Dealing with Bad Data in Linguistic Theory Workshop Amsterdam, The Netherlands
Meertens Institute March 18, 2016

Parsed corpora of vernacular speech: challenges and prospects for the study of syntax

Christina Tortora
City University of New York
(College of Staten Island and The Graduate Center)
ctortora@gc.cuny.edu

[*Work with Anthony Kroch and Beatrice Santorini, U. of Pennsylvania*]

0. Overview of talk

- Brief review of the *Audio-Aligned and Parsed Corpus of Appalachian English* [AAPCAppE] (section 1)
- Discussion of the AAPCAppE as a case study in:
    - challenges in dealing with this type of data, both in the creation and in the use of the corpus (section 2)
    - opportunities for the study of linguistic structure not provided by other forms of data (section 3)

1. AAPCAppE, http://csivc.csi.cuny.edu/aapcappe/ (projected completion date: December 2016)

- The AAPCAppE is a ~1,000,000-word corpus of transcribed vernacular speech which is syntactically annotated.
- The ***underlying speech signal*** is from oral history recordings housed at various colleges and institutions in the Appalachian region:

I. Dante Oral History Project (DOHP). Collection of interviews on cassette tape with residents of Dante, VA (recorded 1997-98). Recordings are housed at, and curated by, the Archives of Appalachia at ETSU.

II. Joseph Hall Collection (JHall). Interviews with residents of the Great Smoky Mountains in Tennessee and North Carolina (1939); collector: Joseph Hall.

III. Appalachian Oral History Project (AOHP_I) at Alice Lloyd College, in Pippa Passes, KY. This history project was conducted from 1971-75 and its materials are housed in the library at Alice Lloyd College, Pippa Passes, Kentucky.

IV. Appalachian Oral History Project (AOHP_II) at Appalachian State University, in Boone, NC. This history project was conducted from the 1960s through the 1980s, and its materials are housed in the library at Appalachian State, in Boone, NC.

V. The Appalachian Archive (SKCTC) at Southeast Kentucky Community and Technical College, in Cumberland, KY. This history project was conducted from the 1960s through the 1980s, and its materials are housed in the library at Southeast Kentucky Community and Technical College, in Cumberland, KY.

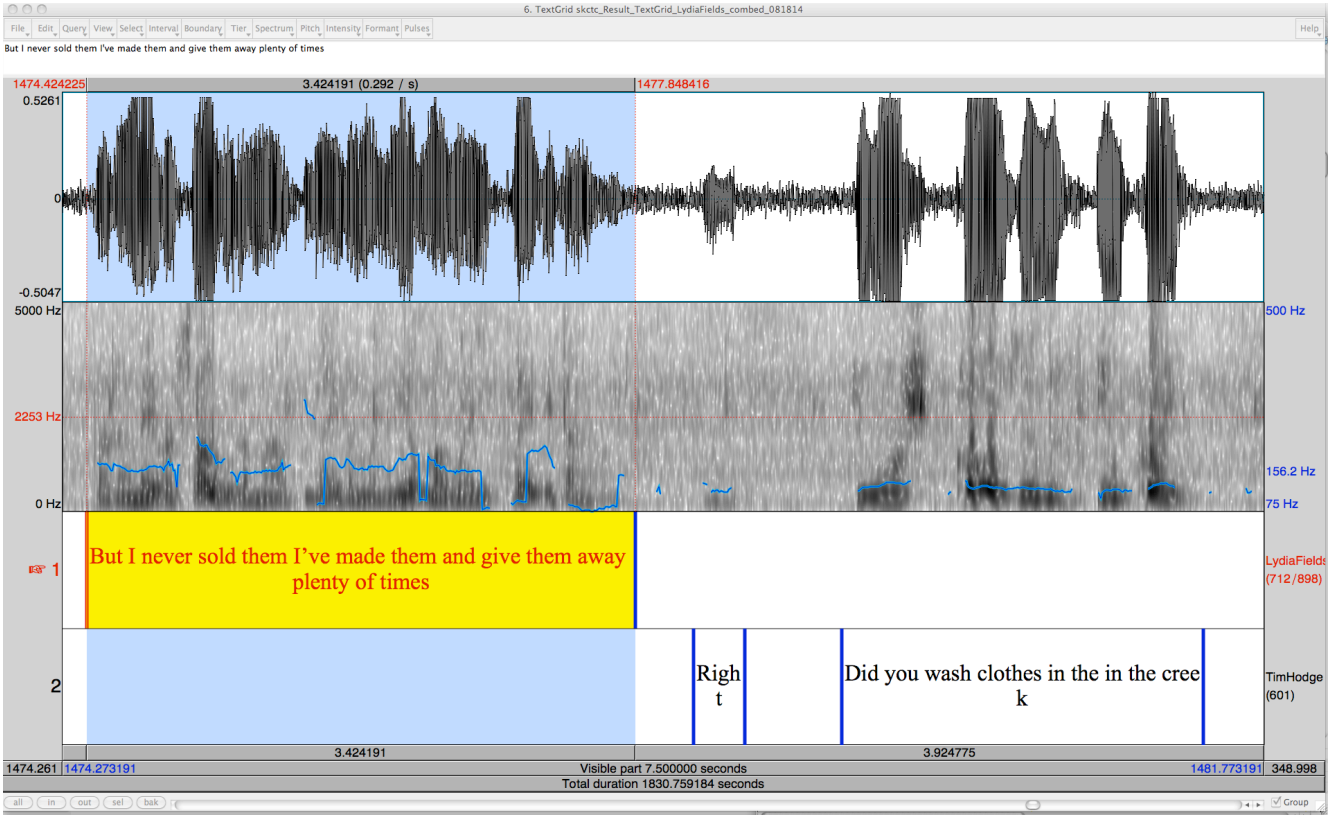
Principal parts: The AAPCAppE consists of the following components:

[i] sound files of the underlying speech signal

[ii] transcripts which are time-aligned with the speech signal (text-searchable in Praat/ELAN)

[iii] a Part-of-Speech tagged and parsed version of the transcribed text

[i-ii] sound files with time-aligned transcripts:



[iii] Part-of-Speech tagged and parsed version of the transcribed text (parsing method based that used for *Penn Parsed Corpora of Historical English*)

(1)

```
( (IP-MAT (CODE <LydiaFields_xmin=1474.42>)
        (CONJ-TEMP But)
        (NP (PRO I))
        (VP (ADVP (ADV never))
           (VBD sold)
           (NP (PRO them)))
        (PUNC .))
  (ID DS,.38181))

( (IP-MAT (NP (PRO I@))
        (VP (HVP @'ve)
           (VP (VP (VBN made)
                 (NP (PRO them)))
              (CONJP (CONJ and)
                 (VP (VBN give)
                    (NP (PRO them))
                    (ADVP (ADV away))
                    (NP-TMP (N plenty)
                       (PP (P of)
                          (NP (NS times))))))))
        (PUNC .)
        (CODE <$$LydiaFields_xmax=1477.85>))
  (ID DS,.38182))
```

2. Some of the challenges in dealing with this type of data, both in the creation and in the use of the corpus

- Well-known issues with corpora, generally speaking (no negative evidence; no access to speaker intent / intended interpretations)
- Issues in working with a speech signal
- Issues with *transcription as theory* and *annotation as theory*

2.1 Semantic interpretation / ambiguity (and speaker intent)

2.1.1 Case study 1. "Elided *have*" in *have*+$V_{participle}$ ("infinitival perfect") structures embedded under modals and infinitival *to*.

General Vernacular English:

(2) a. We coul*d* [əv] expanded the business.
b. She would [əv] liked to [əv] had a big family.

[as an aside: not clear [əv] is the auxiliary verb *have*; as per Kayne 1997, it may be the preposition *of*]

Appalachian English:

Appalachian speech exhibits "elided *have*" (noted by Montgomery & Hall 2004) in the presence of modals and infinitival *to* (exx. from AAPCAppE):

Modals:

(3) a. I **must __ been** four or five years old. (2_dohp)
b. [There] **may __ been** a good reason [for that]. (2_dohp)
c. I said, No, you the one **should __ said** something, cuz I wasn't going to say nothing… (2_dohp)
d. But he **shouldn't** never **__ told** me like he did. (2_dohp)
e. You **could __ heared** a pin drop. (1_dohp)
f. I don't know what in the world I **would __ done** if it wasn't for Ginny. (1_dohp)
g. You didn't see no little skimpy clothes back then; they **wouldn't __ knowed** what to done with the underclothes they've got today. (2_skctc)

Infinitival-*to*

(4) a. And, uh, that was supposed **to __ been** a rare seed. (2_skctc)
(= "It's supposed to be the case that that was a rare seed")
b. Some woman'əz supposed **to __ killed** her husband there. (2_dohp)
(*about a haunted house*; "It's supposed to be the case that some woman killed her husband there")
c. You didn't see no little skimpy clothes back then; they wouldn't knowed what **to __ done** with the underclothes they've got today. (2_skctc)
(*about how robust clothing was back in the day*)
d. But the men should [əv] refused **to __ went** in. (2_dohp)
(*about a mine explosion*)
e. (*How long did it take the train to go from here to Harlan?*) I don't know but they'əz stops, you see, so many places – it wouldn't [əv] took them long **to __ went** straight; but they stopped. (2_skctc)
f. Course if it'd [əv] lasted much longer, he would [əv] had **to __ went** [into the military]. (2_skctc)
(*about the war*)
g. We didn't none– didn't have no clothes, but we had plenty to eat; but we'd work **to __ made** it. (ALC)

The use of "elided *have*" varies with the pronounced-*have* variant. However: are they true variants? Or do the elided *have* structures (as in (3) and (4)) have a distinct interpretation than those with pronounced-*have*?

Need a native speaker to determine this.

Note however that many of the examples with infinitival-*to* would sound unacceptable (to my ears) with [əv]:

(4') c. ' *They wouldn't knowed what **to** [əv] **done** with the underclothes they've got today.
ok: They wouldn't knowed what **to do** with the underclothes they've got today.

d.' *But the men should [əv] refused **to** [əv] **went** in.
ok: But the men should [əv] refused **to go** in.

g.' *But we'd work **to** [əv] **made** it.
ok: But we'd work **to make** it.

Related to following problem:

(5) a. He would've had **to have left** at 3:00pm
b. He would've had **to leave** at 3:00pm

2.1.2 Case study 2. "Preterite *had*" in vernacular Englishes (Rickford & Rafal 1996): the form looks like a pluperfect (*had*+$V_{participle}$), but it's interpreted as a simple past / preterite:

(6) He **had** just **called** you. (= He just **called** you)

Sometimes the context helps you determine that a particular example should be analyzed as preterite *had*:

AAPCAppE, JHC
(7) …and (he) **jumped** the fence and **lost** his shoe-heel, and **run** on to the house, and he **wanted** the folks to go on back and help him; and **they'əd told him** that it was nothing but the boys with a bear skin, but he **said** no, he **said** that it **wasn't** that… (1_jhc)

But sometimes the context doesn't help.

2.1.3 Case study 3. The verb form that is spelled *'d* [d] is ambiguous between modal *would* and aux *had*

AAPCAppE, SKCTC:

(8) Speaker A: They'd never have their teeth pulled when [the] sign was in the head.
Speaker B: Why?
Speaker A: Because they**'d bled** real bad.

If *'d* in (8) is the modal *would* (Hypothesis 1), we get the structure in (8a); if it's *had* (Hypothesis 2), we get the structure in (8b) (where the __ indicates null-*have*):

(8a) they would __ bled real bad (Hypothesis 1, 'd = would)
(8b) they had bled real bad (Hypothesis 2, 'd = had)

In the case of (8), context helps us choose Hypothesis 1.

Same with the following:
(9) So, if I was thirty-three, he**'d __ been** thirty-five [years old]. (2_dohp)

But it's not always clear:

(10) But now if she**'d** [əv] been a little closter, maybe behind it, that blade come down and hit her on top of the head and killed her, maybe.

(10a) if she would [əv] been a little closter… (Hypothesis 1, 'd = would)
(10b) if she had [əv] been a little closter… (Hypothesis 2, 'd = had; but then [əv] has to be analyzed as "spurious-*əv*," as in (12))

(11) 'Course, if it**'d** [əv] lasted much longer, he would [əv] had to went.

(11a) if it would [əv] lasted… (Hypothesis 1, 'd = would)
(11b) if it had [əv] lasted… (Hypothesis 2, 'd = had; but then [əv] has to be analyzed as "spurious-*əv*," as in (12))

spurious-*əv*:
(12) And I guess if she hadn't **[əv]** done it, one of the others would [əv]. (2_skctc)

2.1.4 Case study 4. [ə] is many ways ambiguous

(13) a. hesitation *uh*
b. a-prefix
c. reduced form of [əv]

Are these a-prefixes, or hesitations?
(14) a. Well back when I were just a young man [ə] growing up, why we didn't have any advantage much of schools.
b. There's one [ə] hanging right out there in the back of that shed.

Is this reduced [əv], or is it a hesitation?
c. He shouldn't [ə] done it.

2.1.5 Case study 5. non-present verbs (simple past verbs and verbs in compound tenses like perfect and passive, e.g.) often take the bare form:

(15) *put, come, give, run, take, eat, begin, bring, say, tell*…

(16) So then they come here and say to me…

Is this the simple past, or the historical present (which gets disambiguated in the 3rd singular, as in (16))?

(17) a. So then he come**s** here and say**s** to me… (historical present)
b. So then he come here and say to me… (past tense)

2.2 Speech signal not always clear, and in crucial places
In (18a), the [t] is audible; in (18b), it's not clear, given the following word *together*, beginning in [t]:

(18) a. So I went back in again, right up uh, **closte** as ever. (1_jhc)
b. And then you cover the- the pen a- over with fence rails laid **close|closte t**ogether, all over. (1_jhc)

Example (18) maybe no big deal; the difference doesn't affect the structure. (Maybe!) However the following choice, which is equally unclear, makes a difference w.r.t. structural analysis:

(19) I'd **like|liked t**o had a big family; I'd **like|liked t**o had s- eight or nine children. (2_skctc)

Assuming that *'d* is the modal *would* (and that the __ below represents elided *have*):

(19a) I'd like to … (Hypothesis 1, *like* as a bare infinitive: "I would like to…")
(19b) I'd __ liked to … (Hypothesis 2, *liked* as a participle: "I would [əv] liked to…")

In (19a) we do not assume an elided *have* after the modal *'d*; in (19b) we would have to, given the non-present form of the verb. This is all complicated by the fact that often bare forms are used with "participial functions" ((15) above)!! So another possibility is the following:

(19c) I'd __ like to … (Hypothesis 3, *like* as a participle "I would [əv] like to…")

2.3 Transcription as theory and annotation as theory
A. Transcription as theory
Ochs (1979): our choices for conventions in transcription both reflect a prejudice for how to analyze the data, and also obscure the true nature of the data.

(i) The case of *have*:

(2a) We could ***have*** expanded the business.
(10) But now if she'd ***have*** been a little closter…
(12) And I guess if she hadn't ***have*** done it, one of the others would ***have***.

(ii) The case of contracted *was*:
(4b) Some woman **was** supposed to killed her husband there.

(iii) The case of *oughta, supposta, useta, wanna, gonna*

Conventional orthography can mislead us into being blind to potentially important syntactic patterns.

B. Annotation as theory
While syntactic annotation is precisely what we need to transform the corpus into a powerful tool for the study of syntactic patterning, we must never forget that the tool itself has theory built into it, and the user should not be lulled into thinking that the parses given are the "correct" parses (i.e., the ones intended by the speaker)

Ex: given the ambiguity of the form *give* (can be a present form, or simple past, or past participle), there are multiple possibilities for the parse:

(20) *I've made them and **give** them away plenty of times*.

a. I [have **made**$_{\textbf{participle}}$ them] and [SILENT-HAVE **give**$_{\textbf{participle}}$ them away]
OR:

b. [I [have **made**$_{\textbf{participle}}$ them]] and [NULL-SUBJECT [**give**$_{\textbf{present}}$ them away]]
OR:

c. [I [have **made**$_{\textbf{participle}}$ them]] and [NULL-SUBJECT [**give**$_{\textbf{past}}$ them away]]

3. Prospects for the study of syntax

Despite these problems: this form of data (= **parsed corpus of vernacular speech**) presents opportunities for the study of linguistic structure which are not provided by other forms of data. As such, it is good data.

i. Availability of time-aligned speech signal allows access to information about prosodic patterning in syntactic structures;
ii. Availability of time-aligned speech signal allows access to information about pronunciations that are not subject to the normative pressures of orthography (e.g., the [əv] data);
iii. A sufficiently large corpus can provide frequency information that still allows us to infer or test for grammatical properties that underlie the usage patterns, as the large size lets us discard problematic cases without doing much damage;
iv. Relatively infrequent constructions are observable in corpora that are sufficiently large;
v. Constructions denied by speakers are observable in corpora of vernacular speech (e.g., negative concord, amalgams; elided *have*);
vi. Constructions not present in writing are observable in corpora of vernacular speech (e.g., amalgams; elided *have*)
vii. Allows for large-scale quantitative studies that can reveal patterns of variation that would not otherwise be revealed through experiments like grammaticality judgment tasks (e.g., use of variant non-present forms like *saw, seen, seed, seened, see* in simple past vs. compound tense contexts)

3.1 Case study 1. Relatively infrequent construction: elided-*have*

Returning to the case of elided-*have*:

- the so-called "infinitival perfects" (modal/*to* + *have* + participle) are relatively rare structures in English to begin with;
- elided-*have* is not attested in Englishes outside of Appalachia;
- when asked, Appalachian speakers do no admit to allowing elided-*have* (so, judgments difficult);
- elided-*have* is not a form found in writing;
- this is why virtually nothing has been reported on it (except for a brief note in Montgomery & Hall 2004)

The AAPCAppE allows us to at least begin to study its nature and pursue questions previously not made possible, because

- having the speech signal time-aligned with the text allows us to check for pronunciation (thus, subsequent researchers do not need to rely on our claims about the data);
- tagging and parsing of a very large corpus allows us to study this relatively infrequent construction

Table 1: Infinitival perfects with modals/*to*, with pronounced-*have* vs. elided-*have*

# of words: 860,275	**modal + [*əv*]**	**modal + Ø$_{əv}$**	***modal total***	***to* + [*əv*]**	***to* + Ø$_{əv}$**	***to* total**
1_alc (104,943)	37	7		1	7	
2_aohp (62,992)	0	0		0	1	
1_dohp (281,148)	83	21		2	11	
2_dohp (172,445)	140	12		5	6	
1_jhc (52,718)	9	1		0	0	
2_jhc (6,917)	0	0		0	0	
2_skctc (179,112)	113	18		10	10	
TOTAL	**382 (87%)**	**59 (13%)**	**441**	**18 (34%)**	**35 (66%)**	**53**

Previously unnoted for Appalachian English:

- with infinitival perfects, there's a difference between modals and *to*
- despite their differences in absolute frequency (infinitival perfects with modals are more frequent than infinitival perfects with *to*), their behavior w.r.t elided-*have* seems to be mirror-image:
    - elided-*have* is less frequent with modals
    - elided-*have* is more frequent with infinitival *to*
- this may point to the following analysis:
    - elided-*have* with modals is a true variant of the pronounced-*have* construction, whereas
    - elided-*have* with infinitival *to* is not a semantically equivalent variant (as noted in 2.1.1 above)
- this possibly has implications for our understanding of the syntax and semantics of infinitival perfects in Englishes more generally (especially vis-à-vis Kayne 1997, where [əv] with modals is not an auxiliary, but rather a complementizer)
- calls for a closer study of the semantics of "infinitival perfects" in Englishes more generally (where the $have_{inf}$+participle string may be structurally ambiguous)

3.2 Case study 2. Variation in non-present verb forms

Some basic points:

- hypothesis: the apparent "simple past" (*walked*) in English is not finite; it is actually a participle (Solà 1996, Tortora 2014)
- Solà: apparent problem: distinction between past forms (simple and participial) with irregulars: *sang/sung, gave/given, ate/eaten*, etc.
- but how do we know that this is a valid distinction for all Englishes? Labov et al. (1968) and Wolfram & Fasold (1974) e.g. have noted a levelling of forms
- a rigorous study of the types of simple past / participial forms in vernacular speech, and their relative frequencies in the different syntactic contexts (simple past / compound tense) can be done with the AAPCAppE
- perhaps we will find that the variant forms (e.g. *gave*/*given* etc.) do not specialize for simple past vs. compound tense
- this would further dismantle the view that there is a distinction between simple past and past participle, and align with the view that the simple past is none other than the participle

Tortora et al. (2015): for many verb roots: speakers exhibit **more than two** forms in simple past and compound tense contexts in Appalachian English (see Appendix):

(21) a. *saw* / *seed* / *seen*
b. *saw* / *seened* / *seen*
c. *ran* / *run* / *runned*
d. *gave* / *give* / *given*
e. *did* / *done* / *doned*
f. *et* / *eat* / *ate* / *eaten*
g. *taken* / *take* / *takened* / *took*

DOHP, Speaker: TS

(22) a.	Mommy **taken** care of him till he got over it.	past
b.	They **took** him to the dead house and embalmed him.	past
c.	And we **was took** up to the top of the- what they call the slack...	compound tense
d.	And I said it'll **be taken** care of.	compound tense

DOHP, Speaker EW

(23) a. She **teached** over there at Clinchfield. past
b. She **taught** school up there a long time. past

DOHP, Speaker EC

(24) a. And he **went** down there and got a job.
b. And he come down there one Sunday and said if you go a-home with me I'll give you a job, so I **gone**.
c. She knowed that I was supposed **to went** to Kingsport on Monday and this was on Friday.
d. Said **we'd went** down to Uncle Eli's and got to romping down there with them boys and girls.
e. Well Lois **had done gone** back.

Joseph Hall Recordings (Great Smoky Mountains), **verb *see***

(25) **PAST**:

a. ...and I **seen** them coming down through the old field. seen
b. One of my old dogs, he **seen** me and... seen
c. That was the last one I ever **seen**. seen
d. I **seed** some sycamore trees... seed
e. …and that's the last you ever **seed** of any wolf in this country. seed
f. I went just on up to the top of the mountain, till I **seed** the dark'əz on me. seed
g. The last'un ever I **saw** in the woods I killed it with a pocket knife. saw
h. ...and uh he **saw** the bear passing through a little higher up. saw
i. Well, I never **saw** none but the bear. saw

(26) **COMPOUND TENSE:**

a. ...something I never **had seen** before, you know. seen
b. I**'ve seed** a many a bear, and eat the meat of them, coon too. seed
c. I**'ve saw** as high as ten to twenty drunk women the same day, and men in proportion. saw
d. …and it jumped on the biggest bear I**'ve ever saw** in the Smoky Mountain. saw

There is variation in the verb forms in the different grammatical environments; doesn't appear the forms have different syntactic distributions.

Table 2: Number of pasts vs. compound tenses: pasts are far more common than compound tenses

Joseph Hall Collection Appalachian Eng. (Tortora et al. in progress) **7,640 sentence tokens**		**Penn Parsed Corpus of Middle English** (Kroch & Taylor 2004) **81,230 sentence tokens**		**Penn Parsed Corpus of Modern British English** (Kroch, et al. 2010) **57,416 sentence tokens**	
simple past	compound tense	simple past	compound tense	simple past	compound tense
3,080 (**94%** of all non-presents)	195 (**6%** of all non-presents)	29,688 (**87%** of all non-presents)	4,530 (**13%** of all non-presents)	15,771 (**74%** of all non-presents)	5,670 (**26%** of all non-presents)

We looked at 5 speakers from the DOHP, interviewed in early 1990s, over age 65.

- All speakers had variant types.
- Variants occurred more in past contexts than in compound tense contexts, reflecting the fact that the corpus data contains more past contexts than compound tenses overall (as in Table 2).
- All speakers displayed variant forms that occurred in both past and compound tense contexts

To answer the question of whether the relative frequency of a given variant within a set is similar in past and compound tense contexts, we tallied up the number of tokens of each variant in a set (in each non-present environment). Here are the results:

Table 3: Distribution of morphological variants by context

variant type	***simple past***	***compound***	***total***
majority variant	1150 (94%)	65 (77%)	1215 (93%)
minority variants	76 (6%)	19 (23%)	95 (7%)
total	1226	84	1310

- Table 3 shows the distribution of morphological variants by syntactic context (simple past vs. compound tense).
- Note that for each verb root (e.g., *see* or *run*), there is a set of two or more variants (e.g., *seen*, *saw*, *seed*, *seened*, or *run*, *runned*), whereby one type in this "variant set" occurs more frequently; the term "majority variant" refers to this more frequent form, while "minority variants" refers to the variant or variants which are less frequent.

- The table shows that simple past contexts favor majority variants, relative to compound tense contexts (94% vs. 77%, $\chi 2(1)=29.12$, $p<.0001$).
- In this data, then, context (simple past vs. compound) does have some effect on variant selection, but a much weaker one than would be expected on standard accounts.
- In compound tense contexts, like in past tense contexts, majority forms are strongly favored relative to minority variants (77% vs. 23%), indicating much greater tendency toward a levelled tense paradigm.
- Therefore, syntactic environment (simple past vs. compound tense) doesn't seem to fully condition the frequency of co-variants; overall co-variants don't seem to be specialized for past vs. compound tense.

Summary
Based on these data, we can come to the following preliminary conclusion:

- wherever there is more than one form for the NON-PRESENT, (at least for some verbs) it seems that neither form is specialized for simple past vs. compound tense.
- this further dismantles the view that there is a distinction between simple past and past participle, and aligns with the view that the simple past is none other than the participle (Solà 1996; Tortora 2014)
- this kind of study is unique to a parsed corpus of vernacular speech

3.3 Case study 3. Promise for the study of prosody in amalgams

Amalgams, overview:

- Amalgams are strings which have previously been characterized as (i) anacoluthic, (ii) disfluent, (iii) the product of performance error, and (iv) ungrammatical
- Amalgams are not typically found in writing; they are the product of colloquial / vernacular speech, and are highly robust in spoken corpora
- It is misguided to assume that their study cannot illuminate our understanding of clausal architecture

O'Neill's (2015) extensive study of English amalgams shows that there are many different types (which may arise from different syntactic strategies / structures)

(27) a What she wanted was she wanted coffee
(what she wanted was ~~she wanted~~ coffee)

b. Her main interests are she likes square dancing and she enjoys birdwatching.
(her main interests are ~~she likes~~ square dancing and ~~she enjoys~~ birdwatching)

c. That's the only thing they do is fight
(that's the only thing; the only thing they do is fight)

d. They need a break is what they need.
(they need a break; a break is what they need)

e. That's what bothers me is he can't help us.
(that's what bothers me; what bothers me is he can't help us)

Prosody:

- To date, the naturally occurring prosody of the different kinds of amalgams has not been studied;
- This is an oversight, given that prosody can reveal aspects of syntactic structure that merely written samples cannot reveal;
- A study of the prosody of amalgams is not possible without a large audio-aligned and parsed corpus of vernacular speech, because:
  - amalgams are infrequent (or non-existent) in writing, and
  - only a parsed corpus of vernacular speech which is audio-aligned with the speech signal can allow for the study of the prosodic features of these forms

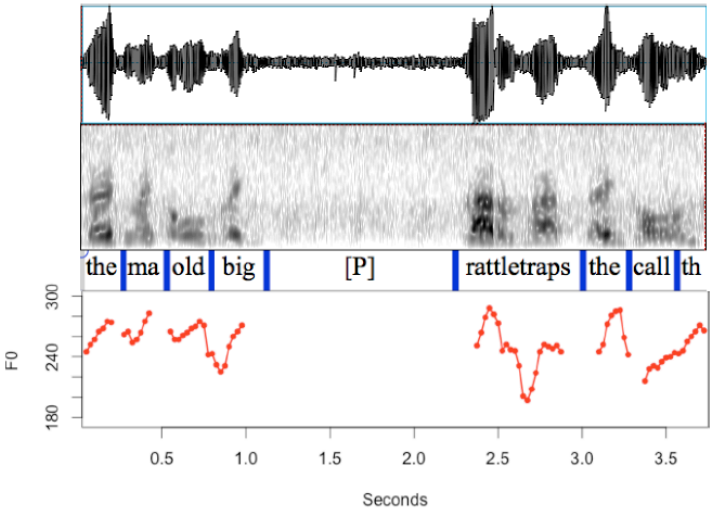
Richter (2015): a preliminary pursuit of amalgams in the AAPCAppE (~420,000 words from AOHP, DOHP, JHC, and SKCTC)

TYPE 1: verb+*call* type

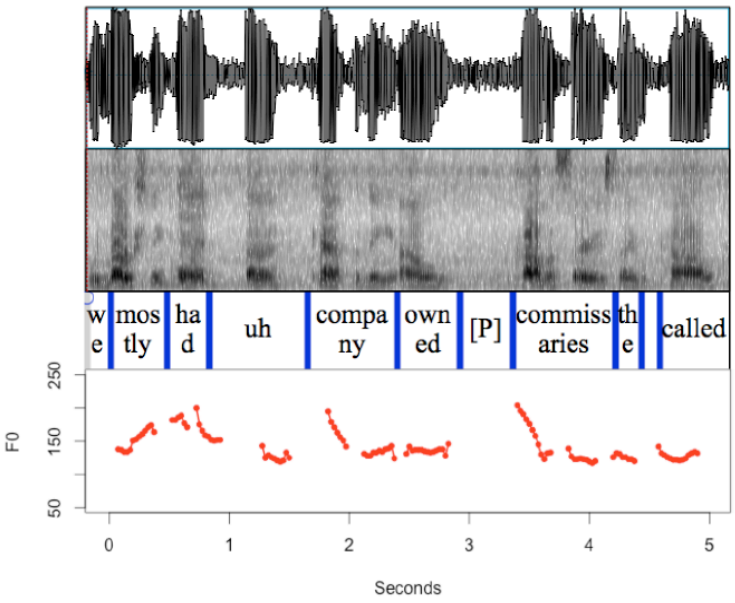
(28) a. They made old big rattletraps they called them.
b. That was the little store they called it.
c. We mostly had uh company owned commissaries they were called.
d. She told us to go to the uh smokehouse they called it.
e. They had a uh a speech I guess you'd call it by the director of the school.
f. The stock part ~~they called it~~ was made out of white oak.

According to Richter, a typical pattern of these sentences has "a pause right before the shared material," but the following three examples (which she states are typical of this structure) show a pause before the head N of the shared material:

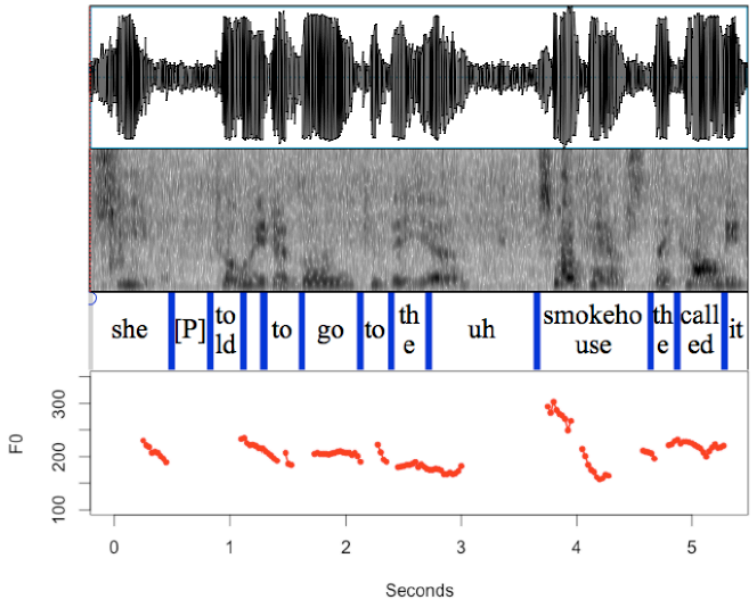
(28a') They made old big rattletraps they called them.



(28c') We mostly had uh company owned commissaries they were called.



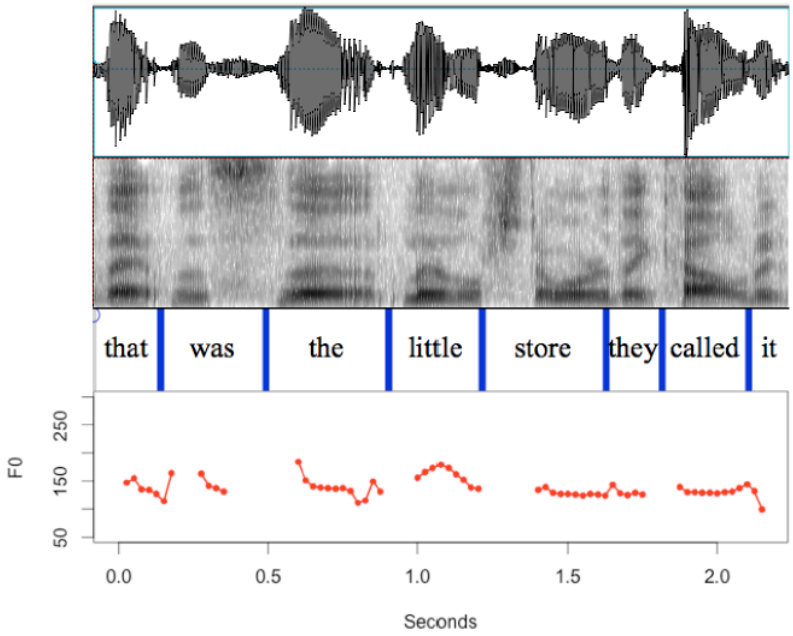
(28d') She told us to go to the uh smokehouse they called it.



Richter: if the NPs *old big rattletraps*, *company owned commissaries*, and *the smokehouse* were syntactic constituents only of the preceding sentence, or only of the following sentence, the you would expect to find a pause either before the entire NP or after the entire NP (but not in the middle).

Additionally, there are a number of examples where there is no pause:

(28b') That was the little store they called it.



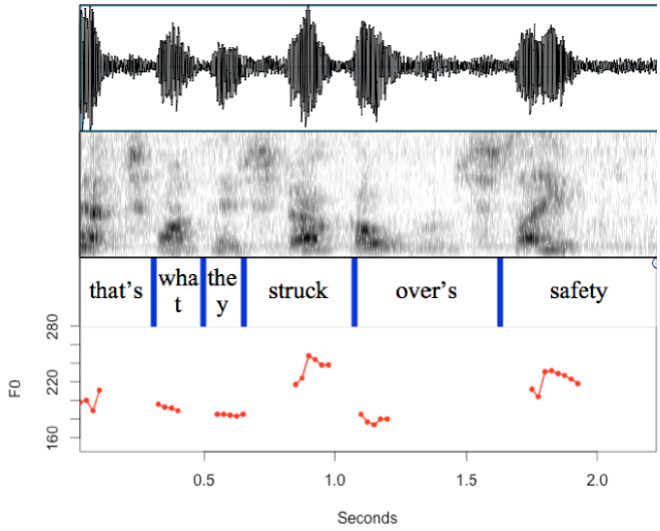
TYPE 2: *be+be* type (most common type observed in the 420,000 words searched: 54 examples)

(29) a. That's my hobby is cooking.
b. That's what I want to do is travel.
c. That's what we used to call them a long time ago was Joe wink.
d. That's what they struck over's safety.
e. That's all we had to depend on is what we raised.
f. It'əz Waukesha was the name of the town.
g. That's the most you could make was 50 cents?

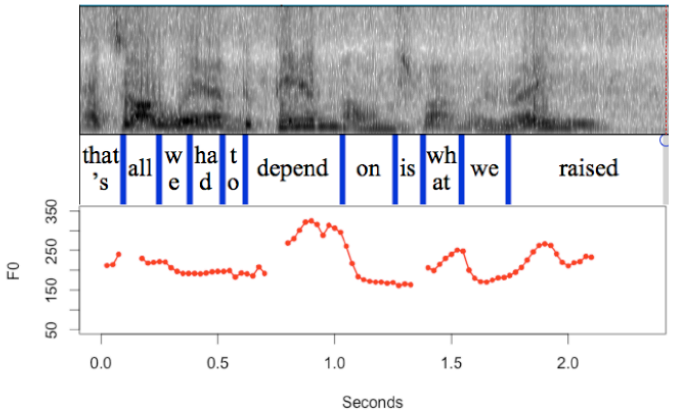
(and similarly, other types of presentational sentences)
h. We moved here to Coxton is where we came.
i. They made chairs some of them made.

Richter notes that it is rare to find any signs of disfluency in this type (no pauses, no hesitations, no fillers), making a disfluency account of these sentences unviable.

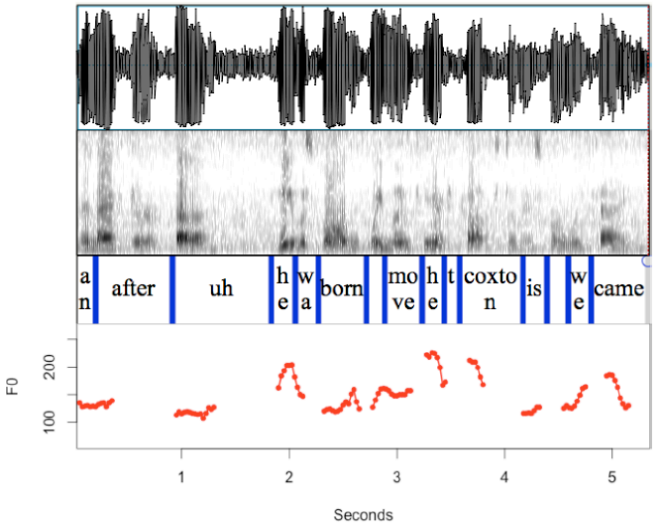
(29d') That's what they struck over's safety.



(29e') That's all we had to depend on is what we raised.



(29h') We moved here to Coxton is where we came.



Final comments

- The commonality of the *be+be* type of amalgam, and the lack of any disfluencies found in this type, support arguments that these structures are not the result of performance disfluencies
- The pauses found in the *call* type are in mid-NP-constituent, so not where we might expect them to be
- Clearly, much work needs to be done in terms of comparing these kinds of examples to base-line, non-amalgam sentences, both in terms of pauses, and in terms of prosodic contours
- This is only possible with an audio-aligned and parsed corpus of vernacular speech

APPENDIX: NON-PRESENT variants for five speakers from DOHP_II (from Tortora, Blanchette, O'Neill, & Arriaga 2015)

began, begin
bring, brought
brought, brung
burned, burnt
came, come
catch, caught
cause, caused
did, done
done, doned
drill, drilled
drop, dropped
get, got
give, given
gone, went
go, gone, went
hand, handed
heard, heared
held, helt
keep, kept
knew, knowed
laid, lay
learned, learnt
load, loaded
lose, lost
lost, losted
made, make
open, opened
paid, pay
push, pushed
ran, run
ran, run, runned
run, runned
rent, rented
sang, sung
saw, seen
saw, seen, seened
scald, scalded
start, started
send, sent
set, sit
start, started
swore, sworn
taken, takened, took
taken, took
take, took
taught, teached
tell, told
turn, turned
walk, walked
want, wanted
work, worked

Acknowledgments:
Thanks go to Jianjing Kuang and Caitlin Richter (for sharing their preliminary look at the prosody of amalgams). For the building of the AAPCAppE, the people we wish to thank are quite numerous; we list some of the names here, but fuller acknowledgment is given on the AAPCAppE website: Alexia Ault, Greta Browning, Kathleen Currie Hall, Ariel Diertani, Marcel den Dikken, Aaron Ecay, Janet Fodor, Robert Gipe, Fred Hay, Anton Karl Ingason, Tyler Kendall, Larry Lafollette, Michael Montgomery, Paul Reed, John Shean, Edward Snajdr, Laura Smith, Doug Whalen, Tiffany Williams, Walt Wolfram, Jiahong Yuan, Raffaella Zanuttini. The research done thus far has been supported by NSF Grants #BCS-0617197 (Tortora) and #BCS-0616573 (Den Dikken); by an NSF REG Award made to Blanchette on NSF Grant #BCS-0963950 (Snajdr); by a RISLUS Fellowship (Blanchette); by a Graduate Center Fellowship (Blanchette); by the College of Staten Island's *Provost Research Scholarship* (2010-11); by an NEH 2011-12 Fellowship (Tortora); and by an NEH Digital Humanities Start-Up Grant (2012-14) #HD-51543-12 (Tortora). Continued research is supported by National Science Foundation BCS Awards #BCS-1152148 (PI: Tortora) and #BCS-1151630 (PI: Santorini), project period: 2012-16.

Selected Bibliography:

Blanchette, F. 2015. *English Negative Concord, Negative Polarity, and Double Negation*. PhD thesis, CUNY. http://ling.auf.net/lingbuzz/002654

Boersma, Paul, and David Weenink. 2011. *Praat: doing phonetics by computer*, version 5.2.32.

Harris, J. 1984. "Syntactic Variation and Dialect Divergence," *Journal of Linguistics* 20: 307-327.

Hindle, D. 1983. "Deterministic parsing of syntactic non-fluencies," in *Proceedings of the 21st Annual Meeting of Association for Computational Linguistics*, pp. 123-128.

Kayne, R. 1997. "The English Complementizer of," *Journal of Comparative Germanic Linguistics* 1: 43-54. also in Kayne 2002, Parameters & Universals, OUP..

Kroch, A. 1989. "Reflexes of Grammar in Patterns of Language Change," *LVC* 1.3:199-244.

Kroch, A. 1994. "Morphosyntactic Variation," in K. Beals et al. (eds.) *Proceedings of the 30th Annual Meeting of the Chicago Linguistics Society* (Parasession on Variation and Linguistic Theory.), vol 2, pp.180-201.

Kroch, A. & Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, 2nd edition, (http://www.ling.upenn.edu/hist-corpora/).

Kroch, A., Beatrice Santorini, and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, (http://www.ling.upenn.edu/hist-corpora/).

Kroch, A., Beatrice Santorini, and Ariel Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English* (PPCMBE). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, (http://www.ling.upenn.edu/hist-corpora/).

Labov, W. P. Cohen, C. Robins, & J. Lewis. 1968. *A study of the nonstandard English of Negro and Puerto Rican speakers in New York City*. Final Report, Cooperative Project No. 3288, United States Office of Education.

Marcus, M., B. Santorini, & M. Marcinkiewicz. 1993. "Building a large annotated corpus of English: The Penn Treebank," Computational Linguistics 19.2:313-330.

Montgomery, M. & J.S. Hall. 2004. *A Dictionary of Smoky Mountain English*. Knoxville: UT Press.

Ochs, E. 1979. "Transcription as theory," in E. Ochs & B. Schieffelin (eds.) *Developmental Pragmatics*, pp. 43-72. New York: Academic Press.

O'Neill, T. 2015. *The domain of Finiteness: Anchoring without Tense in copular amalgam sentences*. PhD dissertation, The CUNY Graduate Center.

Randall, Beth. 2009. CorpusSearch 2: A tool for linguistic research. Includes CorpusDraw, a graphical interface for displaying and correcting parsed corpora. http://corpussearch.sourceforge.net/

Richter, C. 2015. "That's the key to syntactic amalgams is prosody," Ms. UPenn.

Rickford, J. & C. Rafal. 1996. "Preterite *had*+V-ed in the narratives of African American preadolescents," *American Speech* 71.3: 227-254.

Santorini, B. 1990. "Part-of-speech tagging guidelines for the Penn Treebank Project," Department of Computer and Information Science, University of Pennsylvania, Technical Report MS-CIS-90-47.

Solà, J. 1996. "Morphology and Word Order in Germanic Languages," in W. Abraham (ed.) Minimal Ideas: Syntactic Studies in the Minimalist Framework, pp. 217-251. John Benjamins.

Taylor, A. 1994. "Variation in Past Tense Formation in the History of English," in R. Izvorski, M. Meyerhoff, B. Reynolds, & V. Tredinnick (eds.), UPenn Working Papers in Linguistics 1, pp. 143-159.

Tortora, C. 2014. "Evidence for the non-finiteness of English 'present' and 'past' verb forms," talk given at the NYU Syntax Brown Bag series, February 28, 2014.

Tortora, C., F. Blanchette, T. O'Neill, & S. Arriaga. 2015. "Variation in Appalachian non-present forms," talk given at the *Second Formal Ways of Analyzing Variation* Workshop, May 2015, Reykjavik, Iceland.

Tortora, C., B. Santorini, & F. Blanchette. in progress. *The Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE). http://csivc.csi.cuny.edu/aapcappe/

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. *ELAN: a Professional Framework for Multimodality Research*. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Wolfram, W. & R. Fasold. 1974. *The study of social dialects in American English*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.