Phonology (2026), **Vol. 43**:e20 1–11
doi:10.1017/S0952675726100451

PHONOLOGY

INTRODUCTION

Corpus phonology as a way of doing phonology

Marie-Hélène Côté [1] and Julien Eychenne [2]

Guest editors, Corpus Phonology collection
[1]Section des sciences du langage et de l'information, Université de Lausanne, Switzerland
[2]Département des arts, langues et littératures, Université de Sherbrooke, Canada

Corresponding author: Julien Eychenne; Email: julien.eychenne@usherbrooke.ca

1. The emergence and institutionalisation of corpus phonology

In their introduction to the *Oxford Handbook of Corpus Phonology,* Durand *et al.* (2014a: 1) write that 'corpus phonology is a new interdisciplinary field of research that has only begun to emerge during the last few years'. Is it true to say that no work done until then could be described as pertaining to corpus phonology? This question warrants an important distinction between corpus phonology as an institutionalised subfield of phonology (and of corpus linguistics) and corpus phonology as a particular way of doing phonology. The institutionalisation of corpus phonology is indeed recent. While the word *corpus* in its modern linguistic sense dates from the 1950s (McCarthy & O'Keeffe 2010: 5), corpus linguistics as a discipline really took off in the 1980s, but at a different pace across different areas within linguistics. Phonetics and phonology were initially largely absent, as illustrated by the content of the *Routledge Handbook of Corpus Linguistics* (O'Keeffe & McCarthy 2010). This reference work includes 31 chapters devoted to the use of corpora in different areas of linguistics, including morphosyntax, sociolinguistics, pragmatics, discourse, lexicology, language teaching and translation; significantly enough, not a single one touches on phonetic or phonological issues. The same applies five and even ten years later (Biber & Reppen 2015; Stefanowitsch 2020), although Zufferey (2020) includes a small section on the use of corpora in phonetics and phonology.

On the other hand, corpus phonology as a practice has a long history, whose main elements are well described in recent overviews of corpus phonology (Durand 2017; Eychenne 2022; Hall 2022; see also Laks 2008; Ernestus & Baayen 2011). More or less distant sources include early 20th-century dialectology, grammatical description and the building of reference corpora, American structuralism (Bloomfield, Harris, Hockett), work in linguistic typology, and variationist sociolinguistics. Eychenne (2022) mentions a lesser-known work in German from the 1930s (Zwirner & Zwirner 1936), whose phonometrics anticipates modern corpus phonology in many ways. In fact, as Laks (2008) argues, corpus-based linguistic work has been the norm throughout the history of linguistics. It is rather the final period of the 20th century, corresponding to the dominance of generative grammar, that appears exceptional in its marginalisation of the corpus as the basis of linguistic investigation. In that sense, the rise of corpus linguistics as a separate branch within the field may be interpreted as a response to the generative approach.

We will not rehearse this history here, but we would like to briefly address a more specific question. Why did phonology jump on the corpus linguistics bandwagon later than other subfields? Or, in other words, what made the time ripe for corpus phonology to emerge in the last 15 years, but not earlier? We see a combination of three factors: technical, methodological and conceptual. First, the early development of corpus linguistics was heavily based on written corpora. But applying the methods of corpus linguistics to speech involved the construction and use of spoken corpora, whose effective harnessing required additional technical tools, in particular alignment between sound and transcription or annotation. It is thus not a coincidence that the rise of corpus phonology coincides with the development and wider availability of such tools.

© The Author(s), 2026. Published by Cambridge University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

But for corpus-based research in phonology to develop, it must not only be technically accessible, but also methodologically and conceptually relevant to the rest of the field. Methodologically, corpus linguistics relies heavily on quantitative approaches. But quantitative methods were not part of phonologists' standard toolkit until the development of laboratory phonology in the 1990s (Pierrehumbert *et al.* 2000; Pierrehumbert 2022). The quantitative turn was also pursued in constraint-based frameworks with the development of numerical versions and offshoots, using weighted constraints and noisy constraint evaluation (Stochastic Optimality Theory, Harmonic Grammar/MaxEnt; Boersma & Hayes 2001; Pater 2009).

Conceptually, the use of corpora and quantitative methods is naturally associated with the study of variation. But phonological variation was not considered a central topic in phonological theory until the 1990s. It was, rather, part of the agenda of variationist sociolinguistics. It is sociolinguistics, for example, that developed the concept of variable rule, which integrated internal and external factors conditioning the application of a given rule (Labov 1969; Cedergren & Sankoff 1974). Guy (1991a,b) later combined the variable rule framework with the formal architecture of Lexical Phonology to account for the variable process of *-t,d* deletion in English. However, this quantitative application of a formal phonological model remained isolated and confined to sociolinguistic circles until the advent of Optimality Theory (OT), which was instrumental in promoting the study of variation within phonological theory. From the very beginning, OT's potential to account for variation became apparent. The concept of partial ordering of constraints, in particular, was developed as an attempt to bridge the gap between phonology and sociolinguistics (Kiparsky 1993; Reynolds 1994; Anttila 1997). As Anttila & Cho (1998: 54) put it: 'What is new is the way Optimality Theory permits one to embed synchronic variation, diachronic change, usage statistics and the familiar categorical regularities in one homogeneous formal model.' The theoretical promotion of variation in phonology is directly reflected in the content of reference works: while Goldsmith (1995) contained no chapter on variation (out of 31), all subsequent handbooks have included one, starting with Anttila (1997) and, notably, Goldsmith *et al*.'s (2011) second edition of Goldsmith (1995).

With the necessary technical and quantitative tools, and with variation now established as an important research topic in phonology, the ingredients for the rise of corpus phonology as a legitimate branch of phonology were in place. More generally, corpus phonology has been part of a broader effort to cross disciplinary boundaries and bring different methods and theoretical frameworks to bear on the analysis of sound systems: 'phonology, both diachronic and synchronic, phonetics, corpus linguistics, speech technology, information technology and computer science, mathematics, and statistics' (Durand *et al.* 2014a: 1).

As we will see, the seven articles in this special issue highlight this blurring of disciplinary boundaries and showcase the diverse methods and approaches used in corpus phonology today, along with the range of theoretical questions they make it possible to explore.

2. Corpora in corpus phonology

The development of corpus phonology as a field raises the fundamental question of what defines a corpus in phonology. Indeed, definitions vary widely. We can start by distinguishing between two general visions of the object 'corpus' in corpus phonology: type-based and purpose-based. The dominant approach is the former: a corpus is defined by specific characteristics, more or less inclusive depending on the author. Gut & Voormann's (2014: 16) oft-cited version lies on the strict side. They define a phonological corpus as:

> a representative sample of language that contains
>
> - primary data in the form of audio or video data;
> - phonological annotations that refer to the raw data by time information (time-alignment); and
> - metadata about the recordings, speakers and corpus as a whole.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

This definition excludes written corpora that could be used to investigate phonological phenomena to the extent that these are reflected in the orthography (e.g., Anttila 1997), as well as corpora of phonetic transcriptions not associated with audio data. The inclusion of time-alignment highlights the importance of prior technical innovations. In addition, Gut & Voormann (2014) impose as a minimal requirement two levels of annotation: an orthographic annotation and at least another level of time-aligned phonological annotation. So a corpus of transcribed and time-aligned speech is not a phonological corpus unless it includes some sort of phonetic or phonological analysis, in the form of transcriptions or annotation of specific phonological phenomena. Note, however, that any speech style may be considered, from the reading of word lists to free conversations. Finally, a phonological corpus must contain metadata about the recordings (date, place and situation) and speakers (age and background). Thus, it appears that a corpus of audio data gathered from open-line radio shows, where the speakers' characteristics are not known, would not qualify as a phonological corpus. One of the earliest corpora that satisfies Gut & Voormann's conditions is the Switchboard-1 Telephone Speech Corpus, which contains about 250 hours of transcribed and time-aligned telephone conversations between about 500 speakers, along with some metadata about speakers and calls (Godfrey *et al.* 1992). Another good example of a corpus that fulfils these requirements is the *Phonologie du français contemporain* database (PFC; Durand *et al.* 2002). PFC contains audio recordings of conversations, transcribed and time-aligned in Praat (Boersma & Weenink 2025) and annotated for two central phenomena of French phonology: schwa and liaison. Each speaker and conversation is associated with a list of biographical and situational metadata.

Eychenne (2022: 323) provides a more inclusive definition of a phonological corpus, described as 'any structured dataset, built according to explicit principles, which contains at least coded data resulting from an interpretative process at the phonological level'. Here, a phonological corpus is defined by the presence of phonological annotations, but neither audio data nor metadata are required, as long as the principles underlying the corpus's constitution are known. Eychenne distinguishes two types of phonological corpus: SPEECH CORPORA, which correspond to Gut & Voormann's definition presented above, and PHONOLOGICAL DATABASES, usually taking the form of large phonetised lexicons containing a variable quantity of other linguistic information, such as frequency or part of speech. For example, Wedel *et al.* (2013) examined phonological databases from nine languages and found that contrasts with a low functional load (i.e., ones that do little work in the system) were more likely to merge. Another example of the relevance of phonological databases can be found in Frisch *et al*.'s (2004) study of the OCP in triconsonantal Arabic roots, based on a phonological lexicon of standard Arabic roots. They conclude that adjacent consonants sharing the same place of articulation are indeed subject to avoidance.

While Eychenne's definition of phonological corpora is more inclusive than Gut & Voormann's, it still excludes some corpora that may be used to investigate phonological issues, notably written corpora or spoken corpora containing only orthographic transcriptions. Indeed, many phonological phenomena are represented in spelling, and it is not clear why a systematic study of an appropriate written corpus, especially in a language where the mapping between graphemes and phonemes is sufficiently transparent, would not be characterised as pertaining to corpus phonology. This is where the second, purpose-based, approach to the corpus appears relevant. According to Hall (2022: 619), 'all empirical phonology could be considered corpus phonology, as it draws from some collection of data, be it from fieldwork, introspection, experimentation, or what might be considered a more typical corpus, i.e., the large, less controlled type'. This is the broadest possible sense of a phonological corpus – indeed, it is so broad that it precludes any conception of corpus phonology as an approach grounded in a specific type of data. In fact, Hall (2022: 620) does not define corpus phonology by the type of empirical data it is based on, but by the initial purpose for which the data was collected; corpus phonology can be conceived as the *post hoc* use of a set of data that was not collected to answer the research question under consideration:

> Corpus-based linguistics is *post hoc* in the following sense: a researcher develops a question, and then attempts to answer it using some pre-existing collection of data that has enough information of the right type to do so, rather than collecting new data specifically to address the question.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

This approach allows us to conclude that the use of written corpora for phonological purposes is indeed corpus phonology, and it resonates with Liberman's (2019) observation that found data is playing an increasing role in linguistic research. Nonetheless, limiting corpus phonology to the use of pre-existing data is probably too restrictive. A conversational corpus may be used to investigate a large variety of phonological phenomena. So a researcher may well collect a conversational corpus with a specific research question in mind and still be doing corpus phonology. We believe that what is relevant here is not the pre-existence of the corpus, but the fact that it was not specifically designed to answer a given research question. For instance, the PFC project mentioned above was, from the outset, interested in the analysis of schwa and liaison, two variable phenomena in French whose complexity has defied all theoretical frameworks. Annotation schemes were thus developed for schwa and liaison and applied to the recorded conversations included in the PFC corpus. But nothing in the recording protocol of these conversations is likely to affect the realisation of these two variables, and so the analysis of schwa and liaison in the PFC conversations may still, in that sense, be considered post hoc. However, the PFC corpus includes the reading of a word list and a text, which were designed to examine specific phenomena, including schwa and liaison. This part of the methodology is thus targeted, yet it would be difficult to argue that the conversations fall under corpus phonology and the reading tasks do not. Together they provide the stylistic diversity that a proper understanding of schwa and liaison requires. It would also be problematic to argue that the same data set may or may not belong to corpus phonology, depending on the extent to which it was designed to address specific phenomena: following this line of reasoning, the PFC texts would constitute a phonological corpus for the study of prosody (Lacheret *et al.* 2011) and mid vowels (Courdès-Murphy 2018), because they were not specifically designed to target these phenomena, whereas they would not for schwa and liaison.

Such considerations underline the challenges, if not the impossibility, of formulating an unambiguous definition of a phonological corpus or corpus phonology. Instead, we prefer to conceive of corpus phonology as one end of a continuum defined by methodological targetedness, with experimental phonology at the other end. We follow Hall (2022) in distinguishing between corpus phonology and the use of phonological corpora. Corpus phonology is not simply phonology based on phonological corpora, however defined, but phonology based on any corpus that was not specifically designed to address the research questions under consideration. While it is certainly useful to distinguish different types of corpora and understand their limits and possibilities, it is no longer necessary to draw a specific list of characteristics that a corpus must have to qualify for corpus phonology. But, crucially, methodological targetedness is not a binary concept, and studies may fall in between corpus and experimental phonology. Thus the study of schwa and liaison would fall closer to the corpus end of the spectrum in conversations than in read texts. Methodological targetedness, however, suggests inherent limits to the type of data that may be used in corpus phonology. For example, it is difficult to imagine that acceptability judgments may be used for purposes other than the one that led to their collection; see Stefanowitsch (2020) for an argument for excluding acceptability judgments from the realm of corpus linguistics (applied to syntax).

Beyond the issue of what defines a phonological corpus, another set of questions concerns the properties of a *good* corpus. Size and representativeness, in terms of speech style or speaker background, are often included in definitions of a corpus for the purposes of corpus linguistics. Stefanowitsch (2020), for instance, defines a corpus as an authentic, representative and large collection of samples of language use. We could also think of other requisites, such as acoustic quality for speech corpora. But what is considered representative, large enough or acoustically good enough depends entirely on the research question and the linguistic phenomena under investigation – their frequency and distribution in discourse, their variability across different speech styles and categories of speakers, and their acoustic and perceptual salience, as well as their recoverability from the speech signal. Properties such as size and representativeness are then better seen as important elements in the evaluation of specific research designs rather than as part of the definition of corpus phonology as such. Barth & Schnell (2021) adopt a similar point of view for corpus linguistics more generally, but they insist on the importance of metadata

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

to situate the content of the corpus. In phonology, see Hall (2022) and Durand (2017) for discussions of corpus size and representativeness, respectively.

All the articles in this collection are *post hoc* analyses in Hall's sense, in that they make use of existing corpora to address phonological questions. All the corpora are also methodologically untargeted, as their design had no bearing on these questions. Three articles are clear examples of studies that rely on speech corpora as defined by Gut & Voormann (2014). Russell (2025) uses a corpus of 26 sociolinguistic interviews with Paraguayan Guaraní speakers, about 1 hour each in length, to investigate whether affixes undergo nasal harmony. Sano & Guillemot's (2025) article on hypercontrastive articulation uses the Corpus of Spontaneous Japanese to investigate whether acoustic cues to the voicing contrast tend to be hyper-articulated when a word has a lexical competitor for that contrast (i.e., the two words form a minimal pair for that contrast). Lamontagne (2025) also uses data from a speech corpus, PFC, but complements it with data from parliamentary debates to investigate high vowel laxing in non-final position in Laurentian French. Two other articles, Andersson (2026) and Dabouis & Fournier (2025), use phonological databases, in Eychenne's (2022) sense. They extract dictionary data containing phonological information to investigate the realisation of stress in Abkhaz and vowel reduction in British English, respectively. These dictionaries both include information about lexical stress and, in the case of English, contain full phonetic transcriptions, which allow the authors to uncover new patterns in the realisation of stress and vowel reduction. Having some form of pre-existing phonological annotation is critical in the cases of Abkhaz and English, since stress is not entirely predictable in these languages. However, there are languages where the mapping between phonology and spelling is sufficiently transparent to allow the study of a phonological phenomenon based on orthography alone, which makes it possible to use (generic) written corpora. As we have argued above, such work is *bona fide* corpus phonology. Mayer (2025) and Kukhto & Piperski (2026) provide two nice illustrations of this approach. Mayer investigates opacity in the realisation of backness harmony in Uyghur. This language is written in three different scripts (Perso-Arabic, Cyrillic and Latin), but the phonological alternation between back and front vowels is reflected in all three of them. This allows Mayer to pool data from three written sources to build a large phonological database of harmonising roots. Kukhto & Piperski investigate stress-conditioned variation in the realisation of synthetic adjectival comparative forms in Russian. The fact that the deletion of the genitive suffix final vowel is reflected in the spelling (e.g., *sil'**née*** ~ *sil'**néj** sloná* 'stronger than an elephant') made it possible for the authors to use a very large written corpus (*Araneum Russicum Maius,* which contains one billion words; Benko 2014) to investigate this relatively rare phenomenon. They also used a smaller speech corpus to cross-validate their findings.

Not only do the articles in this collection exemplify different types of corpora (speech corpora, phonological databases, written corpora), they also address a diverse set of languages. These include Indo-European languages representing the Germanic, Romance and Slavic groups (English, French, Russian), as well as four typologically distinct languages: Japanese (Japonic), Guaraní (Tupian), Abkhaz (Northwest Caucasian) and Uyghur (Turkic). This linguistic diversity shows that the practice of corpus phonology need not be restricted to the better-documented languages, which benefit from more numerous and diversified language resources.

3. Tools and techniques

As we mentioned previously, the availability of appropriate tools has played an important role in the development of modern corpus phonology. Among the first and most widely used of these tools is Praat (Boersma & Weenink 2025), an open-source and cross-platform computer programme which provides features for the annotation of speech data, advanced acoustic analysis and even some features for phonological modelling (Stochastic OT and neural networks). A wide array of tools has since been developed to cater to the needs of specific communities. For example, ELAN (Wittenburg *et al.* 2006) is a widely adopted tool for creating, editing, and searching time-aligned annotations for both audio and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

video data, with support for multi-level and multi-participant annotation. Phon (Hedlund & Rose 2020) was originally developed for the analysis of child phonological acquisition but can be used for a broad range of tasks related to the phonological annotation and analysis of time-aligned data (audio and video). Phonometrica (Eychenne & Courdès-Murphy 2019) is another tool that facilitates working with time-aligned annotations (audio only); it is particularly well-suited for the extraction of information from symbolic annotations, with a focus on ease of use and performance. Finally, Phonological CorpusTools (Hall *et al.* 2019) offers an array of features for the analysis of phenomena such as neighbourhood density and functional load in phonological databases.

In addition to such specialised tools, corpus phonologists have benefited from the availability of more general tools such as the Python programming language, which excels at data manipulation and transformation, and makes it easy to develop novel algorithms. (See Andersson 2026, where several competing models of stress in Abkhaz verbs are compared using an algorithm developed in Python.) In the same vein, the influence of the R environment, which has become a *de facto* standard for data analysis and visualisation, cannot be overstated, since it has opened up access to the latest statistical methods. For example, regression modelling is a powerful framework for examining the relationship between multiple predictors and a response variable, and it is now a widely used tool in corpus phonology. Building on this, mixed-effects models (also known as hierarchical or multilevel models; see Gelman & Hill 2007) have been especially influential, as they allow researchers to account for structured variation in the data – for example, by modelling random effects associated with individual subjects or specific words. Broadly speaking, two major paradigms underpin statistical inference in such models: the frequentist and Bayesian approaches. These approaches differ in both philosophical foundations and technical implementation, differences that go beyond the scope of this introduction (see, e.g., Garcia 2021: ch. 10). In brief, frequentist statistics treat data as a random sample from a larger population, while model parameters are considered fixed but unknown quantities estimated solely from the observed data. In contrast, Bayesian statistics treats the data as fixed and model parameters as random variables, whose distributions reflect both the data and prior beliefs about plausible values. These prior distributions may be UNINFORMATIVE, incorporating minimal assumptions, or INFORMATIVE, drawing on expert knowledge often based on prior research (e.g., plausible ranges for voice onset time in plosives or formants in vowels).

All the articles in this special collection make use of quantitative methods. Andersson uses descriptive statistics to decide between competing models of Abkhaz stress. The other six articles use regression methods (often in combination with descriptive statistics) to model their corpus data: Russell, Dabouis & Fournier, Kukhto & Piperski and Sano & Guillemot adopt the 'mainstream' frequentist framework, whereas Mayer and Lamontagne use a Bayesian approach. The latter two authors both point out that Bayesian models are usually more robust when parameter estimation is difficult (or fails altogether) in the presence of sparse data, since they rely on prior information in addition to the data itself. Mayer stresses another advantage of the Bayesian paradigm: because it treats parameters as random variables, it allows for a direct probabilistic interpretation of parameter estimates and their uncertainty. There is a growing concern in the literature about over-reliance on null hypothesis significance testing (van Zyl 2018), which evaluates whether an effect (e.g., the influence of lexical frequency on vowel reduction) is statistically different from zero. Bayesian statistics, by contrast, provides full posterior distributions (i.e., updated probabilities after observing the data), which offer a richer picture of the plausible values a parameter can take, rather than reducing inference to a binary decision about significance. It is not our aim to adjudicate between the frequentist and Bayesian paradigms, nor to claim the superiority of one over the other. Both approaches have their strengths and appropriate applications. It should also be pointed out that, in practice, Bayesian analyses often rely on uninformative or weakly informative priors, which frequently results in inferences that closely resemble those obtained through frequentist methods.

In addition to quantitative methods, phonologists sometimes make opportunistic use of pre-existing computational tools to process their phonological corpora. For example, Mayer used a morphological transducer for Uyghur to separate roots and suffixes in the raw textual data from which he built his

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

corpus. Another increasingly common approach is the use of a forced phonetic aligner, such as the Penn Phonetics Lab Forced Aligner (Yuan & Liberman 2008), WebMAUS (Kisler *et al.* 2017) or the Montreal Forced Aligner (McAuliffe *et al.* 2017), to obtain a phone-level alignment of a corpus, which greatly facilitates large-scale analysis of phonological (and phonetic) phenomena in speech corpora. Lamontagne illustrates this trend by leveraging a speaker-adaptive forced aligner, coupled with a classification algorithm, to automatically identify tense and lax vowels in non-final position in Laurentian French.

4. Corpus phonology and phonological theory

A criticism that is sometimes raised against corpus phonology is that it promotes a type of empirical research devoid of theoretical content. As we pointed out earlier, corpus phonology partly developed in reaction to the excesses of the second half of the 20th century, where second-hand data was routinely used to motivate a theoretical innovation, without always ensuring that the data were valid or representative. Morin (1987), for example, offers a sharp critique of how a certain strand of generative phonology has relied on idealised, and sometimes inaccurate, data to develop theoretical analyses of some aspects of the phonology of French. Although there exists a certain strand of work that seems to have gone fully in the other direction, adopting a form of neo-structuralist, data-driven empiricism in the study of language (e.g., Goldsmith & Xanthos 2009), most work we are aware of in corpus phonology falls somewhere between the two extremes, and the articles in this volume are a good illustration of the range of approaches that are encountered today. While most indeed eschew formalisation, all of them clearly address theoretical questions, which underlines the importance of the distinction between theorisation and formalisation.

First of all, two articles in this collection adopt one of the dominant theoretical frameworks, namely Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater & Johnson 2003; Hayes & Wilson 2008), a probabilistic version of Harmonic Grammar (Smolensky 1986; Legendre *et al.* 1990; Pater 2009), itself a variant of OT in which constraints are weighted rather than ranked. Mayer (2025) investigates a case of phonological opacity in Uyghur where backness harmony interacts with vowel reduction, a process which can turn harmonising vowels into neutral ones. His corpus analysis reveals that both the surface-apparent pattern (whereby reduced vowels behave as neutral) and the opaque pattern (whereby reduced vowels harmonise according to their underlying form) are attested, although the opaque pattern appears to be dominant. This type of opacity is known to be problematic in standard parallel constraint-based frameworks. In order to account for the observed variation, Mayer proposes a (modified) constraint, HarmonicUniformity, whose weight depends on the certainty of the harmonic class of the root (back or front). Roots described in the previous literature as 'vacillating' in their harmonic behaviour are treated by Mayer as having low degrees of certainty, which allows other constraints to play a larger role in determining the vowels that appear in their suffixes. Of course, the question of what drives certainty in the harmonic class in the first place now arises. This is where the corpus analysis plays a crucial role: in addition to the distribution of the root and its tendency to co-occur with front or back suffixes, Mayer identifies several phonotactic and morphological factors, as well as a potential general bias for back harmony (which is the default in Uyghur), as predictors of a root's harmonic class.

Kukhto & Piperski (2026) also adopt MaxEnt to account for the role of stress clashes in the selection of genitive suffixes in comparative adjective–noun sequences in Russian, furthering our understanding of postlexical phonology. A common approach in corpus phonology is to first fit a mixed-effects logistic model to some corpus data to identify relevant factors, and to then turn these factors into phonological constraints and incorporate them into a MaxEnt model that infers the constraints' weights based on the corpus data (see, e.g., Griffiths 2022). As Kukhto & Piperski point out, there are important formal connections between MaxEnt and (multinomial) logistic regression, since the former is in fact a restricted form of the latter. They choose to bypass the first, heuristic step, and use a mixed-effect

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

model as a generalised form of MaxEnt, directly incorporating constraints into their model to predict the alternation between the long and short forms of synthetic adjectival comparatives. Although this is not the focus of their article, their approach raises an important theoretical question: which grammar(s) are we in fact modelling, exactly? Mixed-effects models are used to incorporate subject-level variation (and possibly lexical idiosyncrasy), in such a way that the resulting model represents a form of weighted average, known as PARTIAL POOLING, of the sample's grammar, based on the relative contribution of each subject (and/or lexical item). By contrast, MaxEnt does not incorporate subject or item-level variation, and when all the data is pooled together (COMPLETE POOLING), as is often done in practice, the resulting model represents a form of community grammar in which subject-level variation is ignored.

Although he does not use MaxEnt, Lamontagne (2025) directly tackles the issue of community- *vs.* individual-level grammars. His corpus analysis of variable high-vowel laxing in non-final syllables in Laurentian French (e.g., *mythique* /mitik/ → [mi͡tsɪk] ~ [mɪ͡tsɪk] 'mythical') first reveals that there are more laxing patterns than have been previously reported in the literature. In addition, this study suggests that individual speakers of Laurentian French exhibit marked, sometimes opposing, grammatical patterns, forming distinct clusters; despite this variation, the overall distribution of the weights of the factors considered in the analysis tends to be unimodal, implying that individual grammars are variants of a shared community grammar rather than instances of competing grammars within the community. To the extent that the issue of community *vs.* individual grammars has traditionally been addressed by sociolinguists, Lamontagne offers a good illustration of how corpus phonology can be a bridge between phonology and sociolinguistics.

Russell (2025) takes on the issue of progressive and regressive nasal harmony in Paraguayan Guaraní, which has been argued to operate bidirectionally. Her analysis of a conversational corpus shows that they in fact behave differently with respect to both native roots and Spanish borrowings. While regressive nasal harmony appears to be synchronically active, progressive nasalisation seems to be the relic of diachronic nasalisation processes that result synchronically in suppletive allomorphy. This analysis helps explain why regressive harmony has been extended to Spanish lexical items whereas progressive harmony has not.

Andersson (2026) and Dabouis & Fournier (2025) both extend the range of factors that predict stress in languages where stress is not entirely predictable. Andersson examines stress in Abkhaz verbs, which display complex interactions between prosodic structure and morphology. A previous stress rule first proposed by Dybo (1977) relies on underlying accents and states that stress falls on the leftmost accent not immediately followed by an accent. While this rule is relatively successful, Andersson finds several verb classes where it fails altogether. The author establishes a number of allomorphic patterns, some of which have not previously been identified in the literature, where affixes that are normally unaccented receive stress. Furthermore, they argue that some verbs are not subject to Dybo's Rule at all, instead placing stress immediately before the root. This pre-stressing pattern depends on a mix of phonological factors (such as root length and accentual pattern) and morphological ones (person markers, preverbs and causativisation).

Dabouis & Fournier (2025) test a large range of factors to predict vowel reduction in British English in initial pretonic and intertonic positions. In addition to confirming the role of well-established factors such as syllabic structure, the position of the vowel with respect to stress, morphological structure, the semantic opacity of prefixes and lexical frequency, they show that vowel backness, but not height, influences reduction (front vowels are more prone to reduction than back vowels). Foreign words also reduce less than native ones, even after taking into account lexical frequency. Furthermore, the authors identify an intriguing correlation between spelling and phonology. This study also contributes to the debate on the syllabification of /sC/ clusters, showing that variable vowel reduction before these clusters reflects different syllabifications of the base form.

Finally, Sano & Guillemot (2025) adopt a variationist/usage-based/evolutionary framework, an approach that has been slowly but steadily gaining wider acceptance among phonologists. They offer a replication of previous findings in English, showing that the existence of a lexical competitor for a given contrast favours hyperarticulation in Japanese as it does in English, two languages which are

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

genetically and typologically very different. Speech corpora lend themselves particularly well to this type of usage-based/exemplarist study, which goes beyond the traditional dichotomy between lexical and postlexical phonology and provides further evidence that paradigmatic relationships in the mental lexicon can subtly affect the realisation of lexical items in discourse.

5. Concluding remarks

As the discussion above and the breadth of articles in this special collection show, corpus phonology is a broad subfield at the intersection between phonology and corpus linguistics. In trying to pinpoint why corpora are essential in phonological research today, Ernestus & Baayen (2011: 374–375) point out that

> corpora help us bridge the gap between the analyst's conception of the data and the actual data themselves. Phonologists have formulated generalizations, some of which, as we know now, thanks to corpus-based research, do not do full justice to the data. Language appears to be much more complex than is generally assumed and this complexity is important for theories of phonology as well as for theories of speech production and comprehension.

Despite the variety of corpora and the range of methodological approaches and theoretical orientations, we believe that each article in this issue contributes in its own way to bridging this gap.

Acknowledgements. We are grateful to the editors for their valuable feedback and suggestions on an earlier version of this work. Any remaining errors or shortcomings are our own.

Competing interests. The authors declare no competing interests.

References

Andersson, Samuel (2026). Stress assignment in Abkhaz verbs: a corpus-based approach. *Phonology* **43**, article no. e6 (26 pp.).
Anttila, Arto (1997). Deriving variation from grammar. In Frans Hinskens, Roeland van Hout & Leo Wetzels (eds.) *Variation, change and phonological theory*. Amsterdam: Benjamins, 35–68.
Anttila, Arto (1997). Variation and optionality. In Paul de Lacy (ed.) *The Cambridge handbook of phonology*. Cambridge: Cambridge University Press, 519–536.
Anttila, Arto & Young-mee Yu Cho (1998). Variation and change in Optimality Theory. *Lingua* **104**, 31–56.
Barth, Danielle & Stefan Schnell (2021). *Understanding corpus linguistics*. London: Routledge.
Benko, Vladimir (2014). Aranea: yet another family of (comparable) Web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.) *Text, Speech and Dialogue: 17th international conference, TSD 2014*. Cham: Springer, 247–256.
Biber, Douglas & Randy Reppen (eds.) (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.
Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **32**, 45–86.
Boersma, Paul & David Weenink (2025). Praat: doing phonetics by computer. Computer software. Version 6.4.46. https://praat.org.
Cedergren, Henrietta J. & David Sankoff (1974). Variable rules: performance as a statistical reflection on competence. *Lg* **50**, 333–355.
Courdès-Murphy, Léa (2018). *Nivellement et sociophonologie de deux grands centres urbains: le système vocalique de Toulouse et de Marseille*. PhD dissertation, Université Toulouse-Jean Jaurès.
Dabouis, Quentin & Jean-Michel Fournier (2025). An empirical study of vowel reduction and preservation in British English. *Phonology* **42**, article no. e15 (32 pp.).
Dresher, B. Elan & Harry van der Hulst (eds.) (2022). *The Oxford history of phonology*. Oxford: Oxford University Press.
Durand, Jacques (2017). Corpus phonology. In Mark Aronoff (ed.) *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press, not paginated. https://doi.org/10.1093/acrefore/9780199384655.013.145.
Durand, Jacques, Ulrike Gut & Gjert Kristoffersen (2014a). Introduction. In Durand *et al.* (2014b), 1–10.
Durand, Jacques, Ulrike Gut & Gjert Kristoffersen (eds.) (2014b). *The Oxford handbook of corpus phonology*. Oxford: Oxford University Press.
Durand, Jacques, Bernard Laks & Chantal Lyche (2002). La phonologie du français contemporain: usages, variétés et structure. In Claus D. Pusch & Wolfgang Raible (eds.) *Romance corpus linguistics: corpora and spoken language*. Tübingen: Gunter Narr, 93–106.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

Dybo, Vladimir Antonovič (1977). Zapadnokavkazskaja akcentnaja sistema i problema eë proisxoždenija [The West Caucasian accentual system and the problem of its origin]. In Vyacheslav V. Ivanov (ed.) *Nostratičeskie jazyki i nostratičeskoe jazykoznanie [The Nostratic language and Nostratic linguistics]*. Moscow: Institute of Slavic and Balkan Studies of the USSR Academy of Sciences, 41–45.

Ernestus, Mirjam & R. Harald Baayen (2011). Corpora and exemplars in phonology. In Goldsmith *et al.* (2011), 374–400.

Eychenne, Julien (2022). Corpus phonology. In Christoph Gabriel, Randall Gess & Trudel Meisenburg (eds.) *Manual of Romance phonetics and phonology*. Berlin: De Gruyter, 319–341.

Eychenne, Julien & Léa Courdès-Murphy (2019). Phonometrica: an open platform for the analysis of speech corpora. In *Proceedings of the Seoul International Conference on Speech Sciences 2019*. Seoul: Korean Society of Speech Sciences, 107–108.

Frisch, Stefan A., Janet B. Pierrehumbert & Michael B. Broe (2004). Similarity avoidance and the OCP. *NLLT* **22**, 179–228.

Garcia, Guilherme D. (2021). *Data visualization and analysis in second language research*. New York: Routledge.

Gelman, Andrew & Jennifer Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Godfrey, John J., Edward C. Holliman & Jane McDaniel (1992). SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. New York: IEEE, 517–520.

Goldsmith, John A. (ed.) (1995). *The handbook of phonological theory*. 1st edition. Oxford: Blackwell.

Goldsmith, John A., Jason Riggle & Alan C. L. Yu (eds.) (2011). *The handbook of phonological theory*. 2nd edition. Oxford: Wiley-Blackwell.

Goldsmith, John A. & Aris Xanthos (2009). Learning phonological categories. *Lg* **85**, 4–38.

Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University, 111–120.

Griffiths, Joshua M. (2022). Competing repair strategies for word-final obstruent–liquid clusters in Northern Metropolitan French. *Journal of French Language Studies* **32**, 1–24.

Gut, Ulrike & Holger Voormann (2014). Corpus design. In Durand *et al.* (2014b), 13–26.

Guy, Gregory R. (1991a). Contextual conditioning in variable lexical phonology. *Language Variation and Change* **3**, 223–239.

Guy, Gregory R. (1991b). Explanation in variable phonology: an exponential model of morphological constraints. *Language Variation and Change* **3**, 1–22.

Hall, Kathleen Currie (2022). Corpora and phonological analysis. In Dresher & van der Hulst (2022), 619–638.

Hall, Kathleen Currie, J. Scott McKie & Roger Yu-Hsiang Lo (2019). Phonological CorpusTools: software for doing phonological analysis on transcribed corpora. *International Journal of Corpus Linguistics* **24**, 522–535.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**, 379–440.

Hedlund, Gregory & Yvan Rose (2020). Phon 3.1. Computer software. https://phon.ca.

Kiparsky, Paul (1993). An OT perspective on phonological variation. Paper presented at the first Rutgers Optimality Workshop (ROW I), Rutgers University, October 1993, and at the 23rd Conference on New Ways of Analyzing Variation (NWAV 23), Stanford University, October 1994. https://web.stanford.edu/~kiparsky/Papers/nwave94.pdf.

Kisler, Thomas, Uwe Reichel & Florian Schiel (2017). Multilingual processing of speech via web services. *Computer Speech & Language* **45**, 326–347.

Kukhto, Anton & Alexander Piperski (2026). Rhythmic optimisation in the Russian comparative: a corpus study. *Phonology* **43**, article no. e19 (28 pp.).

Labov, William (1969). Contraction, deletion and inherent variability of the English copula. *Lg* **45**, 713–762.

Lacheret, Anne, Chantal Lyche & Atanas Tchobanov (2011). Schwa et position initiale revisités: l'éclairage de la prosodie en phonologie du français contemporain. *Langue française* **169**, 137–158.

Laks, Bernard (2008). Pour une phonologie de corpus. *Journal of French Language Studies* **18**, 3–32.

Lamontagne, Jeffrey (2025). Undergeneration and overgeneration in phonological analysis: revisiting individual and community grammars of Laurentian French high-vowel laxing. *Phonology* **42**, article no. e27 (30 pp.).

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. Technical Report 90-5, Institute of Cognitive Science, University of Colorado, Boulder.

Liberman, Mark Y. (2019). Corpus phonetics. *Annual Review of Linguistics* **5**, 91–107.

Mayer, Connor (2025). A large-scale corpus study of phonological opacity in Uyghur. *Phonology* **42**, article no. e24 (31 pp.).

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger (2017). Montreal Forced Aligner: trainable text–speech alignment using Kaldi. In *Interspeech 2017*. Stockholm: International Speech Communication Association, 498–502.

McCarthy, Michael & Anne O'Keeffe (2010). Historical perspective: what are corpora and how have they evolved? In O'Keeffe & McCarthy (2010), 3–13.

Morin, Yves-Charles (1987). French data and phonological theory. *Linguistics* **25**, 815–843.

O'Keeffe, Anne & Michael McCarthy (eds.) (2010). *The Routledge handbook of corpus linguistics*. London: Routledge.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451

Pater, Joe (2009). Weighted constraints in generative linguistics. *Cognitive Science* **33**, 999–1035.
Pierrehumbert, Janet B. (2022). More than seventy years of probabilistic phonology. In Dresher & van der Hulst (2022), 639–655.
Pierrehumbert, Janet B., Mary E. Beckman & D. Robert Ladd (2000). Conceptual foundations of phonology as a laboratory science. In Noel Burton-Roberts, Philip Carr & Gerard Docherty (eds.) *Phonological knowledge: conceptual and empirical issues*. Oxford: Oxford University Press, 273–304.
Reynolds, William T. (1994). *Variation and phonological theory*. PhD dissertation, University of Pennsylvania.
Russell, Katherine R. (2025). A corpus-based study of variation in and extension of two Paraguayan Guaraní nasalisation patterns. *Phonology* **42**, article no. e20 (34 pp.).
Sano, Shinichiro & Céleste Guillemot (2025). Cue-specificity of contrastive hyperarticulation: evidence from the voicing contrast in Japanese. *Phonology* **42**, article no. e13 (17 pp.).
Smolensky, Paul (1986). Information processing in dynamical systems: foundations of harmony theory. In David E. Rumelhart, James L. McClelland & PDP Research Group (eds.) *Parallel distributed processing: explorations in the microstructure of cognition*, volume 1. Cambridge, MA: MIT Press, 194–281.
Stefanowitsch, Anatol (2020). *Corpus linguistics: a guide to the methodology*. Berlin: Language Science Press.
Wedel, Andrew, Abby Kaplan & Scott Jackson (2013). High functional load inhibits phonological contrast loss: a corpus study. *Cognition* **128**, 179–186.
Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes (2006). ELAN: a professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.) *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa: European Language Resources Association, 1556–1559.
Yuan, Jiahong & Mark Liberman (2008). Speaker identification on the SCOTUS corpus. *JASA* **123**, article no. 3878.
Zufferey, Sandrine (2020). *Introduction to corpus linguistics*. London: ISTE.
Zwirner, Eberhard & Kurt Zwirner (1936). *Grundfragen der Phonometrie*. Berlin: Metten.
van Zyl, Casper J. J. (2018). Frequentist and Bayesian inference: a conceptual primer. *New Ideas in Psychology* **51**, 44–49.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 19:43:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0952675726100451