Workshop on Databases and Corpora in Linguistics Stony Brook, NY
Stony Brook University, Department of Linguistics October 17, 2014

The Audio-Aligned and Parsed Corpus of Appalachian English: Design and Use

Christina Tortora, City University of New York (CSI and The Graduate Center) (ctortora@gc.cuny.edu)
Beatrice Santorini, University of Pennsylvania (beatrice@sas.upenn.edu)
Frances Blanchette, City University of New York (The Graduate Center) (fblanchette@gc.cuny.edu)

0.1 What is the Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE)? http://csivc.csi.cuny.edu/aapcappe/

- Ultimate product: an online, freely accessible ~1,000,000-word corpus of Appalachian English, which will be:
  - syntactically annotated (searchable by any standard tree query language e.g., *CorpusSearch*, Randall 2009)
  - accompanied by a full set of digitized recordings of the underlying speech signal, in the form of *.wav* files (text-searchable using Praat (Boersma and Weenink 2011) and ELAN files (Wittenburg et al. 2006))
- Will be a database that will further research in the various sub-disciplines of Linguistics, and afford novel approaches to the analysis of English dialect data.

0.2 Overview of discussion

DESIGN of the AAPCAppE (section 1)

- I will cover some basic background on **design** (sections 1.1 and 1.2)
- I will address aspects of the design which relate to the following workshop question (section 1.3):
  - *How much input analysis should be included*?

USE of the AAPCAppE (section 2)

- I will go over one basic example of the AAPCAppE's potential as a tool for addressing theoretical questions; this will relate to the following workshop questions:
  - *Who do we think might use the corpus?*
  - *What makes a database or linguistic corpus usable*?
  - *How can we make them better and more interactive*?
  - *What works, and what doesn't?* (related to the question of some corpora requiring "*...considerable training in order to access their wealth of information...*")

1. Some basic facts about the AAPCAppE (Design)

1.1 What is the AAPCAppE?

- *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE): a database that will further research in the various sub-disciplines of Linguistics, and afford novel approaches to the analysis of English dialect data.

- Ultimate product: an online, freely accessible, ~1,000,000-word corpus of Appalachian English, which will be:
    - syntactically annotated, or "parsed" (searchable by any standard tree query language e.g., *CorpusSearch*, Randall 2009)
    - accompanied by a full set of digitized recordings of the underlying speech signal, in the form of *.wav* files (text-searchable using Praat (Boersma and Weenink 2011) and ELAN (Wittenburg et al. 2006))

The AAPCAppE is based on the speech from oral history project 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 East Tennessee State University (ETSU; http://www.etsu.edu/cass/archives/). Approximately 150,000 words generated using Kathy Shearer's transcriptions as a base; approximately 250,000 words generated using Montgomery's ATASC (Archive of Traditional Appalachian Speech and Culture) as a base.

II. Joseph Hall Collection (JHall). Interviews with residents of the Great Smoky Mountains in Tennessee and North Carolina (1939); collector: Joseph Hall. Approximately 60,000 words generated using Montgomery's ATASC as a base; further transcripts to be added.

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. The speech is from Central Eastern Kentucky. Approximately 115,000 words generated using Montgomery's ATASC as a base.

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. The speech is from Western North Carolina. Approximately 200,000 words.

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. The speech is from Eastern Kentucky. Approximately 200,000 words.

1.2 AAPCAppE principal parts and basic procedures

The AAPCAppE will consist of the following components, all of which will be made available for research:

[i] *.wav* files of the underlying speech signal
[ii] TextGrids — i.e., transcripts which are "time-aligned" with the speech signal (usable in Praat/ELAN)
[iii] A Part-of-Speech tagged version of the transcribed text
[iv] A Parsed (or, syntactically annotated) version of the text
[v] A complete, basic transcript, for those not interested in the fancy stuff in [i]–[iv]

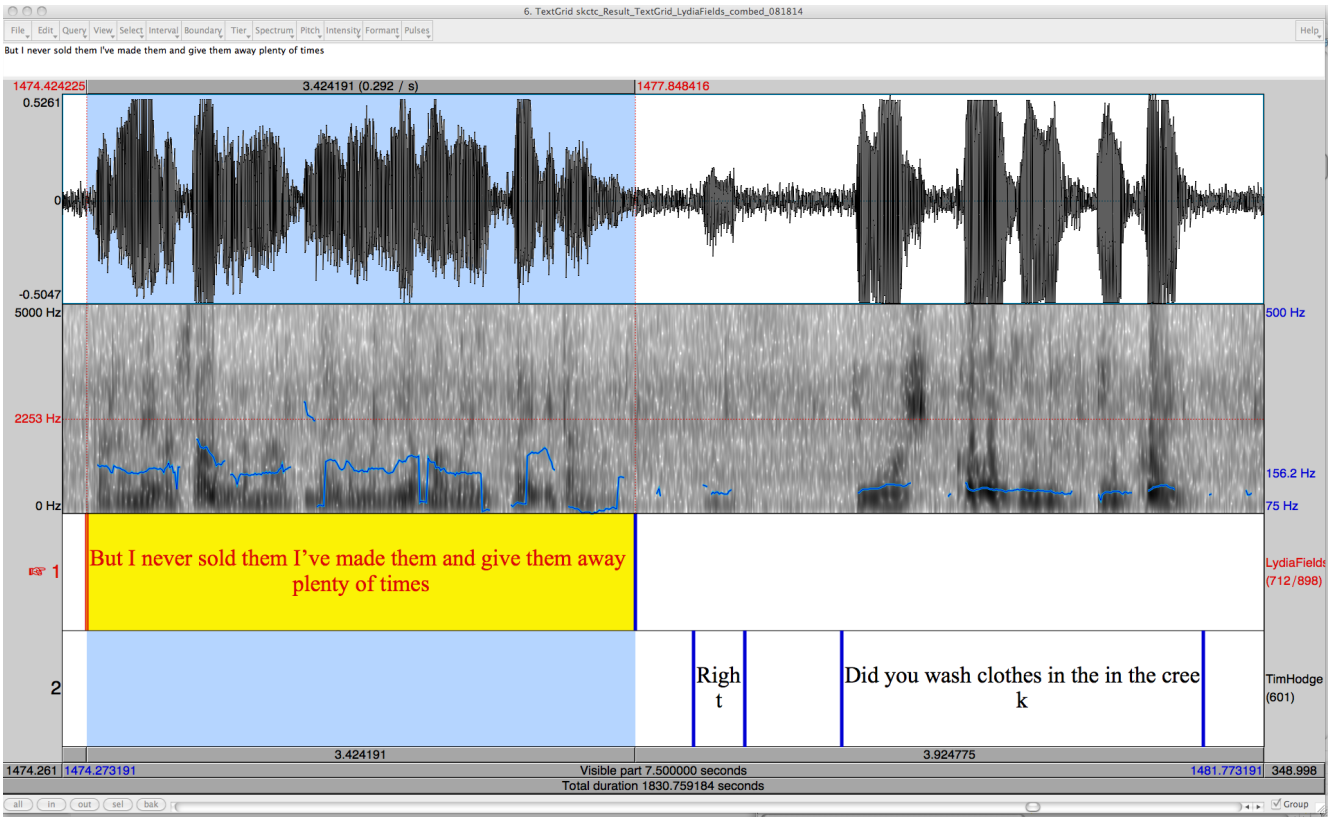
1.2.1 The Audio Part: [i] & [ii]

Regarding [i] & [ii]: We use Praat to create time-aligned transcripts for the *.wav* files — the TextGrids.

We have two ways of creating the Praat TextGrids:

- Type in the transcript by hand (when we don't have a pre-existing transcript), or
- Use the *PPL Forced Aligner*, developed by Jiahong Yuan at the U. of Pennsylvania (this is only an option if you already have a transcript of the recording) https://webann.ldc.upenn.edu/

Example (1)



1.2.2 The Parsed Part: [iii] & [iv]

Regarding [iii] & [iv]: To make the syntactically annotated version of the corpus, we take the TextGrid version of the transcript (seen in Example (1)) and transform it into to a file that can be annotated with Part of Speech tags, and syntactic tags which give basic syntactic structure. The TextGrid in Example (1) looks like this in a text editor:

Example (2):

intervals [710]:
xmin = 1469.5896698724632
xmax = 1473.4784586359294
text = "No I never did I always with- kept them I'd give away a lot of them"
intervals [711]:
xmin = 1473.4784586359294
xmax = 1474.4242247353188
Text = ""
intervals [712]:
xmin = 1474.4242247353188
xmax = 1477.8484160026892
text = "But I never sold them I've made them and give them away plenty of times"
intervals [713]:
xmin = 1477.8484160026892
xmax = 1482.5328252601214

How do we go from what we have in (2) to something a parser can process?

Step 1: We run a script which gives us an output that looks like this (script designed by Tyler Kendall, U. of Oregon; http://pages.uoregon.edu/tsk/):

Example (3) Our so-called "interleaved" files

1207	TimHodge:	[1466.11]	Did you ever sell any kind of quilts or socks that you made [1468.99]
1209	LydiaFields:	[1469.59]	No I never did I always with- kept them I'd give away a lot of them [1473.48]
1211	TimHodge:	[1473.74]	Mm [1474.09]
1213	**LydiaFields:**	**[1474.42]**	**But I never sold them I've made them and give them away plenty of times [1477.85]**
1215	**TimHodge:**	**[1478.21]**	**Right [1478.53]**
1216	**TimHodge:**	**[1478.53]**	**(pause 0.60) [1479.14]**
1217	**TimHodge:**	**[1479.14]**	**Did you wash clothes in the in the creek [1481.39]**
1219	LydiaFields:	[1482.53]	Yeah we went to the creek and built us up a big fire and [1485.54]
1220	LydiaFields:	[1485.54]	(pause 0.92) [1486.46]

Step 2: We run another script on (3) which gives an output that looks like what you see in Example (4) (script designed by Tyler Kendall; based on the Penn Treebank *make-tok* script):

Example (4)

```
<LydiaFields_xmin=1474.42>
But
I
never
sold
them
I@
@'ve
made
them
and
give
them
away
plenty
of
times
<$$LydiaFields_xmax=1477.85>
<TimHodge_xmin=1478.21>
Right
<$$TimHodge_xmax=1478.53>
<TimHodge_xmin=1479.14>
Did
you
wash
clothes
in
the
in
the
creek
<$$TimHodge_xmax=1481.39>
```

Two problems with (4):

- Problem 1: To do a decent job, the parser needs some initial help from a human, who can recognize what a *sentence token* is. As it stands, the string in (4) has zero structure; to be effectively automatically parsed, the parser should at least know what counts as a domain of parsing.
- Problem 2: This is speech — not the kind of grammatical text that the parsers are accustomed to, with written text. Thus, a human is needed to identify disfluencies associated with speech (see Hindle 1983):
  - Repetitions
  - False Starts
  - Elaborations
  - Parenthetical statements
  - Breaks in the sentence

Step 3: To address these problems, we have a human (i) provide structure by identifying sentence tokens, and (ii) identify disfluencies using a manual coding system we devised. Compare (4) with (5):

Example (4)

<LydiaFields_xmin=1474.42>
But
I
never
sold
them
I@
@'ve
made
them
and
give
them
away
plenty
of
times
<$$LydiaFields_xmax=1477.85>
<TimHodge_xmin=1478.21>
Right
<$$TimHodge_xmax=1478.53>
<TimHodge_xmin=1479.14>
Did
you
wash
clothes
in
the
in
the
creek
<$$TimHodge_xmax=1481.39>

Example (5)

<LydiaFields_xmin=1474.42>
But
I
never
sold
them
.

I@
@'ve
made
them
and
give
them
away
plenty
of
times
.
<$$LydiaFields_xmax=1477.85>

<TimHodge_xmin=1478.21>
Right
.
<$$TimHodge_xmax=1478.53>

<TimHodge_xmin=1479.14>
Did
you
wash
clothes
in
the
<REP>
in
the
<$$REP>
creek
?
<$$TimHodge_xmax=1481.39>

Step 4: Regarding the disfluencies: we then run a script which temporarily removes them, so that the parser is dealing just with grammatical sentence tokens to work off of. (Script designed by Anton Ingason.) The disfluencies are re-inserted after parsing and parsing correction is done.

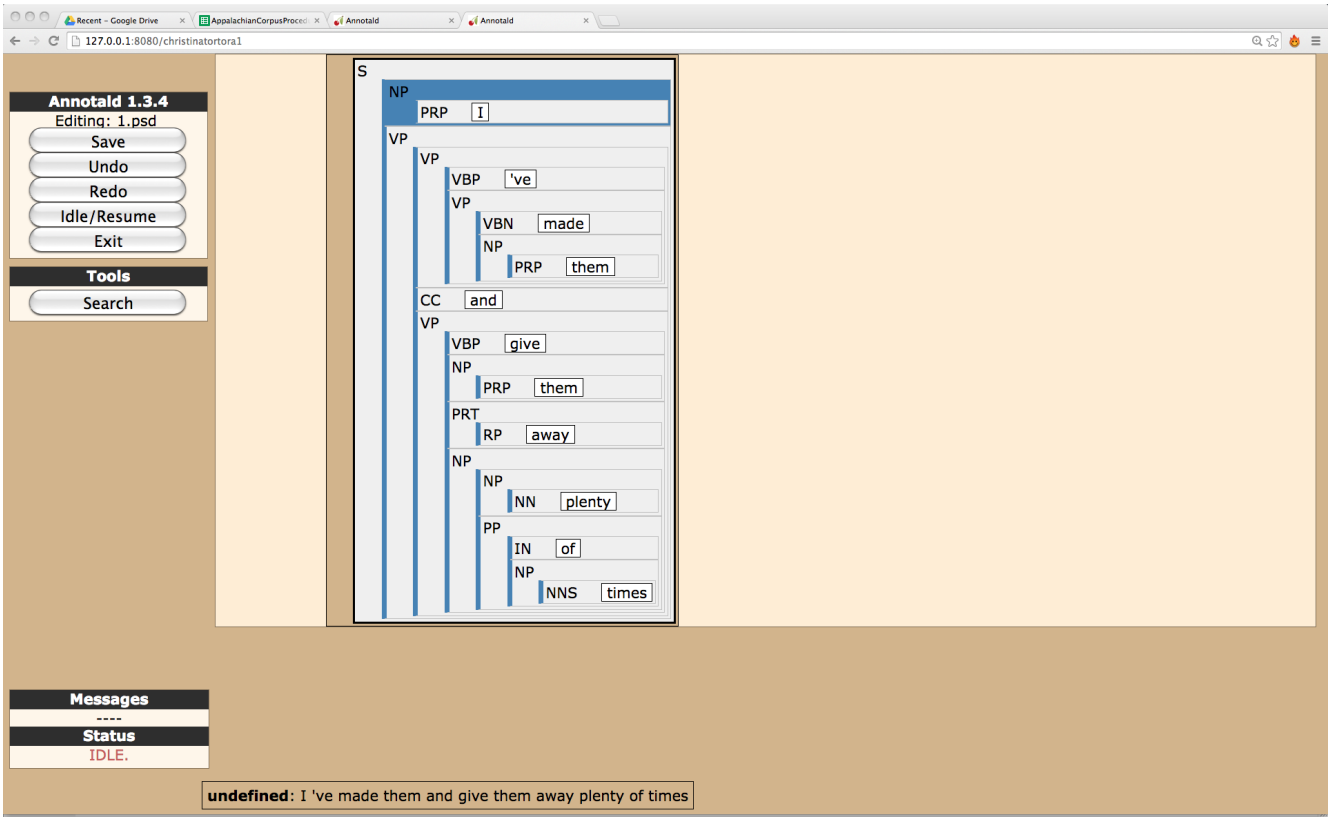
Step 5: The files (represented by Example (5) but without the disfluencies) are parsed, and then the parsing is corrected and perfected by Beatrice.

We are currently preparing to work with *Annotald* on the correction procedures (http://annotald.github.io/).

Annotald is a program for annotating parsed corpora in the Penn Treebank format. Annotald was originally written by Anton Ingason as part of the Icelandic Parsed Historical Corpus project. It is currently being developed by him along with Jana Beck and **Aaron Ecay**. (From the *Annotald* website.)

This is what the Annotald user interface looks like:

Example (6)



1.3 How much input analysis should be included? (A workshop question)

The verb form *give* in examples (1) through (6) — and the consequent syntactic analysis — together raise all kinds of issues regarding the question of *input analysis*.

1.3.1 General issue of input analysis

- On the one hand, an annotated corpus should be as atheoretical as possible, so as to allow the user to do searches for items and structures of interest without being held to an organization of data that is influenced by the corpus creators' own pre-conceived hypotheses regarding certain aspects of syntax;
- On the other hand, it's impossible to create a tagged / parsed corpus without injecting some theory into its structure.

EXAMPLE:

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

VBN = past participle
VBP = present tense
VBD = simple past
VB = bare form

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

OR:

b. [I [have **made**$_{\textbf{VBN}}$ them]] and [NULL-SUBJECT [**give**$_{\textbf{VBP}}$ them away]]

OR:

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

This question arises in part because in Appalachian speech (as with many Englishes), there great variability in the use of **bare** and **non-present** verb forms. (Note that this issue would arise even with the uncontroversial form *put*; it's just that in Appalachian speech, the range of possibilities for each verb is greater.)

****Need SOME input analysis.** A choice has to be made based on (a) some level of theory, and (b) a judgment call based on e.g. context.

(SEE APPENDIX for review of variant non-present forms for five speakers from a portion of the AAPCAppE.)

OTHER EXAMPLES:

Some more obvious issues regarding choice of Part of Speech tags

(8) I **up** and told him.
What tag to give ***up***? Like all other POS tags, must have some theory of what this is.

(9) I **got** a car.
What tag to give ***got***? Like many other cases, must provide an unambiguous analysis (present poss. or past tense of *get*?)

(10) Y'all **have** a very hard time policing?
What tag to give *have*? "bare infinitive" if missing *do*, "simple present" if this is formally a declarative (must give an unambiguous analysis)

(11) Didn't he **used** to do that?
What tag to give ***used***? Not a past tense, functionally, but tagged as a past tense verb in *He **used** to do that*.

Side note:
This problem is nevertheless very interesting, as being forced to provide each word with a tag raises important theoretical questions which in turn underscore the problems which arise in part from using Standard English orthography. For example, what is *used to* (*usta*)?

Not a real modal like *will* / *would* /etc., which invert in interrogatives:
(12a) *Usta he do that?
(cf. *Will he do that*?)

Not a modal like *need*, which does not invert, but which is inflected:
(12b) I used to / usta...
He used to / usta / *ustas...
(cf. *I need* vs. *He needs* / *Does he need?*)

Perhaps like invariable/habitual *be* in AAE (Green 2002):
(12c) Dee **be** waiting for the bus.
(12d) Dee **don't be** waiting for the bus.

Used to / usta:
(12e) Dee **used to** (/ **usta**) wait for the bus.
(12f) Dee **didn't used to** (/ **usta**) wait for the bus.

1.3.2 The case of verb forms

As noted above: many English speakers exhibit robust variability in verb forms, both in simple past contexts, and in compound tense contexts (for discussion, see e.g. Wolfram & Fasold 1974; Taylor 1994; Montgomery & Hall 2004; Tortora et al. 2013)

(13) a. I been there (for two years).
b. I was there (for two years).

(14) a. I seen him (this morning).
b. I saw him (this morning).
c. I've seen him five times today already.
d. I've saw him five times today already.

VBN = past participle
VBP = present tense
VBD = simple past
VB = bare form

What tag to give ***seen***? Simple past (VBD) or past participle (VBN)?
What tag to give ***saw***? Simple past (VBD) or past participle (VBN)?

It might be tempting to tag forms like *been* and *seen* in (13a) and (14a) as VBN (participle) because they look like participles to you (based on your own English). In fact, in discussing this work, we've been asked many times the question "How do you know there isn't just a missing *have* in these cases such as (13a) and (14a)?

Problems with assigning VBN tag based on the "look" of the form alone:

- sometimes the context makes a *have+seen* analysis unlikely
- often, things that might look like a VBD to you (e.g., *saw*) are used with aux *have* (as in (14d), or as in (15)

(15) He'd **went** down the mountain. (JHall, tape5side2)

Thus, given (14d) and (15): if you're going to tag *seen* as a VBN no matter what — as if it were preceded by a null *have* — then what would be your reasoning for not doing the same in the following case?

(16) They went down the mountain.

In other words, why not posit a null *have* in (16), and tag *went* here as a VBN?

(NB: In Tortora 2014, I actually do posit that *went* in (16), like all apparent simple past forms, are in fact past participles. But note that this is already too much personal theory for me to put into this corpus. **<u>This is too much input analysis</u>.)

The decisions based on such considerations are too arbitrary. The more traditional analysis involves a less arbitrary, more straightforward approach:

- compound tense (i.e., preceded by a form of *have*), verb form is VBN
- simple tense (i.e., not preceded by a form of *have*), verb form is VBD

This makes searches straightforward, and possible to do even on just the POS tagged files. (**Illustration**)

2. Use of the AAPCAppE

2.1 Ultimate plan

- to have an online user interface that allows for searches through the annotated part of the corpus, along the lines of what is available for the Penn Parsed Corpora. E.g.: http://csearch2.ling.upenn.edu/APPALACHIAN/querypos.shtml
- (though we're presently in conversation with Aaron Ecay to devise a user interface exploiting Annotald.)
- Any hit will have a link that can play the associated portion of the .wav file, so that users can hear the sentence token in question, should they wish to check up on our decisions (this provides greater transparency)

2.2 Some further workshop questions

- *Who do we think might want to use the corpus?*

  People like us, with our kinds of research questions. For example, consider the hypothesis from Tortora et al. 2013:

  [[General non-present] hypothesis]: Wherever a speaker exhibits more than one form for the **non-present** (e.g., *saw*, *seen*, *seed*, *see*), no one form specializes for one syntactic context (simple past) vs. the other (compound tense).

  This can be tested in the AAPCAppE by counting things. Things to count:
  (a) How frequent is the simple past and how frequent are compound tenses?
  (b) How frequently is each form (e.g. *saw* vs. *seen*) used in each context?

  Problems for I-grammar study: each speaker produces not enough words in an interview. Need fancy statistics to help us draw conclusions for individuals, based on behavior of group. Working with Aaron Ecay on this.

- *What makes a database or linguistic corpus usable? How can we make them better and more interactive? What works, and what doesn't? (related to the issue of some corpora requiring "...considerable training in order to access their wealth of information...")*

In part, what will make the corpus usable is:

(a) its accessibility, and

(b) the user-friendliness of the GUI interface. Consider e.g. the user interface for the Penn Parsed Corpora of Historical English (**Demo**). Must have knowledge of meaning of tags.

Regarding (a): what does "freely accessible" mean? Who pays for upkeep?
Regarding (b): requires constant refinements in response to user feedback

APPENDIX: NON-PRESENT variants for five speakers from DOHP_II

(from Tortora, Blanchette, & O'Neill 2013)

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:
*The people we wish to thank are quite numerous; we list the names here, but fuller acknowledgment is given on the AAPCAppE website: Alexia Ault, Claire Bowern, Dianne Bradley, Greta Browning, Gabriel Cynowicz, Kathleen Currie Hall, Marcel den Dikken, Aaron Ecay, Janet Fodor, Robert Gipe, Fred Hay, Anton Karl Ingason, Dan Kaufman, Tyler Kendall, Mike Kress, Tony Kroch, Larry Lafollette, Tom Lauria, Betsy Layman, Mark Lewental, Michael Montgomery, Paul Muzio, Ricardo Otheguy, 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: 2012-15. And finally, thanks to Lori Repetti and Francisco Ordóñez, for the opportunity of presenting at this Workshop.

Selected Bibliography:

Blanchette, F. 2013. Negative Concord in English. *Linguistic Variation* 13.1: 1-47.

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

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

Hackenberg, Robert G. 1972. "Appalachian English: A Socio-linguistic Study." Ph.D. thesis, Georgetown.

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.

Icelandic Parsed Historical Corpus, creation in progress, by Joel Wallenberg, Anton Karl Ingason, and Einar Freyr Sigurðsson (http://www.ling.upenn.edu/~joelcw/papers/iclt2010treebank.pdf)

Kendall, Tyler (2007). "Enhancing Sociolinguistic Data Collections: The North Carolina Sociolinguistic Archive and Analysis Project," *Penn Working Papers in Linguistics* 13.2: 15-26.

Kim, J., S. Pinker, A. Prince, & S. Prasada. 1991. "Why no mere mortal has ever flown out to center field," *Cognitive Science* 15: 173-218.

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, second 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/).

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.

Munn, A. 2013. "Some observations on participle levelling," talk given in The CUNY Graduate Center Syntax Supper series, December 3, 2013.

Newman, A., M. Ullman, R. Pancheva, D. Waligura, H. Neville. 2007. "An ERP study of regular and irregular English past tense inflection," *NeuroImage* 34: 435-445.

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/

Reaser, Jeffrey. 2007. "Evaluating and Improving High School Students' Folk Perceptions of Dialects," in L. Benitz & T. Cook (eds.) Penn Working Papers in Linguistics 13.2: 179-192.

Reaser, Jeffrey. 2010. "Developing Sociolinguistic Curricula that Help Teachers Meet Standards," in K. Denham & A. Lobeck (eds.) Linguistics at School: Language Awareness in Primary and Secondary Education, pp. 91-105. Cambridge: Cambridge University Press.

Reaser, J., & C. Adger. 2008. "Dialect Diversity in the Classroom: Research and Development," in B. Spolsky & F.M. Hult (eds.) Handbook of Educational Linguistics, pp. 161-173. Malden, MA: Blackwell.

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.

SLAAP The Sociolinguistic Archive and Analysis Project (http://ncslaap.lib.ncsu.edu/), NC State.

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

Tortora, C., F. Blanchette, & T. O'Neill. 2013. "Variation in Appalachian verb forms: evidence for a general past," talk given at the Fifth International Conference on the Linguistics of Contemporary English (ICLCE5), University of Texas, Austin.

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., 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. 1984. "Is there an Appalachian English?" *Appalachian Journal* 11:215-226.

Wolfram, W. & D. Christian. 1976. *Appalachian Speech*.

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