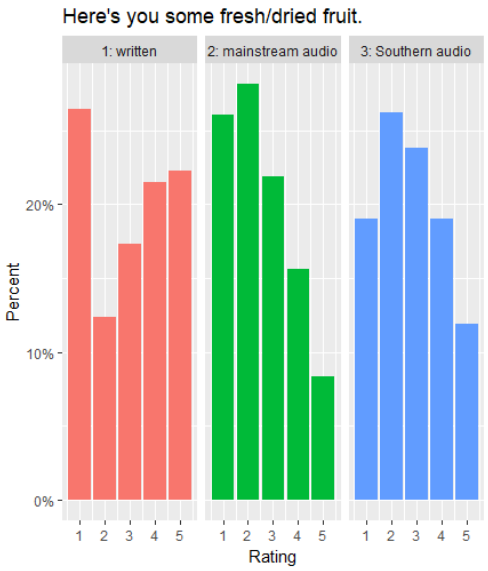
> This survey is part of a linguistics experiment whose primary goal is to compare acceptability judgments for dative presentatives in written and spoken test sentences. Dative presentatives, seen in the heres you portion of sentences like Heres you some tea, are widely accepted in the southeast United States.
>
> Stigma against regional language variation in America can complicate this research. People may have a hard time giving accurate judgments on whether a certain feature of language is acceptable to them, since they may have been criticized for using regional language. Its even possible that someone might say that a certain feature is unacceptable to them, even if they use it when they speak.
>
> Asking for judgments on written sentences may not be the best way to conduct research on acceptability of regional language, since the use of standard/proper English is stressed so highly in writing, more so than in speech. Instead, maybe judgments on spoken sentences will be a more accurate measurement of acceptability, since spoken sentence may evoke a more informal, conversational setting. The aim of this experiment is to examine if there is a significant difference in how participants judge the same sentences, depending on if they are presented in written or spoken format.

8.2 Appendix B: Survey sentences and all graphs

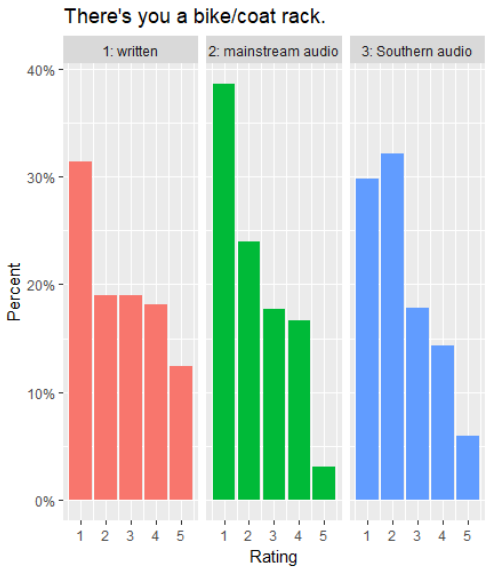
Below are listed all sentences included in the surveys of this experiment, with items containing /aj/ diphthong indicated in *italics*. Type A surveys contained primary test sentences 1001, 1003, 1005, 1007, 1010, 1012, 1014, and 1016. Type B surveys contained primary test sentences 1002, 1004, 1006, 1008, 1009, 1011, 1013, and 1015. Type A and B surveys contained all grammatical and ungrammatical control sentences.

Type	No.	Sentence
Primary	1001	Here's you some *dried* fruit.
Primary	1002	Here's you some fresh fruit.
Primary	1003	There's you a *bike* rack.
Primary	1004	There's you a coat rack.
Primary	1005	There's you a *white* jacket.
Primary	1006	There's you a blue jacket.
Primary	1007	Here's you some *live* bait.
Primary	1008	Here's you some fish bait.
Primary	1009	Now there's you a *prize* hog.
Primary	1010	Now there's you a large hog.
Primary	1011	Here's you a *nice* coffee.
Primary	1012	Here's you a hot coffee.
Primary	1013	Now there's you a *fine* story.
Primary	1014	Now there's you a funny story.
Primary	1015	Here's you a *pile* of papers.
Primary	1016	Here's you a stack of papers.
Control (Grammatical)	1017	Here, have some *ice cream*.
Control (Grammatical)	1018	Over here is where Martha keeps the *timer*.
Control (Grammatical)	1019	Are your *eyes* feeling okay today?
Control (Grammatical)	1020	We need to go *hiking* together.
Control (Grammatical)	1021	What do you need more of?
Control (Grammatical)	1022	Where is my goldfish at?
Control (Grammatical)	1023	There's never enough to do around here.
Control (Grammatical)	1024	Here's that book you asked for.
Control (Ungrammatical)	1025	They decided would need *limes*.
Control (Ungrammatical)	1026	Nicole whispered me that we should *drive* away from here.
Control (Ungrammatical)	1027	Did *Mike* wonder whether had broken the rules?
Control (Ungrammatical)	1028	That man *likes* to you.
Control (Ungrammatical)	1029	She your present put over there.
Control (Ungrammatical)	1030	He seems that is a dishonest person.
Control (Ungrammatical)	1031	That's when she scared me of ghosts.
Control (Ungrammatical)	1032	The loud noise startled she.

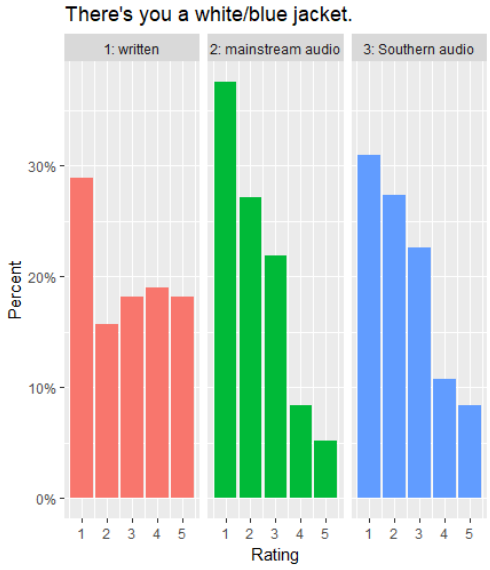
These surveys also included one of two *even* sentences to assist with research for my colleague Katie Martin's senior thesis (Martin 2018). Type A surveys contained *Does even the professor know what she's talking about?* and Type B surveys contained *Does even Mark Zuckerberg know what Facebook is?*.



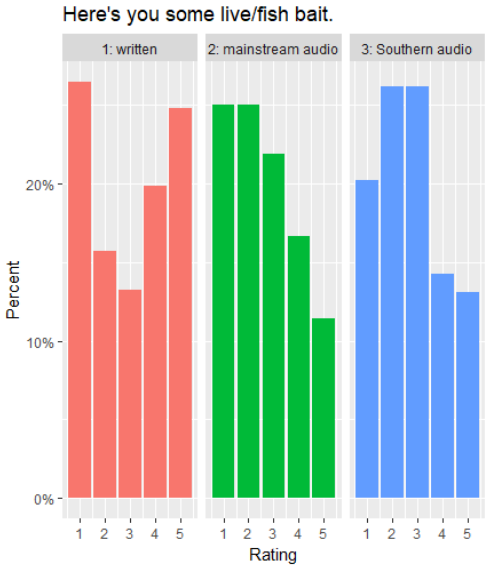
P1001/P1002



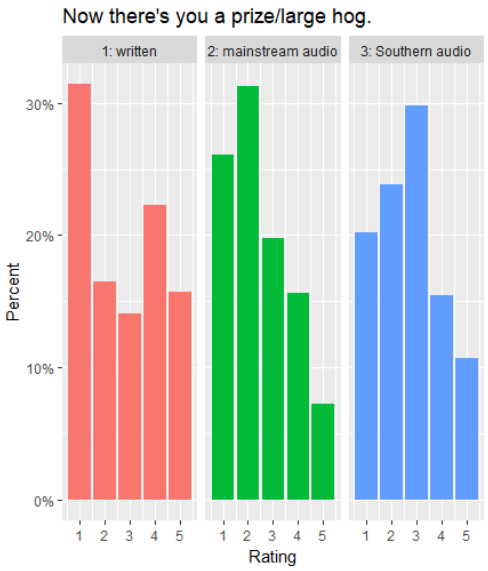
P1003/P1004



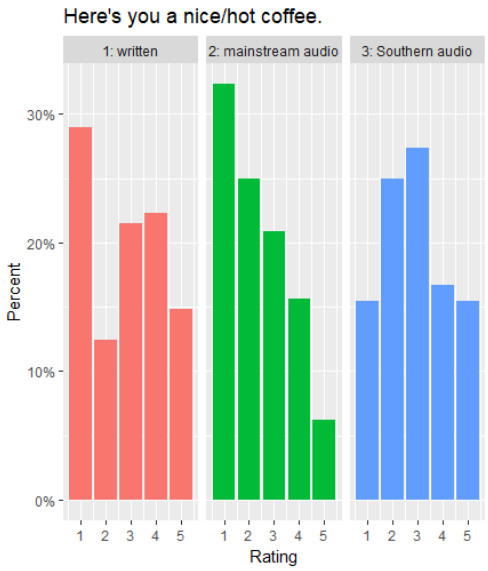
P1005/P1006



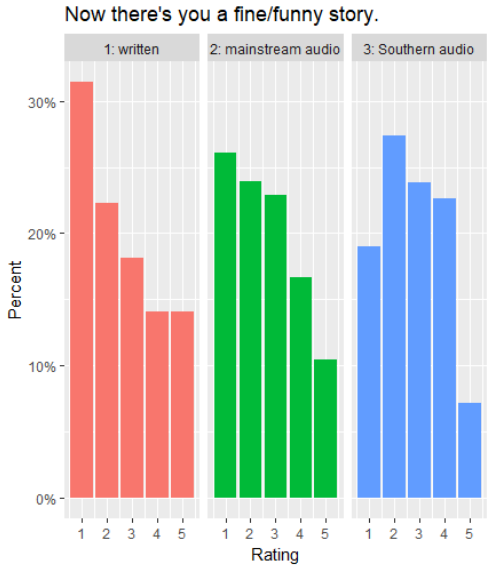
P1007/P1008



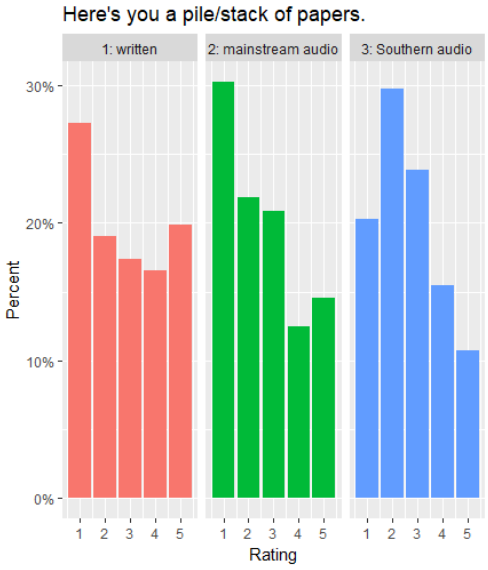
P1009/P1010



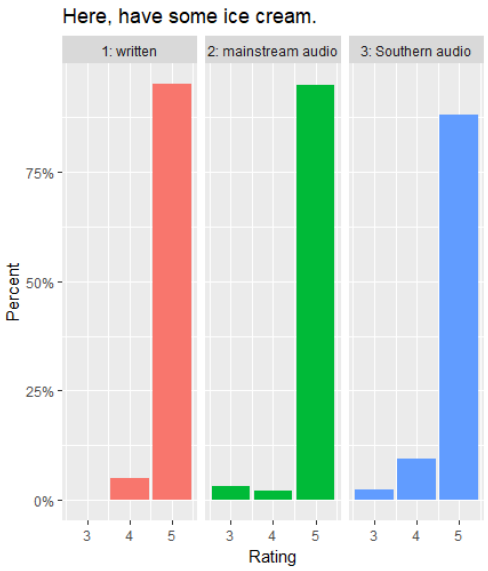
P1011/P1012



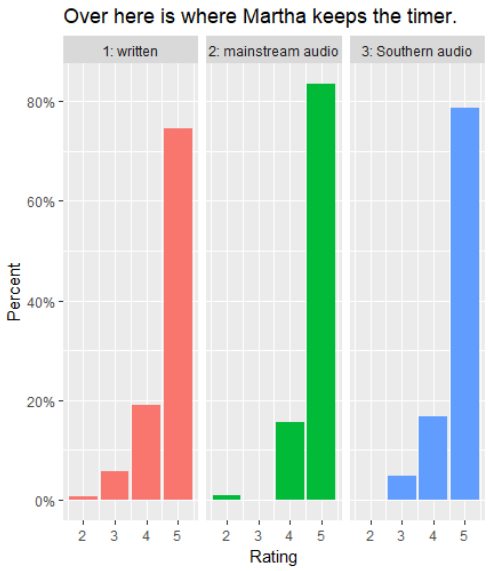
P1013/P1014



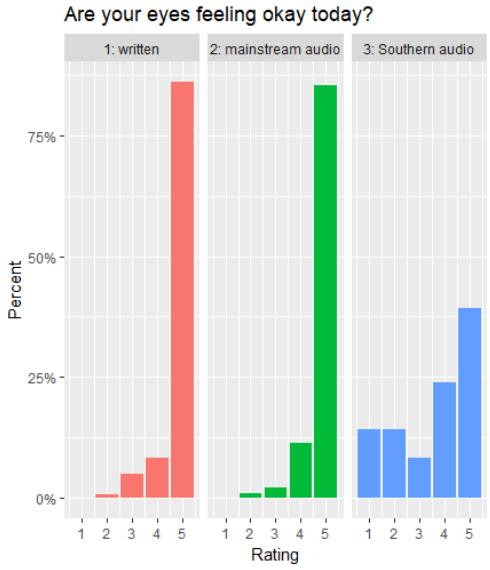
P1015/P1016



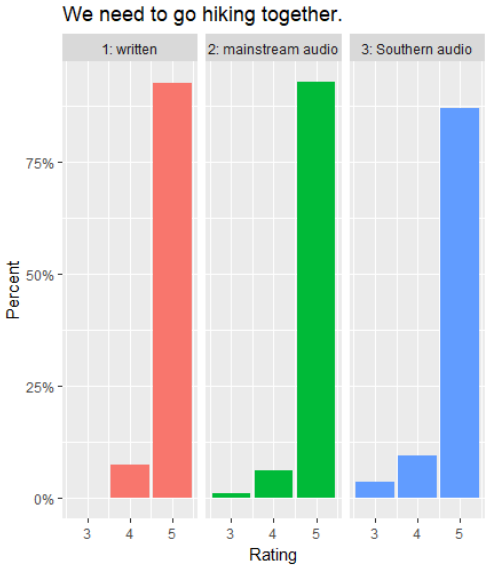
G1017



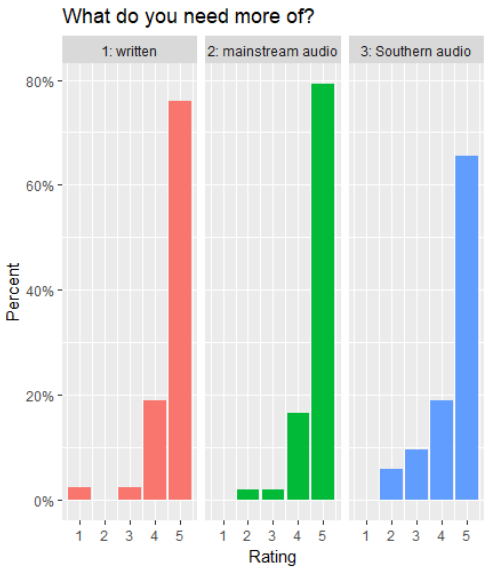
G1018



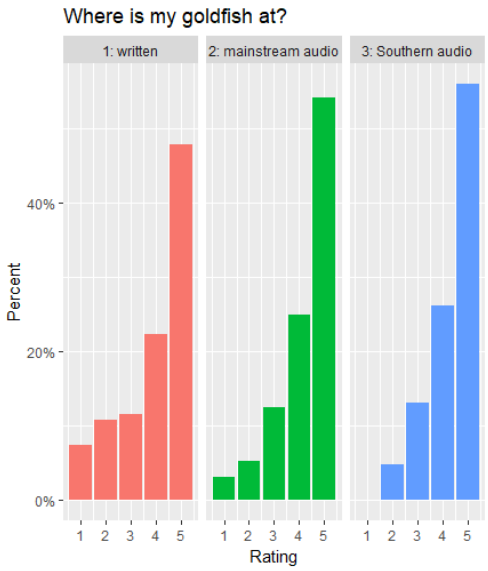
G1019



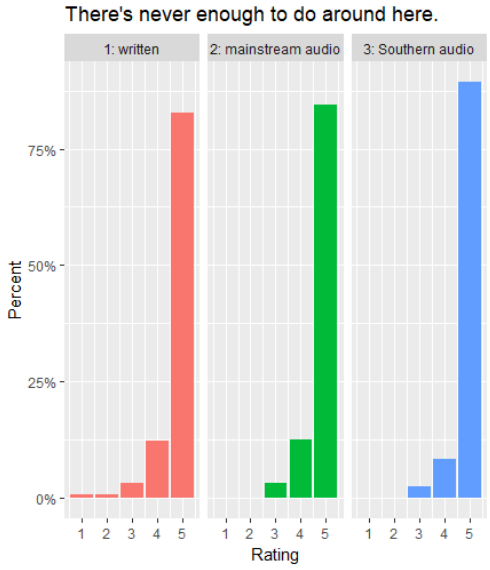
G1020



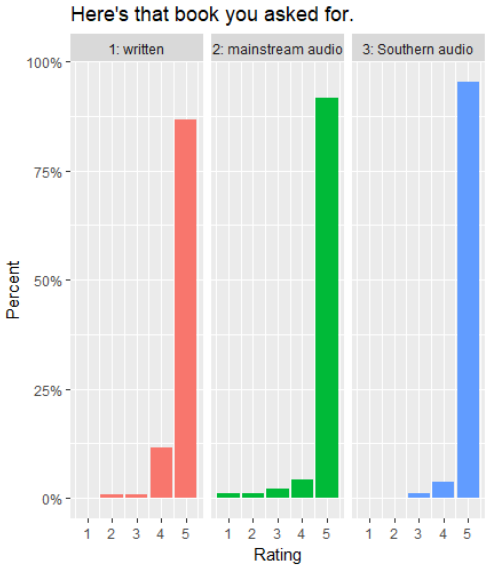
G1021



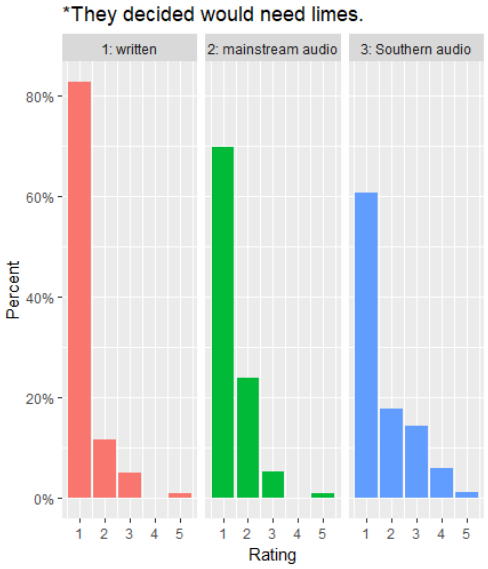
G1022



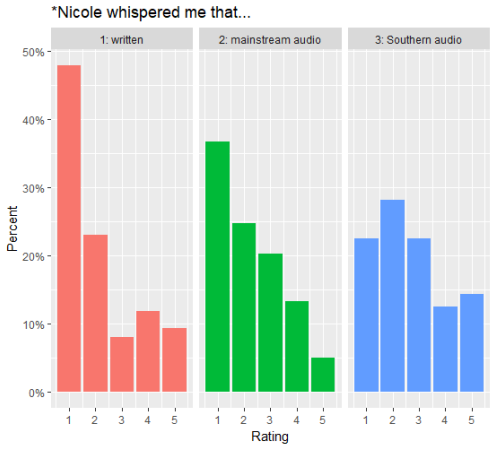
G1023



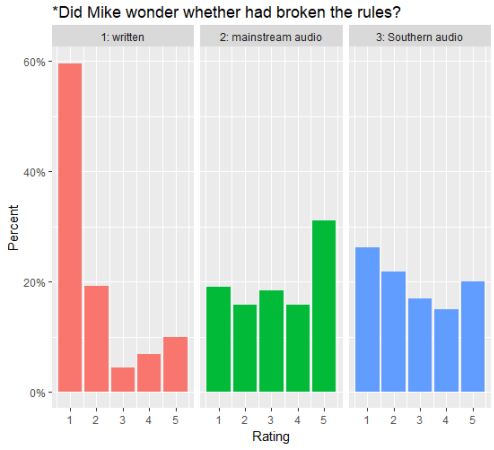
G1024



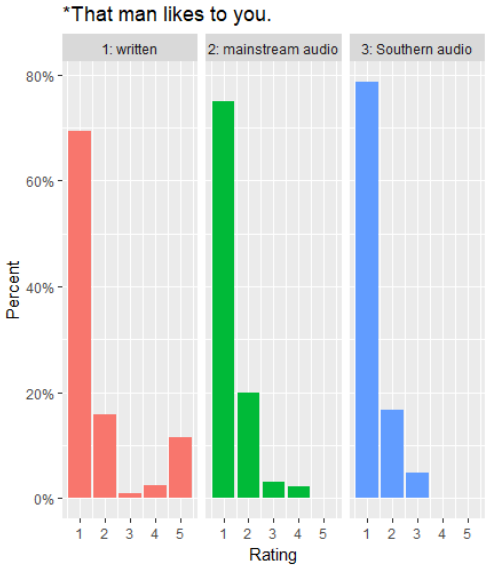
U1025



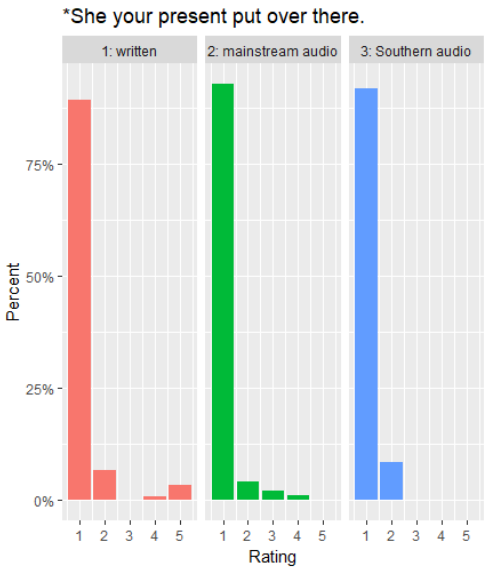
U1026



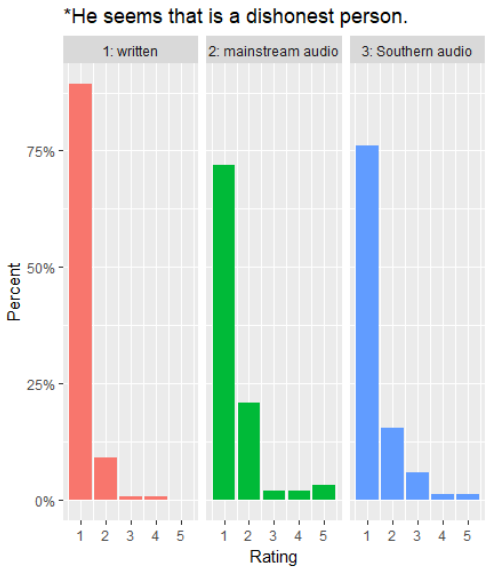
U1027



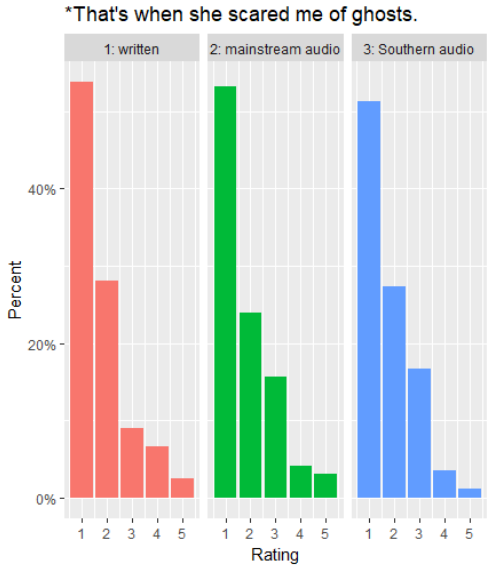
U1028



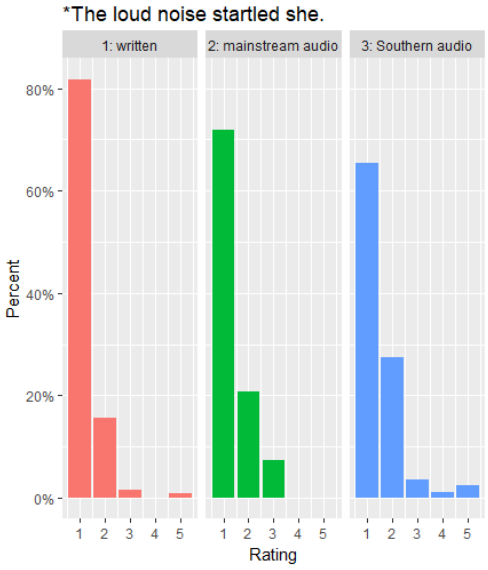
U1029



U1030



U1031



U1032

8.3 Appendix C: Tables with acceptance/rejection rates for significant test sentences

Figure 27: Combinatory acceptance/rejection rates for P1005/P1006

written judgment	N	%		combined N	combined %
1	35	29.41%	REJECT	54	45.38%
2	19	15.97%			
3	22	18.49%			
4	23	19.33%	ACCEPT	45	37.82%
5	22	18.49%			
TOTAL	121				

mainstream audio judgment	N	%		combined N	combined %
1	36	38.71%	REJECT	62	66.67%
2	26	27.96%			
3	21	22.58%			
4	8	8.6%	ACCEPT	13	13.98%
5	5	5.38%			
TOTAL	96				

Southern audio judgment	N	%		combined N	combined %
1	26	30.95%	REJECT	49	58.33%
2	23	27.38%			
3	19	22.62%			
4	9	10.71%	ACCEPT	16	19.04%
5	7	8.33%			
TOTAL	84				

Figure 28: Combinatory acceptance/rejection rates for P1011/P1012

written judgment	N	%		combined N	combined %
1	35	29.41%	REJECT	50	42.02%
2	15	12.61%			
3	26	21.85%			
4	27	22.69%	ACCEPT	45	37.82%
5	18	15.13%			
TOTAL	121				

mainstream audio judgment	N	%		combined N	combined %
1	31	33.33%	REJECT	55	59.14%
2	24	25.81%			
3	20	21.51%			
4	15	16.13%	ACCEPT	21	22.58%
5	6	6.45%			
TOTAL	96				

Southern audio judgment	N	%		combined N	combined %
1	13	15.48%	REJECT	34	40.48%
2	21	25%			
3	23	27.38%			
4	14	16.67%	ACCEPT	27	32.15%
5	13	15.48%			
TOTAL	84				

Bibliography

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2).

Bybel, Kali & Greg Johnson. 2014. The syntax of 'have yet to'. In *Paper presented at the 81st Southeastern Conference on Linguistics, march 27-29*, Coastal Carolina University.

Cowart, Wayne. 1997. *Experimental syntax: applying objective methods to sentence judgments.* Thousand Oaks, Calif.: Sage Publications.

Dudley, Fred A. 1946. 'Swarp' and some other Kentucky words. *American Speech* 21(4). 270–273. doi:10.2307/487323.

Edelman, Shimon & Morten Christiansen. 2003. How seriously should we take Minimalist syntax? *Trends in cognitive sciences* 7. 60–61. doi:10.1016/S1364-6613(02)00045-1.

Feagin, Crawford. 1979. *Variation and change in Alabama English: a sociolinguistic study of the White community.* Georgetown University Press.

Gibson, Edward & Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14(6). 233–234. doi:10.1016/j.tics.2010.03.005. `http://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(10)00052-5`.

Gibson, Edward, Steve Piantadosi & Kristina Fedorenko. 2011. Using Mechanical Turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass* 5(8). 509–524. doi:10.1111/j.1749-818X.2011.00295.x.

Green, Lisa J. 2002. *African American English: A Linguistic Introduction.* Cambridge University Press.

Harves, Stephanie & Neil Myler. 2014. Licensing NPIs and licensing silence: have/be yet to in English. *Lingua,* 148. 213–239.

Horn, Laurence R. 2008. 'I Love Me Some Him': the landscape of non-argument datives. In Oliver Bonami & Patricia Cabredo Hofherr (eds.), *Empirical issues in formal syntax and semantics 7: Papers from CSSP 2007*, Paris: Colloque de Syntaxe et Sémantique à Paris.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2013. *eWAVE.* Leipzig: Max Planck Institute for Evolutionary Anthropology. `http://ewave-atlas.org/`.

Labov, William. 1996. When intuitions fail. In Lisa McNair (ed.), *Papers from the parasession on theory and data in linguistics* CLS 32, Chicago: University of Chicago.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool.* Berlin ; New York: Mouton de Gruyter.

Martin, Katie. 2018. *Even and negative bias in polar questions.* New Haven, CT: Yale University undergraduate thesis.

Montgomery, Michael B. & Joseph S. Hall. 2004. *Dictionary of Smoky Mountain English.* Knoxville: Univ of Tennessee Pr first edition edition edn.

Purnell, Thomas, William Idsardi & John Baugh. 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18. 10–30. doi:10.1177/0261927X99018001002.

Rockwood Tennessee Police. 2016. Here are you some statistics - Facebook post 26/08/16 13:21. `https://www.facebook.com/Rockwoodpolice/posts/857067441060756`.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167. doi:10.3758/s13428-010-0039-7.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48(03). 609–652. doi:10.1017/S0022226712000011.

Sprouse, Jon & Diogo Almeida. 2017. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences* 40. doi:10.1017/S0140525X17000590.

Walker, Abby. 2008. *Phonetic detail and grammaticality judgments*. Christchurch, New Zealand: University of Canterbury dissertation. `https://ir.canterbury.ac.nz/handle/10092/2179`.

Webelhuth, Gert & Clare J. Dannenberg. 2006. Southern American English Personal Datives: the theoretical significance of dialectal variation. *American Speech* 81(1). 33–55. doi:10.1215/00031283-2006-002.

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273. doi:10.1353/lan.2011.0041.

Wolfram, Walt & Natalie Schilling-Estes. 2005. *American English: Dialects and Variation, 2nd Edition*. Malden, MA: Blackwell Publishing 2nd edn.

Wood, Jim. 2005. Dative presentatives. *Yale Grammatical Diversity Project* `https://ygdp.yale.edu/phenomena/dative-presentatives`.

Wood, Jim. Submitted. Quantifying geographical variation in acceptability judgments in regional american english dialect syntax.

Wood, Jim, Laurence Horn, Raffaella Zanuttini & Luke Lindemann. 2015. The Southern Dative Presentative meets Mechanical Turk. *American Speech* 90(3). doi:10.1215/00031283-3324487.

Wood, Jim & Matt Tyler. 2018. Microvariation in the Have Yet To Construction. *Linguistic Variation,* to appear. `https://ling.auf.net/lingbuzz/003950`.

Wood, Jim, Raffaella Zanuttini, Laurence Horn & Jason Zentz. Submitted. Dative country: Markedness and geographic variation in Southern dative constructions.

Zanuttini, Raffaella, Jim Wood, Jason Zentz & Laurence Horn. 2018. The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English. *Linguistics Vanguard* 4(1). 1–15. doi:10.1515/lingvan-2016-0070.