

## Collaborative Research: A syntactically annotated corpus of Appalachian English

### 0. Introduction

This research project aims to create an innovative database that will further research in phonetics, phonology, morphology, syntax, and sociolinguistics. Specifically, we seek funds to create an online, freely accessible, 1,000,000-word syntactically annotated (or “parsed”) corpus of Appalachian speech. The proposed parsed corpus will be accompanied by a full set of digitized recordings of the underlying speech, in the form of .wav files. The .wav files will be text-searchable in Praat (Boersma and Weenink 2011) as a result of force-aligning the transcripts with the speech signal, using text-to-speech alignment technology developed and made available by the University of Pennsylvania ([martinet.sas.upenn.edu/PPLClient/](http://martinet.sas.upenn.edu/PPLClient/)). The syntactically annotated text files will be searchable by any standard tree query language (e.g., CorpusSearch, Randall 2009).

The aligned parsed corpus will afford an unprecedented resource for the analysis of synchronic English dialect data, allowing both large-scale quantitative studies of the dialect data themselves and cross-dialectal and diachronic comparison with other parsed corpora of English. Such studies will allow researchers to address overarching and general theoretical questions regarding dialect variation and language change that would otherwise remain unanswered. Because the text-to-speech alignment process forces every aspect of the utterances to be recorded with the highest degree of accuracy, the corpus will be of particular interest to discourse analysts, as it will include every last hesitation, pause filler, and interruption, as well as time-stamps for all overlapping speech. For all users, the text-searchability of the speech signal will render the corpus transparent in that they will be able to conduct ordinary text searches to jump directly to desired points in the speech signal and to verify the transcription and the annotation.

The proposed project, which aims to create the first such corpus of Appalachian speech (or of any English dialect for that matter), will be based on Michael Montgomery’s *Archive of Traditional Appalachian Speech and Culture*, an unparsed corpus of speech that Prof. Montgomery has transcribed from recordings from ten parts of the Appalachian region in six different states. Additionally, it will include a smaller unparsed corpus of recorded speech, collected under PI Tortora’s previous NSF grant (see Section 4). We will use resources at the College of Staten Island to archive the files of the parsed corpus (see *Data Management Plan*). All component files will be made publicly available.

The project banks on four key elements: (1) Prof. Montgomery has over 30 years of research on Appalachian English, including over 20 years of collecting and transcribing recorded interviews in Appalachia; (2) co-PI Santorini has over 20 years of experience in constructing large parsed corpora and in refining the requisite computational tools; (3) co-PI Tortora has nine years of research on Appalachian syntax and has forged strong relationships with students and scholars at colleges in the Appalachian region and with native speaker informants in Northeastern Tennessee; and (4) co-PI Tortora has spent over two years of pilot work mastering the computational tools necessary to complete the project and training a now highly specialized research assistant (see Section 2).

The remainder of the project description is organized as follows. Section 1 reviews the theoretical and practical motivation for creating a parsed corpus of Appalachian speech. Section 2 provides a detailed account of the staff involved and the work plan, presenting the expertise of the team, the methods for creating the corpus, a description and time accounting of the tasks to be undertaken, and a plan for data management (elaborated in the *Data Management Plan*). Explication of the *Intellectual Merit* of the proposed project forms an integral part of this entire narrative. Section 3 provides a summary statement of the proposal’s *Broader Impacts*. Section 4 summarizes results of prior NSF funding.

### 1. Motivation for Present Project

A syntactically annotated (or parsed) corpus is a corpus of written text or transcribed speech that has been annotated with syntactic tags (e.g., NP-SBJ for “Subject Noun Phrase,” CP-REL for “Relative Clause,” and so forth). The corpus chosen for parsing could be anything from a corpus of historical texts, to a corpus of recorded telephone conversations, and it could be from any language; the choice of corpus is influenced in part by the creator’s focus on particular research questions. Over the past decade or so, many extensive parsed corpora have come into existence, including the 7.5-million-word corpus of historical

English produced at the University of Pennsylvania and the University of York, England (see Section 1.5), the 1.4-million-word Switchboard corpus of American speech (Godfrey & Holliman 1997), and a 600,000-word corpus of dialectal European Portuguese (Martins et al. 2010). Even while under construction, such corpora have given rise to novel results (see Section 1.5). There is, however, as yet no parsed corpus of American dialect speech. Section 1.1 motivates the creation of such a corpus (here we make arguments for how the proposed corpus will provide an unprecedented database for innovative research in synchronic syntactic variation; for a full discussion, see Tortora under review); Sections 1.2 through 1.5 discuss additional motivations and justification, and the ways in which the usefulness of the corpus will be enhanced.

### 1.1 Motivation from syntax: Intra-speaker variability and parametric theory

**1.1.1 Parametric Theory.** An important goal of current syntactic theory is to identify abstract properties, so-called parameters, that link clusters of syntactic phenomena, often superficially unrelated. A well-known example of such a parameter is the so-called verb movement parameter (Pollock 1989). Specifically, we observe that lexical verbs obligatorily follow sentence adverbs like *never* in Modern English, as is evident from the contrast in (1).

- (1) a. Mary never smokes.      b. \*Mary smokes never.

An apparently unrelated fact concerns the obligatoriness of *do*-support in questions, as shown in (2).

- (2) a. Does Mary smoke?      b. \*Smokes Mary?

The facts in (1) and (2) are superficially unrelated; for instance, even so astute and talented a traditional syntactician as Ellegård never explicitly connected the two sets of facts in his 1953 study of *do*-support (Anthony Kroch, pers. comm.). Nevertheless, modern generativists derive both sets of facts from a single underlying parameter. The parameter has been characterized in subtly different ways (see, e.g., Chomsky 1995); for present purposes, we will formulate it in terms of the strength of Infl, the inflectional head of a clause. If Infl is “strong,” it attracts the verb, which then overtly moves from its underlying position to adjoin to it. This movement yields verb-adverb orders as in (1b) and subject-verb inversion as in (2b). If Infl is “weak” (as is the case in Modern English), it does not attract the verb, and the verb remains within VP, following sentence adverbs as in (1a) and necessitating *do*-support as in (2a). Apart from details of implementation, this analysis of the data in (1) and (2) is widely accepted in the syntactic literature.

**1.1.2 Proving the parameter: diachronic syntax.** Strong independent empirical support for the theoretical claim that the patterns in (1) and (2) are attributable to a single parameter setting comes from comparative data. This data can be synchronic or diachronic. Synchronic comparative evidence for the parameter comes from the fact that various languages (French, insular Scandinavian) exhibit the converse of (1) and (2); in other words, the counterparts of (a) are ungrammatical, whereas the counterparts of (b) are grammatical in those languages. Diachronic evidence comes from the parallel loss of the (b) variants in English. In particular, Kroch (1989) provides quantitative evidence that (on the logistic scale that is appropriate for the replacement phenomena under consideration) the rate at which *do* support (2a) replaces the older verb movement option (2b) is the same as the rate at which adverb-verb order (1a) replaces the older verb-adverb order (1b), exactly as would be expected if the two phenomena are reflexes of a single parameter setting. We briefly review Kroch’s methodology here; unfortunately, for space reasons the presentation is oversimplified, but the main point should remain clear.

Kroch (1989) looked at gradual syntactic change of the above phenomena. In the earliest English texts (Early Middle English), he observed that the structures in (1a) and (2a) were not possible; instead, the (now ungrammatical) forms in (1b) and (2b) were the rule. However, at some point, a change began, with the introduction of (1a) and (2a) as apparent options; subsequently, for a period of hundreds of years, all of the forms seen in (1) and (2) were possible. With the passage of time, the grammatical options in (1a) and (2a) occurred more and more frequently, while the original forms in (1b) and (2b) occurred less and less frequently. Putting aside the details of the statistics, of most significance in this study is the following: the increase of *do*-support in interrogatives over time (seen in (2a)) was at the same rate as the in-

crease of the adverb-verb order over time (seen in (1a)), regardless of the fact that at any given time, one construction might have appeared in the text more frequently than the other. This is Kroch's "Constant Rate Effect." But why would these two apparently unrelated phenomena increase in use at the same rate over time? As Kroch argues, the Constant Rate Effect reflects the increased choice, over time, of a single grammatical option (our "weak Infl" mentioned earlier). That is, the increased choice of a single parameter will give rise to the increased use of all of the syntactic phenomena that are the surface reflexes of this choice (and of course, this entails that their increased use will be at the same rate). For the case at hand, this also means that, even though there must have been English speakers who at some point in history allowed both (2a) and (2b) — i.e., they exhibited "intra-speaker variability" — these two options did not belong to the same grammar; rather, they represented two different parameter settings, or, "two different grammars." In addition to providing independent evidence for the notion of parameters, the diachronic work also established that it is possible for speakers to have access to competing parameter settings in usage. Though this is not generally the case, it is well attested in situations of diachronic change.

It is important to be clear regarding the assumptions embedded in the present discussion. Note that the findings of Kroch's study strongly support the hypothesis that intra-speaker variability (or, "mixed phenomena") reflects the juggling, on the part of the speaker, of "two different grammars." This "syntactic diglossia" (Kroch 2001), i.e., a speaker's use of two different syntactic structures which are semantically identical, can be captured by the theory in the following way, in the case of the verb movement parameter: a mono-dialectal speaker of Modern English would only have in his lexicon the weak Inflectional head; likewise, a mono-dialectal speaker of Early Middle English who only exhibits lexical verb movement would only have in his lexicon the strong Inflectional head. This is illustrated in (3) and (4):

- (3) Lexicon of mono-dialectal Modern English speaker: **weak Infl** (yields no overt verb movement)  
 (4) Lexicon of mono-dialectal Early Middle English speaker: **strong Infl** (yields overt verb movement)

However, a bi-dialectal speaker of Middle English who exhibits mixed phenomena would have both (3) and (4) in his lexicon. For such a speaker, the question of whether the lexical verb moves or not is just a question of whether s/he chooses (3) or (4) at any given relevant moment. So the existence of two different Inflectional heads in this speaker's lexicon is the sense in which the speaker has access to "two different grammars." Note that the notion of "two grammars" as defined here does not entail a complete mental state of bilingualism (whatever *that* means); rather, for a particular point of variation, such as that of a bi-dialectal speaker of Middle English who allows both (2a) and (2b), we can hypothesize that the speaker has in his/her lexicon two different functional heads (as in (3) and (4)), which yield two different grammatical outputs. The juggling of the two different functional heads in the lexicon can be thought of on analogy with the juggling of two different lexical items, such as *pocketbook* and *purse*, or, *sofa* and *couch*. The difference of course is that when a Middle English speaker chooses one functional head vs. the other ((3) vs. (4)), there are direct computational consequences.

**1.1.3 The complication of intra-speaker variability and synchronic studies.** The import of Kroch's study lies in the fact that it provides robust empirical support for a parameter which was hypothesized to exist for independent reasons (both empirical and theory-internal). What makes studies like this possible, however, is (a) the existence of the parsed corpus of historical English texts, and (b) the very fact that we can look back at the change over time, and that we know how the story ends (i.e., we know what English is like today). But what if there had been a linguist alive at the time Middle English was spoken, who wanted to understand the nature of the intra-speaker variability observed? Such a linguist would have found him/herself in the middle of the change, and as such, would have been witness only to the existing variation, i.e., the fact that speakers used all four grammatical options seen in (1) and (2). That is, the data that linguist would have faced would have been the equivalent of the following:

- (1') a. Mary never smokes.      b. Mary smokes never.  
 (2') a. Does Mary smoke?      b. Smokes Mary?

This linguist would not have had the advantage of observing the trajectory of the change, or its conclusion. And although it is true that the Modern English judgments in (1) and (2) alone do not serve as solid

empirical support for the hypothesis that (1a) and (2a) cluster, the data in (1) and (2) at least do not *preclude* a clustering hypothesis, since we at least know that (1a) and (2a) are possible in the same system. However, our linguist alive during the Middle English period would not have had this advantage; witnessing the fact that speakers use all four grammatical options, as in (1') and (2'), s/he would not have had any way of knowing whether the clustering hypothesis was even precluded, given that there would have been no way to know whether (1'a) and (2'a) were even possible in the same system. That is, given that all four possibilities exist for a single speaker of Middle English, *prima facie* it would be just as reasonable to hypothesize that (1'a) and (2'b) belong to the same system, or that (1'b) and (2'a) belong to the same system, or that all four belong to four different systems, for that matter. It thus becomes far less clear what kind of indications the linguist could have used to suggest a direction for hypothesis formation regarding clustering. And even if some kind of sociolinguistic indications were available, there would still be the problem of finding empirical support for the clustering hypothesis suggested by the sociolinguistic data.

While all of the above is only an imagined problem for the data in (1') and (2'), linguists today are faced with this exact problem regarding any number of other analogous points of intra-speaker variability. Consider, in this regard, two cases from present-day Appalachian English (for a more thorough discussion of these and similar cases, see Tortora 2006, Tortora & den Dikken 2010, and Tortora under review).

As is well known (see e.g. Montgomery 1997a, Wolfram & Christian 1976; Hazen 1996, 2000a), Appalachian speakers exhibit variation in subject-verb agreement with third-person plural subjects, allowing both standard concord as in (5a) and so-called singular concord as in (5b).

- (5) a. The boys plow their corn in June.                      b. The boys plows their corn in June.

Tortora & den Dikken (2010) hypothesize, following Henry (1995) for Belfast English (which has a slightly different type of singular concord), that these two different forms reflect different possibilities for syntactic placement of the subject and that these different subject positions might reflect two different (but semantically equivalent) grammatical options — that is, “two different grammars,” in the sense discussed at the end of Section 1.1.2.

In addition to the variation concerning subject-verb agreement, Appalachian speakers also exhibit variation with respect to the superficially unrelated phenomenon of subject relative clauses. In addition to a standard form like (6a), they produce *true subject contact relative clauses* (SCRs); these are subject relatives without a relative pronoun or complementizer, as illustrated in (6b) (from the *Dante Oral History Project*). (The Appalachian construction differs from *pseudo subject contact relative clauses* of Belfast English, analyzed by Henry 1995 as Topic-Comment structures.)

- (6) a. But he tied the company up some way to get a royalty off the timber that was cut for the mines.  
 b. But he tied the company up some way to get a royalty off the timber \_\_\_ was cut for the mines.

Thus, Appalachian speakers exhibit intra-speaker variability with respect to all of the forms in (5) and (6). In other words, they use all four forms, and admit all four in grammaticality judgment tasks (see Section 4). Moreover, although singular concord and true SCRs are superficially unrelated, there is good reason to believe that they are the reflexes of a single parameter (see Section 4 and den Dikken 2006).

The problem now, however, is the following: regardless of the theory-internal reasons one might have for hypothesizing that these two apparently unrelated phenomena in (5a) and (6a) cluster, the independent question of how one would find empirical support for this hypothesis arises. We cannot simply ask speakers if they think that (5a) and (6a) arise from the same parameter, because human beings do not have such intuitions (if we did, the job of the syntactician would be easy). Furthermore, given the intra-speaker variability observed, we cannot even be sure whether the clustering hypothesis is precluded at all, as we cannot rely on speaker judgments to tell us whether (5a) and (6a) even belong to the same system (i.e., dialect/register). As far as the speakers are concerned, all the sentences in (5) and (6) are possible, and there is no *prima facie* empirical indication that any two of these four possibilities are part of the same system. It may be tempting for a non-Appalachian speaker to hypothesize that (5a) and (6a) belong to the same system, or dialect, given that both are perceivable as non-mainstream, in contrast with their “standard-seeming” counterparts in (5b) and (6b), but there are two observations to make in this regard. First, the

linguist's sense of what counts as non-mainstream may not reliably line up with the speaker's sense of what counts as non-mainstream (an issue addressed by Buchstaller & Corrigan 2008).

Second, in any case, there are numerous other examples of sets of semantically equivalent syntactic variants in Appalachian English, where both members of the set are non-mainstream. Consider in this regard the examples of existential sentences (7) (Tortora 2006, adapted from Montgomery & Hall 2004):

- (7) a. They is something wrong with him. [= There is something wrong with him.]  
 b. They are something wrong with him. [= There is something wrong with him.]

The semantically equivalent syntactic variants in (7) raise many important questions regarding the syntax of agreement, the nature of existential *they*, and the source of the plural form in (7b) (all are issues discussed in detail in Tortora 2006). As Tortora (2006) argues, there are theory-internal reasons to hypothesize that the form *are* in (7b) is, contrary to appearances, a singular form of the verb. The possibility in Appalachian for (8b), alongside (8a), might in fact lead us to conclude that this hypothesis is on the right track (data adapted from Montgomery & Hall 2004):

- (8) a. He is better. b. He are better.

That is, speakers' use of the structure in (8b) suggests that the form *are* in (7b) is singular, contrary to appearances; if this were the case, then a theory holding that expletive *they* in (7) does not trigger plural agreement would be supported. The problem is, as things stand, we do not have any independent empirical evidence which would confirm that (7b) and (8b) even belong to the same system (and therefore, that they could be linked), despite the fact that a speaker may use both. Thus, while there may be theoretical motivation for claiming that the form in (7b) is singular (as in Tortora 2006), there is no empirical support for this claim. And in any case, we can imagine a competing hypothesis with different theory-internal motivation, which holds that the form *are* in (7b) arises from some form of plural agreement — triggered by a formally plural existential *they*. But the lack of an empirical means to decide the issue leaves us at an impasse. And crucially, to make the point we set out to make here, we cannot simply claim that the non-mainstream (7b) “matches up” with the non-mainstream (8b), because in this case, (7a) and (7b) are both non-mainstream forms. As Cornips (2006) notes within the context of the study of Dutch dialects, “...clear-cut judgments between the local dialect and the standard variety are not attainable at all.” In the Appalachian case at hand, the complexity becomes even greater, as neither of the varying forms in (7a) and (7b) belong to the standard; here, the speaker is juggling two alternating non-standard forms (see Tortora under review, especially Sections 3.2 and 3.3).

The question of how to test the clustering hypothesis for such cases of intra-speaker variability is thus a non-trivial one, and syntacticians have made little progress on addressing the issue, and on finding forms of empirical support for clustering hypotheses for the numerous cases of intra-speaker variability found in the languages of the world. In the following sub-section, we claim that building parsed corpora of varieties like Appalachian would constitute a big step towards making progress to this end.

#### 1.1.4 Simulating the Constant Rate Effect in Synchrony: one motivation for the present project.

Given the success of Kroch's (1989) diachronic study, we would like to look to that methodology as a model for finding empirical support for our hypothesis that the apparently unrelated (5a) and (6a) are in fact linked via an abstract underlying parameter. But how can we possibly apply a methodology which was designed for diachronic (i.e., historical) change, to a synchronic (i.e., current) linguistic situation?

We would like to suggest the possibility of an approach which would capitalize on the following well-known and accepted observation regarding dialectal variation: closely related dialects (such as Appalachian English and another variety of English, or one variety of Piedmontese and another) synchronically represent “stages of change” of a language. This observation goes back at least to Gaston Paris (1888), and is discussed explicitly by e.g. Labov (1975) and Wolfram (1984). As a very simple illustration, take the past participial form of the verb *help*. In Modern Standard English, the form is *helped*, as in *I have helped him many times*; for some speakers in Appalachia (see Montgomery & Hall 2004), the form is *holpen*, as in *I have holpen him many times*, just as it was in Old English. The form *helped* (which is also used by speakers in Appalachia) is thus a modern innovation, representing a change that took place in the

evolution of the grammar of some speakers. However, the fact that the form *holpen* still exists for some speakers in Appalachia suggests that the change has not yet entirely taken place for all. That is, even though varieties of Modern Appalachian English and Modern Standard English are contemporaries, in the former we find grammatical forms that represent an older stage of the ancestral language of both these dialects. So in this one small corner of the grammar, with this one small example, we can say that Grammar A represents an older stage of Grammar B, even though the two are contemporaries.

What this means is that contemporary, closely related dialects are equivalent to different stages of change. What this means in turn is that we can capitalize on this well-established observation of Paris/Labov/Wolfram, and apply Kroch's diachronic methodology to synchronic situations. Kroch looked at different stages of change in the English language to establish that the adverb-verb order seen in (1a) was (and is) inextricably linked to the phenomenon of *do*-support seen in (2a). Specifically, he observed that at different stages of change, the use of adverb-verb order increased at the same rate as the use of *do*-support (the Constant Rate Effect). We propose that in a similar way, we can study the corpora of different but related dialects of English, and count the frequency of occurrence of the contemporary syntactic variants, such as those seen in (5) and (6) (or in (7) and (8)). If these different but related dialects — e.g. (5a)/(6a) vs. (5b)/(6b) — represent different stages of change, then given the hypothesis that (5a) and (6a) are linked via a single underlying parameter, we predict the following: the frequencies at which the forms in (5a) and (6a) occur in the different dialect corpora should differ at the same rate — regardless of whether they occur at different frequencies in the overall data. In other words, even if (5a) occurs in the corpus of Dialect A much more frequently than (6a), the frequency of both should change from Dialect A to Dialect B to Dialect C at the same rate (with A, B, & C representing different “stages”).

We would like to stress furthermore that the syntactic variants seen in (5/6) and (7/8) represent but a small set of relevant examples, which we appeal to here for pure illustration. Of course, there are many, many more points of syntactic variation in the English of Appalachian speakers which raise questions in relation to theories of micro-parametric variation (as discussed for example in Hackenberg 1972, Wolfram & Christian 1976, and Montgomery & Hall 2004). All of these features of Appalachian could be tapped for analysis of statistical tendencies through the creation of a parsed corpus. Just to mention a few: (i) negative concord (*I didn't see nobody*); (ii) negative inversion (*Wasn't nobody here*); (iii) non-inverted negative concord (*Nobody wasn't here*); (iv) complex floating quantifiers (*We can every one sing*); (v) variable simple past and past participial verb forms (*known/knowed, seen/seed*); (vi) variable use of the perfect construction to semantically encode simple past for the verb *be* (...*that's been a car* [= *that was a car*]); (vii) variable use of the pronoun *hit* [= *it*]; (viii) variable use of existential *it* (vs. *they* and *there*, which are also used variably); (ix) transitive expletive constructions (*There can't many people say that*); variable use of morphologically complex nominative pronouns (*we'uns* vs. *we*). See Section 1.5 for comments on the presence of (iv), (ix), and subject contact relatives in the *Penn Corpora*.

**1.1.5 Conclusion of arguments in this section.** To conclude: herein thus lies one of the motivations for the creation of a syntactically annotated corpus. All of these questions discussed in this section (1.1) regarding synchronic intra-speaker variability in syntax, and regarding clustering and parametric theory, can only be addressed by the creation of a syntactically annotated corpus. As noted by Wallenberg et al. (2010), parsed corpora allow us to form and test hypotheses about statistical tendencies in morpho-syntax, and to study syntactic variation; and of course, the empirical value of the corpus is strengthened by the fact that the results of studies based on it can be easily replicated by future researchers. Our main point here is that while this has become accepted as a truth for diachronic variation and change, it is time to pursue studies of synchronic variation in English in the same way. The approach we advocate in Section 1.1.4 above offers a way to provide empirical support for claims regarding micro-parametric variation and the clustering of properties for synchronic situations which involve intra-speaker variability. Clearly, the motivation is driven by particular theoretical considerations, claims, and assumptions, which derive from the Principles and Parameters framework, and in particular, the “two grammars” hypothesis. We are aware, however, that there are researchers who do not adopt the “two grammars hypothesis,” as described above, but rather, take intra-speaker variability to reflect options available within a single grammar; see

for example Anttila 1997; Guy 1991; Guy and Boberg 1996; Reynolds 1994; Reynolds and Nagy 1994, among others. Furthermore, there are non-generativists who do not subscribe to the view of grammar outlined above, in more general terms. We hope it is nevertheless clear that the corpus we propose would be of use to these linguists as well. In fact, because the corpus is theory-neutral in this sense, it promises to serve as a testing ground for the various predictions made by the different theories.

**1.2 Why the English of Appalachian speakers, as opposed to any other corpus of speech?** Another question which might arise is why we propose to create a parsed corpus of *Appalachian English*, in particular. Given Labov's (1972:109) observation that intra-speaker variability is a universal rule, then wouldn't a parsed corpus of the speech of *any* community or region in the world serve the same purpose, in terms of addressing the theoretical questions outlined in Section 1.1 above? Of course, the answer is "Yes." However, there are a number of practical reasons for using Appalachian English for this purpose.

**1.2.1 There is a ready-made, 1,000,000-word unparsed, transcribed corpus available as a base.** Fortunately for the field of Appalachian studies, Distinguished Professor Emeritus Michael Montgomery (Univ. of South Carolina) realized long ago that as far as regional speech goes, there is nothing available which remotely rivals the amount of recorded material that exists for Appalachia. Through many years of research and careful weeding through materials, he has collected large amounts of recorded speech from numerous collections already in existence in the Appalachian region; his recordings come from ten parts of the Appalachian region in six different states. Here are some examples: (i) the *Appalachian Oral History Project* housed at Alice Lloyd College in Pippa Passes, KY, and at Appalachian State University in Boone, NC, which represents traditional speech from Eastern Kentucky and Northwestern North Carolina; (ii) the *Dante Oral History Project* housed in the Archives of Appalachia at East Tennessee State University, which represents speech from Dante, in Southwestern Virginia; (iii) the Joseph Hall Tapes, on which Montgomery's *Smoky Mountain English Dictionary* is based. These are just a few of the collections of recorded interviews from different parts of the Southern Appalachians underlying Montgomery's *Archive of Traditional Appalachian Speech and Culture* (ATASC), which contains over 1,000,000 words of speech that Montgomery has already transcribed. These recordings are of great value, in part because of the different areas of Appalachia that they represent, and in part because some of the interviewees were born as far back as the 1850s, which will allow us to compare earlier stages of the dialect with present-day speech; furthermore, social information about each speaker is available, something which enhances the corpus' usefulness for sociolinguistic research. Together, all of these transcribed corpora represent a ready-made base from which to create a syntactically annotated corpus, and we have been collaborating with Prof. Montgomery over the past 21 months; in addition, we have recordings which were collected and transcribed under a previous NSF grant (see Section 4). Given this accessibility, Appalachian speech is a good place to start, to establish a model which can eventually be used for other varieties.

**1.2.2 There are ready-made hypotheses to test, once the corpus is completed.** As already discussed in Section 1.1.3, we have done enough previous work on Appalachian syntax to have well-motivated and ready-made hypotheses to test on the parsed corpus; see Section 4, "Prior NSF Support" as well as Tortora & den Dikken (2010), Tortora (2006), and Tortora (under review). Thus, we are poised to statistically study a number of questions regarding syntactic variation in Appalachian English.

**1.2.3 There is a ready-made text-to-speech aligner already in existence for English.** As highlighted in the introduction, and as we motivate in Section 1.3 immediately below, the proposed parsed corpus will be accompanied by a full set of digitized, text-searchable recordings of the speech from which the corpus is transcribed, in the form of .wav files. The .wav files on which the parsed corpus is based will be text-searchable in *Praat* as a result of force-aligning the transcripts with the speech signal, using the "text-to-speech" alignment technology developed by Jiahong Yuan at the University of Pennsylvania. Yuan's PPL Forced Aligner Online Processing System (<http://martinet.sas.upenn.edu/PPLClient/>) creates a so-called text grid, which can be viewed in alignment with a spectral representation of the speech signal, using the freely available software *Praat*. What the text alignment ultimately creates is a configuration in *Praat* in which each word of the speech signal is aligned with each word of the transcript; see the "screenshot" in

Appendices 1A and 1B in *Supplementary Documents* for an illustration.

Importantly, the Penn PPL *Forced Aligner Online Processing System* works off of an English dictionary. Thus, we can take advantage of its existence to align Montgomery's transcripts with the digitized recordings, by virtue of the fact Appalachian English is English; that is, most of the words in the Appalachian corpus are in the PPL aligner's dictionary already. Preliminary experimental work and consultation with Jiahong Yuan has revealed it to be necessary to add some words to the PPL dictionary, in order to make it more Appalachian-friendly; for example, since the verb form *knowed* comes up frequently in the Appalachian text, it is necessary that the PPL aligner have this form in its dictionary. Nevertheless, this modification to the dictionary requires comparatively little work, and thanks to the collegiality of Prof. Yuan, the dictionary has, over the past 21 months, been modified to accommodate Appalachian speech.

**1.3 Why forced text aligning?** The reader might wonder why we have chosen to include, as an additional feature of the syntactically annotated corpus, a full set of digitized, text-searchable recordings of the speech on which the corpus will be based. As we discuss in Section 2 below, the process of correcting the results of the forced-alignment is time-consuming, and will represent a good part of the labor in the building of the entire corpus; nevertheless, we show that there is good motivation for including this feature.

First, as we have learned, the process of forced alignment coerces a level of accuracy in the transcription which is simply not achievable by relying solely on the human ear and human memory. This is in part because the results of the forced alignment process allow the researcher to view each word in the transcript together with its sound-spectral representation. Such analysis of the speech signal in tandem with the text forces the transcription process to go beyond the gist of what the speaker said, and to include every last hesitation, false start, and pause-filler, as precisely as possible. In this regard, it is important to acknowledge that as hard-working and precise as a transcriber wants to be, the unconscious tendency to provide subjective interpretations of what was heard is always a danger. However, the mechanical, concomitant listening and viewing of the speech signal allows us to overcome this subjective tendency to a significant extent. Furthermore, it is equally important to note that the spectral representation of the speech signal aids tremendously in the researcher's ability to identify the word being uttered. Simply put, a parsed corpus of Appalachian speech will only be valid if the text reproduces the speech as accurately as both humanly *and* technologically possible. Since our experience over the past 21 months shows that forced alignment yields a much more accurate transcript than that produced by the human ear alone, we are compelled to make this a part of the project.

Fortunately, a second and equally attractive consequence of the forced alignment process is the following: we will make the product of this text-to-speech alignment available to researchers, along with the parsed corpus itself. Given that the text-grids allow for text-searchability of the speech signal, users will be able to do global searches for features of interest, and jump directly to the points in the *.wav* file containing these features. For example, if a phonetician wants to view and simultaneously listen to the speech signal associated with all instances of the string "born" in the 1,000,000-word speech signal, this can be done via a garden-variety text search. This open availability and accessibility of the data in all its aspects will permanently allow future generations to readily review, with single key-strokes, the actual speech signal corresponding to any part of the corpus. And since future phoneticians are sure to develop evermore sophisticated tools of analysis, this will provide ongoing opportunities to perfect the product (for example, in ambiguous cases which current technology has difficulty resolving). Furthermore, the chance for error will be reduced, because any researcher can check whether they agree with the accuracy of the transcription. To give just one example: in many cases, it is difficult to tell in our recordings whether a speaker has uttered *they's* or *there's* or *they'uz* or *there'uz* (as in, *There's a guy down the street*). Since we do not feel in a position to make a final judgment regarding what has been uttered in such cases, our set of guidelines for use of the corpus can include a list of these "known issues," so that any researcher interested in checking up on the accuracy of our transcription in these cases can easily do so. In this regard, we have already begun to develop a system of notation which signals to the user, for the ambiguous cases, that there is some doubt regarding the true nature of the utterance (e.g., we notate the doubtful item with the pipe symbol, i.e., [they's | there's]). In this way, users can do searches and provide their own interpre-

tations, capitalizing on levels of expertise that the corpus creators themselves may not have. Thus, creating and providing files of the transcripts that are force-aligned with the .wav files of the speech signal will enhance this project's usefulness for both phoneticians and phonologists, and increase replicability.

Finally, because the text-to-speech alignment process forces the written registration of every aspect of the utterances with the highest degree of accuracy, discourse analysts will find the corpus useful as well, given that it will include every last hesitation, pause filler, and interruption, as well as time-stamps for all overlapping speech. In this last regard, it is important to note that traditional transcription procedures do not provide a ready way of indicating when and how two speakers are overlapping. One of the features of the forced alignment correction process in *Praat* is that it allows for the creation of independent "tiers" of speech, each of which is time-stamped; these time stamps can in turn be provided in the final, linear transcript. There are thus multiple reasons why force-aligning the speech signal with the text is an important component in the process of creating a parsed corpus of Appalachian speech.

**1.4 What about the quality of the recordings?** Given that some of the recordings are older (most notably, the Joseph Hall tapes which were made in the 1930s; see Appendix 1B), the question might arise as to whether it is worthwhile to include the .wav files of these recordings for phonetic research. We have consulted with phoneticians (Jiahong Yuan of U. of Pennsylvania and Kathleen Currie Hall of CUNY), and both have indicated that even the worst quality recordings in the ATASC collection contain enough information available in the signal to make them worthwhile for acoustic analysis. Although analysis of consonants such as fricatives would be problematic on the poorest recordings, they would still be useful for phonetic analysis of vowels, duration, and intonation; for example, generally speaking, the first two — and often three — vowel formants (F1, F2, and F3) are clear on the poorest of these recordings; furthermore, noise reduction options available on *Praat* and *Audacity* render speech more audible, and as such can be a useful option employed by researchers in tandem with viewing the speech signal.

**1.5 Why 1,000,000 words?** Although there exist online corpora of dialectal English (e.g., the 400,000-word dialectal part of the Helsinki Corpus, described in Ihalainen 1990), to our knowledge, none of them is as large as the corpus we propose, and none has been parsed. The closest counterparts to our proposed corpus are the parsed historical corpora that have been built over the past decade for various stages of English (Kroch & Taylor 2000; Kroch et al. 2004, Kroch et al. 2010; Pintzuk & Plug 2001; Taylor et al. 2003; Taylor et al. 2006), Old and Middle French (Martineau et al. 2010), Icelandic (Wallenberg et al. 2011), and Portuguese (Britto et al. 2010). These are all within the same order of magnitude as the corpus we propose (1-2 million words). A 1-million-word corpus can be constructed within a reasonable length of time; at the same time, it allows the discovery of novel results and supports research into relatively rare constructions. For example, the 1-million-word *Penn Parsed Corpus of Middle English* (Kroch & Taylor 2000) was the basis of several results that were previously unknown (Kroch & Taylor 1997; Kroch et al. 2000; Speyer 2010; Trips 2002; Wallenberg 2009). Given that our proposed corpus is synchronic, its size will support research into rarer constructions than are supported by the currently available historical corpora (e.g., the incidence of zero-marked SCRs or doubly-marked *wh*-constructions). Thus, previous research has shown that results are obtained most readily on corpora that are at least 1-million words. This is again something which makes Montgomery's ATASC attractive as a base for a corpus of parsed American speech: it provides a large enough body of material to make a parsed corpus worthwhile.

As preliminary empirical support for the claim that the proposed corpus is appropriately sized, consider the fact that three of the variable phenomena of interest to the proposed project (see paragraph immediately preceding section 1.1.5) occur reasonably often in the historical corpora of English, and we can therefore use them to provide a rough estimate of their incidence in our proposed corpus. Complex floating quantifiers are not attested in the 1.3-million-word PPCME2 (Kroch & Taylor 2000), but occur 54 times in the 1.8-million-word PPCEME (Kroch et al. 2004) and 33 times in the 1-million-word PPCMBE (Kroch et al. 2010). If we assume these rates of occurrence, we predict 30-33 complex floating quantifiers in the proposed corpus. Transitive expletives occur 150 times in the PPCME2, but decline in frequency after 1500, occurring 47 and 49 times in the PPCEME and the PPCMBE, respectively. If we assume the post-1500 rates of occurrence, we predict 25-50 transitive expletives in the proposed corpus. Subject con-

tact relatives occur 76 times in the PPCME2, 203 times in the PPCEME, and 23 times in the PPCMBE. Assuming that the large drop between the PPCEME and the PPCMBE indicates stigmatization in the standard language, and excluding the data from the PPCMBE on those grounds, we predict 58-113 subject contact relatives in the proposed corpus (including the PPMCBE predicts 23-113). All of the estimated Ns for the three variables fall well within the range of previous quantitative corpus-based research, suggesting that the proposed corpus size is reasonable.

## 2. Methodology / Work Plan / Staff and Consultants

**2.1 Project Team.** Before laying out our work plan/methodology, we summarize the qualifications of the project team.

**2.1.1 PI Christina Tortora (Lead Institution).** As discussed in Section 1.1, this project is in part motivated by years of research done on Appalachian syntax by PI Tortora. Section 4 summarizes some of the results from her previous collaborative NSF award (*The Comparative Morphosyntax of Appalachian English*), 2006-2010, and the *References Cited* lists numerous articles published in journals and books, and numerous invited talks given, on syntactic theory and Appalachian. These articles and lectures have resulted from nine years of research on Appalachian syntax and of travel to the Appalachian region to work at local colleges and universities, and to conduct fieldwork. Since 2008, Tortora has been regularly visiting two life-long natives (now in their 90s) in Kyles Ford, Hancock County (in Northeastern Tennessee, north of Clinch Mountain). She has developed lasting relationships with them, and is in regular phone contact; this lasting and continuous connection to the local community allows for frequent data checking with current speakers. In addition, she has developed a professional relationship with Tiffany Williams of Letcher County, KY (see 2.1.5), who has served as a regional consultant to both her and Montgomery for the last 3 years; she has also been working with Claude Crum, VP for Academic Affairs at Alice Lloyd College, in Pippa Passes, KY (see 2.1.6), to coordinate efforts on organization of the *Appalachian Oral History Project* (AOHP), one of the components of Montgomery's ATASC. In addition, Tortora has engaged in over two years of pilot work on the present project, to master software and techniques and develop a clear idea of the necessary division of labor and timeline to achieve the goals of the overall project, and to train a now highly specialized research assistant (see Section 2.1.4); she has also been working with PI Santorini to develop a protocol for the creation of the proposed corpus, and has been collaborating with Montgomery in the gathering of materials and interpretation of the data.

**2.1.2 PI Beatrice Santorini.** Although several parsed corpora of different languages already exist and are available to the public, the original, most well-known, and most highly emulated are the *Penn Parsed Corpora of Historical English*, created by University of Pennsylvania researchers. PI Santorini has been involved in the creation of about 3 million words worth of these parsed corpora (1 million: *Early Modern English*; 1.8 million: *Modern British English*), and 1 million words worth of the *Old and Middle French* corpus. This 4 million words of experience not only speaks to her expertise, but also to the fact that she can get the job done. Santorini also has years of experience writing and customizing scripts for unique situations such as this (see Step 4 below; it should be noted that Santorini 1990, and Marcus, Santorini, & Marcinkiewicz 1993 are the most cited works in the parsed corpus creation industry). For the past 21 months, Santorini has been working with Tortora to develop a protocol for the creation of the proposed corpus, and has been experimenting in writing scripts which will expedite numerous steps in the process of creation (e.g., scripts which will convert preprocessed transcripts from Word format to plain text files suitable for input to the PPL forced alignment algorithm, and which will convert the output of forced alignment algorithm (text grids) to input for automatic POS taggers). She has also been adapting and creating software to partially automate POS tagging correction.

**2.1.3 Consultant Michael Montgomery.** Prof. Montgomery, a native of Appalachia, has over 30 years of experience in researching and writing on Appalachian English, and his *Dictionary of Smoky Mountain English* includes the most extensive morphosyntactic description of this variety. Montgomery has used his expertise to prepare the transcripts which are part of his *Archive of Traditional Appalachian Speech* (ATASC), and which form the base of the proposed corpus, and which are essential to expedite the

forced-alignment and thus the syntactic parsing. Given the extent of Montgomery's expertise, he will prove invaluable also in assisting in decision-making in the tagging of the proposed corpus.

**2.1.4 Research Assistant Frances Blanchette.** Blanchette, a graduate student in the Linguistics program at the CUNY Graduate Center, has been Tortora's research assistant on this project since June 2010. Blanchette distinguishes herself as an extremely intelligent, devoted, flexible, efficient, analytical, methodical, meticulous, and competent assistant, and is in fact one of the principle reasons we have been seeking funding to support this project. We would like to underscore the timeliness of Blanchette's interest in this work, and we cannot afford to let too much time pass without securing additional funds to pay her; we have invested an enormous amount of time in training her in various aspects of the corpus protocol (see especially Steps 1-3 below), and she now stands as an expert in her own right. Her assistance on this project has also prompted her to do her own research on Appalachian syntax; she is currently working on a project entitled *Levelling in the presence of n't in varieties of English*, for which she was recommended for an NSF *Research Experiences for Graduates* (REG) award in June 2011. This concomitant development as a scholar on the topic makes her the perfect graduate student assistant and colleague. In terms of authorship of the proposed corpus, her contribution has already been so great, that we plan to include her in the list of co-authors (Tortora, Santorini, Montgomery, and Blanchette).

**2.1.5 Tiffany Williams.** PI Tortora met Tiffany Williams at East Tennessee State University (ETSU) in 2008. Williams is a graduate of ETSU, with a background in Appalachian Studies and Linguistics (she was a student of Steven Gross). She is also a native of Letcher County, Eastern KY, and a native speaker of the dialect. For the past 3 years, Tortora and Williams have been in touch, and Williams has also been assisting Montgomery in the collection of materials from Alice Lloyd College in Eastern Kentucky. She has proven herself to be an invaluable contact and local consultant.

**2.1.6 Other participants and consultants.** Tortora has been in contact with Claude Crum, VP for Academic Affairs and English Professor at Alice Lloyd College, in Pippa Passes, KY, regarding the AOHP. Dr. Crum is committed to collaborating with our team in any way that he can, and is currently supervising two students at Alice Lloyd, Jamie Hold and Justin Maynard, who are in the process of organizing the AOHP. Because ALC can accept private donations, in Summer 2010 Tortora donated \$200 of personal funds to support this student work on the AOHP. Unfortunately, Alice Lloyd "...does not accept direct state or federal funding from the federal, state, or local government..." ([alc.edu/future\\_students/quick\\_facts.php](http://alc.edu/future_students/quick_facts.php)), so our negotiations to work the ALC students into our project budget failed. Nevertheless, Crum and his students are committed and delighted to work with us. In return, Tortora plans to give workshops on the completed parsed corpus at ALC, so that the students will learn how their local oral history collection relates to Linguistics (see Section 3); the students will also be acknowledged in our manual on the corpus.

## 2.2 Work Plan

In this section, we describe the general protocol we have devised for the creation of our corpus (given that the series of tasks are too numerous to mention, we only describe each task in general terms). Although we describe each task in terms of "steps," the reader should keep in mind that in many cases the tasks involved will not necessarily be done strictly sequentially.

**Step 1:** All of the cassette tapes that form the basis of the *Archive of Traditional Appalachian Speech and Culture* (ATASC; see Section 1.2.1) have to be digitized. Based on preliminary work we have done over the past 21 months on the Joseph Hall tapes (which represent approximately 50,000 words of text), we estimate that it takes approximately 30 hours to digitize 50,000 words of recordings. It would take 20 times that amount of time to digitize 1,000,000 words worth of recordings, which amounts to 600 hours, just to digitize. Since 50,000 words have already been digitized, we can reduce that number to **570 hours**. However, in addition to digitization, the resultant .wav files have to be what we could term "streamlined." That is, all extraneous recorded material — e.g., music and reading (or anything which is not part of the text of the corpus) — has to be chopped out of the .wav file. This is because the *PPL Forced Aligner* (see Section 1.2.3) cannot match the transcripts up with the speech signal if there is intervening extraneous

audio. Our preliminary work reveals that it takes approximately 20 hours to streamline the *.wav* files of 50,000 words, so that 400 hours will be necessary to do this for 1,000,000 words. Since 50,000 words have already been streamlined, we can reduce this number to **380 hours**. It will thus take a **total of 950 hours** to make the 1,000,000 words worth of alignable *.wav* files. Since it would be most efficient to accomplish this as early on in the grant period as possible, we propose to hire two graduate research assistants who will each work for 475 hours in the first 25 weeks of the grant period; we have not yet identified the exact individuals who will do this work.

**Step 2:** The *.wav* files as prepared above have to be uploaded, together with pure-text transcriptions, onto the *PPL Forced Aligner Online Processing System*. While this is also a time-consuming process, during the semester preceding the proposed grant period (Fall 2011), PI Tortora will be on a leave of absence for this project, supported by an NEH Fellowship award; as such, preparing pure-text transcripts from the *.doc* files provided by Montgomery and running the alignment are two of the many tasks which she can perform, almost simultaneously with Step 1 (in a semi-serial / semi-parallel fashion). The PPL Aligner produces what are called “text-grids.” In *Praat*, the text-grids appear as text lined up under the *.wav* file; this can be seen in Appendices 1A and 1B in the *Supplementary Documents*. In a text-editor like Emacs, the text-grids appear as a list of words, with various kinds of coding, including time-stamps. This can be seen in Appendix 2.

**Step 3:** As the so-called “text-grids” (i.e., the Results of the PPL Forced Alignment Process) are produced, Blanchette will check and correct these text-grids, using *Praat*. This is not only a very time-consuming process, but it also requires experience and a good understanding of Appalachian grammar and phonology, which is why we propose to employ Blanchette, who has been being trained over the last 21 months. This task is not trivial: the alignment process is not 100% accurate, in part because the transcripts which act as the input have a degree of human error built into them, as discussed in Section 1.3 above. This causes the aligner to misalign parts of the speech signal with the transcript. As noted above, correcting the results to create an accurate “text-grid”— which will serve as the basis of the text used for the parsed corpus transcript — is a time-consuming activity; we have determined that it takes one hour to correct about 550 words of a *.wav* file which has been force-aligned with the transcript it is matched up with; this means it will take approximately **1800 hours** to correct the forced-alignment of 1,000,000 words. Tortora and Blanchette will divide this workload, as we would like her to also be involved in the Part of Speech Tagging and Parsing processes (Steps 5 & 6).

**Step 4:** As noted just above, the corrected text-grids will serve as the basis for the text used for the parsed corpus transcript. This is a domain where PI Santorini plays an important role. As can be seen in Appendix 2 in the *Supplementary Documents*, the text-grids are interpreted in a text-editor as a list of words with various forms of coding. Santorini will work on how to best convert this configuration into something that can serve as input to a Part of Speech tagger. The problem is, again, not a trivial one, as a number of issues need to be taken into consideration, including, but not limited to: (i) the re-introduction of punctuation (which is stripped from the transcripts used as input by the PPL aligner — see Step 2); (ii) the stripping of irrelevant coding such as time-stamps, while taking into account the need to keep time-stamps for the overlapping speech; (iii) the reintegration of the overlapping speech (which is stripped apart as separate text-grids by the *Praat* correction process described in Step 3); and (iv) the preservation of the system of notation used for doubtful items (see Section 1.3).

**Step 5:** This brings us to the actual “Part of Speech” (POS) tagging of the text. POS tagging is not as simple as feeding text into a machine. Once the corrected text is sent through a “tagger,” which labels each and every word with a “Part of Speech” tag, it takes time for the researcher to correct all of the mistakes, as tagging is not 100% accurate. This step will again involve Santorini, in part for training both Tortora and Blanchette, who will share this workload. We must not only learn how to run the tagger, but also how to correct the mistakes. It takes on average one hour to correct 2,000 words worth of POS-tagged text. Thus, this activity will take approximately 500 hours for 1,000,000 words.

**Step 6:** Once the “POS tagging” mistakes are corrected, the text is then sent through a “parser,” which assigns syntactic structure to the text. Again, here, Santorini will play a large role, as we learn how to use the parser and correct the mistakes in the output. It takes on average one hour to correct 1,000 words

worth of parsed text. Thus, this activity will take 1,000 hours, for 1,000,000 words. There are thus **1,500** labor hours involved in Steps 5 and 6, which in part why the proposed project period is 3 years. Tortora and Blanchette will again divide the labor. In addition, for both Steps 5 & 6, Montgomery's expertise on Appalachian lexicon and grammar and native familiarity with the region's speech will be invaluable.

**Other activities:** There will be other necessary activities, in addition to those outlined in Steps 1 through 6. For example, Montgomery notes that there are a number of place names throughout the ATASC which still need to be identified; since many of these place names do not appear on any local maps, we will have to retain the services of approximately 6 different regional consultants from 6 different localities in Appalachia (one of these will be Tiffany Williams; see 2.1.5). Five of these local consultants will be contacted by Montgomery directly, and all 6 will listen to the portions of the recordings where speakers use these place names, in order to provide identifications. It will be necessary to transcribe these place names before we match the transcripts up with the .wav files using the PPL Forced Aligner, as such missing information otherwise promises to present significant problems for the Aligner. In addition, Tortora will manage and archive all of the relevant files so that they can be accessed as a coherent corpus by the general public. Computing and internet resources at the College of Staten Island will be used to archive these files; they will include: (a) the .wav files of Appalachian speech, from which the corpus is transcribed; (b) the "text grids," which are aligned with these .wav files with *Praat*; (c) ordinary text files of syntactically annotated text, in the style of the *Penn Parsed Corpora*; (d) ordinary text files of the *unparsed* text, which can be re-purposed by anyone in any way; and finally, (e) a manual, in the style of the *Penn Parsed Corpora*, laying out the following: (i) a step by step protocol for how to create a corpus like this one, (ii) instructions on how to use the corpus (with instructions for how to use *Praat* and *CorpusSearch2*, which are both free), (iii) an inventory of the known problematic issues, (iv) plans for updating and improving the corpus, (v) social information about each speaker (something which enhances the corpus' usefulness for sociolinguistic research), and (vi) an invitation to users to contact us regarding mistakes and problems (more details are provided in the *Data Management Plan*).

### 3. Broader impacts of the proposed activity

As outlined throughout this narrative, the proposed corpus will be of use to scholars in many sub-disciplines of Linguistics, as well as to anyone interested in analyzing the vocabulary, grammar, and text of American speech. Given the innovative nature of the corpus, the protocol for its creation will also serve as a model for researchers who wish to create similar corpora for other dialects. Furthermore, PI Tortora will use the corpus as a teaching tool in her undergraduate Linguistics courses, where students are training to become English teachers in the New York City public school system; similarly, students at two different colleges in the Appalachian region with an *Appalachian Studies* component will be trained to use the corpus (we have already coordinated with Claude Crum of Alice Lloyd College in Pippa Passes, KY, and with Ted Olson of East Tennessee State University in Johnson City, TN). The corpus will thus serve as a novel and highly useful analytical tool for future educators and their public school students, providing them with a dynamic context in which they can learn about the grammar of a historically important but highly stigmatized American dialect. The opportunity for analysis of a stigmatized dialect such as Appalachian English, in an educational setting — both within and outside the region, promises to broaden a younger generation's understanding of America's linguistic heritage, and to promote an appreciation of the value of regional culture and language.

In terms of the nature of the lesson plans, workshops, and training to be developed for these students at CUNY and at Alice Lloyd and ETSU (in Appalachia), this project will follow the best practices recommended in Reaser (2007; 2010). There are two kinds of lesson plan which come to mind (though we are certain that the ideas for lesson plans and materials for workshops and training will become clearer once the corpus is built). One lesson plan will focus on the concept that dialect grammars are rule-governed. In this regard, our proposed parsed corpus would be the perfect tool to investigate, for example, the syntax of complex floating quantifiers (*We can every one sing*), mentioned at the end of Section 1.1.4, and in Section 1.5. Montgomery & Hall (2004) provide a number of examples (*The Queen family was all of them good to sing; They would nary one of them go*), and examples in the corpus abound. Given that each

of the floated quantifiers would be syntactically annotated as such, students could be guided through a search for these forms, and in the process, would not only learn about the systematicity of the syntax of these floated quantifiers in Appalachian, but also about the contrasting syntax of quantifiers in mainstream English, as well as about the structure of sentences more generally. Another lesson plan could focus on micro-comparative morpho-syntactic variation in verb forms. Here, a lesson on the comparison of the use of *ain't* in Staten Island English vs. Appalachian English would be very useful. Students would learn that in the former variety, *ain't* only corresponds to present tense negative *be*, but in the latter, it also corresponds to present tense negative *have*. This could further be compared with African American English, where *ain't* is *also* used for the present and past tense of negative *do*. Here, students would learn about tense, implicational relations, and again, the rule-governed behavior of dialects.

In addition to the above, the research team will also pursue various ways of making the linguistics community aware of this project's final product, by presenting at various conferences on language variation and change, and submitting a project report to the journal *Language Variation and Change*. Talks will be of two types: those that will provide information about the corpus and its uses, and those that will discuss the results of our research, which we will conduct based on data gleaned from the corpus.

#### 4. Results from Prior NSF Support

**4.1 PI Tortora.** In 2006, Tortora received NSF award #BCS-0617197 (\$29,898), for a project entitled *Collaborative Research: The Comparative Morpho-Syntax of Appalachian English*. This project had the College of Staten Island as the "Lead Institution," and was linked with three other awards, as follows: Bernstein (#BCS-0617210, \$33,180); den Dikken (#BCS-0616573, \$89,652), and Zanuttini (#BCS-0617133, \$55,307). The total amount of the Collaborative Award was thus \$208,037. The initial award was for two-years, from 2006-2008. Bernstein and Zanuttini received one no-cost extension (for 2008-2009), with their award ending in August 2009; den Dikken and Tortora received two no-cost extensions (2008-2009; 2009-2010), with their awards ending in 2010. See references for Project Description.

During the first two years of the grant period (2006-2008), all members of the research team focussed exclusively on subject-verb agreement in Appalachian English, with an eye towards testing hypotheses outlined in our Project Description. In the first year we ran various sets of questionnaires. Our first questionnaires were run on a total of eight informants, four in each of two fieldwork locations (Dante in Southwestern Virginia, and Mountain City in Northeastern Tennessee). These questionnaires concentrated on differences in agreement behavior between the various types of subjects, and in different sentence-types, with the goal of establishing whether subject agreement in Appalachian is influenced by the position of the finite verb with respect to the subject. Our second set of questionnaires was run on a subset of the original informants; these questionnaires revisited some of the test items from the first questionnaire, and also included several new test items developed for the first phase of the project. In the second year, we continued to focus on subject-verb agreement, and interviewed a greater number of participants in more communities. We administered the first year's questionnaire to a new group of eight participants (Rogersville, TN); we also continued to administer to all participants a second token set of sentences to ensure greater reliability of findings. A second set of questionnaires was developed, and focused on other phenomena relevant to the issue of subject-verb agreement, namely: (a) whether the agreement form of the verb influences the scope of the plural subject with respect to sentential negation scope; (b) whether focal stress of a subject pronoun influences the form of the accompanying verb; (c) whether (and if so, how) the distribution of subject-contact relatives is affected by the agreement form of the verb. In order to test the sentences involving scope, we developed a series of picture-prompts illustrating the various possible sentence interpretations. These were presented to participants as they listened to the sentences being read by regional consultants. In this second year, the research team also organized a special Symposium on *Phi-feature Inflection* at the Annual Meeting of the *Linguistic Society of America* in Chicago, January 2008. The Symposium brought together four specialists on phi-feature inflection, each covering a different range of expertise (morphology, syntax, functionalist approaches, sentence processing and production), and was moderated by the PIs. We are currently co-editing a special volume of *Natural Language and Linguistic Theory* which features papers from this Symposium, as well as response papers.

The findings of the funded research activities in these first two years were mixed. On the one hand, certain patterns discussed in the literature seemed to be robustly confirmed by the data; specifically, we found that the form of the subject does indeed influence the distribution of agreement and “singular concord.” But the results of the first questionnaire shed no immediate light on the question of whether the differential behavior of subject types with respect to agreement is a consequence of (a) their structural complexity, (b) their prosodic strength or weakness, (c) their (in)explicit nominativity, or (d) their closeness to the verb, all hypotheses we originally set out to test. The results of the second questionnaire study suggested rejecting the hypothesis that the differential behavior of subject types is due to factors (b) and (c). Unfortunately, we found that participants were generally unable to perform the picture-prompt task; we therefore achieved no concrete results to report on whether subject-verb agreement influences scope. Furthermore, it was not clear whether participants were unable to perform the task by virtue of its specific nature, or whether the problem was related to a more general one that the speakers had, in understanding the nature of a judgment task. The team was concerned in this regard that in general, the results of judgment elicitation were not consistently reliable (with many speakers contradicting judgments given previously, on the same sentences). This called into question, for example, the validity of judgments given on data provided by only one speaker (such as our only speaker who seemed to admit *we 'uns*).

PI Tortora thus advocated the pursuit of a different methodology, one which involved spending much more time with the informants than we had previously done. She travelled to the region herself, and questioned whether some of the informants were representative of the linguistic features we were interested in. Thus, towards the end of the second year, and into the first no-cost extension year, she made several trips alone, seeking alternative informants in other counties. One result of this was the development of personal relationships with speakers in Hancock County, TN (on the north side of Clinch Mountain), who were much older and much more traditional than those we were originally dealing with in Hawkins County, TN (on the south side of Clinch Mountain). She invested a lot of time visiting with these speakers and recording them in 2008-2009, and in 2009-2010 (the second no-cost extension), she explored more actively the possibility of using corpora to crack the tough nut of the clustering question. In the last year of the grant, she spent most of her time working with Michael Montgomery and Beatrice Santorini.

Despite its problems, this previous NSF project had many positive outcomes and benefits, such as training four different graduate students at two different institutions in research methods, and developing more than 70 recorded audio files, which will be used as part of the present proposed parsed corpus. In addition, the team produced a number of publications discussing research results, including: den Dikken et al. (2007), den Dikken & Tortora (2010), den Dikken et al. (to appear), Bernstein & Zanuttini (to appear), Bernstein & Zanuttini (under review (a/b)), Tortora (under review), and Benincà & Tortora (to appear) (see References Cited). We also gave numerous talks, disseminating information about our work (including Tortora 2006a/b/c; 2008a/b; Tortora 2011). In 2007, Tortora taught a Graduate Seminar at the CUNY Graduate Center on English dialect syntax, and she continues to use data from this project in her undergraduate linguistics classes at the College of Staten Island. Furthermore, many of our participants truly became interested in what we were doing, and understood that their participation was crucial. Over time, these native speaker participants have become our most important advocates; in the past 6 months Tortora made two trips back to Hancock County to visit with the participants she developed a relationship with, and it is clear that she has become trusted and welcome by the local community, and that there is a “buzz” about the idea that their variety of English is important.

**4.2 PI Santorini.** In 2003, Santorini received NSF award #BCS-0317826 (\$299,853), for a project entitled *Querying Linguistic Databases* (with Mark Liberman, Steven Bird, Susan Davidson and Michael Maxwell); project period: August 1, 2003 – January 31, 2008. Her contribution to the grant focused on real-world syntactic queries. The grant has produced a graphical query-by-example tool based on LPath (<http://projects ldc.upenn.edu/QLDB/lpath-plus.pdf>), which is available with the *Natural Language Toolkit* (<http://www.nltk.org/Home>), and which continues to be improved. It resulted in 13 papers ([nsf.gov/awardsearch/showAward.do?AwardNumber=0317826](http://nsf.gov/awardsearch/showAward.do?AwardNumber=0317826)), and at least 8 presentations (<http://projects ldc.upenn.edu/QLDB/>).