Laboratory Phonology
Journal of the Association for Laboratory Phonology

Ferreira, F. (2026). Prosody, Syntax, and Unrehearsed Language. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 26(1), pp. 1–20. DOI: https://doi.org/10.16995/labphon.20016

Open Library of Humanities

Prosody, Syntax, and Unrehearsed Language

Fernanda Ferreira, University of California, Davis, fferreira@ucdavis.edu

Traditional theories of language processing, syntax, and prosody often stem from written language conventions, but spontaneous speech differs substantially in its frequent corrections, simpler syntactic structures, and shorter utterances. Recent advances in data analysis enable large-scale examination of unrehearsed speech, revealing that syntactic and prosodic boundaries only moderately align—indeed, nearly half of the words analyzed show mismatches between prosody and syntax in boundary strength. These findings challenge categorical, competence-based theories of prosody and instead support an ecologically valid, resource-based perspective, wherein prosodic features fluctuate dynamically as speech unfolds.

Laboratory Phonology: Journal of the Association for Laboratory Phonology is a peer-reviewed open access journal published by the Open Library of Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/. Articles published by the Open Library of Humanities may also contain other CC-licensed material, or material used or reproduced under an exception or limitation to copyright. Such material is not subject to the article's license. It is either marked with its own license, or marked that it has been reproduced with permission or under fair-use.
OPEN ACCESS

1. Introduction

Prosody is not merely ornamental; it is a rich source of information for grouping words into linguistic constituents, the fundamental units of language processing. Decades of careful research using artificial materials and tightly controlled experimental methods have established some important relationships between prosody and syntax. For example, word lengthening and pauses often signal the boundary of a syntactic constituent, and the more prominent the syntactic boundary, the stronger these prosodic cues tend to be. In this article, I make the case that, despite these important findings, our understanding of prosody, its relationship to syntax, and prosody's role in language processing are still underdeveloped because the field has been dominated by what I will term a "competence-based approach": The theories from which experimental predictions have been derived assume a key role for linguistic constituency and downplay the potential role of performance-based factors that likely play a key role in determining prosodic features. In addition, the linguistic stimuli that have been tested together with the artificial nature of the tasks used to elicit prosodic renditions have led to an incomplete picture of what prosody looks like in unrehearsed language. Investigations of prosody that employ more naturalistic stimuli reveal the importance of performance-based factors such as those related to working memory constraints and speakers' need to recover from having engaged in linguistic work and their need to plan upcoming material. The findings from such studies diverge to a surprising degree from findings derived from studies rooted in the more competence-based tradition that focus primarily on linguistic representations.

The article is organized as follows. First, I will specify how prosody is defined for the purposes of this paper and why prosody is important in psycholinguistics and in the language sciences more generally. Then I will discuss how prosodic phrasing is treated in some of the classic literature on prosody that adopts a competence-based approach according to which prosodic features emerge from the organization of linguistic constituents. From this section I will consider more performance-based approaches, which naturally transitions to a discussion of the findings that have emerged from some studies that examine spontaneous, unrehearsed language. This will not be an exhaustive review of the literature, but instead will highlight some recent work from our laboratory that we believe exemplifies the strengths of adopting a more performance-based approach. The implications of these performance-based effects are then briefly considered in the final section, with an emphasis on what they imply for theories of language production and spoken language comprehension.

2. What is Prosody and Why is it Important?

We will begin with a brief definition of prosody, sufficient to allow us to make some points regarding its importance, and then return to a more detailed characterization of prosody and how it is generated. We can define prosody as the variations in rhythm, pitch, and speech rate

that communicate information beyond lexical identity: “Prosody is an intrinsic determinant of the form of spoken language. The prosodic structure of an utterance exercises effects on the timing, amplitude, and frequency spectrum of the utterance, and these are the dimensions of sound itself” (Cutler, Dahan, & van Donselaar, 1997, p. 141). For example, the word at the end of an utterance will tend to have a longer duration than if it occurred phrase-medially, and it will often be followed by a pause.

Prosody is of interest for a number of reasons. First, prosody is a feature of every single utterance we produce. Indeed, when in a lecture or conversation we wish to elicit an intuition about how valuable prosody is, we will often attempt to produce an utterance with no prosody at all, in what is informally referred to as a “monotone voice”, and what we discover is that it simply can’t be done. An utterance that attempts to be monotone will always have some variations in duration, pitch, intensity, and so on—those features would just be different from what they would be if the utterance were said normally. Any theory of language and language processing, then, must account for these prosodic variations and attempt to explain why they are generated. A second reason to study prosody is because, when it is in fact absent—say in artificial speech— the resulting utterances sound awful and may even be unintelligible. This tells us that prosody plays a major role in language understanding and that its presence is bound up with what we expect to hear when humans talk. Third, prosody links multiple levels of linguistic representations to each other, and this is because it reflects the pragmatic, semantic, and syntactic content of the speaker’s utterance, in addition to performance factors. And, finally, prosody allows researchers to investigate the relationship between linguistic representations and performance. In syntax, a noun phrase will always be a noun phrase regardless of whether the speaker produces the utterance while relaxing in a chair or running with a jogging partner. In contrast, the form and content of a prosodic phrase will vary depending on performance factors: For example, if a person attempts to speak when somewhat out of breath, the prosodic phrases they generate will usually be shorter, for example—perhaps only one to two words long. Thus, if our goal is to understand how performance factors have shaped the grammar and the overall architecture of the human language system, prosody is an excellent domain to explore.

But what are these prosodic features that constitute the prosody of an utterance? Another way to pose this question is to ask, what do we measure when we investigate prosody? Typically, researchers would consider the following variables. First are variables that relate to rhythm, including word duration, intensity or stress, and pausing. Some syllables in an utterance will be louder than others, and others will have a longer duration. (We tend to use the term *word duration* but what we are typically referring to is the duration of the last one or sometimes two syllables of the word.) One study demonstrated that a word at the end of a phrase has a longer duration than that same word in the middle of a phrase, and if that word contains a lax vowel (e.g., the ɛ in *bet*), it will also tend to be followed by a pause (Ferreira, 1993). This pattern of

results has been interpreted to mean that speakers attempt to maintain a rhythm that can be captured via a phonological representation known as a metrical grid, a structure that specifies syllable stress and duration for a word or sequence of words (Selkirk, 1984). The metrical grid assigns a word at the end of a phrase or utterance greater stress and more time; if a word contains a lax vowel and is therefore not easy to elongate, a speaker may resort to a pause to fill out the interval specified in the metrical grid.

Second, let us consider pitch. Pitch variations convey important postlexical information. For example, pitch communicates whether an utterance is a statement or question, and can also signal surprise, excitement, or boredom. In addition, the ends of phrases tend to be marked with certain pitch features such as a falling tone. These boundary tones, together with features such as word lengthening and pausing, signal the right edge of a prosodic phrase. To constrain the discussion, in this paper we will focus primarily on the prosodic features that combine to signal a phrase boundary, and our focus will be on the identification of prosodic boundaries in speech, as signaled by features such as phrase-final lengthening, pausing, and the presence of boundary tones.

3. Prosodic Phrasing: Competence-Based Approaches

Next we turn to a consideration of more traditional approaches to prosody that characterize the presence and absence of prosodic features in terms of linguistic representations. We will use the term "competence-based approach" even though it is perhaps a bit clumsy, but it is meant to capture the basic idea behind theories that explain prosodic features with respect to some type of linguistic constituency, either syntactic or prosodic. Theories that assume only syntactic constituency and that derive prosodic features from the characteristics of the syntactic structure are often termed "direct theories". In contrast, theories that posit a layer of prosodic constituency derived from syntax are often called "indirect theories", as prosodic features are only indirectly related to syntax, via a prosodic constituency that is derived from syntax but is not identical to it (Bennett & Elfner, 2019). Although proponents of direct views still exist, most phonologists subscribe to the indirect view.

Prosodic constituency differs from syntactic constituency in many important ways. First, of course the constituents themselves are different: Instead of syntactic constituents such as noun phrase and verb phrase, prosody posits prosodic constituents such as phonological phrase (henceforth, PPh) and intonational phrase (henceforth, IPh). These prosodic constituents are arranged hierarchically: Prosodic words are contained inside PPhs, PPhs are contained inside IPhs, and IPhs are contained inside utterances. Unlike syntax, these constituent levels are "strictly layered", meaning that a higher-level constituent can never be contained in a lower-level unit. For example, an IPh cannot be inserted into a prosodic word or a PPh. Note that this constraint differs from what we find with syntax: For example, a clause may occur inside a noun phrase,

as we see in relative clauses, for instance, where a clause modifies a noun. Prosodic constituents obey the strict layering constraint whereas syntactic constituents do not.

An IPh will often correspond to a single syntactic clause, although it is possible for a speaker to generate other intonational phrasings—for example, one that makes the subject and verb of a sentence one IPh and the verb's object a separate IPh, as in *(The teenager texted) (after the party)(yesterday)* (parentheses indicate IPhs). Skipping the level of the PPh for just a moment, prosodic words are similar to ordinary words except that they group highly reduced words such as determiners with their hosts, as in *the bet*, which is one prosodic word. What about PPhs, then? In one prominent theory, a sequence of words up to the boundary of a syntactic phrase such as Noun Phrase, Prepositional Phrase, or Verb Phrase constitutes one PPh (Selkirk, 1986). For example, in *The teenager texted after the party yesterday*, *the teenager* would be a single PPh because a syntactic phrase terminates on *teenager*—namely, a Noun Phrase, as shown below.

(1) $[[\text{The teenager}]_{NP}\ [\text{texted})\ [\text{after}\ [\text{the party}]_{NP}]_{PP}\ [\text{yesterday}]_{AdvP}]_{VP}]_{S}$

The sequence *texted after the party* would constitute another PPh since the next phrase boundary that is encountered is the one following *party*, which means everything from the end of the previous phonological phrase to the end of this one becomes part of the second PPh. Notice that the fact that two syntactic phrases terminate on *party* makes no difference: A single syntactic phrase boundary is sufficient to establish the right edge of the PPh boundary, and additional phrase boundaries have no prosodic effect. Thus, the prosodic constituent structure for this sentence might be something like this (omitting the layer of prosodic words, which are not at issue in this article):

(2) $((\text{The teenager})_{PPh}\ (\text{texted})_{IPh}\ (\text{after the party})_{PPh})_{IPh}\ ((\text{yesterday})_{PPh})_{IPh}$

The result is a structure that differs in significant ways from the syntactic structure of the sentence, even though it is derived from it. At the same time, it is important also to recognize the syntactic information this prosodic constituency still manages to convey despite the lack of isomorphism. Because PPhs mandatorily terminate at the right edge of a syntactic phrase, the prosodic structure preserves the information about syntactic boundaries. And, of course, on a direct theory which assumes no distinct prosodic constituency but instead attempts to derive prosodic effects such as word lengthening directly from syntax, that same information about syntactic phrasing will be transparently preserved. It is in part because of this marking of syntactic information that prosody has played a major role in psycholinguistic discussions of language comprehension, especially around questions of parsing. If prosody marks syntactic boundaries (either directly or indirectly), then a listener does not have to rely solely on abstract, silent syntactic knowledge to parse a sentence into syntactic constituents; instead, the listener

will literally "hear" the boundary in the prosody of the sentence, presumably facilitating the task of parsing and comprehension (see Section 6).

Having now described the competence-based approach, what kind of evidence supports it? One set of classic studies, summarized in Cooper and Paccia-Cooper (1980), is illustrative. In one of their experiments, the investigators tested whether a stronger syntactic boundary would lead to longer phrase-final word durations and longer following pauses. They compared the duration of *class* and the pause after it in a sentence such as *Our Sunday school class will ask for support* and *Our Sunday school class Will asked for support*, and the investigators reported that both were longer in the second case due to the stronger syntactic boundary that terminates on that word. A complementary effect from another study made use of the rule of external sandhi known as "flapping", which is a rule that operates across words if there is only a weak syntactic boundary between them. The investigators compared cases such as *The woman with the winning bet agreed…* and *The woman who had won the bet agreed…* and reported more flapping of the /t/ in *bet* in the former case due to the weaker syntactic boundary associated with modifying prepositional phrases versus relative clauses. Many similar effects were reported in the 1970s and 1980s. My own later work attempted to establish that lengthening and pausing were attributable not to syntactic but to prosodic structure (Ferreira, 1993). I showed that a phrase-medial word such as *cop* had a short duration compared to that same word when it occurred phrase-finally, but there was no difference in phrase-final *cop* duration depending on whether the phrase terminated on just a simple noun phrase, a more complex noun phrase with a modifying prepositional phrase, or an even more complex noun phrase containing a modifying relative clause. Because the direct view predicted longer durations with greater syntactic complexity but the indirect view did not, I argued for prosodic constituency as the source of metrical (i.e., durational) effects.

These studies and most others from the same era can be viewed as emerging from a competence-based approach because constituent structure is the primary focus of the hypotheses concerning prosodic effects like word duration or pausing. A speaker's need to recover or plan an upcoming constituent are not relevant to predicting whether a word will be lengthened or a pause will be required—all that is taken into account is the constituent structure, whether it be prosodic or syntactic. An approach that focuses on linguistic structure and does not consider these performance-based factors will also naturally tend to ignore phenomena such as the production of fillers such as *uh* and *um* and will likely also ignore variations in speech rate. A speaker will usually produce an *uh* or an *um* in order to gain more planning time, and speech rate may vary to reflect the speaker's need to plan or the speaker's ability to execute a sequence of words fluently, but those factors do not belong to the prosody itself. Indeed, an ideal rendition of the sentence would not contain any fillers and would be spoken at a fairly constant speech rate.

It is useful to review the methods used in these studies to shed light on why the influence of performance factors might not have appeared important. As the examples provided in this section illustrate, the stimuli were sentences created specifically for the purposes of testing the study's hypotheses. To obtain the experimental control required to ensure the same word was compared across conditions—a consideration that is critically important in controlled experiments given that different words have different intrinsic durations—carefully crafted materials had to be the linguistic items on which theories are tested. And how would researchers get speakers to say these unusual sentences? For example, in what typical situation might a person utter *Our Sunday school class Will asked for* support? The solution in most studies is to ask the speaker simply to read the sentence out loud, first by practicing it and then producing it for recording purposes once they were confident they could generate what intuitively seemed like an idealized rendition. If the speaker failed to say the sentence fluently—for example, if they included a filler such as *uh*—the recording was usually discarded and they were asked to try again. (In Ferreira, 1993, I introduced a memorization paradigm aiming for somewhat more natural renditions, but the stimuli were still fairly unnatural.) This classic procedure has speakers generating sentences under conditions very different from natural production. Given incremental language production (Levelt, 1989), in normal situations speakers do not plan an entire sentence in advance of starting to say it, unlike the situation in these experiments in which the sentence is printed, practiced, and said as a whole. In addition, the sentences are rather different from what is observed in unrehearsed speech. Recently, it has been reported that conversational utterances in Dutch, German, and English are usually very short, often only one to two words long, and syntactically rather simple (Corps et al., 2022).

This procedure for eliciting sentences and assessing the prosody that would normally accompany them has yielded a great deal of important information, but it is important also to recognize the limitations. A speaker who is presented with a complete, long, and complex sentence and asked to pronounce it "as if it were being spoken naturally" is essentially providing a type of grammaticality judgment: The speaker is inclined to approach the situation as if their task is to produce what they imagine to be the ideal prosodic form for the sentence. It is as if they are answering the question "What would be the best or ideal way to say this sentence?" Again, this method is an important tool in the psycholinguist's or phonologist's toolkit, but it leaves unanswered the important question of what prosody looks like when people generate language in real time, typically resulting in short, syntactically simple utterances which often will contain disfluencies and outright errors. To begin to answer this question, let us next consider a more performance-based approach.

What we have seen, then, is that competence-based approaches rely on somewhat unnatural stimuli and methods to characterize prosody, including its relationship to syntactic structure. Next we turn to approaches that make use of more naturalistic stimuli and analysis

techniques and emphasize how cognitive constraints such as working memory capacity influence prosodic form.

4. Prosodic Phrasing: Performance-Based Approaches

Performance-based approaches emphasize factors related to the physiological and cognitive constraints associated with speaking, and the ways those factors might influence features such as word durations, pausing, and speech rate. For example, imagine a speaker wishes to utter the two-clause sentence *I planned to make tiramisu for dessert but the store was out of mascarpone*. As would be expected given incremental planning, which refers to the idea that speakers begin to speak before they have fully planned every detail of their upcoming utterance, the speaker might start by articulating the first clause before having worked out the details of the second. In that situation, they might pause between the two clauses to plan that upcoming clause, and they might slow down as they approach that unplanned segment, leading to longer end-of-clause word durations. This type of performance effect—considered "performance" because it relates to the speaker's need to manage their linguistic and cognitive resources during speaking rather than arising from a constituent structure representation—would lead to word lengthening and pausing, the same acoustic features associated with the effects of prosodic structure.

What are some of these performance-based factors? One that has already been mentioned is the need to slow down or stop in order to plan an upcoming segment of speech. Another is the need to recover from planning and articulating the previous sequence of words—in other words to take a brief rest. In addition, speakers must coordinate their breathing with their speaking. In normal speaking situations, speakers on average take a breath every three to four seconds, and the more syllables in the speech segment just produced, the longer the breath will tend to be (Kallay, Mayr, & Redford, 2019). Of course, in aerobically taxing situations such as running or cycling, breaths tend to be shorter, resulting in shorter utterances. This can be viewed as a biological constraint, since it arises from the body's need to manage respiration. A more cognitively based constraint arises from the limits of working memory, which has a severely limited capacity—on some theories, accommodating only two to four distinct items (e.g., Oberauer et al., 2016). Consistent with this idea, work on recall of random digit sequences from working memory shows that people tend to recall digits in three-item chunks, pausing for about a second in-between (Cowan & Elliott, 2022). And at the level of conceptual planning, our own research demonstrates that when people describe a complex scene, they tend to describe objects from the same semantic category (e.g., all the small appliances in a kitchen scene) before moving to another semantic grouping (e.g., all the cooking utensils). When they shift from one semantic cluster to another, they tend to draw out the two words on each side of that semantic boundary and they insert a pause. Fillers such as *uh* and *um* are likely as well. Thus, the need to plan, to recover from planning, to take a breath, to manage working memory resources, and to

organize thoughts into appropriate semantic groupings affects acoustic features in ways that are similar to what emerges from prosodic constituency, and those features may also be localized to the same parts of utterances (e.g., the boundary between two clauses).

Investigations of prosody in human language must confront a major challenge, which is that the same acoustic features that are associated with linguistic representations may be confounded with those that arise from performance—from the need to manage cognitive and linguistic resources during online production. To spell this out, according to theories of prosodic or syntactic constituency, a word at the end of a phrase will tend to be lengthened and may be followed by a pause. It will also be the locus of what is sometimes termed Nuclear Stress (Chomsky & Halle, 1968): That word will also be the most stressed. According to performance-based theories, a word at the end of a phrase will tend to be lengthened and may be followed by a pause because of the need to recover from the generation of that phrase and the need to plan the next one. How can these explanations for the same phenomena be disentangled? One way to disentangle the influences of competence and performance on these variables such as word duration and pausing is to move away from artificial stimuli and paradigms to naturalistic language. If the previous analysis is correct, perhaps speakers generate classic prosodic renditions of the sort that would be expected according to theories of syntactic and prosodic constituency because they recognize that, in an ideal world, that is how the sentences would be said. But what does real language look like—that is, what do the utterances people actually say in the situations in which they typically say them sound like? For example, do we observe the same alignment between syntactic boundaries and prosodic marking? We turn to this question in the next section.

5. Prosody and Naturalistic Language

As already discussed, in the field of prosody it has generally assumed that, to study prosody properly, it is necessary to develop tightly controlled stimuli with specific characteristics to allow the relevant variables to be isolated and to control for nuisance influences such as intrinsic word duration. For example, to determine whether a word has undergone phrase-final lengthening, it is necessary to compare that word's duration when it occurs phrase-finally to its duration when it occurs in some other position within the phrase or utterance. The field has assumed that it would be problematic to compare the durations of different words to each other because all words have distinct intrinsic durations, which means the investigator would end up comparing apples and oranges—in other words, the analyses would confound lexical and postlexical variables. And because sentences designed to test effects of syntactic or prosodic constituency must have highly specific content, they are unlikely to arise spontaneously—for example, how would a researcher get a speaker to spontaneously utter a sentence such as *The woman who won the bet agreed to give us the money*? It would seem that the only way to elicit these utterances is to show them to a speaker and ask them to say the sentence out loud, either by reading the sentence from a

piece of paper or a screen, or by studying the sentence and then repeating it from memory. An additional feature of such studies is that the amount of data they generate will tend to be small, since the sample sizes are usually around 10 to 20 subjects and the speakers can produce only so many sentences during the length of a standard experiment (30-60 minutes). In addition, acoustic analyses are normally done manually, which further limits the amount of data that can be processed and interpreted in a single study since those prosodic analyses are notoriously laborious and time-consuming.

Is any attempt to understand prosody and its basis in human performance doomed, then, since both the stimuli and the experimental context must be tightly controlled to allow for interpretable results? Fortunately, today the answer is no: New algorithms for measuring syntactic structure and prosodic features allow syntactic and prosodic representations to be extracted automatically from large amounts of unrehearsed language, without the need for manual analyses—amounts of data so large any variations such as those related to intrinsic durations and other nuisance factors can be treated as statistical noise. One way to think about this approach is to see it as the researcher placing the scientific controls at the data analysis end of the research pipeline rather than at the stage of data collection. Of course, for the effects of variables such as intrinsic duration to "wash out", as it were, a great deal of data must be analyzed. The good news is that researchers now have access to large corpora of speech, and they also can create their own, as we have done in our own laboratory.

A recent study conducted to evaluate to what extent prosodic boundaries facilitate the brain's decoding of syntactic constituents provides an excellent example of this approach, and indeed inspired our own research program on naturalistic prosody (Degano et al., 2024). The investigators' goal was to use relatively naturalistic speech to assess to what extent prosody facilitates the identification of syntactic boundaries. For this purpose, the researchers analyzed four TED talks in which well-known researchers spoke on a specific topic. Although TED talks are prepared ahead of time and rehearsed, they are delivered in the moment, and the speech is in no way related to the testing of linguistic hypotheses.

To perform the analyses in the Degano et al. (2024) study, the online transcriptions of the TED talks were syntactically analyzed using the spaCy parser (Honnibal & Montani, 2017). The parser provides a count of opening and closing syntactic nodes, which can serve as a measure of the strength of a syntactic boundary.

To determine the strength of the prosodic boundaries, the features word duration, energy, and pitch were calculated for each word. These analyses was performed automatically using the Wavelet Prosody Toolkit (https://github.com/asuni/wavelet_prosody_toolkit). The toolkit uses continuous wavelet transformation to calculate the pitch, energy, and duration values for each word in a given audio file and corresponding Praat text grid, and then sums those values to provide a measure of prosodic boundary strength. The range of possible pitch values is adjusted

depending on the gender of the speaker (males: 50–350Hz; females: 100–400Hz). The computed values for pitch, energy, and duration are weighted, with the default settings being 1 for pitch, 1 for energy, and .5 for duration, but these can be adjusted. In Degano et al. (2024) as well as the research described in Section 5, the default settings were used.

The result of the automatic syntactic and prosodic analyses is a measure of the strength of the syntactic and the prosodic boundary for every word in the corpus. The two boundary strength types were then dichotomized, leading to the creation of four experimental "conditions". Two of the conditions included the words for which prosody and syntax aligned: that is a weak syntactic boundary was associated with a weak prosodic boundary, and a strong syntactic boundary was associated with a strong prosodic boundary. The two other conditions are the misaligned renditions: the cases in which a weak syntactic boundary occurred with a strong prosodic boundary, and the ones in which a strong syntactic boundary was not prosodically marked.

The purpose of the Degano et al. (2024) study was to use magnetoencephalography (MEG) to determine whether the neural response to a syntactic boundary was more pronounced when it was marked by a prosodic boundary: that is, when the two types of boundaries align, and particularly when the end of a syntactic phrase is prosodically marked. The investigators found stronger MEG responses to prosodically marked syntactic boundaries, allowing them to conclude that the brain decodes syntax more easily when prosody provides cues to the syntactic structure. But, for the purposes of assessing the relationship between syntax and prosody in more naturalistic language samples, our focus in this article, the interesting data are the numbers concerning the four linguistic conditions mentioned above. That is, the data allow us to estimate to what extent prosody and syntax align, as would be expected according to *both* direct and indirect theories of the prosody-syntax relationship, because of course, even on the indirect view, a syntactic right boundary should always be prosodically marked because it defines the right boundary of a PPh. The data also allow the extent of misalignments between prosody and syntax to be assessed.

The linguistic analyses Degano et al. (2024) performed as preparation for their MEG study reveal the following with respect to prosody-syntax alignment. For all the words in this corpus of four TED talks that they analyzed, 56% of them were coherent in that a weak syntactic boundary was associated with a weak prosodic boundary, and 72% were coherent in that a strong syntactic boundary was associated with a strong prosodic boundary. As the data were exhaustively dichotomized, the reader can infer the extent of misalignment: Since 56% of data showed a weak prosody-weak syntax alignment, that implies that, for 44% of the words, a strong syntactic boundary was not prosodically marked. Similarly, since 72% of the data showed a strong-strong alignment, 28% of the words were ones in which a prosodic boundary occurred with no syntactic boundary. This latter misalignment is not problematic for any theory of the prosody-syntax relation, because prosodic boundaries may arise from a number of sources beyond syntax. For example, an IPh boundary may occur after the main verb of a sentence even

though there is no closing syntactic phrase (Selkirk, 1984), as illustrated in (2). A pitch accent on a word generated for emphasis (e.g., to focus the word) will also often result in a prosodic phrase boundary (Ferreira, 1993; Selkirk, 1984). But the complementary case is highly problematic for competence-based theories of prosodic phrasing, because those are words that end on a syntactic phrase where that boundary is not prosodically marked. Recall that earlier we considered the importance of the prosody-syntax relationship for theories of language comprehension, noting that prosodic marking of syntactic phrases might allow the parser to perform its job more easily. Indeed, again, the Degano et al. (2024) study reveals that the brain more easily decodes syntactic phrases when they are prosodically marked. But the 56%/44% split reveals that it is almost a coin toss whether a syntactic phrase will be prosodically marked or not. This pattern would suggest that prosody is a far less reliable cue to syntax than has been assumed. It is also problematic for theories of the prosody-syntax relationship that assume a method for generating PPhs that depends on aligning the boundaries of syntactic phrases with the right edges of PPhs algorithmically.

A skeptic could quibble with these results. Although TED talks are somewhat naturalistic in that they are language samples produced for purposes other than testing linguistic hypotheses, they are certainly not spontaneous or unrehearsed. (Indeed, the TED website provides advice to speakers on how to practice their talks to ensure they are delivered fluently and engagingly: https://www.ted.com/pages/rehearsals.) However, this observation only makes the concern about the extent to which prosody and syntax align in naturalistic speech more severe: Since the talks are practiced and are meant to be clear and aesthetically appealing, the productions are probably closer to the competence-based ideal than is typically found in more spontaneous situations. Practice would lead one to expect stronger prosody-syntax alignment in the resulting speech samples, not weaker. It would therefore seem useful to examine unprepared, spontaneous speech, as doing so might provide an even better (and likely lower) estimate of the extent to which prosody and syntax align naturally.

6. Investigations of Naturalistic Language from our Laboratory

Our laboratory has recently embarked on a research program to determine the prosodic features of naturalistic speech, focusing specifically on prosodic phrasing and its alignment with syntactic structure. As part of another project examining how speakers decide how to linearize their ideas into a sequence of orderly utterances—that is, how they solve the linearization problem (Ferreira & Rehrig, 2019; Levelt, 1981)—we have collected a large corpus of scene descriptions. Our participants were shown photographs of real-world indoor and outdoor scenes and were given 30 seconds to describe each one. Each participant described 30 scenes, and 30 participants were tested in each study. Thus far four such studies have been conducted: In one, 30 college students simply described each of the 30 scenes; in a second, 30 college students did the same task but

were required to begin generating their description as quickly as possible (thus depriving them of initial planning time); in a third, 30 college students described what actions a person might perform in the scene; and in the fourth, 30 older adults aged 60-85 described the scenes in whatever way they wished. Thus, with four studies in which 30 subjects each described 30 scenes for 30 seconds each, we have a corpus of 1800 minutes of unrehearsed, spontaneously generated speech.

We must acknowledge that the speech in this corpus is not conversational: that is, the speakers did not interact with any sort of interlocutor. In future work we hope to apply these same analyses to conversational, interactive speech. However, what is useful about these speech samples is that the 30 second stretches of speech allow for long enough utterances to permit some degree of syntactic variation and complexity to emerge (this is especially true for the study in which speakers described actions one might perform in a scene). A recent investigation of a Dutch, German, and American corpus of conversational speech has revealed that turns in interactive contexts are typically brief, often as short as one or two words (Corps et al., 2022). Such short utterances would seem to allow for little syntactic structure, making it difficult to evaluate any prosody-syntax relationship. We therefore think our scene descriptions are a good place to begin with this research program, and we hope that in the future we will be able to examine conversational language in contexts that encourage longer turns (and, we hope, more syntactic variation), as might be found in conversations about science or popular culture.

Returning to our investigations of prosody-syntax alignment, our aim was to conduct the same analyses as those reported by Degano et al. (2024). Our first step was to transcribe all the utterances from the scene descriptions naïve subjects generated using the tool Whisper (https://openai.com/index/whisper/). The transcriptions contained no punctuation. Next, we used spaCy to parse the speech into syntactic constituents, and we used the Wavelet Prosody Toolkit to analyze the prosodic features associated with every word of the descriptions. Then, exactly as in Degano et al., for each utterance produced in each of the four studies, we dichotomized the words that were produced according to the representations generated from Spacy and the Wavelet Prosody Toolkit. For syntactic features, words were binarized into two types, words with weak syntactic boundaries (defined as in Degano et al. as words with no closing right bracket, based on the Spacy parses) or words with one or more syntactic boundaries. We dichotomized the data for the prosodic features using the method reported in Degano et al., which involved taking the prosodic distribution of the data and using a two-Gamma mixture model to establish two latent distributions, those with a weak prosodic boundary and those with a strong prosodic boundary. These procedures yielded the same four conditions as in Degano et al.: cases in which prosody and syntax aligned (weak syntactic and weak prosodic boundary; strong syntactic and strong prosodic boundary), and cases in which they did not (weak syntactic boundary with a strong prosodic boundary, and strong syntactic boundary with weak prosodic boundary). As mentioned

above, it is the cases in which we find a syntactic boundary that is not marked prosodically that present a challenge for classic theories of the prosody-syntax relationship. Note that although these analyses treated the two variables as dichotomous, we also can treat prosody and syntax as continuous variables.

The results of our analyses for the four studies are as follows. For the study in which college students simply described the scenes for 30 seconds, we observed that 53% of words showed the expected prosody-syntax alignment: weak syntactic boundaries with weak prosodic boundaries and strong syntactic boundaries with strong prosodic boundaries. Therefore, 47% of the time, prosody and syntax were misaligned. This means that almost half of all words ending on a strong syntactic boundary were not prosodically marked. For the study in which college students did the same task but were required to begin speaking quickly, the results were essentially the same: 55% and 54% of words showed alignment, meaning that 45% and 46% of words were misaligned. For the study in which subjects provided action descriptions of the same scenes, 55% and 57% of words were aligned (weak syntax with weak prosody and strong syntax with strong prosody, respectively), and so 45% and 43% of words were misaligned. And when older adults described the scenes in any way they wished, 52% and 54% of words showed the expected prosody-syntax alignment, meaning that 48% and 46% of words did not. We thus show a pattern that is remarkably similar to what Degano et al. (2024) reported: Nearly half the time, words with a strong syntactic boundary were not prosodically marked.

This section as well as Section 4 describing the Degano et al. (2024) study suggest that the alignment between prosody and syntax is not as strong as we might expect it to be based on traditional, competence-based theories. Specifically, only about half of all words that close out a major syntactic phrase are prosodically marked so as to align with the boundary of a prosodic phrase. In the next section, we consider how this pattern might arise given performance constraints that influence language production.

7. Implications for Language Production and Language Comprehension

It appears that, if we adopt a more performance-based view of the prosody-syntax relation, and particularly if we examine unrehearsed, naturalistic language samples, we find results that are difficult to reconcile with either the direct or the indirect view. In our analyses, nearly half of all words that ended a syntactic phrase were not prosodically marked. On both the direct and indirect views, this pattern should not be possible because the right bracket of a syntactic phrase marks the right edge of a PPh, and a PPh is associated with the set of features including lengthening, energy, and pitch variation that is associated with prosodic boundaries. It appears, then, that at least these particular competence-based theories that derive predictions about prosodic features from syntactic or prosodic constituency alone are not adequate to explain the prosodic patterns that are observed in naturalistic language.

An alternative to both the direct and indirect views would be to adopt a more performance-based approach that assumes prosody reflects the speaker's need to plan and recover during normal language production, rather than any type of linguistic structure. As discussed in Section 3, speakers have to breathe, they need to manage their working memory resources, and they must take the time required to translate their thoughts into words and settle on an ordering for those words. It is likely that breaking speech up into small groups would help to coordinate these processes. For example, a speaker may decide to produce a description of a dining table that has been set for dinner, but the actual plan loaded into working memory can accommodate perhaps only seven plus or minus two words. To produce the next plan, working memory must be cleared and another set of words and phrases loaded in. This process of clearing and loading working memory may require more time than the time it takes to articulate the current linguistic chunk, and so to obtain more time, the speaker may slow down at the end of that production chunk. This slowing down would result in lengthening of the final syllable or two, and perhaps also the insertion of a pause (and possibly also a filler such as *um*). Of course, sometimes the working memory chunks will correspond to linguistic constituents such as syntactic phrases, but because language is produced incrementally, often speakers will find themselves having to insert a break in a less than optimal grammatical location. These performance factors may be what underlies the classic production of sentences such as *This is the cat that chased the rat that ate the cheese*, where it has been noted since Chomsky and Halle (1968) that the preferred "prosodic" or performance-based phrasing *(This is the cat) (that chased the rat) (that ate the cheese)* does not align with the sentence's syntactic structure. One explanation for this preferred prosodic phrasing may be that the speaker's working memory can accommodate only one proposition at a time (that there is a cat; that the cat chased a rat; that the rat ate the cheese, etc.), and after each proposition is linguistically encoded and articulated, the speaker needs time to purge working memory and load in new content. If the speaker had the ability to plan the entire utterance and accommodate all the associated linguistic representations in working memory, then it is less likely a boundary of any type would be required, which would allow the prosody to more accurately mirror the right-branching syntactic structure. As we know, however, working memory is severely constrained, and therefore most people in most communicative situations will lack the capacity to engage in long-range language planning, at least to this degree. As a result, a prosody-syntax mismatch will tend to occur.

Note that including performance constraints in our theories of the prosody-syntax relation does not require the rejection of all competence-based approaches. It is quite possible and indeed likely that both linguistic constituency and performance constraints play a role in determining the prosody of an utterance. One consideration is that the notion "syntactic phrase" is fairly flexible: a sentence such as *The hungry cat chased the rat* can be viewed as one large constituent, the entire clause; but at the other extreme, the word *hungry* can be parsed as an adjectival

phrase, which would allow for a phrase boundary after that single word. Thus, even a phrasing as awkward as *(The hungry) (cat) (chased the rat*) could be seen as derived from a PPh parsing derived from the syntactic structure. In response to this argument, someone might point out that most theories of the syntax-prosody interface do not treat prenominal adjectives as "full" syntactic phrases (e.g., Selkirk, 1984), but that is a stipulation imposed to prevent the kind of phrasing highlighted here, since it sounds so off. In addition, because the boundaries of IPhs need not coincide with syntactic phrases, allowing for intonational phrasing such as *(Marijus ate) (all of the chocolates)* (Ladd, 1996; Selkirk, 1984), the presence of prosodic features typically associated with the right edges of phonological constituents such as PPhs or IPhs need not imply a disrespect for linguistic structure. In other words, the processing system that must manage constraints related to performance constraints may have the capacity to take advantage of this linguistic flexibility to insure that the break locations that are generated during production do end up aligning with some type of linguistic and even syntactic boundary.

What remains problematic, however, is the finding from our studies and Degano et al. demonstrating that about half of all syntactic right boundaries are not prosodically marked—that is, the words at the right edges of those syntactic and PPhs have a prosodic boundary strength of zero or near zero. As mentioned earlier, the prosody-syntax misalignments can be of two types: a prosodic boundary is inserted with no syntactic boundary, and a prosodic boundary is omitted despite a syntactic boundary. The first type of misalignment is really not a misalignment at all, since a syntactic right boundary is a sufficient condition for a prosodic boundary, but it is not necessary—a prosodic boundary could be inserted for reasons other than syntax (e.g., when a word is focused and receives a pitch accent). It is the second type of misalignment that presents a serious challenge, because if the language production system obeys the algorithms for generating PPhs from syntactic structure, it must insert a boundary at the right edge of a syntactic phrase. As we have seen, nearly half the time, it does not.

This specific form of misalignment between prosody and syntax is also problematic for theories of language comprehension which assume that prosodic information is used to parse sentences—that is, that the acoustic features associated with prosodic boundaries allow the parser to accurately posit syntactic phrases. In response to this concern, some might point to the literature appearing to demonstrate that a prosodic boundary allows the parser to avoid a syntactic misanalyses—for example, to avoid so-called "garden paths", where the parser assumes a wrong syntactic parse which must be repaired. In one comprehension study (Engelhardt, Ferreira, & Patsenko, 2010), the ambiguity illustrated in *While the woman cleaned the dog that was big and brown stood in the yard* was assessed to determine whether prosody could prevent listeners from experiencing a garden-path—that is, misanalyzing the phrase *the dog that was big and brown* as the object of *cleaned*, requiring a reanalysis of that phrase as the subject of *stood* once that main clause verb was encountered. Intuition tells us that a prosodic boundary after *cleaned*

would help make clear that *cleaned* terminates a major syntactic and prosodic phrase (a clause and a PPh/IPh respectively), which drastically lowers the temptation to treat *the dog…* as the object of *cleaned*. The results supported this intuition: Listeners understood the sentences better with a prosodic boundary after *cleaned* than without one, and their pupil diameters showed less dilation, indicating the sentence with the boundary took less cognitive effort to interpret.

This result (and many others like it) would seem to establish clearly that prosody is used by listeners to establish syntactic boundaries. However, an earlier report puts an important qualification on this conclusion based on the investigators' consideration of both production and comprehension in the same study (Snedeker & Trueswell, 2003). The researchers did demonstrate that listeners used prosody to help them disambiguate sentences such as *Tap the frog with the flower*, where the noun phrase could refer either to a frog holding a flower or to an empty-handed frog that the listener was supposed to tap using a flower as the tapping instrument. Lengthening of *frog* plus a pause after it tended to encourage the latter interpretation, as revealed in eye movements to various objects, including frogs with and without flowers. However, the investigators also explored to what extent speakers provided those prosodic cues, and they reported that speakers did so only when they were made aware of the syntactic ambiguity. This study is an excellent demonstration of one of the themes of this article: If speakers are asked to produce an ideal rendition of a sentence, and, as was done in this study, are explicitly told about the ambiguity and asked to make an effort to eliminate it prosodically, they certainly can do so, in line with the linguistic competence that they possess. But if the speaker generates the sentences without awareness of the ambiguity, those cues are often not included. This result reinforces the idea that language generated in situations that do not draw specific attention to sentence form often results in productions that do not conform to what we'd expect to see according to more competence-based theories. In other words, people can intentionally prosodically disambiguate a sequence of words if they are asked to do so, but left to their own devices and in situations that are less conducive to meta-awareness, they often do not.

8. Conclusions

Although the investigation of prosody in language production and comprehension is a topic whose significance to linguistics and psycholinguistics is acknowledged by almost everyone working in the field, the literature on the topic remains sparse. One major reason why prosody has received less attention than, for example, syntax, is that it appears to be much harder to study prosody due to the necessity to perform almost all data analyses manually. This constraint also meant that studies tended to be small-scale, with data usually coming from practiced and trained speakers whose experimental task was to produce idealized renditions of sentences created for the purposes of testing linguistic hypotheses. The stimuli and experimental situations were fairly unnatural, and the productions likely reflected speakers' intuitions about how a sentence should

be said under idealized conditions, and with the intention of producing versions whose meaning would be as clear as possible to any listener. Because the experimental contexts in which these methods are used differ substantially from normal speaking situations, the prosodies elicited using this empirical paradigm tend not to reflect performance factors, but instead highlight the influence of syntactic and prosodic constituent structure.

The advent of new methods for creating large corpora, automatically transcribing speech, parsing speech into syntactic constituents, and automatically extracting prosodic information liberates researchers from these constraints. Using these new methods, it is possible to examine how people generating unrehearsed language in normal speaking situations produce prosody, and those methods permit the investigation of the role of performance factors related to the speaker's need to manage biological and cognitive resources. The argument in this article is that analyses of these more naturalistic stimuli reveal patterns that differ in startling ways from those that emerged from classic studies focused on the role of constituent structure. The assumption that speakers reliably mark syntactic boundaries—an assumption made by linguists and psycholinguists, and by researchers investigating both language production and comprehension—is called into serious question. Based on the data our lab has analyzed, it appears that about half of all syntactic boundaries are not marked, which means that speakers do not seem to obey the mandates of syntactic or prosodic constituency in their productions, and that listeners do not have access to the acoustic cues to syntactic structure that have been widely assumed to exist and to be exploited during comprehension.

What comes next is to determine why these discrepancies from what would be predicted from competence-based models are observed and what they mean for both linguistic theories of prosody and psycholinguistic theories of processing. Key issues include determining under what conditions syntactic boundaries are marked and under what conditions they are not, and figuring out how listeners adapt to the variability in syntactic phrase marking via prosody. In other words, these findings open up more avenues for research, which should help the field move forward in its quest to develop a more ecologically valid psycholinguistics that reflects how people really talk and how people understand real language.

Funding Information

This work was funded by the National Institute of Health/National Institute of Child and Human Development (grant number R01 HD100516) awarded to Fernanda Ferreira (Principal Investigator).

Competing Interests

The author has no competing interests to declare.

References

Bennett, R., & Elfner, E. (2019). The Syntax–Prosody Interface. *Annual Review of Linguistics*, *5*(1), 151–171. https://10.1146/annurev-linguistics-011718-012503

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* Harper & Row.

Cooper, W. E, & Paccia-Cooper, J. (1980). *Syntax and speech.* Harvard University Press. https://doi.org/10.4159/harvard.9780674283947

Corps, R. E., Knudsen, B., & Meyer, A. S. (2022). Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, *223*, 105037. https://doi.org/10.1016/j.cognition.2022.105037

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Degano, G., Donhauser, P. W., Gwilliams, L., Merlo, P., & Golestani, N. (2024). Speech prosody enhances the neural processing of syntax. *Communications Biology*, *7*(1), 1–10. https://doi.org/10.1038/s42003-024-06444-7

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology*, *63*(4), 639–645. https://doi.org/10.1080/17470210903469864

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological review*, *100*(2), 233-253. https://doi.org/10.1037/0033-295X.100.2.233

Ferreira, F., & Rehrig, G. (2019). Linearisation during language production: Evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, *34*(9), 1129–1139. https://doi.org/10.1080/23273798.2019.1566562

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

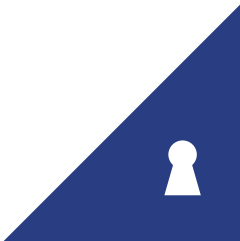
Kallay, J. E., Mayr, U., & Redford, M. A. (2019). Characterizing the coordination of speech production and breathing. *International Congress of Phonetic Sciences*, *2019*, 1412–1416. https://doi.org/10.1121/1.5067628

Ladd, D. R. (1996). *Intonational phonology*. Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. MIT Press.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405. https://doi.org/10.1017/s0952675700000695

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130. https://doi.org/10.1016/s0749-596x(02)00519-3