INTRODUCTION

Romance Parsed Corpora

Editors' introduction

Christina Tortora, Beatrice Santorini and Frances Blanchette
The Graduate Center, CUNY | University of Pennsylvania | Penn State Center for Language Science

1. Introductory remarks

The construction of grammatically annotated corpora has been a steadily growing area of research activity in linguistics. The present Special Issue highlights this research area by featuring five synchronic and diachronic studies of variation and change using grammatically annotated corpora of Romance languages. The studies are very different one from the other, in terms of the specific research questions they address, the languages they study, and the type of corpus they are based on; at the same time, taken together, they deepen our understanding of corpus-based linguistic research in a very focused way. Before turning to a detailed presentation of each of the contributions, we first give a basic idea of the nature of a parsed text and how it can be used for linguistic research, for those readers who are unfamiliar with parsed corpora.

Any text can be grammatically annotated, from a historical or contemporary piece of fiction or non-fiction, to a body of newspaper articles, to a corpus of transcribed recorded interviews; the choice of language and text is driven mainly by the corpus creator's research questions. Within the generative tradition, the *Penn Parsed Corpora of Historical English* (PPCHE; Kroch & Taylor 2000; Kroch et al. 2004; Kroch et al. 2016) have served as an influential model for many other parsed corpus projects, in terms of (i) the method of annotation,[1] (ii) their availability to all researchers, and (iii) the ways in which they have been used to investigate theoretical questions related to syntactic variation and change. Further Germanic

[1] The PPCHE and their offspring use a standard data format known as the Penn Treebank format (Marcus et al. 1993), but the substance of the annotation guidelines differs from the Penn Treebank, and also from corpus to corpus, depending on language-specific exigencies and other factors. Parsed corpora conforming to the Penn Treebank format are searchable with CorpusSearch (Randall 2009), a freely available program.

https://doi.org/10.1075/lv.00002.tor
Linguistic Variation 18:1 (2018), pp. 1–22. issn 2211-6834 | e-issn 2211-6842

corpora using roughly the same annotation guidelines include the *Penn Parsed Corpus of Yiddish* (Santorini 1997), the *Icelandic Parsed Historical Corpus* (Wallenberg et al. 2011), and the *Parsed Corpus of Early New High German* (Light 2011). For Romance, notable corpora include *Modéliser le changement: Les voies du français* (MCVF; Martineau 2009), the *Tycho Brahe Parsed Corpus of Historical Portuguese* (TBC; Galves & Faria 2010; Galves et al. 2017), and the *Syntax-oriented Corpus of Portuguese Dialects* (CORDIAL-SIN; Martins 2010). The MCVF and the TBC (which are used in three of the articles in this Special Issue) are historical corpora; the CORDIAL-SIN is a corpus of contemporary Portuguese dialects.

To get a sense of what a parsed text looks like, let us review an example of a sentence token from the *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE; Tortora et al. 2017), a corpus consisting of approximately 1 million words which follows the annotation philosophy of the PPCHE. Consider the non-annotated text in (1a), and the parsed text in (1b):

(1) a. *It wouldn't əv took them long to went straight.*[2]
(AAPCAppE, ID SKCTC_GD_1,.745)

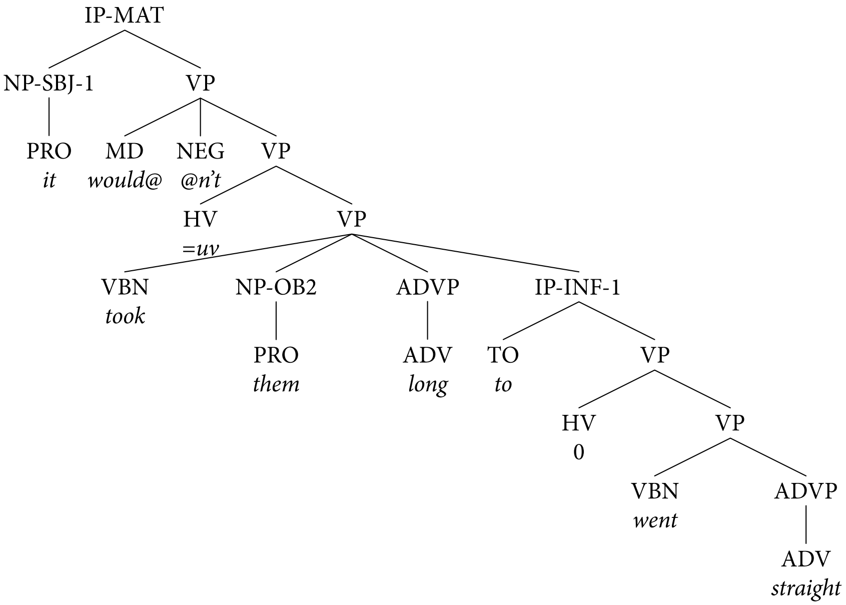
b.

```
(IP-MAT (NP-SBJ-1 (PRO It))
        (VP (MD would@)
            (NEG @n't)
            (VP (HV =uv)
                (VP (VBN took)
                    (NP-OB2 (PRO them))
                    (ADVP (ADV long))
                    (IP-INF-1 (TO to)
                              (VP (HV 0)
                                  (VP (VBN went)
                                  (ADVP (ADV straight)))))))))
```

To illustrate graphically, a tree version of this structure is provided in (1c):

[2] In this interview, the speaker discusses train travel along a local line. The context for (1) suggests that the speaker is communicating the idea that if a certain train had just gone directly to its destination without making all the local stops, the trip would not have taken so long. The "ID" SKCTC_GD_1,.745 uniquely identifies this sentence token within the corpus. The prefix "SKCTC" refers to the particular collection that this interview comes from, namely the *Southeast Kentucky Community and Technical College Appalachian Archives* (see References).

(1) c.



As can be seen in (1b, c), the annotation consists of part-of-speech (POS) tags for individual words, as in (2), and syntactic constituent tags, as in (3):[3]

(2) a. PRO (= pronoun)
b. MD (= modal)
c. NEG (= negative marker *n't* or *not*)
d. HV (= bare infinitive *have*, as opposed to e.g. present tense *have* (HVP))
e. VBN (= past participle, as opposed to e.g. simple past (VBD))
f. ADV (= adverb)
g. TO (= infinitival *to*, as opposed to the preposition *to*)

(3) a. IP-MAT (= matrix sentence)
b. NP-SBJ (= subject noun phrase, here co-indexed with IP-INF)
c. IP-INF (= infinitival clause, here co-indexed with NP-SBJ)
d. VP (= verb phrase)
e. NP-OB2 (= indirect object noun phrase)

It is important to note that grammatically annotated corpora are not intended to provide the user with theoretical analyses. It is true that the structural features of

[3] To use parsed corpora, researchers must familiarize themselves with the system of annotation. Annotation guidelines for the PPCHE and the AAPCAppE (including a catalog of the POS and syntactic tags) are available at http://www.ling.upenn.edu/~beatrice/annotation/index.html and http://www.ling.upenn.edu/~beatrice/annotation-for-audio-aligned-corpora/index.html, respectively.

a particular parsed corpus are driven by the theoretical inclinations and research interests of the corpus creators, and there is no such thing as a theory-free grammatically annotated corpus. A completely atheoretical annotation system cannot be achieved, much in the same way that transcription—be it with a standard system of orthography or with a phonetic alphabet—cannot be anything other than a particular transcriber's theory of the linguistic structure underlying the speech signal being transcribed (see Ochs 1979).[4] Like the act of transcribing a speech signal, grammatical annotation of text by definition involves a theory of language. Nevertheless, it is reasonable for corpus creators to have as their goal an annotation method that is as simple and "theory independent" (or "theory neutral") as possible, while still remaining useful.[5]

In this regard, the annotation scheme applied to (1b,c) reflects its phrase structure grammar origins (and as Stein's contribution to this volume shows, this is not a logical necessity in the annotation of a corpus). At the same time, a close look at the structure reveals its "theory independence," for example in its allowing n-ary branching, contrary to the binary-branching requirement assumed in current generative syntax. Note too in this structure that the IP-INF-1 node (i.e., daughter of the third-highest VP) does not immediately dominate a phonologically null NP subject, as it would in most generative theories. In sum, we can see that the annotation scheme dispenses with certain structural features otherwise widely assumed in contemporary generative models.

As three of the contributions to this volume reveal (Stein, Bullock et al., and Estigarribia & Wilkins), the question of how much (and what kind of) "input analysis" a system of annotation should involve is not always straightforward for corpus creators.[6] Note in this regard that for cases like (1), the creators of the AAPCAppE made a policy decision to include a null infinitival *have* wherever one might expect an overt *have* in such infinitival perfect constructions in mainstream varieties of English (infinitival *to HV 0 went straight* < *to have gone straight*).[7] This is independent of the fact that the right analysis of Appalachian "null *have*" con-

[4] We would like to thank Tyler Kendall for a rewarding discussion on the topic of "transcription as theory," and for directing us to Ochs (1979).

[5] The requirement of simplicity is in part dictated by the need to design user-friendly programs for querying the corpora.

[6] This question was one of the main topics of discussion at the *Workshop on Databases and Corpora in Linguistics*, held at Stony Brook University on October 17, 2014. URL: https://linguistics.stonybrook.edu/events/conference/2014/10/17/workshop.databases.and.corpora.linguistics

[7] The main reason for this policy decision was to make the CorpusSearch (Randall 2009) queries for infinitival perfects as short as possible. Representing the absence of *have* using a silent node "0" makes it possible to search for the two variants *have* and 0 with queries con-

structions such as (1) might very well involve a radically missing *have*. Furthermore, the AAPCAppE contains other structures without HV 0 (and with the main verb easily analyzable as an infinitive), which the researcher might take to be just as relevant to the theoretical analysis as structures like that in (1) are. Consider the sentence in (4):

(4) *And we'əz supposed* ***to put*** *eighty cars in a lead track over there at Elkhorn.* (DOHP_EW_2,.652)

The string *to put* in (4) looks at first glance like an innocent, standard-issue *to*-infinitival. And indeed, this is how the corpus creators annotated this sentence token, as can be seen in the bolded text in (5):

(5)
```
(IP-MAT (CONJ and)
        (NP-SBJ (PRO we))
        (VP (BED =uz)
             (VP (VAN supposed)
            (IP-INF (TO to)
                    (VP (VB put)
                        (NP-OB1 (NUMP (NUM eighty))
                                (NS cars))
                                         …
```

But as the reader may have realized, there are two issues which call into question the theoretical analysis of *to put* in (4) as a standard-issue infinitival: (i) the morphological form *put* is ambiguous between a bare infinitive and a past participle, and (ii) we can see from numerous examples in the AAPCAppE (such as that in (1)) that structures which are otherwise analyzable as infinitival perfects in mainstream American English (*to have gone*) often exhibit a null *have*.

This raises the question of whether the AAPCAppE creators should have annotated (4) not as in (5), but rather along the lines of (1), as illustrated in (6):

sisting of a single clause. In particular, the queries in (i.a) and (i.b) yield the variants with and without *have*, respectively.

(i) a. (HV iDoms !0)
b. (HV iDoms 0)

Omitting null *have* from the representation entirely would require queries with two clauses instead of one, as illustrated in (ii), which again retrieve the variants with and without *have*, respectively.

(ii) a. (VP iDoms HV) AND (VP iDoms VBN)
b. (VP iDoms !HV) AND (VP iDoms VBN)

(6)

```
(IP-MAT (CONJ and)
        (NP-SBJ (PRO we))
        (VP (BED =uz)
            (VP (VAN supposed)
            (IP-INF (TO to)
                    (VP (HV 0)
                        (VP (VBN put)
                                     …
```

The point, however, is this: the parse is not intended to be a correct representation of the structure of the sentence; rather, it is intended as a maximally convenient tool for retrieval. Therefore, for any case where the main verb is ambiguous between an infinitive and a past participle, the corpus user should consistently expect the parse in (5) as the default. More generally, it is up to the user to form a clear idea of which parses in the corpus will be relevant to their study, and to create their search queries accordingly.

The example above is but one illustration of the various kinds of questions which arise in this research area: the corpus creator must make policy decisions on parses which by definition exclude other logically possible parses, and the corpus user must be aware of the decisions made with respect to the constructions under study, in order to utilize the corpus in a maximally profitable way. But regardless of this fact, there is no doubt that the benefits of annotated corpora far outweigh the unavoidable difficulties introduced for the creators and for the users. As can be seen in (1b) and (5), grammatical annotation gives the text structure, and as we will now see with the contributions to this volume, such structure makes it possible for researchers to abstract away from particular word, phrase, or sentence tokens, to identify general grammatical properties and patterns in written or spoken language, and to find and analyze syntactic structures of any type.

Our interest in these matters is what led us to focus our attention on corpus-based research, as organizers of the *43rd Linguistic Symposium on Romance Languages* (LSRL43). The five papers collected for this volume are peer-reviewed developments of those LSRL43 presentations which collectively treat what we feel are the most urgent and interesting problems and questions in this research area. The volume addresses two principal issues, namely, what it means to build a grammatically annotated corpus, and how such corpora allow researchers to fruitfully address long-standing theoretical questions related to variation and change through the use of structural frequency data which is made available only with this kind of tool. The organization we chose for the papers reflects our own sense of the most compelling way to proceed through the volume.

2. Contributions to this volume

The volume begins with a study by **Monique Dufresne, Mireille Tremblay** and **Rose-Marie Dechaine**, who make novel use of texts from *Modéliser le changement: Les voies du français* corpus (MCVF; Martineau 2009), in order to tackle a number of interconnected empirical and theoretical questions concerning the rise of overt determiners (overt D) in Old French (OF).

Modern French argument noun phrases (DPs) require an overt D, but historically this was not the case: OF (12th–13th century) exhibited overt D with arguments only variably. While this change in French grammar—from OF "bare N arguments" (= covert D) to Modern French obligatory overt D—has been treated previously in the literature, the present study offers a quantitative analysis of OF data from the MCVF which allows us to evaluate two points clearly and in detail: exactly which conditioning factors licensed or inhibited bare N arguments, and how these conditioning factors interacted with one another, as use of overt D increased over time. The study further shows that it is possible to make sense of the changing impact of the different conditioning factors through use of a theoretical model of two different possible feature organizations in the D paradigm.

For DP arguments, Dufresne et al. hypothesize five different factors predicting the likelihood of covert D in languages that have it, namely: (i) grammatical function (object vs. subject); (ii) position with respect to verb (post- or preverbal); (iii) mass/abstract noun vs. count noun; (iv) definiteness (indefinite vs. definite); and (v) number (plural vs. singular).[8] Dufresne et al. explore the interaction of these conditions in the rise of overt D in French, offering a fine-grained analysis of the relative rankings of these constraints over time, and examining the exact "tipping point" in the history of French which led to the eventual take-over of overt D with arguments. Let us review this more closely.

The authors access two Anglo-Norman texts from the MCVF, with an eye towards ensuring homogeneity in form across texts (both are prose), and also zeroing in on two different time periods that are no more than two generations apart: the estimated date of composition is 1106–1121 for the first text, and 1154–1189 for the second. Using stepwise regression analysis on the above five factors, they are able to measure the tendency of argument bare Ns to appear in both texts, revealing the relative impact of each factor and also the impact of all of the factors when considered simultaneously. This approach to their data leads

[8] For the purposes of this discussion we put aside the question of predicates, which Dufresne et al. show are in a class by themselves: in contrast with arguments, predicates are robustly bare throughout the history of French, a fact which supports the hypothesis that unlike arguments, predicate noun phrases do not have a DP layer.

to a number of novel findings, a few of which we summarize here. First, they find that grammatical function (i.e., whether the argument is an object or a subject) remains a significant factor in the likelihood of bare N arguments of all types over the time period under consideration: objects in OF were more likely than subjects to appear with covert D, in both texts. Second, taking argument count nouns by themselves, they find that the three factors of (in)definiteness, number (plural/singular), and grammatical function (object vs. subject) also remain constant over time—with (in)definiteness outranking number and grammatical function as the strongest factor. Specifically, indefinite count nouns (and most especially plural indefinite objects) in OF were always more likely than definite nouns to appear with covert D.[9] Nevertheless, there is a decrease in bare indefinite count nouns from the earlier text to the later one, which is reflected in an increased use of *des* as a plural indefinite determiner in the later text. Third (and related to this previous finding in a novel theoretical way, as we will see shortly), taking argument definite count nouns by themselves, the authors find a difference in the strength of the conditioning factors for bare N between the two texts: in particular, the strongest factor in the earlier text is grammatical function, whereas in the later text, it is number.

Why should number emerge over time as the most important factor conditioning overt D with definite count nouns, especially in the absence of any change in the actual inventory of definite determiner morphemes (*li, le, la, les*) in both texts? The authors propose that this change in the importance of the number factor reflects a shift in the grammatical organization of the definite determiner system—specifically, a change in the featural composition (though not the surface forms) of the items involved. They argue that in the grammar represented in the earlier text, the forms are distinguished according to a case contrast, which sets *li* (which is [+NOM]) apart from *le, la*, and *les* [*u* NOM]; however, in the grammar represented in the later text, the forms are distinguished according to a number contrast, which sets *les* [+PL] apart from *le* / *la* [−PL] and *li* [*u* PL] (cf. Dechaine et al. 2014).

The question arises of what change in the input would have prompted learners to analyze definite determiners with this new organization of the featural system. The answer is to do with the second finding discussed above, namely the increased use of *des* as a plural indefinite determiner. Essentially, more frequent use of *-s* as a marker of plurality in the indefinite domain triggered a re-analysis of *-s* in the definite domain, such that it was no longer a marker of a case distinction (i.e., [*u* NOM]), but rather of a plural. Importantly, this is supported not only by the frequency data exhibited by argument indefinite and definite count nouns, but

[9] Already in the earlier text we see almost complete obligatoriness of overt D with definites.

also by another phenomenon in the data, namely, that while grammatical function remains a significant constraint throughout the time period under study (as noted earlier), semantic class comes to outrank grammatical function over time, as the strongest constraint conditioning bare N arguments. In other words, in the earlier text, the argument's status as an object or a subject is the primary factor determining whether it will be bare; in the later text, however, it is the argument's status as mass/abstract (i.e., non-count) that is the primary factor determining whether it will be bare. Specifically, Dufresne et al. find a surprising increase in abstract bare N arguments in the later text. In the absence of any theory of the grammatical reorganization of the D paradigm, this shift might seem counterintuitive, especially considering the larger picture of the history of French (through the medieval and classical periods, and into modern times), where overt D eventually takes over altogether, even with non-count arguments. The authors argue, however, that the observed phenomenon of change between the two texts is actually predicted under their model of the grammatical reorganization of the D paradigm, where number becomes the defining feature for overt D. Under the view that non-count nouns resist individuation, it follows that they would resist an overt D that is marked for number (as in the later text)—which imposes individuation (hence the increase in non-count Ns that are bare). This is in contrast with the earlier text, where overt D is hypothesized to be marked for case (and not number), and therefore does not impose individuation on non-count arguments.

In sum, Dufresne et al.'s contribution serves as a compelling illustration of the ways in which the frequency data from a parsed corpus can be exploited to further our understanding of syntactic variation and change with very specific phenomena, such as the variable presence of determiners with predicates and arguments. The two texts used for this study reflect 43,860 words from the MCVF corpus (10,829 words from the earlier text, and 33,031 words from the later text); this is not the entire corpus, but it is enough to ensure reliable quantitative analysis. In addition, this study represents a model lesson in how to treat such corpus data. It is important to note in this regard that the hits returned by the authors' CorpusSearch queries (namely, nominative subjects and accusative objects with and without overt D) had to be further hand-vetted, given the fact that the corpus itself does not provide a fine-grained theoretical analysis capturing all possible semantic nuances, as discussed in Section 1 above. For example, the annotation scheme does not differentiate between mass, abstract, and count nouns, which thus had to also be hand coded by the authors. Likewise, only a human researcher can evaluate on a case by case basis which of the bare Ns are interpretable as either definite or as indefinite. Dufresne et al. spell out the structure of their CorpusSearch queries and their criteria for all of their hand-coding choices, and therefore provide an exemplary model for how to treat parsed corpus data in a way that allows for replicability.

The contribution by **Charlotte Galves** is similar in nature to Dufresne et al.'s paper. Galves' study makes use of the *Tycho Brahe Corpus of Historical Portuguese* (TBC; Galves & Faria 2010),[10] which, like the MCVF corpus, uses the Penn Treebank format. Instead of looking at variation and change in the domain of the noun phrase, however, Galves examines various word order facts and changes within the sentence (CP), providing novel insights which tie together an array of otherwise seemingly independent phenomena, attributing them to a single parametric change.

Galves' principal concern in this work is to make sense of a striking word order change in Portuguese syntax at the turn of the 18th century (i.e., the end of the Classical Portuguese (ClP) period). Specifically, right after 1700, the TBC data exhibit a notable increase in the frequency of pre-verbal subjects, with a concomitant drop in frequency of post-verbal subjects; this contrasts with what is observed during the ClP period, which exhibits significantly greater frequency of post-verbal subjects, including VOS and VSO orders (in the case of transitive verbs). To account for this change in word order, Galves proposes that ClP was a verb second language (V2). The idea is straightforward: if the verb moves to C in ClP, then we expect a more frequent occurrence of post-verbal subjects, as the pre-verbal position inside CP is not one that is specifically dedicated to subjects. Likewise, loss of T-to-C movement into the 18th century would entail an increase in the frequency of pre-verbal subjects. In other words, loss of V2 is responsible for loss of post-verbal subjects. Galves convincingly shows that all of the quantitative and qualitative facts under discussion can be tied back to this single hypothesis.

As Galves notes, under the view that V2 must strictly be understood literally in terms of "verb second," there is controversy in the literature as to whether Classical Portuguese (ClP) can rightly be classified as a V2 language, with previous authors citing examples of V1 and V3 in ClP. However, Galves offers a unique mix of theoretical insight and empirical fact to support the claim that ClP was indeed V2. First, during the ClP period, V2 is the most frequent order, accounting for 60% of the cases. Second, during the same period, V3 is a comparatively infrequent word order, accounting only for 10% of the cases; furthermore, V3 arises only when the verb is preceded by two adjuncts or by an adjunct plus the subject (demonstrating that it is not a highly productive configuration). Third, while V1 accounts for the remaining 30% of the cases (still less frequent than V2), it is important to note that many cases of V1 are due to null subjects.

[10] Since the time the data for Galves was collected, the TBC has been extended and revised; the citation for the current version of the corpus is Galves et al. (2017).

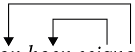
This is not to say that Galves dismisses the reality of V1 and V3; instead, she takes these word order possibilities in ClP as an opportunity to clarify an important point about "verb second," which she capitalizes on in order to explain some differences between Romance V2 and Germanic V2. Specifically, she notes that the phenomenon of verb second can be broken down into two component parts: (i) T-to-C movement, and (ii) movement of a constituent to the left of the verb in C. If we take (i) to be a defining characteristic of "V2," then ClP is by definition a V2 language. The question is why (ii) is apparently not at play in ClP to the extent that it is in German. Here, she considers a hypothesis put forth by Light (2012), namely that there are two kinds of XP-movement to the left of C in V2 languages: "formal movement" and "True A-bar Movement." While German exhibits both types (as argued by Light), Galves argues that ClP only exhibits "True A-bar Movement," which is a movement driven by discourse structural considerations. The fact that ClP contrasts with German with respect to the frequency of subjects in pre-verbal position is supported by this hypothesis. Specifically, in ClP only 20–40% of the pre-verbal constituents are subjects (in contrast with the case of German, where a full 70% of the pre-verbal XPs in V2 structures are subjects, as per Lightfoot 1997). It also explains the cases of V1: if there is no discourse-driven reason to move a phrase to the left of the verb, the verb will be the first element in the string.

Continuing on the theme of frequency data, there are several additional quantitative facts which support a V2 analysis of ClP, as Galves demonstrates. For example, ClP embedded clauses have a much higher rate of SVO word order than do matrix clauses (which is expected under the view that T-to-C does not obtain in embedded CP). In addition, in contrast with modern European Portuguese, ClP exhibits numerous cases of subjects appearing to the left of VP-adjoined adverbs (which is expected under the view that V2 entails a post-verbal subject that itself occupies a relatively high position). Likewise, the VOS and VSO word orders in ClP reflect wide focus interpretations (which contrasts with what we find in modern European Portuguese, where such word orders give rise to narrow focus on the subject); this is expected under the view that the post-verbal position of the subject is unmarked, reflecting a V2 grammar. As an aside, it is worth noting in this regard that—as we saw earlier in our discussion of the AAPCAppE and Dufresne et al.'s use of the MCVF—a parsed corpus can provide only so much analysis. Ultimately, it has to be up to the researcher to determine and hand-code those features of the data that involve nuances of interpretation that cannot possibly be part of the corpus annotation. In this particular case, it is clear that the determination of the information status of each sentence required careful attention to the context in which the sentences in question appear, something which is precisely possible with those corpora that make the entire text available to the user, as Galves rightly notes.

Of particular note is this work's rigorous and copious use of frequency data of all different kinds, with structures that taken by themselves might otherwise seem unrelated to one another. Each bit of quantitative analysis tells a story which anecdotal references to single examples alone are not capable of speaking to. Indeed, without the frequency data, the use of single examples to make a particular point can prove dangerous, as it can ascribe undue importance to a particular syntactic configuration which a more finely-grained understanding of its nature and infrequency would actually belie (an important point made in explicit detail in the contribution by Bullock et al., discussed further below). We see this for example with Galves' careful attention to the frequency distribution of V1 and V3 in ClP, accompanied by her qualitative analysis of these data, where we see for example that V3 is not possible with just any two kinds of XP in pre-verbal position. Both the quantitative and qualitative analysis in this regard weaken to a significant degree any study which refers only to single, uncontextualized examples of V1 and V3 as evidence against a V2 analysis of ClP. Also noteworthy is the wide variety of syntactic structures taken from the TBC for the purposes of this study, such as non-subject XPs in preverbal position; pre-verbal subjects; post-verbal subjects; SVO; VSO; VOS; XVOS; XVSO; null subject structures; structures with pre-verbal adverbs versus post-verbal adverbs; enclisis with SV order; proclisis with SV order (a non-exhaustive list). In sum, Galves' study, as she herself notes, clearly illustrates how parsed corpora make it possible in practice to investigate simple hypotheses—here, the idea that several word order changes reflect the loss of a single syntactic change, the loss of V2—based on what on the surface appears to be a veritable welter of variable data.

The first two contributions to this Special Issue involve quantitative studies based on frequency data gleaned from parsed corpora using the Penn Treebank format. Our third contribution by **Achim Stein** instead shifts our attention to the issue of annotation schemes. His detailed comparison of two different parsed corpora of Old French, namely the MCVF (which we saw utilized in the Dufresne et al. study) and the *Syntactic Reference Corpus of Medieval French* (SRCMF; Prévost & Stein 2013), highlights the role that "input analysis" plays in the research techniques that a corpus user must develop in order to access the desired data. Stein discusses the difference in annotation schemes between the MCVF and the SRCMF, noting that "[u]nlike lemmatization and more so than POS annotation, syntactic annotation involves a commitment to a particular theory." As we have already seen, the Penn Treebank format of the MCVF reflects its phrase structure grammar origins; in contrast, the SRCMF was annotated according to a dependency-based grammar model. This non-constituency-based grammar model relies on a hierarchy of functions, each of which creates a connected, directed link from one (governing) word in a sentence to another (dependent) word, capturing a dependency relation for each word. Consider for example the following string (adapted from Figure 1 and Example (3) in Stein):

(7)



il ait son boen seignor ocis
it has his good lord killed
'it killed his good Lord.'

The dependency grammar model on which Stein bases his SRCMF annotation scheme involves a number of functions. For example, there is a function *ModA* ('attached modifier') which in Example (7) creates a directed link from the governing node *seignor* 'lord' to the dependent node *boen* 'good.' This same function (*ModA*) also creates a link from the governing node *seignor* 'lord' to the dependent node *son* 'his.' The nodes *seignor* and the two nodes it governs (*boen* and *soen*) define a structure; the catalogue of functions and structures in turn provide the basis of the SRCMF's annotation scheme. Corpus users can avail themselves of TigerSearch (Lezius 2002), a software package designed to facilitate searches in the SRCMF, much as CorpusSearch can be used to create search queries for parsed corpora conforming to the Penn Treebank format, like the MCVF.

Using null subjects and cleft sentences as case studies, Stein explores the ways in which the two different annotation schemes—constituency based and dependency based—dictate the kinds of search strategies needed for accessing syntactic information from the MCVF and the SRCMF. Regarding null subjects, Stein notes that because dependency grammars do not represent null categories, researchers using the SRCMF corpus with an interest in identifying null subjects cannot extract the relevant data simply by searching for a category tagged as such, because by definition the category is not present in the data. At first glance, a TigerSearch search query which would specify those structures containing verbs not governing a subject would seem to do the trick. However, as Stein notes, given the way that the negative operator is interpreted with this particular user interface, such a search query would return non-relevant tokens, with verbs which govern any non-subject. On the other hand, retrieval of null subject structures from the MCVF requires users to be aware of the fact that null subjects (represented as **pro**) are always structurally represented in pre-verbal position, even though overt subjects are frequently found in post-verbal position in historical French, before the 15th century. Thus, if MCVF users wish to study both null and overt subjects, distinct CorpusSearch queries which take into account the different word orders must be constructed accordingly.

Stein discusses similar issues of data retrieval with respect to cleft constructions in the two different corpora. As Stein notes, cleft constructions are a particularly delicate matter, given their variation in both form and informational status throughout the history of French. In order to retrieve the relevant data from the MCVF, Stein shows that structures annotated with CP-REL (i.e., relative clauses) are arguably analyzable as clefts, and as such are relevant to the study of diachronic variation and change with clefts. Corpus users interested in this question should therefore not limit themselves to search queries only targeting CP-CLF (i.e., cleft) structures. This issue is reminiscent of the one we raised in Section 1 above, regarding structures like that in (4): although *to put* is represented structurally as an infinitive in (5), the corpus user must not take this annotation to be a theory of the structure, most especially in light of the existence of infinitival perfect structures like that in (1), with a null *have*. Researchers must be aware of these issues, create their search queries accordingly (remaining aware of the annotation conventions of each corpus), and subsequently hand-code data according to semantic interpretation and information status—something that both the Dufresne et. al and Galves contributions showed was necessary for their own studies. In this regard, Stein's contribution provides us with a very precise series of illustrations of how users must be aware of the conventions and limits of corpus annotation schemes, in relation to our theoretical motivations for using these corpora.

The issue of both "input analysis" (as discussed in Stein's contribution) and quantitative analysis (as discussed in Dufresne et al.'s and Galves's contributions) come together in one package, in our fourth contribution by **Barbara Bullock, Jacqueline Serigos, Almeida Jacqueline Toribio & Arthur Wendorf**. In contrast with the previous three studies, however—which are all based on written corpora of historical text—Bullock et al.'s study is based on the *Spanish in Texas* corpus (SpinTX; Bullock & Toribio 2013), a contemporary corpus of oral bilingual text. The underlying speech signal on which the corpus text is based, namely a language-mixed vernacular, gives users the unique opportunity to study aspects of natural language in contact situations which are not detected in more formal or standardized versions of linguistic behavior. As with the MCVF, the TBC, and the SRCMF, the publicly available SpinTx was created to make possible replicable, quantitative research in language variation and change.

Bullock et al. provide several illustrations of the usefulness of such a corpus in resolving long-standing theoretical questions related to language contact and code-switching. As they note, one of the most pressing theoretical issues in this research area is the question of the source of particular linguistic features observed in contact varieties. Consider for example use of the form *para atrás* (or *patrás*) 'back' in the Spanish of Spanish-English bilinguals, in phrases such as *estoy esperando que* ***comience para atrás*** 'I'm waiting for it [i.e., school] to **start**

back'. While some researchers have claimed that this particular use of *para atrás* by Spanish-English bilinguals derives from contact with English, others have argued against this claim, noting examples which show that this use of *para atrás* has always been a Spanish-internal feature. Thus, the question arises as to how one can show whether a particular feature observed in a contact variety (a) was brought about as the result of language contact, or (b) has always been present, independent of contact. As Bullock et al. rightly note in this regard, citing single examples from contemporary or historical non-contact Spanish is not enough to decide the question, as the use of anecdotal evidence to make claims about contact features runs the risk of placing disproportionate importance on structures which may actually be highly infrequent, and therefore insignificant. Instead, the kinds of competing claims under examination here cry out for adjudication through use of frequency data available only with large, annotated corpora. As way of illustration, they compare the frequency *para atrás* in the SpinTX with frequency data from the *Corpus del Español* (Davies 2002), showing that *para atrás* is nine times more frequent in the SpinTX corpus. As they argue, a natural interpretation of this finding is that use of *para atrás* by the SpinTX speakers is the result of contact with English. In a much more detailed way, they make a similar point regarding use of the form *agarrar* 'to get', in collocations such as *agarrar ayuda* 'get help' or *agarrar crédito* 'get credit'. While there is evidence that *agarrar* is used in non-contact Spanish with the same meaning that we see in the SpinTX, the question arises as to whether its use in either Spanish variety has the same source. Again, we see that corpus data play an important role in identifying significant qualitative and quantitative differences between non-contact and contact varieties. In quantitative terms, *agarrar* in its English-like use occurs significantly more frequently in the contact variety, than in the non-contact variety. Furthermore, qualitatively speaking, a comparison of the contact data from SpinTX and the non-contact data from *Corpus del Español* reveals important differences in the kinds of complements that this verb can take; in particular, the contact variety exhibits a significant use of abstract noun phrase complements, in contrast with the non-contact variety, which exhibits use of *agarrar* only with concrete noun phrase complements.

In addition to offering a number of illustrations of the ways in which a corpus like the SpinTX can be used to address theoretical questions important to the field of contact linguistics, Bullock et al. also discuss the challenges they faced in the creation of an annotated corpus of language-mixed text. As we will see again below in our discussion of Estigarribia & Wilkins' study of Jopara, one of the biggest challenges in annotating language-mixed data is determining language tags, that is, establishing whether a particular word in the corpus is English or Spanish (for the SpinTX). Putting aside the question of how to automate such a tagging procedure (which the authors also address in detail), the tagging of a particular word as Spanish or English in a code-switching context requires the corpus creator to make decisions regarding mixed-language words, like for example *pushear* 'to push'. While it is clear that the root of this word is based on English *push*, the morphological form suggests this is a Spanish word. However, tagging this lexical item as either Spanish or English by definition entails a theory of which linguistic elements count as belonging to Language X or Language Y, in language-mixed data. If we take use of *pushear* to indicate a fully-integrated borrowing of the word *push* into Spanish, then *pushear* can be rightly tagged as a Spanish word. However, if this form represents a code-switch point, then perhaps we are dealing with an English word. But the question of whether forms like *pushear* represent switch points or fully integrated borrowings is important from a theoretical perspective, and corpus creators aiming to minimize input analysis will prefer to avoid suggesting a pre-determined answer with their annotation scheme. Bullock et al. provide very useful discussion on difficult questions like this one, which serves at least two purposes: (a) to make corpus users aware of the problems with certain tagging decisions, and the extent to which the researcher must hand vet examples relevant to their own research questions (an issue we have seen arise numerous times already, throughout this introduction), and (b) to underscore the importance of frequency data. In this latter regard, corpus users can in fact take advantage of such data in order to determine whether a particular form should be analyzed as a single word switch (i.e., a nonce borrowing), or whether it should be considered to be a fully integrated lexical item in the language.

Bruno Estigarribia & Zachary Wilkins' (E&W) study has elements which make it both similar to and different from the other studies, adding uniquely to the profile of corpus approaches to Romance. Like the Bullock et al. study (and unlike the other three contributions to this volume), E&W investigate code switching in a contemporary vernacular language, in this case *Jopara*, a "mixed language" spoken in Paraguay. Unlike the SpinTX corpus, however, the corpus used by E&W is based on a mix of two typologically very distinct languages (Guarani and Spanish), and is not based on transcribed vernacular speech. Instead, like the studies by Dufresne at al., Galves, and Stein, it is based on a written text, in this case Ayala de Michelagnoli's (1989) novel *Ramona Quebranto*.

Strictly speaking, E&W's corpus is not grammatically annotated, but the way in which they utilize Muysken's (1997, 2000, 2013) taxonomy to identify three different kinds of code switching in Jopara opens a very promising door for a more general application of such coding, in language-mixed texts. Their method of manual coding involved (i) identifying a set of clauses as the object of study, (ii) coding each for either *non-mixed* or *mixed* (i.e., exclusively Guarani or exclusively Spanish, versus mixed Guarani-Spanish), and (iii) identifying switch points in the mixed text. The method in (ii) allows E&W to quantitatively establish a distribution of word and morpheme tokens, which reveals Spanish as the "primary lexical contributor," and therefore the "base language of the novel." This conclusion, based on their own objective, quantitative methods, is corroborated by independent claims that the novel is *Castení* (i.e., a Jopara that is more Spanish than Guarani), as opposed to *Guarañol* (i.e., a Jopara that is more Guarani than Spanish). Thus, while E&W's corpus does not follow any established set of annotation guidelines shared by other corpus creators, their use of a taxonomic scheme allows for more objective, quantitative analysis of the text.

In coding the text from *Ramona Quebranto*, we see that E&W face challenges similar to those discussed in Bullock et al. with respect to the oral SpinTX corpus. As we saw for Bullock et al.'s mixed Spanish-English, the decision on the part of the corpus creator to tag particular lexical items (such as *pushear* 'to push') as Spanish, or as a "borrowing" from English, or as "mixed language" is not without problems, as the question of an item's linguistic nature in this regard is a theoretical one; therefore, any decision on the part of the corpus creator in such cases runs the risk of over-stepping the bounds of theory-neutrality. In E&W's study of Guarani-Spanish mixing, the question of how much input analysis counts as "too much" rears its head again: in their case, we see (for example) that the question of whether a particular lexical item seemingly from Language X should be coded as (a) a fully-integrated borrowing from Language X into Language Y, or (b) as a code-switch point, does not have a clear answer (though see Section 5 of the Bullock et al. contribution for a discussion of the potential that frequency data has, for addressing this question). Furthermore, as E&W argue, it is not even obvious whether it is legitimate to make such a sharp distinction, as that which might be conceptualized as "borrowing" by some theorists may instead—in a code-switching context—simply reflect one end of the continuum of language mixing. Under this view, consider a single Jopara sentence in which the only word of Spanish origin is *trabaja* 'work': the question of whether *trabaja* represents a borrowing, or whether it represents a switch point, could depend on how dominant a given speaker is in Spanish or Guarani; whether Spanish is the speaker's L2 or an L1 on a par with Guarani; and so on. E&W are correct for raising this point, as it reminds us of our true object of study, which is not the text, but the individual mind: the ultimate concern should be whether identification of *trabaja* as a borrowing (on the one hand) or as a switch point (on the other) accurately captures the psychological reality of the individual speaker.

One of the principal questions E&W ask in their work is whether theoretical "models designed for spontaneous speech apply to contemporary literary texts." If they can show that they do, then we have some evidence that, despite the non-spontaneous nature of the text (in relation to vernacular speech, like what we find in Bullock et al.), such literary texts can be taken as legitimate objects of linguistic study. As noted earlier, the particular theoretical model they apply to *Ramona Quebranto* is Muysken's taxonomy of code switching, focusing on *insertions* versus *backflagging* versus *alternations*. For each of these categories of code switching, E&W review examples taken from the novel, with an eye towards illustrating that the vernacular constructed by the author exhibits all of the properties of authentic code switching. In addition to arguing for the text as a legitimate object of linguistic study, E&W also argue that Jopara adds to the empirical base for code-switching studies, in particular calling for more finely-grained characterizations of different kinds of code switching. For example, they introduce the concept of *backflagging by bound morphology*. For Muysken, backflagging is a type of code switch whereby a sentence which is otherwise purely in Language A contains a single discourse marker from Language B, which carries with it "a clear ethnic connotation." E&W argue, however, that the agglutinative and polysynthetic nature of Guarani requires extending the theoretical notion of backflagging to account for the behavior of Guarani bound morphemes in Jopara. Thus, in addition to showing that a code-switching model based on spontaneous, oral linguistic behavior can be applied to a written text, E&W show that their corpus of written text can in turn inform models of code-switching.

3. Final remarks

Regarding the concept of "authenticity" for written texts: E&W note that the various properties of written sources (such as careful planning and editing, and increased prescriptive pressure) does not necessarily entail that such sources are "unavailable for linguistic analysis," especially when they are rich in conversational passages, as is the case with *Ramona Quebranto*. The authors further note that planning, editing, and increased prescriptive pressure are properties that historical linguists must also work with using historical texts. And as we have seen in connection with the contributions by Dufresne et al., Galves, and Stein, these properties "do not automatically make such productions unavailable for linguistic analysis."

On the other hand, Kroch (1994: fn. 6) raises the question of whether linguists should assume that certain phenomena uncovered in the study of historical texts reflect genuine linguistic processes, without seeking ways to independently corroborate their findings. Specifically, his investigation of syntactic variation and grammar competition in historical texts—which is similar to the kind of change-in-progress studies offered by Dufresne et al. and Galves in this volume—leads him to question whether the observed effects "reflect stylistic options limited to the written language, with its known peculiarities and tendencies to linguistic unnaturalness." He considers the distinct possibility that historical texts may reflect "competition between the grammar of a spoken language of a given time and an archaic but still influential literary standard," and as he notes, if this turns out to be true, then the grammar competition under investigation "will have no purely linguistic significance." He takes the position that "[o]nly work on possible cases of competition in living languages can determine whether it exists in unreflecting vernacular speech." The implication is that if corpus-based studies on unreflecting vernacular speech can uncover the same kinds of patterns of variation and processes of change as those which are found in studies on historical corpora, then we can become more confident that written texts are equally legitimate objects of linguistic inquiry, i.e., that they have "linguistic significance." We believe that the unique mix of contributions offered in this volume take a positive step in that direction.

Acknowledgments

The idea for a volume dedicated to Romance Parsed Corpora arose out of a Special Session on the topic, held at the 43rd *Linguistic Symposium on Romance Languages* (LSRL43) on April 18, 2013. Given this volume's origins, we would like to thank Anthony Kroch and Ana Maria Martins, two invited speakers at that Special Session who are not represented in this volume, but whose participation contributed greatly to its spirit. We would also like to thank Gita Martohardjono (Linguistics Program Executive Officer of the CUNY Graduate Center), for her support in the organization of the conference, and William Fritz, President (and Geologist) of the College of Staten Island (CUNY), for his opening remarks; his discussion of the similarities between geology and linguistics introduced a sense of warmth and collegiality that had a lasting positive effect throughout the conference. We likewise thank LSRL43 co-organizers Marcel den Dikken, Ignacio Montoya, Teresa O'Neill, and Emily Wilson, for their generosity of spirit and for ensuring that the conference was a success. In addition, we thank Lori Repetti, for giving us an opportunity to discuss some of the ideas presented here at Stony Brook University's *Workshop on Databases and Corpora in Linguistics* (October 17, 2014), and likewise, we thank Sjef Barbiers and Hans Bennis for a similar opportunity at the *Dealing with bad data in linguistic theory* workshop at the Meertens Institute, Amsterdam (March 18, 2016). And finally, we owe a debt of gratitude to the ten anonymous reviewers of the papers in this volume for their careful work, to Anke de Looper for her assistance in the early stages of development of this volume, and most importantly, to Jeroen van Craenenbroeck, for all of his help, patience, and goodwill. LSRL43's Special Session on Romance Parsed Corpora, and the work leading to the creation of this Special Issue, was supported by National Science Foundation awards BCS-1256700, BCS-1151630, and BCS-1152148; a National Endowment for the Humanities Fellowship (FB-55760-11); and a National Endowment for the Humanities Digital Humanities Start-Up Grant (HD-51543).

References

Ayala de Michelagnoli, Margot. 1989. *Ramona Quebranto*. Asunción, Paraguay: Editorial Arandurã.

Bullock, Barbara E. & A. Jacqueline Toribio. 2013. *The Spanish in Texas Corpus project*. Center for Open Education Resources and Language Learning (COERLL), the University of Texas at Austin. http://www.spanishintexas.org.

Davies, Mark. 2002. Corpus del Español: 100 million words, 1200s–1900s. www.corpusdelespanol.org

Déchaine, Rose-Marie, Raphaël Girard, Calisto Mudzingwa & Martina Wiltschko. 2014. The internal syntax of Shona class prefixes. *Language Sciences* 43. 18–46. https://doi.org/10.1016/j.langsci.2013.10.008

DOHP: *Dante Oral History Project*. Archives of Appalachia, East Tennessee State University, Johnson City, TN.

Galves, Charlotte & Pablo Faria. 2010. *Tycho Brahe Parsed Corpus of Historical Portuguese*. http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Galves, Charlotte, Aroldo Leal de Andrade & Pablo Faria. 2017. *Tycho Brahe Parsed Corpus of Historical Portuguese*. http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip

Kroch, Anthony. 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, Anthony & 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, Anthony, Beatrice Santorini & 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, Anthony, Beatrice Santorini & Ariel Diertani. 2016. *The Penn Parsed Corpus of Modern British English (PPCMBE2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1

Lezius, Wolfgang. 2002. Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German) University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 8, no. 4. Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS). http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html

Light, Caitlin. 2011. *Parsed Corpus of Early New High German*. http://enhgcorpus.wikispaces.com

Light, Caitlin. 2012. *The syntax and pragmatics of fronting in Germanic*. Philadelphia, PA: University of Pennsylvania dissertation.

Lightfoot, David. 1997. Catastrophic change and learning theory. *Lingua* 100. 171–192 https://doi.org/10.1016/S0024-3841(93)00030-C

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.2. 313–330. Reprinted in Susan Armstrong (ed.), 1994, *Using large corpora*, 273–290. Cambridge, MA: MIT Press.

Martineau, France (ed.). 2009. *Le corpus MCVF. Modéliser le changement: Les voies du français*. Ottawa: Université d'Ottawa. http://www.voies.uottawa.ca/corpus_pg_fr.html

Martins, Ana Maria (ed.). [2000-] 2010. *CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects*. Lisboa, Centro de Linguística da Universidade de Lisboa. http://www.clul.ulisboa.pt/en/10-research/314-cordial-sin-corpus

Muysken, Pieter. 1997. Code-switching processes: alternation, insertion, congruent lexicalization. In M. Pütz (ed.), *Language choices: Conditions, constraints, and consequences*, 361–380. Amsterdam: John Benjamins. https://doi.org/10.1075/impact.1.25muy

Muysken, Pieter. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge, UK: Cambridge University Press.

Muysken, Pieter. 2013. *Language contact outcomes as the result of bilingual optimization strategies*. Bilingualism: *Language and Cognition* 16(04). 709–730.

Ochs, Elinor. 1979. Transcription as theory. In Elinor Ochs & Bambi B. Schieffelin (eds.), *Developmental pragmatics*, 43–72. New York: Academic Press.

Prévost, Sophie & Achim Stein (eds.). 2013. *Syntactic reference corpus of Medieval French (SRCMF)*. Lyon & Stuttgart: ENS de Lyon; Lattice & Paris; Universität Stuttgart. http://srcmf.org.

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/

Santorini, Beatrice. 1997. *Penn Parsed Corpus of Yiddish*. Department of Linguistics, University of Pennsylvania.

SKCTC: *Southeast Kentucky Community and Technical College Appalachian Archives*. Cumberland, KY.

Tortora, Christina, Beatrice Santorini, Frances Blanchette & C. E. A. Diertani. 2017. *The Audio-Aligned and Parsed Corpus of Appalachian English*. https://csivc.csi.cuny.edu/aapcappe/

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson & Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.9. http://www.linguist.is/icelandic_treebank

Address for correspondence

Christina Tortora
Linguistics Program
The Graduate Center, CUNY
365 Fifth Avenue, Room 7407
New York, NY 10016
USA

ctortora@gc.cuny.edu

Co-author information

Beatrice Santorini
Department of Linguistics
University of Pennsylvania
beatrice@sas.upenn.edu

Frances Blanchette
Penn State Center for Language Science
fkb1@psu.edu