Co- and pro-speech integration: The parsing hypothesis*

Robert Pasternak
Leibniz-Center for General Linguistics (ZAS)
mail@robertpasternak.com

April 23, 2021

A growing body of research investigates the semantic and pragmatic contributions of *co-speech gestures* (gestures coinciding with spoken utterances) and *pro-speech gestures* (gestures occupying the role of normally spoken constituents). A critical observation about such gestures is that their semantic content is *integrated* into the meanings of spoken utterances: gestures can perform much of the same kinds of semantic work as speech, including scoping under spoken logical operators. One possible way of effecting this semantic integration is by treating gestures as part of the grammar: gestures are semantically integrated because they appear in syntactic representations (the **Grammatical Hypothesis (GH)**). However, I argue that GH requires some strange and undesirable stipulations about the nature and function of the grammar, in particular when considering the substantial variety of meaningful content beyond gestures that can be similarly integrated. I propose an alternative to GH, the **Parsing Hypothesis (PH)**, according to which gestures are not a part of the grammar of a language like English; instead, it is the human parser that is responsible for integrating co- and pro-speech content. I then provide a toy formal grammar and parser illustrating what a PH-friendly theory of co- and pro-speech integration could look like.

Keywords: gesture; co-speech content; pro-speech content; integration; parsing

1 Introduction

Recent years have seen a growing interest among formal semanticists in the meaningful contributions of *co-speech gestures*, semantically contentful gestures aligned with speech (see, e.g., Lascarides & Stone 2009; Ebert & Ebert 2014; Anvari 2017; Tieu et al. 2017, 2018; Schlenker 2018a,b; Esipova 2018, 2019; Zlogar & Davidson 2018; Hunter 2019). This is exemplified in (1), in which the upward-pointing gesture UP coincides with the verb phrase *used the stairs*, generating an inference that Mary used the stairs in an upward direction.

(1) Mary [used the stairs]$_{\text{UP}}$. (≈ 'Mary went up the stairs.')

Theories vary widely in their accounts of inferences from co-speech gestures, but practically all share the crucial feature that gesture-derived semantic content can be *integrated* into the meaning of the spoken utterance with which it coincides. Thus, UP in (1) appears to semantically function like a VP-adjoined adverbial on a par with spoken *upwards*, and can even take scope under higher operators like negation, modals, and quantificational subjects.

* Acknowledgments to be added later. **Comments are welcome, and much appreciated.**

A more extreme example of gesture integration would be what Schlenker (2018a,b) calls *pro-speech* gestures, in which gestures seem to "stand in" for otherwise spoken constituents. This is illustrated in (2), in which the non-manual disgusted face gesture DISGUST does not *coincide* with a spoken adjective phrase (AdjP), but instead seems to in some sense or another *be* the AdjP:

(2) The soup was DISGUST and smelled like feet.
(≈ 'The soup was disgusting and smelled like feet.')

It is hard to conceive of pro-speech gestures as anything other than integrated. With co-speech gestures, one can typically derive a sensible interpretation while ignoring the gesture entirely: for example, (1) is perfectly fine without UP, and UP only serves to further enrich the interpretation. With pro-speech gestures, however, this is often not the case: DISGUST in (2) plays a crucial role in deriving a well-formed semantic interpretation in the first place.

When presented with the phenomenon of co- and pro-speech integration, the empirical nature of which is discussed in more detail in Section 2, a natural follow-up question would be to ask what the cognitive mechanisms are that facilitate this semantic integration. One route would be to treat gestures as part of the grammar of spoken languages like English: in some form, gestures can participate in the syntax. On this view the semantic integration of gestures is unsurprising and straightforward: in (1) UP does the semantic work of a VP-adjoined modifier because UP *is* a VP-adjoined modifier, and in (2) DISGUST does the semantic work of an AdjP because it *is* an AdjP. The differences between gestures and speech are then mostly reducible to phonology, with some parts of a syntactic structure being phonetically externalized as speech and some being externalized as gesture (see, e.g., Esipova 2019). I will call this the **Grammatical Hypothesis (GH)** of co- and pro-speech integration.

In Section 3 I discuss GH in more detail, including what I take to be its principal shortcoming: namely, that when seen to its natural conclusion it requires strange and undesirable stipulations about the nature and function of the grammar. This is because the class of meaningful content that can be integrated in a gesture-like fashion extends far beyond gestures themselves, including things like pre-recorded sound effects (Pasternak 2019), emoji (Pierini to appear), videos, signage, and even sensations of pain. Much of this content cannot plausibly be treated as the phonetic externalization of some cluster of syntactic features, as would be required by GH: human language (morpho)phonology surely does not take some collection of syntactic features and output, say, a video of an explosion. But this leaves us in a sticky situation, since the grammar seems like the only place where we could reasonably define those compositional relations like modification and scope that co-/pro-speech content clearly takes part in. After all, if these compositional-semantic relations are determined at the syntactic level of Logical Form (LF), then it stands to reason that if co-/pro-speech content can take part in such relations it is because it can be a part of LF representations (and thus syntactic structures more generally).

In the rest of the paper I offer an alternative account of co- and pro-speech integration, which I refer to as the **Parsing Hypothesis (PH)**. According to PH, gestures do not appear in syntactic representations as generated by the grammar proper, at least for a spoken language

like English. Instead, it is the human parser that is responsible for integrating gestures and other content. In the case of co-speech content, as the parser builds a semantic interpretation based on the structure it assigns to the string of words in its input, it simultaneously folds the interpretation of the co-speech content into the constructed meaning. And in the case of pro-speech content, the parser does still more work: while a gesture like DISGUST is not an AdjP as far as the grammar proper is concerned, the parser is capable of treating pro-speech DISGUST *as if* it were an AdjP, and can parse its input accordingly. Viewed from the perspective of PH, the diversity of meaningful content that can be semantically integrated (both co- and pro-speech) should not come as a surprise, as there is no requirement that that content be the phonetic realization of syntactic material: instead, the parser is simply capable of integrating the meanings of various kinds of sensory input into the interpretation it builds over the course of a parse.

After introducing the conceptual basis for the Parsing Hypothesis, I next turn to the task of defining a formal model that instantiates PH. I start with a (gesture-free) formal grammar that generates syntactic structures and their interpretations, as well as a formal parser that takes strings generated by the grammar and determines their structure and semantic interpretation (Section 4). I then turn to co-speech content integration and define what I call an **augmented parser**, which takes an input consisting of both a spoken string and co-speech content and parses the former while semantically integrating the latter (Section 5). In Section 6 I turn to pro-speech content, which requires slight revisions to the augmented parser as introduced in the previous section. It is worth noting that the discussion in these three sections is aimed at readers with no background in formal parsing whatsoever. Readers with more of a background in parsing can likely skim Section 4 with little loss of understanding, in order to get to the more novel material in Sections 5 and 6.

The formal model introduced in Sections 4–6 should be thought of more as a toy model and proof of concept than as a finalized analysis. The point of the exercise is to show, in a manner that is friendly to readers without expertise in formal parsing, that it is possible to define a formally rigorous and predictive theory of co- and pro-speech content integration that does not require that content to appear in the grammar proper, and instead integrates it at the level of the parser. As a result, the analysis offered in this paper includes several substantial simplifications and abstractions that ought to be done away with in future work. I will discuss these simplifications and abstractions both at the point at which they arise and in the conclusion (Section 7), and I will also discuss how future work can do away with them in order to establish a more complete PH-friendly theory that more closely adheres to standard assumptions in the syntactic-semantic literature.

2 What integration looks like

Consider again example (1) with co-speech UP, repeated below:

(1) Mary [used the stairs]$_{\text{UP}}$. (≈ 'Mary went up the stairs.')

The interpretation we end up deriving for (1) is roughly the same as that for the gesture-less

sentence *Mary used the stairs upwards*. Since the only conceivable source for the "upwardness" inference is the gesture UP, it appears that the interpretation of this gesture is somehow integrated into the meaning of the sentence with which it coincides (though later discussion will highlight some trickiness here). Two questions raised by examples like (1) are (I) what is the semantic-pragmatic impact of gestures (e.g., what *kinds* of inferences do they generate?), and (II) what cognitive mechanisms facilitate this integration in the first place? Before discussing (II) it will help to discuss (I), which has received more attention in the formal semantic literature. The point of this discussion is not to reach a definitive conclusion about the semantic-pragmatic contributions of gestures. However, it will be useful to have in the back of our minds a rough picture of the hypothesis space, and along the way we will see some empirical observations that further support the claim that gestural interpretations are integrated into the meanings of spoken utterances.

In light of the convenient paraphrase of (1) as *Mary used the stairs upwards*, a reasonable temptation would be to treat the interpretation of co-speech UP as an at-issue VP modifier akin to the interpretation of *upwards*. However, Ebert & Ebert (2014) observe that the semantic interpretations of gestures are often not-at-issue. Consider (3), where the gesture LARGE is one in which the hands are used to iconically indicate a large (tall) bottle:

(3) I brought [a bottle of water]$_{\text{LARGE}}$.

If the semantic contribution of LARGE were at-issue, one would expect (3) to have roughly the same meaning as (4):

(4) I brought a large bottle of water.

But notice that if I believe that you actually brought a small bottle, I can respond to (4) with a direct denial like (5a), but this is harder with (3), which seems to require a *Wait a minute!*-type discourse-interrupting protest like (5b). This suggests that the contribution of LARGE is in fact a not-at-issue inference.

(5) a. That's not true! You actually brought a small bottle.
b. Wait a minute! You actually brought a small bottle.

Using this and other evidence, Ebert & Ebert (2014) conclude that co-speech gestures behave like supplements (e.g., non-restrictive relative clauses), which similarly introduce not-at-issue inferences:

(6) I brought a bottle of water, which (by the way) was large.

Schlenker (2018a,b) agrees with Ebert & Ebert (2014) that co-speech gestures contribute not-at-issue meanings, but disagrees about the nature of these meanings. Suppose that the interpretation of *use the stairs* is V, and the interpretation of UP is G. Schlenker argues that *[use the stairs]*$_{UP}$ introduces a presupposition of the form $V \Rightarrow G$ (where $\Rightarrow$ indicates type-generalized entailment), which he calls a *cosupposition*. Thus, the assertive component of the meaning of (1) is simply that Mary used the stairs, while the gesture contributes a cosupposition roughly stating that using the stairs entails going upwards. His main argument

for this conclusion comes from projection. (Ebert & Ebert (2014) also consider projection in their argumentation, but they do not consider as wide a range of projection environments as Schlenker does.) Consider, for example, the negated (7):

(7) Mary did not [use the stairs]$_{\text{UP}}$.
$\rightsquigarrow$ If Mary had used the stairs, she would have gone upwards.

The default interpretation of (7) does not seem to be the same as that of (8):

(8) Mary did not go up the stairs.

Whereas (8) is compatible with Mary either going down the stairs or not taking the stairs at all, (7) seems to require that Mary not have taken the stairs full stop, and that *if she had, she would have gone upwards*. This is precisely what one would expect on a cosuppositional account, since the conditional presupposition ($[\![\text{use the stairs}]\!] \Rightarrow [\![\text{UP}]\!]$) would be expected to project through negation. Schlenker goes through other kinds of environments (e.g., the scope of epistemic modals like *might* and quantifiers like *each*, *exactly one*, and *none*), and argues that in each case the conditional inference projects as one would expect if it were a presupposition. Tieu et al. (2017, 2018) provide experimental support for Schlenker's claims, furnishing evidence that linguistically naïve speakers share Schlenker's judgments.

An advantage to treating co-speech gestural inferences as cosuppositional is that presuppositions can be *locally accommodated*—roughly speaking, they are reevaluated as non-projecting and at-issue (Heim 1983). For example, (9a) generates a presupposition that Harvey previously smoked, thanks to the lexical semantics of *stop*. This inference projects through negation, hence (9b) retains the inference that Harvey previously smoked.

(9) a. Harvey stopped smoking.
$\rightsquigarrow$ Harvey previously smoked.
b. Harvey didn't stop smoking.
$\rightsquigarrow$ Harvey previously smoked.

But in (10) we no longer retain this inference: the "used to smoke" inference is rendered at-issue through local accommodation, and then negated.

(10) Harvey didn't stop smoking, since he never smoked to begin with.
$\not\rightsquigarrow$ Harvey previously smoked.

One can get what appear to be locally accommodated readings of co-speech gestures fairly easily, for example in Schlenker's (11):

(11) If you bring a [bottle]$_{\text{SMALL}}$, I'll be disappointed, but if you bring a [bottle]$_{\text{LARGE}}$, I won't be.
(≈ 'If you bring a small bottle, I'll be disappointed, but if you bring a large bottle, I won't be.')

Here the size inferences are presumably locally accommodated because projection leads to contradiction: projection from the first antecedent would lead to an inference that any bottle you bring would be small, while projection from the second antecedent would lead to

an inference that any bottle you bring would be large. To prevent this contradiction, the size inferences are (presumed to be) locally accommodated, leading to the observed reading that whether or not I am disappointed depends on the size of the bottle you bring. While a cosuppositional analysis allows for such local accommodation, this is not available for supplements, as evidenced by the ill-formedness of (12):

(12) # If you bring a bottle, which would be small, I'll be disappointed, but if you bring a bottle, which would be large, I won't be.

With this in mind, Schlenker proposes that co-speech gestures are *weak* presupposition triggers: cosuppositions are easily locally accommodated and thereby interpreted as at-issue. Thus, a cosuppositional analysis seems to make the right predictions about the projection of gestural inferences when those inferences do project, while at the same time correctly allowing for the possibility of an at-issue interpretation by way of local accommodation.

Finally, Esipova (2019) offers an interesting alternative to Schlenker's cosuppositional analysis that makes many of the same predictions in relevant cases, but that avoids the stipulation of a kind of presupposition unique to gestures. In short, Esipova claims that at the level of the compositional semantics, gestures compose in exactly the same way as their spoken counterparts. More specifically, she argues that co-speech gestures that are interpreted as modifying NPs like *bottle* or VPs like *use the stairs* compose like spoken restrictive modifiers: on a purely semantic level, the result of composing ⟦use the stairs⟧ with ⟦UP⟧ is roughly the same as the result of composing ⟦use the stairs⟧ with ⟦upwards⟧. Meanwhile, co-speech gestures that are interpreted as modifying DPs like *the bottle* compose as supplements, much like other DP modifiers like non-restrictive relative clauses.

Initially, the claim that NP- and VP-modifying gestures are compositionally treated as simple restrictive modifiers seems strange in light of the observations about projection and (not-)at-issueness discussed above. After all, didn't we just see evidence that *[used the stairs]*$_{UP}$ and *used the stairs upwards* behave differently with respect to projection and at-issueness? To account for this, Esipova notes—building on prior work by Bolinger (1967), Larson & Marušić (2004), Umbach (2006), Morzycki (2008), and especially Leffel (2014)—that just because a modifier is restrict*ive* does not mean that that modifier is restrict*ing*. Adapting an example of hers, consider the song 'Big Yellow Taxi', which was written by a single, female songwriter (Joni Mitchell). Given this world knowledge, the adjective *female* in the DP *the female songwriter of 'Big Yellow Taxi'* is still restrict*ive*—it still serves the compositional function of intersection with the NP *songwriter of 'Big Yellow Taxi'*—but it is not restrict*ing*, since the result of this intersection is the exact same (singleton) set. Esipova discusses non-restricting restrictive modifiers in various environments, and argues that when a restrictive modifier is intentionally used in a non-restricting fashion, the inference of "non-restrictingness" (e.g., that the sole songwriter of 'Big Yellow Taxi' is female) projects like a presupposition. For reasons of space I will not go through the evidence offered in favor of this claim; instead let us simply accept it as true and see what it can get us.

Esipova argues that in the case of NP- and VP-modifying co-speech gestures the gesture is interpreted as compositionally restrictive, but there is a pragmatic preference in favor

of interpreting the gesture as non-restricting. As an illustration, consider (1) again. What would it mean for the contribution of UP in this example to be non-restricting? It would mean that the truth conditions are the same with or without the semantic contribution of UP. That is, the set of worlds in the context set *C* in which (1) is true must be identical to the set of worlds in *C* in which *Mary used the stairs* is true. This will only be the case if for every world in *C* in which Mary used the stairs, she went up the stairs. But this, of course, is the same inference that Schlenker (2018a,b) referred to as a cosupposition. Moreover, if non-restricting inferences are determined relative to local contexts and project like presuppositions, the same projection behavior is also predicted. As for apparent cases of local accommodation, Esipova argues that this is not actually the result of local accommodation, but rather the presupposition not being triggered in the first place: the preference to interpret gestures as non-restricting is defeasible, and the at-issue interpretation arises by instead interpreting the gesture's contribution as both restrictive and restricting.

Notice that while Ebert & Ebert (2014), Schlenker (2018a,b), and Esipova (2019) disagree about the semantic and pragmatic nature of co-speech gesture inferences, a crucial feature that all three analyses share in common is that they posit that inferences from gestures can project. But projection and scope are of course intimately bound together: the claim that a gestural inference projects from the scope of some operator *O* only makes sense if the gesture makes its semantic contribution within the scope of *O* to begin with. Thus, in order for gestural inferences to project it must be possible for them to scope below logical operators in the spoken sentence with which they coincide. The same is presumably true of cases of *non*-projection like (11), regardless of whether that is through local accommodation or through a lack of triggering in the first place: either way, according to the analyses at hand the gesture makes its semantic contribution within the scope of some operator, but this time the contribution "stays put". Thus, for example, the size inference stays within the scope of the conditional antecedent in (11), rather than projecting.

What these observations suggest is the existence of what I will call **Non-Maximal Gesture Scope (NMGS)**:

(13) **Non-Maximal Gesture Scope (NMGS):**
There are cases in which a gesture G coincides with (some part of) a sentence S, and in which the semantic contribution of G scopes below some operator *O* in S.

NMGS means that gestures can be interpreted in the scope of spoken content. Or, somewhat informally, the semantic contributions of gestures are interwoven with the semantic contributions of spoken constituents. This possibility is an extremely important issue to address within any theory of gesture integration: how can gestures compose with spoken utterances in a way that allows them to take narrow scope?

It is worth noting that there are difficulties involved in definitively proving the truth of NMGS in the case of co-speech gestures. The main reason for this is that it is hard to prove what exactly the semantic interpretation of a gesture like UP is. For example, say that instead of NMGS we wish to defend what I will call the **Dull Hypothesis**, defined in (14):

(14) **Dull Hypothesis:**
For co-speech gesture G aligned with (some part of) spoken sentence S, the interpretation of S and G together is ⟦G⟧(⟦S⟧).

The Dull Hypothesis basically states that gestures can only make their semantic contributions at the (matrix) clausal level: their semantic contributions are not interwoven with those of syntactic constituents, but are merely tacked on at the end. Nonetheless, if we are willing to be sufficiently flexible about what ⟦G⟧ is, we can still derive many of the inferences discussed above. For example, in the simple case of (1) ⟦UP⟧ might be defined as in (15):

(15) a. $[\![\text{UP}_1]\!] = \lambda p.\ p \wedge$ Mary went upwards
b. $[\![\text{UP}_1]\!]([\![\text{Mary used the stairs}]\!]) =$ Mary used the stairs $\wedge$ Mary went upwards

Meanwhile, in the case of negated (8) the purported cosupposition can be simulated by assigning a different interpretation to UP in this environment:

(16) a. $[\![\text{UP}_2]\!] = \lambda p.\ p \wedge$ if Mary had used the stairs, she would have gone upwards
b. $[\![\text{UP}_2]\!]([\![\text{Mary did not use the stairs}]\!]) =$ Mary did not use the stairs $\wedge$ if Mary had used the stairs, she would have gone upwards

Or perhaps instead of modifying ⟦UP⟧ so severely, one could try and keep the interpretation relatively consistent and generate the effects through discourse relations (see, e.g., Hunter 2019). The point is that if one is willing to be flexible enough in the interpretations of gestures or in how they relate to the utterances with which they coincide, one might be able to capture a decent amount of the "projection" data without requiring that the gesture scope below spoken logical operators, thereby making NMGS false and rendering the problem of co-speech gesture integration more or less trivial.

However, there are at least two reasons to believe that the Dull Hypothesis cannot offer a satisfactory account across the board, and that in some cases the interpretation of the gesture really must take non-maximal scope. The first is something that has already been discussed: namely, instances in which gesture-derived inferences do *not* project (regardless of whether this is because of local accommodation or a lack of triggering in the first place). For example, it is difficult to conceive of plausible meanings for SMALL and LARGE that would generate the observed interpretation in (11) without taking non-maximal scope, and in particular without scoping inside the antecedents of their respective conditionals.

The second reason to suspect that the Dull Hypothesis will not cover all instances of purported gesture integration are *pro-speech* gestures, in which the gesture does not align with speech, but essentially plays the role of speech. This is illustrated in (2), repeated below:

(2) The soup was DISGUST and smelled like feet.
(≈ 'The soup was disgusting and smelled like feet.')

With examples like (2) it is hard to see how the Dull Hypothesis even gets off the ground. According to the Dull Hypothesis, we should first derive an interpretation for *The soup was* PAUSE *and smelled like feet* (where PAUSE indicates a pause in speech), and then the facial gesture DISGUST should modify this interpretation. But *The soup was* PAUSE *and smelled like feet*

is of course not a well-formed sentence in English, and it is difficult to see what its interpretation would be, how that interpretation would be derived, and how DISGUST would modify that interpretation in order to derive the meaning observed in (2). Furthermore, pro-speech gestures can compose in the scope of negation and other operators, as illustrated in (17):

(17) a. The soup was **not** DISGUST. It tasted fine.
b. A: Why did Alex return her soup?
B: I don't know. The soup **might** have been DISGUST.
c. **Every** bowl of soup they served was DISGUST.

Thus, by all appearances pro-speech gestures are fully semantically integrated parts of the sentences in which they appear, and can compose below logical operators like negation, modals, and DP quantifiers.

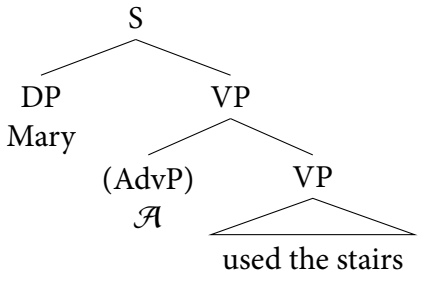
In summary, then, the evidence strongly suggests that the interpretations of gestures can be integrated into linguistic meanings. Moreover, gestures can be semantically integrated in a way that leads to them taking non-maximal scope: they can semantically compose below other operators (Non-Maximal Gesture Scope). Hence, co-speech gestural inferences can arguably project (a notion that only makes sense relative to scope) or contribute at-issue interpretations within the scope of other operators (through local accommodation or simply non-triggering), and pro-speech gestural inferences are fully integrated into the compositional semantics in the same way as the constituents they are "replacing".

3 Integration in the grammar or the parser?

So gestures can be semantically integrated with spoken material, including being able to compose within the scope of spoken operators. The natural follow-up question, and the central concern of this paper, is: how does this semantic integration actually occur?

Based on standard views about the nature of semantic composition, there seems to be only one reasonable answer here. If the semantic contributions of gestures can be compositionally interwoven with the semantic contributions of syntactic constituents, then gestures seemingly have to participate in those syntactic relations and operations that the compositional semantics is sensitive to. Take, for example, (1). It appears that the only way that UP can semantically modify the VP *use the stairs* to the exclusion of higher operators is if there is some bundle of features $\mathcal{A}$ that adjoins to the VP and has the appropriate semantic interpretation, and is such that the phonology, which is presumably modality-blind, pronounces $\mathcal{A}$ + VP as $[\text{VP}]_{\text{UP}}$. This is illustrated in (18):

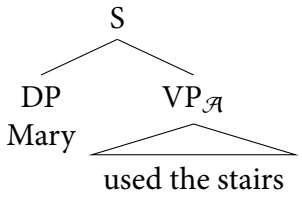
(18)



Phonology: Pronounce this as *Mary [used the stairs]*$_{UP}$

One does not have to encode this syntacticizing of co-speech gestures strictly as adjunction: the marking of the presence of the gesture can be effected in other ways as well. For example, the VP node could simply be "tagged" with the features $\mathcal{A}$:

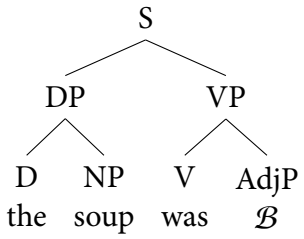
(19)



The phonology would then once again determine that VP$_{\mathcal{A}}$ is pronounced as [VP]$_{UP}$, and the compositional semantics would interpret $[\![\text{VP}_{\mathcal{A}}]\!]$ as $[\![\mathcal{A}]\!] \star [\![\text{VP}]\!]$, where $\star$ is whatever compositional operation one takes to be involved in co-speech gesture interpretation (e.g. supplements, cosuppositions, or "normal" semantic composition). The point is the same: in order for the interpretation of the gesture to participate in the compositional semantics in the ways that it needs to, the gesture has to participate in the syntax as well.

More or less the same thing would have to be done with pro-speech gestures. In the case of (2), the bundle of features $\mathcal{B}$ would be the AdjP (presumably, though other syntactic labels are conceivable), and the phonology would dictate that $\mathcal{B}$ is pronounced as DISGUST:

(20)



Phonology: Pronounce this as *The soup was* DISGUST

One noteworthy benefit of this analysis is that it predicts that pro-speech gestures standing in for some category XP can only appear in environments in which XP could otherwise appear (see, e.g., Esipova 2019). For example, (21), in which the predicative adjective and copula switch places, is not a well-formed sentence of English:

(21) * The soup disgusting was.

Likewise, (2) is ill-formed when the copula and DISGUST switch places:

(22) * The soup DISGUST was.

If $\mathcal{B}$ literally constitutes an AdjP in the syntactic tree, then the observation that DISGUST is linearized like an AdjP comes as no surprise.

Let us refer to this general hypothesis, in which gestures are semantically integrated by being some part of the syntactic representation and/or derivation, as the **Grammatical Hypothesis (GH)**. With the notable exception of Esipova (2019), GH has not to my knowledge been explicitly endorsed in the formal semantic literature on gesture integration. But this is tied to a much larger issue: namely, that the question of how exactly gestures are semantically integrated is often left tacit altogether. Instead, most work seems to simply assume that gestures can somehow be semantically integrated, shifting the focus to what happens post-integration (e.g., projection). Nonetheless, given an architecture of language in which the compositional semantics generates an interpretation from a syntactic structure, the observation that the interpretations of gestures can be interwoven with those of syntactic constituents seems to thereby require that gestures be part of that syntactic structure.

The claim that syntax is modality-blind, and thus that certain bundles of syntactic features are phonologized as gestures rather than as speech, is not inherently implausible. Gestures use the same articulators as sign languages (e.g., hands, mouth, eyebrows), which obviously have their own well-behaved phonology and phonetics. Thus, the notion that an individual's phonology and phonetics can encompass both speech and gesture, rather than just speech, is entirely within the realm of reason.

However, this view of the grammar becomes less plausible when one considers other types of content that are capable of the same kinds of integration as gestures. One example discussed by Pasternak (2019) are *co-speech sound effects*, contentful sound effects coinciding with recorded speech. An example can be seen in (23), where the explosion sound effect EXPLODE coincides with the verb phrase *assassinate his target*, generating an inference that the assassination will be by means of an explosion. (Recordings of all sound effect examples can be found in the supplemental materials at https://bit.ly/2QpfW2Y.)

(23) The soldier will [assassinate his target]$_{\text{EXPLODE}}$.
(≈ 'The soldier will assassinate his target by means of an explosion.')

Moreover, as noted by Pasternak and experimentally corroborated by Pasternak & Tieu (in revision), inferences from co-speech sound effects seem to project in the same fashion as gestures. (24) illustrates with negation: what we end up with is a cosupposition-like inference that *if* the soldier *were* to assassinate his target, it would be by means of an explosion:

(24) The soldier will not [assassinate his target]$_{\text{EXPLODE}}$.
$\rightsquigarrow$ If the soldier were to assassinate his target, it would be by means of an explosion.

Pro-speech sound effects also behave like their gestural counterparts, as illustrated in (25):

(25) The car will EXPLODE and make a mess of the street.
(≈ 'The car will explode and make a mess of the street.')

If GH is to be believed, then in an example like (23) there is some bundle of features $\mathcal{A}$ that is semantically interpreted roughly as "by means of an explosion". Either $\mathcal{A}$ is adjoined in the syntax to *assassinate her target*, or the VP node is "tagged" with $\mathcal{A}$, or something else along those lines. Then, the phonology and phonetics interprets this combination as *[assassinate his target]*$_{EXPLODE}$, i.e., as *assassinate his target* aligned with the pre-recorded sound effect EXPLODE. The same goes for (25): $\mathcal{A}$ is a VP in the syntax, and the phonology and phonetics externalize this as a standalone pre-recorded explosion sound effect. But the claim that some feature bundle $\mathcal{A}$ can be phonetically realized as an audio recording of an explosion is considerably less plausible than in the case of gestures.

Another example comes from co- and pro-text emoji, as discussed in detail by Pierini (to appear). Emoji are small images encoded as text, common in (usually informal) online communication. Consider (26), featuring the toilet emoji 🚽.[1]

(26) The student will
🚽 step out of the classroom 🚽
(≈ 'The student will step out of the classroom to use the toilet.')

(26) leads one to infer that the student will step out of the classroom to use the toilet. And again, as extensively discussed by Pierini (to appear) and experimentally corroborated by Pasternak & Tieu (in revision), the inference from 🚽 projects in the same manner as gestures. This is illustrated in the negated (27); again we derive a conditional inference that *if* the student were to step out of the classroom, it would be to use the toilet.

(27) The student will not
🚽 step out of the classroom 🚽
⇝ If the student were to step out of the classroom, it would be to use the toilet.

And pro-text emoji are just as much of a possibility as pro-speech gestures, as shown in (28):

(28) The student will 🚽 (≈ 'The student will use the toilet.')

Thus, emoji can also be semantically integrated, seemingly in the same ways as gestures. And again, GH would require that some bundle of syntactic features is semantically interpreted roughly as "use the toilet", and is phonetically realized (in text) as 🚽.

A pattern is emerging here. While a grammatical treatment of co- and pro-speech gestures has a certain plausibility to it, since the phonetic realization of syntactic constituents as hand and arm movements and non-manuals is already attested in sign languages, a grammatical treatment of sound effects and emoji is considerably less plausible: a theory of grammar in which syntactic features are phonetically externalized as EXPLODE or 🚽 pushes the

[1] Here I follow Pasternak & Tieu (in revision) in using "emoji bracketing". Pierini's examples involve a single emoji immediately following the sentence; the core observations remain the same. Emoji bracketing is less natural than Pierini's examples, but avoids the possibility of an ambiguity between co- and post-text emoji. (See Schlenker 2018a for discussion of differences between co- and post-speech gestures.)

bounds of credulity. This game can be played into further rounds, with additional kinds of co- and pro-speech content that are increasingly implausible candidates for a treatment as phonetic externalizations of syntactic objects. I will illustrate with pro-speech content, but similar results are obtainable with co-speech content.

Another type of pro-speech content that can be semantically integrated are speaker interactions with nearby objects. This is illustrated in (29), where (29a) features the speaker instructively cutting a flower, and (29b) features the speaker menacingly ripping the head off of a teddy bear.

(29) a. ***Context:*** *A gardening instructor is holding a flower and scissors and teaching students how to preserve their flowers.*
All you have to do is ACTION: CUT-STEM and keep it in a vase with water.

b. ***Context:*** *The villain is menacingly holding a teddy bear.*
If you betray me one more time, then ACTION: RIP-HEAD-OFF.

Videos—without audio, thereby ruling out an analysis in which they are merely sound effects with some accompanying visuals—can also be integrated. This is illustrated with the examples in (30): (30a) features a video of an explosion, while (30b) features a video of flowers blooming. (Files where these videos are integrated with speech can be seen in the supplemental materials at https://bit.ly/2QpfW2Y.)

(30) a. The car will VIDEO: EXPLODE.

b. The flowers will soon VIDEO: BLOOM.

For another example, suppose that we are playing a game where we each have several cardboard signs in front of us with drawings on them, and we are trying to use those signs as frequently as possible in our communication. In this context, either a sign or the act of choosing and holding up that sign can be semantically integrated as well, as illustrated in (31a) with the speaker holding up a sign with a picture of an explosion, and in (31b) with the speaker holding up a sign with someone laughing on it.

(31) a. The car will SIGNAGE: EXPLODE.

b. The student will SIGNAGE: LAUGHING.

Pro-speech integration is not even restricted to audiovisual content. (32) could come straight from an action film, though perhaps not a particularly good one:

(32) ***Context:*** *The villain has installed a neural implant in the protagonist without their knowledge. The apparatus is set up so that when the villain presses a hidden button under their desk, the implant fires an electrical signal that causes the protagonist brief but intense pain. With a finger surreptitiously on the button, the villain says:*

If you betray me again, you just might feel a sudden PAIN: IMPLANT-FIRE.
(≈ 'If you betray me again, you just might feel a sudden, intense pain.')

Here the communicative content is not some audiovisual stimulus, nor the pushing of the button (whose existence the protagonist is not aware of), but rather the sensation of pain that the protagonist feels. Yet once again, this communicative content can be fully semantically integrated, even being modified by an adjective (*sudden*) and scoping below a modal (*might*).

This game can be played more or less *ad infinitum*: practically anything that can be intentionally used to convey information can be semantically integrated with speech. And in many cases, a syntactic analysis along the lines of GH looks bizarre: can features in a syntactic tree really be phonetically externalized as the speaker ripping the head off of a teddy bear, or a sound recording or silent video of an explosion, or a neural implant painfully firing in the addressee's brain?

In summary, then, here is where we stand. Viewed strictly in terms of their capacity for integration with speech, there does not seem to be any motivation for drawing a clean boundary between gestures and other potentially meaningful content. All sorts of content can be integrated in both co- and pro-speech form, including many types of content that any reasonable account should not treat as part of the grammar, and in particular as a potential output of linguistic morphophonology. It therefore appears that an approach in which the semantic integration of co- and pro-speech content is effected through syntactic integration is undesirable. But this leaves us in something of a pickle: if integration does not take place in the syntax, then where *does* it take place? We have already seen evidence against the Dull Hypothesis, leading us to infer that gestures and other content can semantically compose with spoken content prior to the completion of the semantic derivation. It is in this sense that the semantic contributions of gestures and other content are "interwoven" with the contributions of syntactically realized material. But it seems like there is no way for this to actually occur *without* syntactic integration.

However, this puzzle hinges on a very specific and not at all obvious assumption: namely, that the empirical observation that gestures and other content can be semantically integrated tells us something about the grammar, or more generally about human linguistic *competence* (in the sense of Chomsky 1965). In the rest of this paper I will illustrate an alternative picture, the **Parsing Hypothesis (PH)**. According to PH, the observation that gestures and other content can be semantically integrated does not in and of itself tell us much about the nature of the *grammar*, but it does tell us a lot about the nature of the human *parser*. In other words, humans' ability to semantically integrate co- and pro-speech content is not a matter of linguistic *competence*, but of *performance*.

From a big-picture perspective, PH can be thought of in the following way. Roughly speaking, the goal of the human parser is to take some sensory input that includes linguistic material, and to build a syntactic structure and semantic (and possibly pragmatic) interpretation based on that sensory input. For that part of the sensory input that is determined to constitute linguistic material, the parser uses the grammar furnished by the individual's linguistic competence to build the structure and interpretation: put simply, an English-language parser makes use of an English-language grammar. But there is no inherent reason why the parser should confine itself to only the linguistic part of its sensory input when building an interpretation. Consider again the case of (1), *Mary [used the stairs]*$_{UP}$.

According to PH, *Mary used the stairs* is the only part of the sensory input in (1) for which the parser uses the individual's grammar to build a syntactic structure and semantic interpretation. But that does not mean that the parser ignores UP entirely: instead, it folds the communicative content of UP into the semantic interpretation that it builds over the course of a parse. As for pro-speech content, the parser has to do a bit more work: the pro-speech content is standing in the stead of what would otherwise be linguistic content, meaning that the linguistic part of the sensory input is incomplete without the pro-speech content. Thus, according to PH (or at least the version adopted here), when presented with pro-speech content the parser builds a structure and interpretation for its input *as if* the grammar in fact allowed non-linguistic material to serve as a VP, AdjP, etc. In other words, even though the grammar does not generate sentences like *The soup was* DISGUST, the parser can "pretend" that DISGUST is a legitimate AdjP and build a structure and interpretation accordingly. Thus, from the vantage point of PH it is unsurprising that all sorts of non-linguistic content are capable of semantic integration. The parser can semantically integrate more or less whatever meaningful content it wants, even if that content could not be generated by the grammar that the parser uses in analyzing the narrowly linguistic part of its input.

The analysis in this paper is part of a long lineage of attempts to tease apart which judgments about grammaticality and interpretation are due to competence, and which are due to performance. A classic example from Chomsky & Miller (1963) are center-embedded relative clauses, which become increasingly unacceptable with more iterations:

(33) a. The horse [$_{RC_1}$ that the farmer owns] left the stable.
b. ? The horse [$_{RC_1}$ that the farmer [$_{RC_2}$ that the boy greeted] owns] left the stable.
c. * The horse [$_{RC_1}$ that the farmer [$_{RC_2}$ that the boy [$_{RC_3}$ that the teacher likes] greeted] owns] left the stable.

A common view nowadays about the unacceptability of iterated center-embedding is that it is a matter of performance rather than competence: (33c) is a grammatical sentence of English, but it presents an overwhelming burden to the human parser and is thus judged as unacceptable. A more semantically relevant example would be the small but growing body of literature arguing that what have previously been taken to be hard grammatical bounds on quantifier scope are really reflections of prohibitive processing difficulty (see, e.g., Syrett & Lidz 2011; Wurmbrand 2018; Pasternak & Graf 2021). According to this view, a quantifier scoping out of a finite embedded clause (for example) is perfectly fine as far as the grammar proper is concerned—at least so long as independent constraints like islands are respected—but again can present an insurmountable challenge to the parser.

The parsing hypothesis of co- and pro-speech integration falls in the same tradition as this work, but turns the narrative on its head. In the aforementioned work on center-embedding and quantifier scope, the argument is that the grammar can generate more sentences than the parser can process. PH states something in the opposite direction: there are ways in which the parser can accomplish more than what the grammar would allow on its own. More generally, the takeaway from these kinds of arguments is that in order to capture the full range of judgments about syntactic well-formedness and available semantic inter-

pretations, a theory of linguistic competence must be paired with a sufficiently robust theory of linguistic performance, and in particular of human parsing.

It is worth noting that one could in theory adopt a stance that falls somewhere between GH and PH, i.e., a **Mixed Hypothesis (MH)**. According to MH, some co- and pro-speech content (presumably including gestures) is syntactically and semantically integrated in the grammar proper like in GH, while some (e.g., videos) is only integrated in the parser as in PH. But given the similarities in behavior between gestures and other co- and pro-speech content, it would presumably prove somewhat difficult to provide empirical evidence specifically favoring MH over GH or PH. For example, while gestures do have certain traits that distinguish them from other co- and pro-speech content, such as their important role in language acquisition (see, e.g., Iverson & Goldin-Meadow 2005), these unique qualities do not necessarily entail that the particular path to semantic integration in the adult grammar/parser is distinct from that for other, "less special" content. And even if the results of integrating gestures versus other content turn out to be semantically and/or pragmatically distinct—for example, if gestural inferences project differently from video-based inferences—this could just as well stem from differences in *how* the parser integrates gestures and videos, or from differences in the semantic interpretations of gestures and videos, rather than from a difference between integration in the grammar versus in the parser. With this in mind, in the rest of the paper I will assume a more absolutist version of PH that applies to gestures as well as other forms of co- and pro-speech content.

Finally, before moving on to a simple illustration of PH it is worth briefly discussing what the possible truth of PH would mean—and just as importantly, what it would *not* mean—for the future of formal research on co- and pro-speech gestures. Obviously, the claim that gestures are semantically integrated in the parser rather than the grammar does not in the least entail that gestures are not a "proper" object of linguistic study. After all, the parser, while neglected in a great deal of formal theoretical syntactic and semantic research, is just as important an object of linguistic study as the grammar is. Moreover, while according to PH the mere ability to integrate gestures does not in and of itself tell us terribly much about the grammar, *how* gestures are integrated could in fact tell us a great deal about the grammar. Regardless of the path to integration, the effects thereof have the potential to furnish substantial insights about the syntactic configurations and semantic interpretations of those structures into which gestural interpretations are integrated. Thus, PH should be thought of not as an attempt to toss gestures into a vague "parsing wastebasket" that is subsequently ignored by linguists, but rather as an attempt to assign a proper address to gestures *within* the general domain of the human faculty of language.

4 Semantically Interpretable CFGs and bottom-up parsing

Starting in this section I will introduce a toy grammar and parser illustrating what a PH-friendly analysis of co- and pro-speech integration might look like. Out of a desire to keep things as simple and accessible as possible, especially for readers without a background in formal parsing, at several points this toy analysis will adopt simplifying assumptions that

substantially deviate from standard views in research on syntax and/or semantics, or else abstract away from issues that are not of immediate relevance to the matter at hand. I will discuss these simplifications and abstractions as they arise, including some suggestions for how future work could fill in those gaps left by the analysis as presented in this paper.

As discussed above, according to PH the parser takes as its input a mixture of speech and co- and/or pro-speech content, using the grammar to parse the spoken material. A reasonable first step in defining a (toy) parser that meets these conditions is to define a formally explicit grammar, as well as a parser for that grammar that is capable of handling just linguistic material. This parser can then be "augmented" so that it can also integrate co- and pro-speech content. This section is dedicated to this first step, defining a formally explicit toy grammar (Section 4.1) and a parser for this grammar (Section 4.2); starting in the next section I will discuss how the parser can be appropriately augmented.

4.1 Semantically Interpretable CFGs

For our toy implementation I will treat the grammar as an extended version of a context-free grammar (CFG), which I call a **Semantically Interpretable CFG (SI-CFG)**. SI-CFGs are a significant (over)simplification of natural language in multiple respects that warrant discussion. Perhaps the most obvious simplification is in the use of a CFG in the first place. While CFGs are capable of generating phrase structure trees, they are not built to handle things like syntactic movement.[2] Thus, since the work of Chomsky (1957) CFGs have often been argued not to be an explanatorily adequate model for natural language syntax.[3] In addition, contrary to the commonly accepted view since the seminal work of May (1977), SI-CFGs will not generate separate syntactic structures for phonetic (PF) and semantic (LF) interpretation: there will be a single phrase structure that is both pronounced and interpreted.

It is worth emphasizing that these simplifications are a matter of convenience and not a matter of necessity, as those observations about natural language that render (SI-)CFGs inadequate—such as movement dependencies and mismatches between pronounced and interpreted structure—are irrelevant for our relatively modest purposes. More complex formal grammars have been developed that hew closer to standard contemporary syntactic assumptions, with perhaps the closest being Minimalist Grammars (MGs, Stabler 1997). MGs are formal grammars that have well-defined parsers (see, e.g., Harkema 2001) and that adopt core features from Chomsky's (1995) Minimalist Program, including feature-driven structure-building and movement operations. While most work in MGs has stayed away from semantic interpretation, this is not universally the case: for example, Kobele (2006) provides a purely derivational model-theoretic semantics for MGs, while Pasternak & Graf (2021) provide a way of extending MGs and their parsers so that they simultaneously derive

[2] Standard CFGs can be treated interchangeably as generating either tree structures or simply output strings, depending on how one interprets the syntactic rules (as tree branching rules or as string rewriting rules). This choice is irrelevant for CFG parsing, but it matters for SI-CFGs, where semantic interpretation is based on tree structures. I thus assume that (SI-)CFGs generate trees, not just strings.

[3] In fact, there is evidence suggesting that CFGs do not even reach the level of *descriptive* adequacy, as they lack sufficient weak (and thus strong) generative capacity; see, e.g., Shieber 1985 on Swiss German.

PF and Heim & Kratzer (1998)-style LF representations. MGs are thus a promising area for the future development of PH-friendly models that are more syntactically realistic. But for our purposes it will suffice to stick to (SI-)CFGs, whose grammars and parsers are substantially simpler than MGs and their parsers.

Our next task is to actually formally define SI-CFGs:

(34) An **SI-CFG** is an ordered tuple $\langle V_G, \Sigma_G, R_G, \text{S}, \textsc{Sem}, ||\cdot||_G, C_G\rangle$, where:

a. V_G is a finite set of **variables** (e.g., S, DP, VP),

b. Σ_G is a finite set of **terminals** (e.g., Mary), with $V_G \cap \Sigma_G = \emptyset$,

c. R_G is a finite set of **rules** from a variable to a string of variables and terminals (e.g., S $\rightarrow$ DP VP),

d. S $\in V_G$ is the **start variable**,

e. Sem is a set of **semantic objects** (e.g., model-theoretic objects),

f. $||\cdot||_G : \Sigma_G \rightarrow$ Sem is a **lexical interpretation function** from terminals to semantic objects,[4] and

g. $C_G : (\textsc{Sem})^n \rightarrow$ Sem is a **composition function**, taking a non-empty tuple of semantic objects and returning a semantic object.

The first four elements of the tuple are a standard part of CFGs; the final three are novelties allowing structures derived by SI-CFGs to be semantically interpreted. More specifically, the rules for compositional interpretation $[\![\cdot]\!]_G$ can be recursively defined as follows:

(35) **Lexical Interpretation:**
For leaf node a in tree τ generated by G, $[\![a]\!]_G := ||a||_G$.

(36) **Compositional Interpretation:**
For a given subtree $[_{\text{M}}\ b_1 \ldots b_n]$ of τ generable by G, where $n \geq 1$ and $b_1 \ldots b_n$ are the immediate daughters of M, $[\![\text{M}]\!]_G := C_G([\![b_1]\!]_G \ldots [\![b_n]\!]_G)$.

Importantly, for our purposes it does not matter what the objects in Sem are, and likewise for the compositional principles underlying C_G. If we wish to hew closely to standard operating procedure in semantic research, Sem would consist of model-theoretic objects like individuals, predicates, propositions, etc., while C_G would encompass operations like function application and predicate modification (i.e., intersection). But this is not a necessary assumption, and so I will frame things on a more abstract level in order to illustrate the fundamental components of the analysis. More specifically, I will simply use sans serif fonts to indicate objects in Sem in lexical interpretations, so that $||\text{Mary}||_G$ will for example be mary. As for C_G, I will assume throughout this paper that all branching nodes are either unary or binary branching, meaning that the only cases that need to be handled are when C_G receives a single argument (unary branching) and when it receives two arguments (binary branching). For the former I assume that C_G is the identity function, much like Heim & Kratzer's

[4] This definition operates under the empirically inaccurate simplifying assumption that any member of the alphabet Σ_G has a single interpretation, thereby disallowing lexical ambiguity. SI-CFGs can be revised to allow lexical ambiguity, but this requires formal complexities that are unnecessary for the main point of this paper.

(1998) rule of Non-Branching Nodes: for $P \in \textsc{Sem}$, $C_G(P) = P$. For the latter, I will assume a composition operation $* : (\textsc{Sem})^2 \rightarrow \textsc{Sem}$, so that for $P, Q \in \textsc{Sem}$, $C_G(P, Q) = P * Q$. If the analysis were couched in model-theoretic terms, $*$ would encompass both function application and predicate modification; since the two operations have disjoint domains, they do not conflict. I assume that $*$ is commutative ($P * Q = Q * P$ for all $P, Q \in \textsc{Sem}$), so that the compositional semantics is not sensitive to linear order. As a matter of notational convenience I will treat $*$ as right-associative: $P * Q * R = P * (Q * R)$.

As a concrete example, consider the toy grammar G specified in (37):

(37) $G = \langle V_G, \Sigma_G, R_G, \text{S}, \textsc{Sem}, ||\cdot||_G, C_G \rangle$, where:

a. $V_G = \{\text{S}, \text{DP}, \text{VP}, \text{V}, \text{D}, \text{NP}\}$

b. $\Sigma_G = \{\text{Mary}, \text{used}, \text{the}, \text{stairs}\}$

c. R_G contains the following rules:
(1) $\text{S} \rightarrow \text{DP VP}$ (2) $\text{DP} \rightarrow \text{Mary}$ (3) $\text{DP} \rightarrow \text{D NP}$ (4) $\text{D} \rightarrow \text{the}$
(5) $\text{NP} \rightarrow \text{stairs}$ (6) $\text{VP} \rightarrow \text{V DP}$ (7) $\text{V} \rightarrow \text{used}$

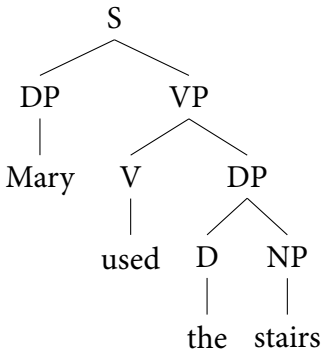
d. $\textsc{Sem} = \{\textsf{mary}, \textsf{used}, \textsf{the}, \textsf{stairs}, \textsf{the} * \textsf{stairs}, \textsf{used} * \textsf{the} * \textsf{stairs}, \textsf{used} * \textsf{mary} \ldots\}$

e. $||\text{Mary}||_G = \textsf{mary}$
$||\text{used}||_G = \textsf{used}$
$||\text{the}||_G = \textsf{the}$
$||\text{stairs}||_G = \textsf{stairs}$

f. C_G is as described above.

As a toy grammar, G has a severely constrained generative capacity: the only sentences it generates are *Mary used the stairs*, *Mary used Mary*, *The stairs used the stairs*, and *The stairs used Mary*. Consider the first of these. The structure that G assigns to *Mary used the stairs* is the one in (38):

(38)



On the semantic end, (39) details the interpretation furnished by $[\![\cdot]\!]_G$:

(39) a. $[\![\text{stairs}]\!]_G = ||\text{stairs}||_G = \textsf{stairs}$

b. $[\![\text{NP}]\!]_G = C_G(\textsf{stairs}) = \textsf{stairs}$

c. $[\![\text{the}]\!]_G = ||\text{the}||_G = \textsf{the}$

d. $[\![\text{D}]\!]_G = C_G(\text{the}) = \text{the}$
e. $[\![\text{Object DP}]\!]_G = C_G(\text{the}, \text{stairs}) = \text{the} * \text{stairs}$
f. $[\![\text{used}]\!]_G = ||\text{used}||_G = \text{used}$
g. $[\![\text{V}]\!]_G = C_G(\text{used}) = \text{used}$
h. $[\![\text{VP}]\!]_G = C_G(\text{used}, \text{the} * \text{stairs}) = \text{used} * (\text{the} * \text{stairs}) = \text{used} * \text{the} * \text{stairs}$
i. $[\![\text{Mary}]\!]_G = ||\text{Mary}||_G = \text{mary}$
j. $[\![\text{Subject DP}]\!]_G = C_G(\text{mary}) = \text{mary}$
k. $[\![\text{S}]\!]_G = C_G(\text{mary}, \text{used} * \text{the} * \text{stairs}) = \text{mary} * (\text{used} * \text{the} * \text{stairs})$
$= \text{mary} * \text{used} * \text{the} * \text{stairs}$

We now have a simple syntactic-semantic formalism in place, along with a toy grammar exemplifying this syntactic formalism. Next I turn to the task of defining the SI-CFG parser.

4.2 Bottom-up parser for SI-CFGs

Recall that for now we wish to define a parser that only parses linguistic material; starting in the next section I will discuss how the parser can be augmented so that it can handle an input consisting of both linguistic material and co- and pro-speech content. Putting the latter aside, the purpose of a parser is to take as its input a string σ, and to determine a possible structure or derivation generating σ within the grammar in question. The particular form of parser to be implemented here is what is known as a *bottom-up parser*. As the name suggests, bottom-up parsers operate by traversing syntactic structures from the bottom upward, first scanning words in the input string and then building up possible structures based on those words that have been scanned.[5] More specifically, the parser I use will be a *depth-first bottom-up parser*: after scanning a word in the input string, the parser will build as much structure as it can using those words that have been scanned thus far before proceeding to scan the next word. Our parser will be different from many other CFG parsers in that rather than simply building the syntactic structure for the input string, it will simultaneously build a syntactic structure and a semantic interpretation.

A particular implementation of a parser can be thought of as consisting of a *parsing schema*, the set of rules determining how the parser can make and confirm predictions; and a *control structure*, the algorithm that determines what prediction-making or -confirming steps to actually take at a given point in a parse. The control structure is naturally most relevant when a parser has multiple options for how to proceed. Take, for example, the sentence *The house I own is big*. When the parser has scanned only *the house*, it must decide whether to assume that this is a complete DP and build the structure accordingly, or to scan more material on the assumption that there will be something like a relative clause that will also be part of the DP. Alternatively, the parser could try both simultaneously and eliminate the

[5] This can for instance be contrasted with *top-down parsers*, which start from the top of the tree and make "downward" predictions, as well as *left corner parsers*, which incorporate features from both top-down and bottom-up parsers.

incorrect option once a crash occurs. This determination is a matter for the control structure: on their own, the rules of the parsing schema leave both doors open, and the control structure decides which door(s) to take.

With all of that being said, a substantial amount of research in formal parsing abstracts away from issues pertaining to the control structure, focusing only on the parsing schema.[6] I will follow suit in this paper, providing only a parsing schema for the bottom-up SI-CFG. As a result, all of the example parses discussed in this paper will be ones in which the parser makes the right decision at every step: we are concerned only with what happens when everything goes right, rather than with how the parser prevents things from going wrong.

As mentioned above, a parsing schema is a set of rules dictating in what ways the parser is able to make predictions and confirm those predictions. This parser will be implemented in a "parsing as deduction" framework (Pereira & Warren 1983): each parse will start with an *axiom*, and there is a particular *goal* that the parser must reach by way of deduction in order for the parse to be successful. Parse rules are then couched as *inference rules* that take us from one step in a deduction (i.e., parse) to the next. The axiom and goal are *parse items* stating the current state of the parser, including what has been scanned and what deductions have been made; the axiom states that nothing has been scanned and no predictions have been made, while the goal states that everything has been scanned and a whole clausal structure has been built. The parse/inference rules then serve to update the current parse item by scanning from the input string or making predictions about what has already been scanned.

To be more specific, a parse item in our parser will be a pair $[\beta,\ j]$. β is a (possibly empty) sequence of pairs (b, P), where $b \in V_G \cup \Sigma_G$ and $P \in$ Sem. In other words, β is a sequence of pairs of nodes and their semantic interpretations. j, meanwhile, is an index, indicating how far into the input string the scanner has scanned. In a parse with input string $\sigma = w_0 w_1 \ldots w_{n-1}$, the axiom is $[,\ 0]$; the sequence of node-interpretation pairs is empty because nothing has been scanned yet, and for the same reason the index is set to 0. The goal, meanwhile, is $[(\mathrm{S}, P),\ n]$; the pair (S, P) indicates that we have constructed a full sentence with some interpretation P, and n indicates that we have scanned the full string (since $n = |\sigma|$, the length of σ).[7] The parse/inference rules then serve to update the axiom by scanning words and making inferences about the structure and interpretation, with the aim being to reach the goal and thereby conclude a successful parse of the input string.

The parser will have two inference rules: **Shift** and **Reduce** (hence why such bottom-up parsers are often referred to as *shift-reduce parsers*). A definition is provided in (40) for the Shift rule, which as its name suggests serves to shift us forward in the input string by one word. In other words, it scans the next word in the input string.

[6] For example, this is a common theme in work using formal parsers to account for human sentence processing results from the psycholinguistic literature; see, e.g., Joshi 1990; Rambow & Joshi 1994; Kobele et al. 2013; Gerth 2015; Graf & Marcinek 2014; Graf et al. 2015, 2017; Zhang 2017; Liu 2018; De Santo 2019; Lee 2019; Pasternak & Graf 2021.

[7] In actuality there is not a single goal item but many, since P can be any object in Sem. The aim of a parse is thus to reach *some* goal item, rather than *the* goal item.

(40) **Shift:**

$$\frac{[\beta,\ j]}{[\beta\ (a, P),\ j+1]} \text{ where } a = w_j \text{ and } ||a||_G = P$$

If a is the next word and has interpretation P, Shift adds the pair (a, P) to the end of the sequence of node-interpretation pairs, and furthermore increments the index to indicate that we have moved forward one spot in the input string. The other rule, Reduce, is defined in (41):

(41) **Reduce:**

$$\frac{[\beta\ (b_1, P_1) \ldots (b_n, P_n),\ j]}{[\beta\ (M, C_G(P_1 \ldots P_n)),\ j]} \text{ where } M \rightarrow b_1 \ldots b_n \in R_G$$

Put simply, if the last n nodes in the sequence of node-interpretation pairs collectively make up the right-hand side of a syntactic rule in R_G, Reduce serves to replace them with the left-hand side of that rule. The interpretation assigned to that higher node is then the result of applying C_G to the tuple of interpretations assigned to its daughters. Notice that unlike Shift, an application of Reduce does not increment the index. This is because a Reduce step does not scan new material from the string, instead building structure based on what has already been scanned.

To see how the parser works, we will go through the parse of the structure in (38) for *Mary used the stairs*, which is written out in Table 1. For readability's sake, rather than writing the parse out as a deduction, I have written it as a table with three columns: the first is the parse/deduction step, the second is the parse item that results from that step, and the third is the deduction rule used to generate that parse item. For parse steps where the Shift rule is used, I have indicated in parentheses the word that was scanned by that shift step; for steps where Reduce is use, I indicate in parentheses the grammatical rule in (37c) used to build the structure. A step-by-step summary of the parse can be seen below:

0. We start with the axiom [, 0], as in any parse.
1. Since there is no material for Reduce to work with yet, we apply Shift in Step 1. This adds the pair (Mary, mary) to the parse item and increments the index from 0 to 1, indicating that we have now scanned the first word of the input string.
2. Since this is a depth-first parser, we next want to build as much structure as we can based on what has been scanned so far. *Mary* is the daughter of a unary branching DP node, meaning that we perform Reduce in Step 2. This involves replacing *Mary* with DP and applying C_G to mary; since C_G is the identity function when it takes one argument (see discussion above), this simply returns mary again.
3. We have now built up as much structure as was possible with the previously scanned material, and so we perform Shift again. This time we add (used, used) to the end of the sequence and increment the index to indicate that we have now scanned the first two words of the input string.

0	[, 0]	Axiom
1	[(Mary, mary), 1]	Shift(Mary)
2	[(DP, mary), 1]	Reduce(2)
3	[(DP, mary) (used, used), 2]	Shift(used)
4	[(DP, mary) (V, used), 2]	Reduce(7)
5	[(DP, mary) (V, used) (the, the), 3]	Shift(the)
6	[(DP, mary) (V, used) (D, the), 3]	Reduce(4)
7	[(DP, mary) (V, used) (D, the) (stairs, stairs), 4]	Shift(stairs)
8	[(DP, mary) (V, used) (D, the) (NP, stairs), 4]	Reduce(5)
9	[(DP, mary) (V, used) (DP, the ∗ stairs), 4]	Reduce(3)
10	[(DP, mary) (VP, used ∗ the ∗ stairs), 4]	Reduce(6)
11	[(S, mary ∗ used ∗ the ∗ stairs), 4]	Reduce(1)

Table 1 Parse for *Mary used the stairs*.

4. Since *used* is the daughter of a unary-branching V node, we can apply Reduce again; this looks much like what occurred in Step 2.
5. Since we have once again built up as much structure as the currently scanned material allows, we scan again with Shift, adding (the, the) to the sequence of node-meaning pairs and incrementing the index to 3.
6. *the* is the daughter of a unary-branching D node, and so we Reduce in the same way as in Steps 2 and 4.
7. Having exhausted our options for structure-building, we again apply Shift, adding (stairs, stairs) to the sequence and incrementing the index a final time to 4.
8. Once again, *stairs* is the daughter of a unary-branching NP node, and so it is reduced in the same manner as in Steps 2, 4, and 6.
9. Our last two node-meaning pairs have the nodes D and NP. Because of the rule DP → D NP, we can now finally perform our first Reduce step involving a binary-branching node. The pairs (D, the) and (NP, stairs) are replaced with a single pair in which the node is DP, and the interpretation is C_G(the, stairs), i.e., the ∗ stairs.
10. Our last two node-meaning pairs have the nodes V and DP. Because of the rule VP → V DP, we can again Reduce to a binary-branching node. (V, used) and (DP, the∗stairs) are replaced with a single pair (VP, C_G(used, the ∗ stairs)), i.e., (VP, used ∗ the ∗ stairs).
11. We have two node-meaning pairs remaining, one with node DP and one with node VP. Thanks to the rule S → DP VP, we can perform one more binary-branching Reduce, replacing these with the pair (S, C_G(mary, used ∗ the ∗ stairs)), i.e., (S, mary ∗ used ∗ the ∗ stairs). This parse item matches the goal conditions for a parse: we have built a complete sentence with an assigned interpretation, and we have scanned the whole input string (since the index 4 is equal to the length of the string).

We have now defined SI-CFGs, which are CFGs that generate semantic interpretations, as well as a (depth-first) bottom-up parser for SI-CFGs, which both builds the structure and assigns a semantic interpretation to that structure. We are now ready for the task of introducing co-speech content into the mix.

5 Adding co-speech integration

With our grammar and linguistic-only parser in place, we can now begin the task of extending the parser so that it can fold in the interpretation of co- and pro-speech content. I start in this section with co-speech content, and in Section 6 I discuss pro-speech content. In accordance with PH, the only revisions that will be made here will be to the parser, and not to the grammar: UP in (1), for example, will not appear in any syntactic representation, but the parser will nonetheless incorporate its semantic content.

The task of extending the parser to co-speech content will take place in two steps. First, in Section 5.1 I define what I refer to as a **co-/pro-speech augmentation** (or simply an **augmentation**), which provides information about possible co-/pro-speech content and its meaning, as well as how co-speech content semantically composes with spoken content. In Section 5.2 I introduce the **augmented parser**, which allows an SI-CFG parser to parse an input consisting of simultaneous linguistic and co-speech content. As a result, co-speech content can be integrated into the parsed meaning of the utterance without being a part of the syntactic structure about which the parser makes deductions

5.1 Defining augmentations

As mentioned above, an **augmentation** essentially provides a co-/pro-speech lexicon: it includes a set of possible co-/pro-speech content, an interpretation function for that content, and a composition operation indicating how co-speech content is to be integrated with spoken content. A formal definition is provided in (42):

(42) An **augmentation** H of SI-CFG G is an ordered triple $\langle \Sigma_H, ||\cdot||_H, \star \rangle$, where:
   a. Σ_H is a **co-/pro-speech alphabet** (with $\Sigma_G \cap \Sigma_H = \emptyset$ and $\varepsilon \notin \Sigma_H$),[8]
   b. $||\cdot||_H : \Sigma_H \rightarrow$ SEM is a **co-/pro-speech interpretation function** from elements of the co-/pro-speech alphabet to semantic objects,[9] and
   c. $\star : (\text{SEM})^2 \rightarrow$ SEM is the **co-speech composition function**.

The co-/pro-speech alphabet would include things like the gesture UP and sound effect EXPLODE. The interpretation function $||\cdot||_H$, meanwhile, would return something like up for UP and explode for EXPLODE. As for $\star$, what this does depends on one's theory of co-speech

[8] ε is the empty string.

[9] Much like what was discussed in fn. 4 for SI-CFGs, this formulation does not allow for ambiguity: UP, for example, must have a single interpretation. Augmentations can easily be revised in order to allow ambiguity, with $||\cdot||_H$ being a function from members of Σ_H to *sets* of members of SEM, and the parser choosing a single member of $||\text{UP}||_H$ to use in a given parse.

content integration. For example, if Ebert & Ebert (2014) are correct, $P \star Q$ (with P being the interpretation of the gestural content) would generate an interpretation of Q with P as a supplement. On Schlenker's (2018a) analysis, meanwhile, $P \star Q$ would generate an interpretation of Q plus a presupposition of the form $Q \Rightarrow P$ (i.e., a cosupposition). On Esipova's (2019) analysis, $\star$ would simply be $*$, since gestures are not compositionally unique.[10] Notice that if we adopt the account of either Ebert & Ebert or Schlenker, $\star$ must be unlike $*$ in being non-commutative: for many P and Q, $P \star Q \neq Q \star P$.

Before moving on to how the parser can make use of these augmentations, some discussion is warranted about two important simplifying assumptions that are being made by defining augmentations the way I have. First, a great deal of co- and pro-speech content is interpreted *iconically*, with a close connection between form and meaning: for example, the sound effect EXPLODE has an interpretation involving explosions because EXPLODE sounds like an explosion. Such iconicity is not manifested in any way in the definition of augmentations in (42): $||\text{EXPLODE}||_H$ is simply defined as being explode. I take the issue of the iconic relationship between form and interpretation of co-/pro-speech content to be orthogonal to the issue at hand of how such interpretations are integrated into the interpretations of spoken utterances. And second, no attempt is made here to distinguishing between different kinds of co-/pro-speech content, e.g., gestures vs. sound effects vs. emoji: all are lumped under the same umbrella. As was discussed above, theoretical discussion from Pasternak (2019) and Pierini (to appear) and experimental evidence from Pasternak & Tieu (in revision) suggest that this may be the right way to go. However, it is plausible that there are areas in which divergence occurs, in which case there might be a need for multiple different augmentations—say, one for gestures and one for sound effects—each of which is integrated into the parser. I leave this issue as a matter for future work.

5.2 Integration into the parser

With co-/pro-speech augmentations defined, we next turn to the task of integrating such augmentations into the parser for the purposes of interpreting co-speech content. That is, for a given SI-CFG G and augmentation H of G, I will define an ***H*-augmented parser for *G***. Because the augmented parser must account for co-speech as well as linguistic content, the input σ will be different from the input for a simple SI-CFG parser: namely, rather than taking in a string of words, the augmented parser will take in a string of pairs (a, A), where $a \in \Sigma_G$ and $A \in \Sigma_H \cup \{\varepsilon\}$. (The inclusion of ε permits cases where a word occurs without any co-speech content.) In other words, each element of the input will be a pair consisting

[10] This may seem odd in light of Esipova's claim that DP-modifying gestures make different contributions from NP- and VP-modifying gestures (supplements vs. restrictive modifiers). So how can a single composition operation $*$ accomplish this? One possibility, and in my mind the preferable one, is to say that the composition operation(s) is/are the same, but DPs and their modifiers simply have different interpretations from NPs/VPs and their modifiers. In this case, things can remain more or less as they are. Another possibility is that the compositional semantics is sensitive to syntactic categories, in which case C_G and $\star$ would have to be suitably revised so that their inputs are not just interpretations, but pairs of interpretations and syntactic categories. SI-CFGs and their (augmented) parsers can easily be revised accordingly.

of one linguistic item that is treated as obeying the syntactic rules of the SI-CFG in question, and (possibly) one piece of co-speech content whose semantic interpretation is folded in during the course of the parse. I will often write such pairs in the form $[a]_A$; furthermore, I will often write $[a]_\varepsilon$ (i.e., a without any co-speech content) as simply a.

At this point it is worth discussing another abstraction being made here. I am assuming that each "word" fed into the parser is a pair of a spoken word and some (possibly empty) co-speech content. Thus, any given bit of co-speech content is treated as co-occurring with a single spoken word. This abstracts away from the rather complicated issue of *alignment*, including how the timing of a given bit of co-speech content determines what syntactic constituent it semantically composes with. Understanding co-speech alignment in general, and its effects on semantic interpretation in particular, is a tricky issue that requires a great deal more empirical research. First and most obviously, a single piece of co-speech content can have a duration spanning multiple words, and it seems likely that this will have compositional repercussions. Second, gestures are themselves not discrete objects and can often be divided into smaller parts like a *pre-preparation position*, a *prestroke hold*, a *stroke*, a *post-stroke hold*, and a *retraction* (see, e.g., McNeill 2005); the alignment of some or all of these parts may also impact the compositional semantics. Third, it appears that the alignment of gestures is in certain ways affected by spoken prosody (see Esipova 2019 and sources therein), and much more work is required in order to fully determine this interaction, as well as to what extent this is shared with other kinds of co-speech content. And fourth, there appear to be ways in which the choice of medium affects alignment: for example, the strictly linear nature of text + emoji means that co-text emoji cannot align with text in the same way that gestures align with speech.

Given the empirical complexities presented by alignment of co-speech content, as well as the relatively modest aims of the implementation in this paper, I have opted for a simplification. More specifically, on the analysis in this paper I will assume that any co-speech content semantically composing with some syntactic constituent X is aligned with the (linearly) first word in X. Or, put another way, co-speech content aligned with a given word can be interpreted as semantically composing with any constituent of which that is the first word. Thus, for (1) I will assume that the input to the parser is the sequence of pairs in (43a), which in our notation can be rewritten as (43b).

(43) a. (Mary, ε) (used, UP) (the, ε) (stairs, ε)
b. Mary $[\text{used}]_{\text{UP}}$ the stairs.

Naturally, as we gain a greater understanding of the nature of co-speech alignment, formally implemented parsers will need to be appropriately refined to capture the subtle ways in which alignment affects interpretation.

With this caveat out of the way, let us now turn to the augmented parser itself. Parse items in the augmented parser are similar, but not quite identical to parse items in the simple SI-CFG parser. Recall that in the simple SI-CFG parser, parse items were of the form $[\beta, j]$; β was a (possibly empty) sequence of pairs (a, P), with $a \in V_G \cup \Sigma_G$ and $P \in$ Sem, and j was an index. The only difference in an augmented parser is that the first element of each pair in

β could also be κ, a designated symbol indicating that the semantic content it is paired with comes from co-speech content. The axiom and goal for the parser are the same as before, namely $[,\ 0]$ and $[(\mathrm{S}, P),\ n]$ (where $n = |\sigma|$), respectively.

This leaves us with the parse rules. Rather than simply having the rules Shift and Reduce, as in the simple SI-CFG parser, the augmented parser has four rules, depending on the presence and manipulation of co-speech content: NCShift (NC = No Co-speech), NCReduce, CShift, and CReduce. We start with NCShift, defined in (44):

(44) **NCShift:**

$$\frac{[\beta,\ j]}{[\beta\ (a, P),\ j+1]} \text{ where } [a]_{\varepsilon} = w_j \text{ and } ||a||_G = P$$

NCShift can only apply if the next element in the input string is a pair of the form (a, ε), i.e., if it is a word without any co-speech content. NCShift then operates identically to the old Shift rule, adding the pair (a, P) (where $||a||_G = P$) to the sequence of form-meaning pairs and incrementing the index. NCReduce, provided in (45), is exactly identical to the old Reduce rule: we move up the tree through some rule in R_G and compose the assigned semantic interpretations, without incrementing the index.

(45) **NCReduce:**

$$\frac{[\beta\ (b_1, P_1) \ldots (b_n, P_n),\ j]}{[\beta\ (M, C_G(P_1 \ldots P_n)),\ j]} \text{ where } M \rightarrow b_1 \ldots b_n \in R_G$$

Next up is CShift, which is the version of Shift that applies when the next element of the input string comes paired with co-speech content:

(46) **CShift:**

$$\frac{[\beta,\ j]}{[\beta\ (\kappa, P)\ (a, Q),\ j+1]} \text{ where } [a]_A = w_j,\ A \in \Sigma_H,\ ||A||_H = P, \text{ and } ||a||_G = Q$$

When CShift applies, instead of a single form-meaning pair being added to the sequence, two are added: the first pair consists of κ and the interpretation of the co-speech content, and the second consists of the spoken content and its interpretation. Once again the index is incremented, since we have moved forward one spot in the input string. Finally, there is CReduce, defined in (47):

(47) **CReduce:**

$$\frac{[\beta\ (\kappa, P)\ (b, Q),\ j]}{[\beta\ (b, P \star Q),\ j]} \text{ where } b \in \Sigma_G \cup V_G$$

CReduce can apply when the penultimate member of the sequence is a pair (κ, P), and the final member is a pair (b, Q), where $b \in V_G \cup \Sigma_G$ (b is linguistic content). When applied, CReduce removes (κ, P) from the sequence, and replaces Q in (b, Q) with $P \star Q$, the result of applying the co-speech composition operation $\star$ to P and Q.

To illustrate the workings of the augmented parser, I will go through the parse of the simplified version of (1) seen in (43). I assume that the grammar itself is exactly the same

0	[, 0]	Axiom
1	[(Mary, mary), 1]	NCShift(Mary)
2	[(DP, mary), 1]	NCReduce(2)
3	[(DP, mary) (κ, up) (used, used), 2]	CShift([used]$_{\text{UP}}$)
4	[(DP, mary) (κ, up) (V, used), 2]	NCReduce(7)
5	[(DP, mary) (κ, up) (V, used) (the, the), 3]	NCShift(the)
6	[(DP, mary) (κ, up) (V, used) (D, the), 3]	NCReduce(4)
7	[(DP, mary) (κ, up) (V, used) (D, the) (stairs, stairs), 4]	NCShift(stairs)
8	[(DP, mary) (κ, up) (V, used) (D, the) (NP, stairs), 4]	NCReduce(5)
9	[(DP, mary) (κ, up) (V, used) (DP, the $*$ stairs), 4]	NCReduce(3)
10	[(DP, mary) (κ, up) (VP, used $*$ the $*$ stairs), 4]	NCReduce(6)
11	[(DP, mary) (VP, up $\star$ (used $*$ the $*$ stairs)), 4]	CReduce
12	[(S, mary $*$ (up $\star$ (used $*$ the $*$ stairs))), 4]	NCReduce(1)

Table 2 Possible parse for *Mary [used]$_{UP}$ the stairs*.

as before, with the same rules in (37c); after all, the whole point is that UP is integrated in the parser, not the grammar. The full parse in the familiar tabular format can be seen in Table 2. Many of the steps are the same as in the parse in Table 1 of the gesture-less *Mary used the stairs*, so I will stick to those points where the two differ. The first point of difference between the two parses lies in Step 3, the point at which the augmented parser scans *[used]$_{UP}$*. Since this point in the string has a non-empty co-speech component, the form of Shift used is CShift. As a result, two new form-meaning pairs are added to the sequence in the parse item: the gesture pair (κ, up) and the linguistic pair (used, used). Notice that once this occurs, the ensuing parse item is already in the domain of CReduce: the penultimate form-meaning pair is of the form (κ, P), and the final pair is a linguistic node and its interpretation. Thus, the parser could decide to perform CReduce now and then proceed with the parse in more or less the same way as in Table 1. In this case, the interpretation would be one in which up only composes with used, as in (48):

(48) mary $*$ (up $\star$ used) $*$ the $*$ stairs

Of course, this interpretation could be deviant for purely semantic or pragmatic reasons, or it could be perfectly well-formed; either way, the parser itself does not prohibit it. However, in the parse in Table 2 I assume that the parser declines to take this option, and instead holds off on performing CReduce until later.

For the next seven steps, the parser proceeds identically to the parse in Table 1, performing NCShift and NCReduce steps until the full VP with associated interpretation is built up after Step 10. Notice that at this point, the parse item we have built is once again in the domain of CReduce: the penultimate pair is (κ, up), and the final pair consists of VP and its interpretation (used $*$ the $*$ stairs). This time the parser does perform CReduce, removing the pair (κ, up) and composing up with the interpretation of the VP with the gesture composition operation $\star$. The resulting interpretation for the VP after the gesture content is folded

in in Step 11 is therefore up ⋆ (used ∗ the ∗ stairs). The final step of reduction then proceeds precisely as before, and the final parse item meets the goal conditions—a sentence with semantic interpretation has been built, and all words have been scanned—meaning that the parse is successful.

Now is a good point to take stock of what the augmented parser has accomplished. First and most importantly, it has successfully parsed the syntactic structure and integrated the interpretation of the gesture into the derived interpretation of the input string. Moreover, in accordance with NMGS the gestural content is integrated in a way that allows us to meaningfully talk about the scope of gestural inferences (and thus, for example, projection). For example, the gesture composed with the verb phrase to the exclusion of the subject, meaning that it can scope below a quantificational subject. The same goes for modals and negation, as illustrated in (49): up composes with use ∗ the ∗ stairs, with both not and might composing later (and thus taking wider scope).

(49) a. Mary might not [use]$_{\text{UP}}$ the stairs.
b. mary ∗ might ∗ not ∗ (up ⋆ (use ∗ the ∗ stairs))

Not only has the augmented parser allowed us to fold gestural inferences into the semantics of the sentences with which they co-occur, but it has done so without requiring the stipulation that gestures and other co-speech content are part of syntax: there was no need to posit that UP, for example, is a syntactic head adjoined to the VP *used the stairs*. Given the enormous variety of content that can be integrated in this manner, much of which cannot be plausibly treated as syntactic objects, this is a substantial advantage over GH.

6 Adding pro-speech integration

Next up is pro-speech content integration like in (25), repeated below:

(25) The car will EXPLODE and make a mess of the street.
(≈ 'The car will explode and make a mess of the street.')

According to the definition of the augmented parser furnished in Section 5, the input to the parser is a string of pairs (a, A), where a is a (linguistic) word and A is either some co-speech content or the empty string ε (indicating a lack of co-speech content). The first member of the pair is then treated by the parser as a syntactic object subject to the rules of the SI-CFG used by the parser, while the second member of the pair (if not ε) is marked with κ to indicate that it is co-speech content and semantically folded in accordingly. However, in the case of pro-speech content some non-linguistic material is treated *as if* it is a syntactic object subject to the same rules as other syntactic objects. Or, put differently, pro-speech content plays the role of the primary, linguistic content, not secondary co-speech content. Thus, the parser needs to be revised in order to allow pro-speech content to be the first member of the (a, A) pair, and not just the second as is the case with co-speech content.

This raises a question: when presented with some sensory input containing a gesture, sound effect, or other content, how does the human parser determine whether that content

should be treated as secondary and co-speech, in which case it is the second member of a pair in the input string, or as primary and pro-speech, in which case it is the first member of such a pair? In this paper I assume that this determination has already been made by the time the parser is presented with its input string of pairs of (pro-)speech and co-speech content. That is, some preliminary analysis has already occurred that has determined that EXPLODE in (25) is pro-speech and not co-speech. The assumption of pre-processing of this sort is commonplace in work on parsing, and is a feature even of the gesture-free SI-CFG parser in Section 4. After all, that parser treated the input as a string of words, but this is not the actual sensory input that humans are regularly confronted with: when you speak and I listen, the input on my end is not a string of words but a complex audio(visual) signal. Therefore, some pre-processing must occur in order to take that complex audio(visual) signal and return an input string that is appropriate for the SI-CFG parser. I assume a similar sort of pre-processing for pro-/co-speech content: the complex sensory input represented by something like (25) has been filtered in advance in such a way that EXPLODE is treated as pro-speech and not co-speech content. How humans actually accomplish this is a potentially fascinating domain of study, but one that is well beyond the scope of this paper.

With this in mind, in order to illustrate how the revised augmented parser will work, I will use (50) as a toy example. (50a) shows the example input in a plain format, while (50b) shows the sequence of pairs fed to the parser. As promised, EXPLODE is the first member of its pair: it is treated as if it is primary linguistic content that is subject to the rules of the grammar. The second member of a given pair is always ε, since the example sentence only has pro-speech content and no co-speech content.

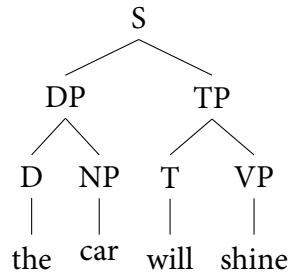
(50) a. The car will EXPLODE.
b. (the, ε) (car, ε) (will, ε) (EXPLODE, ε)

Suppose we have a toy grammar G along the following lines:

(51) $G = \langle V_G, \Sigma_G, R_G, \text{S}, \textsc{Sem}, ||\cdot||_G, C_G \rangle$, where:
a. $V_G = \{\text{S}, \text{DP}, \text{D}, \text{NP}, \text{TP}, \text{T}, \text{VP}\}$
b. $\Sigma_G = \{\text{the}, \text{car}, \text{will}, \text{shine}\}$
c. R_G contains the following rules:
(1) $\text{S} \rightarrow \text{DP TP}$ (2) $\text{DP} \rightarrow \text{D NP}$ (3) $\text{TP} \rightarrow \text{T VP}$ (4) $\text{D} \rightarrow \text{the}$
(5) $\text{NP} \rightarrow \text{car}$ (6) $\text{T} \rightarrow \text{will}$ (7) $\text{VP} \rightarrow \text{shine}$
d. $\textsc{Sem} = \{\mathsf{the}, \mathsf{car}, \mathsf{will}, \mathsf{shine}, \mathsf{the} * \mathsf{car}, \mathsf{will} * \mathsf{shine} \ldots\}$
e. $||\text{the}||_G = \mathsf{the}$
$||\text{car}||_G = \mathsf{car}$
$||\text{will}||_G = \mathsf{will}$
$||\text{shine}||_G = \mathsf{shine}$
f. C_G is as before.

Importantly, again adhering to PH, it remains the case that EXPLODE $\notin \Sigma_G$, i.e., there is no lexical item EXPLODE. This toy grammar will only be able to generate a single sentence: *The car will shine*. This sentence will be assigned the syntactic structure in (52):

(52)



While EXPLODE is still not a word or phrase in the grammar, the parser needs to be revised so that EXPLODE can be treated as if it is a word or phrase. First, as discussed above we must revise our definition of what sorts of input strings can be fed into the augmented parser. Before, the augmented parser took a string of pairs (a, A), where $a \in \Sigma_G$ and $A \in \Sigma_H \cup \{\varepsilon\}$. Now the parser will still take a string σ of such pairs, but this time $a \in \Sigma_G \cup \Sigma_H$. That is, the first member of the pair, which is the primary content that is assumed to be subject to the rules of the SI-CFG, can be either a word or some pro-speech content. Parse items will again be of the form $[\beta, j]$, with j being an index and β being a sequence of pairs (b, P), with $P \in$ SEM. However, we will expand the options of what b can be. Before b could be a word in Σ_G (e.g., *the*), a variable in V_G (e.g., D), or the designated symbol κ indicating co-speech content; now we will add the possibility that b is π, a second designated symbol indicating pro-speech content. The axiom and goal items will be exactly the same as before, i.e. $[, 0]$ and $[(S, P), n]$ (where n is the length of the input string), respectively.

While the two Reduce rules from the prior version of the augmented parser (NCReduce and CReduce) can remain as they are, the two Shift rules (NCShift and CShift) need to be revised to account for the new possibilities in terms of what can be scanned. The revised version of NCShift can be seen in (53):

(53) **NCShift (revised):**

$$\frac{[\beta,\ j]}{[\beta\ (m, P),\ j+1]}$$

where $[a]_\varepsilon = w_j$, and (I) $a \in \Sigma_G$, $m = a$, and $||a||_G = P$; or (II) $a \in \Sigma_H$, $m = \pi$, and $||a||_H = P$

In short, NCShift applies if the next pair in the input string is of the form (b, ε), where $b \in \Sigma_G \cup \Sigma_H$. If $b \in \Sigma_G$, then the pair $(b, ||b||_G)$ is added to the sequence β, as in all prior instances of (NC)Shift. If $b \in \Sigma_H$, then the pair $(\pi, ||b||_H)$ is added.

A similar revision is made to CShift, as can be seen in (54):

(54) **CShift (revised):**

$$\frac{[\beta,\ j]}{[\beta\ (\kappa, P)\ (m, Q),\ j+1]}$$

where $[a]_A = w_j$, $A \in \Sigma_H$, $||A||_H = P$, and (I) $a \in \Sigma_G$, $m = a$, and $||a||_G = Q$; or (II) $a \in \Sigma_H$, $m = \pi$, and $||a||_H = Q$

That is, CShift applies when the next member of the string is (b, A), where $A \neq \varepsilon$. The co-speech content is added to β in the same way as before, while the linguistic or pro-speech content is added in the same way as in NCShift.

This way of revising the augmented parser, and in particular the rule CShift, suggests an interesting possibility: namely, that it is possible for some co-speech content to co-occur and be semantically integrated with pro-speech content. After all, CShift does not make any differentiation as to whether b in the input (b, A) is linguistic or pro-speech content. In fact, evidence from Esipova (2019) suggests that this is a good thing. In her example, the primary gesture is the Russian gesture DRUNK, performed by either flicking one's neck or tapping one's neck with the back of one's hand. The secondary gesture is the non-manual wide-eyed look of surprisal, which Esipova notates as O_O. The relevant example from Esipova (2019, p. 107) can be seen in (55):

(55) Yesterday there was a party, and Mia got $[\text{DRUNK}]_{\text{O\_O}}$.
(≈ 'Mia got drunk to a surprising extent.')

With the parse rules as they currently are, plus the additional new parse rule to be introduced shortly, the parser will have no problem handling an input like (55): it is simply the case that the last item in the input string is the pair (DRUNK, O_O), a possibility that the augmented parser is fully capable of accounting for.

The one brand new parse rule mentioned above allows the parser to treat pro-speech content as properly syntactic material. As things currently stand, when the parser scans some pro-speech content, it introduces a pair of the form (π, P) to the parse item, where P is some semantic content, and π indicates that the semantic interpretation comes from pro-speech content. But in order for the parser to syntactically integrate the pro-speech content, π must be replaced with a variable in V_G: if EXPLODE is to be treated as a VP, then π must be replaced with VP. This is done through the parse rule PReduce (P = Pro-speech), which essentially allows the parser to "pretend" that there is some rule $M \rightarrow \pi$ in R_G for any variable $M \in V_G$, and to reduce accordingly.

(56) **PReduce:**
$$\frac{[\beta\ (\pi, P),\ j]}{[\beta\ (M, P),\ j]} \text{ where } M \in V_G$$

With these revisions to our augmented parser, we can now successfully parse (50). The full parse is provided in Table 3. Parentheses in the right-hand column indicate what is scanned in the case of Shift steps, what syntactic rule in (51c) motivates the reduction in the case of NCReduce steps, and what syntactic variable is assigned to pro-speech content in the case of PReduce steps.

Up until Step 8, this looks like any parse from our prior (augmented) SI-CFG parser: various parts of the input string are scanned with Shift, and then Reduce guesses the syntactic structure and builds an appropriate interpretation. But in Step 8, the fourth element of the input string EXPLODE must be scanned, and here the revised version of NCShift comes into play: a new pair $(\pi, \textsf{explode})$ is added to the parse item (assuming $||\text{EXPLODE}||_H = \textsf{explode}$), and the index is of course also incremented.

0	[, 0]	Axiom
1	[(the, the), 1]	NCShift(the)
2	[(D, the), 1]	NCReduce(4)
3	[(D, the) (car, car), 2]	NCShift(car)
4	[(D, the) (NP, car), 2]	NCReduce(5)
5	[(DP, the ∗ car), 2]	NCReduce(2)
6	[(DP, the ∗ car) (will, will), 3]	NCShift(will)
7	[(DP, the ∗ car) (T, will), 3]	NCReduce(6)
8	[(DP, the ∗ car) (T, will) (π, explode), 4]	NCShift(EXPLODE)
9	[(DP, the ∗ car) (T, will) (VP, explode), 4]	PReduce(VP)
10	[(DP, the ∗ car) (TP, will ∗ explode), 4]	NCReduce(3)
11	[(S, (the ∗ car) ∗ will ∗ explode), 4]	NCReduce(1)

Table 3 Parse for *The car will* EXPLODE.

In Step 9, PReduce occurs. As discussed above, PReduce allows π to be replaced with any variable in V_G, such as VP, DP, S, etc. In this case the parser chooses VP. (Note that as per the discussion above, the parser choosing VP over, say, DP is a matter for the control structure, something that we are abstracting away from.) From here, the two NCReduce steps occur in an entirely pedestrian manner, with the result being a successful parse with a derived interpretation of (the ∗ car) ∗ will ∗ explode. Thus, our revised augmented parser successfully allows us to parse the string *The car will* EXPLODE without requiring that explosion sound effects somehow be part of the grammar of the language in question.

This analysis also makes the empirically sound prediction, discussed previously, that pro-speech content must be syntactically linearized as if it was linguistic content: for example, in English pro-speech content that replaces the VP must appear only where English VPs can appear. This is because while PReduce allows us to pretend that EXPLODE is whatever syntactic category we like, once we have made that determination we must obey all of the grammatical rules required for subsequent reduction steps in order to complete a successful parse. In other words, once we commit to pretending that EXPLODE is a VP, that VP has to obey all of the same syntactic rules as actual VPs in order for the parse to be successful.

Let us once again take stock of where we currently stand. In this paper, we are seeking to furnish a model of co-/pro-speech content integration in which the co-/pro-speech content is not part of the syntactic structure—there are no heads UP or EXPLODE in our syntactic representations—but the parser is capable of integrating co- and pro-speech content into the semantic interpretation that it derives in the course of the parse. In the previous section I provided a very simplistic toy model of how this integration could be effected for co-speech content, using an augmented parser for SI-CFGs. In this section the augmented parser was extended to allow for pro-speech content to be both semantically and syntactically integrated. We thus have a toy model that can handle both co- and pro-speech content in a PH-friendly manner, allowing for both kinds of integration without making undesir-

able claims about the human language faculty, and in particular about the nature of syntactic representations and their relation to overt externalization.

7 Conclusion

In this paper I have argued against a **Grammatical Hypothesis (GH)** of co- and pro-speech gesture integration, on the grounds that it demands too much of the grammar. According to GH, syntax is modality-blind, and gestures are in some form or another part of the syntactic representations generated by the grammar, with the differences between gesture and speech being reduced to phonology. But I have argued that this view is untenable in light of the seemingly enormous variety of content that can be integrated in a gesture-like fashion: an approach to the grammar in which the overt externalization of syntactic structures can include pre-recorded sound effects, emoji, cutting flowers, ripping heads off of teddy bears, (use of) cardboard signs, and the painful firing of neural implants is highly implausible.

Given these arguments against GH, I have argued in favor of an alternative **Parsing Hypothesis (PH)**, in which co- and pro-speech content is integrated not in the grammar, but in the human parser, which folds in the interpretation of co-speech content and can "pretend" that pro-speech content is linguistic material in the relevant sense. I then developed a toy model illustrating what an analysis in accordance with PH might look like. This included a formal grammar and parser for purely linguistic material, an augmented parser that allows for the semantic integration of co-speech content, and a revision to that augmented parser to allow for the possibility of pro-speech integration. The result was an analysis of co- and pro-speech integration that allowed the parser to successfully integrate material without positing that that material somehow be encoded in the grammar proper.

Since the toy grammar and parser were intended more as a proof of concept than as a full-blown implementation of PH, there were numerous areas in which I adopted simplifications or abstracted away from issues that were not directly relevant to the problem at hand. Among the most important of these were the following:

1. The syntax of the grammar in question was treated as (a slightly revised version of) a context-free grammar with no PF/LF distinction, i.e., no distinction between pronounced and semantically interpreted structures.
2. Rather than adopting a model-theoretic semantics of the sort more commonly seen in the semantic literature, I have simply assumed some class of semantic objects Sem, with a binary composition operation $*$ and a (possibly identical, possibly distinct) binary co-speech composition operation $\star$.
3. With respect to the parser, I have ignored issues related to the control structure—that is, the algorithm determining which parse steps to take in those cases where multiple steps are possible—and have stuck to the parsing schema, meaning that all parses in this paper are such that the parser makes the (or a) correct move at every step.
4. I have ignored issues pertaining to the frequent *iconicity* of co-/pro-speech content.

5. I have to a certain extent abstracted away from the complicated issue of *co-speech alignment* by simply assuming that co-speech content always aligns with the first word of the constituent it semantically composes with.
6. I have assumed that the input to the parser is pre-processed, not only in generating a string of words from a complex audiovisual signal, but also in determining which non-linguistic content qualifies as co- versus pro-speech.

I have suggested various ways in which these and other simplifications might be improved upon in future work, especially with respect to bringing the grammar more in line with contemporary syntactic-semantic assumptions. On the semantic side, this would involve model-theoretic implementations with composition operations like function application and predicate modification. On the syntactic side, I have suggested that Minimalist Grammars (MGs, Stabler 1997), and more specifically PF+LF MGs of the sort developed by Pasternak & Graf (2021), could prove useful in their capacity as formally explicit grammars with well-defined parsers, with the grammars themselves more closely adhering to contemporary syntactic assumptions (e.g., feature-driven merge and move operations).

Beyond simply improving upon the basic analysis adopted in this paper, there are various empirical matters that still need to be figured out. For example, as was discussed above, it is entirely compatible with PH that all co- and pro-speech content is semantically integrated in the parser rather than the grammar, but that different types of content are semantically integrated in different ways (e.g., with distinct projection profiles). While theoretical arguments from Pasternak (2019) and Pierini (to appear) and experimental evidence from Pasternak & Tieu (in revision) suggest that sound effects and emoji behave in the same way as gestures, there is a great deal of work left to do in determining to what extent this is shared with other kinds of content. If there are differences, then an explanation is needed as to why, and the model must be revised in order to allow for this possibility.

Sign languages also pose particular problems for the theory of gesture that need to be addressed in future work. While sign languages are clearly full languages with their own syntax, phonology, lexicon, etc., and thus are not reducible to sequences of gestures, utterances in sign languages frequently have very gesture-like components (beyond the obvious use of the visual modality), including iconicity. Given that possible co-sign gestures and signs themselves occupy the same modality and share the same articulators (e.g., hands, arms, eyebrows), in contrast to speech and co-speech gestures, determining which parts of a signed utterance are part of the discrete combinatorial system of language and which (if any) are co-sign gestures is very much not a trivial task. Nonetheless, some scholars have argued that there are indeed co-sign gestures, and that these co-sign gestures share certain core traits in common with their co-speech counterparts (see, e.g., Davidson 2015; Aristodemo 2017; Goldin-Meadow & Brentari 2017; Schlenker 2018b). If co-sign gestures exist as well, then according to PH (or at least the strongest version thereof) they should be handled in much the same way as co-speech gestures, i.e., semantically integrated at the level of the parser rather than at the level of the grammar. But as mentioned above, divvying up signed utterances into linguistic and co-/pro-speech content is not an easy task given how closely the

two are integrated in sign languages, and more work is needed to determine how they can be differentiated in principle, as well as how the parser differentiates them in real time.

Finally, if we accept the premise that certain parts of compositional interpretation arise in the parser rather than the grammar, then there may be other compositional phenomena that have been treated as grammatical in origin, and that would receive a better explanation in a parsing-based account. For example, consider scalar implicatures. One of the arguments in the recent implicature literature favoring a syntactic-semantic approach over a pragmatic approach is the existence of *embedded implicatures*, in which scalar implicatures are generated within the scope of structural operators like negation (Chierchia et al. 2012). But treating scalar implicatures as instead being generated by the parser could equally well account for the presence of embedded implicatures without the existence of an exhaustification head Exh of the sort commonly proposed in this literature, much like how on the analysis in this paper gestures can scope below spoken operators without appearing in the syntactic representation. This is of course not an argument against structural approaches to implicature generation, as the observation that (embedded) implicatures *could* in principle be generated in the parser does not thereby entail that they *are*. However, it is an intriguing possibility, and it would be worthwhile to explore in future work the extent to which these two approaches can be differentiated on empirical and theory-internal grounds.

References

Anvari, Amir. 2017. Dislocated co-suppositions. In Alexandre Cremers, Thom van Gessel & Floris Roelofsen (eds.), *Proceedings of the 21st Amsterdam Colloquium*, 106–114. Amsterdam: ILLC.

Aristodemo, Valentina. 2017. *Gradable constructions in Italian Sign Language*. Paris: École des Hautes Études en Sciences Sociales PhD dissertation.

Bolinger, Dwight. 1967. Adjectives in English: Attribution and predication. *Lingua* 18. 1–34.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Klaus von Heusinger, Claudia Maienborn & Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning, Volume 3*, 2297–2331. Berlin: De Gruyter.

Chomsky, Noam. 1957. *Syntactic Structures*. Berlin: de Gruyter.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam & George A. Miller. 1963. Introduction to the formal analysis of natural languages. In R. Duncan Luce, Robert R. Bush & Eugene Galanter (eds.), *Handbook of Mathematical Psychology, Vol. II*, 269–321. New York, NY: Wiley.

Davidson, Kathryn. 2015. Quotation, demonstration, and iconicity. *Linguistics and Philosophy* 38(6). 477–520.

De Santo, Aniello. 2019. Testing a Minimalist grammar parser on Italian relative clause asymmetries. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL) 2019*, June 6 2019, Minneapolis, Minnesota.

Ebert, Cornelia & Christian Ebert. 2014. Gestures, demonstratives, and the attributive/referential distinction. Slides from a talk given at Semantics and Philosophy in Europe (SPE 7).

Esipova, Maria. 2018. Focus on what's not at issue: Gestures, presuppositions, appositives under

contrastive focus. In Uli Sauerland & Stephanie Solt (eds.), *Proceedings of Sinn und Bedeutung 22*, 385–402. Berlin: ZAS.

Esipova, Maria. 2019. *Composition and projection in speech and gesture*. New York, NY: NYU PhD dissertation.

Gerth, Sabrina. 2015. *Memory limits in sentence comprehension: A structural-based complexity metric of processing difficulty*. Potsdam: Universität Potsdam PhD dissertation.

Goldin-Meadow, Susan & Diane Brentari. 2017. Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences* 40. e46.

Graf, Thomas, Brigitta Fodor, James Monette, Gianpaul Rachiele, Aunika Warren & Chong Zhang. 2015. A refined notion of memory usage for Minimalist parsing. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, 1–14. Chicago, IL: Association for Computational Linguistics.

Graf, Thomas & Bradley Marcinek. 2014. Evaluating evaluation metrics for Minimalist parsing. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, 28–36.

Graf, Thomas, James Monette & Chong Zhang. 2017. Relative clauses as a benchmark for Minimalist parsing. *Journal of Language Modelling* 5(1). 57–106.

Harkema, Henk. 2001. *Parsing Minimalist languages*. Los Angeles, CA: UCLA PhD dissertation.

Heim, Irene. 1983. On the projection problem for presuppositions. In *Proceedings of the Second West Coast Conference on Formal Linguistics*, 114–125. Stanford, CA: Stanford University Press.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Malden, MA: Blackwell.

Hunter, Julie. 2019. Relating gesture to speech: reflections on the role of conditional presuppositions. *Linguistics and Philosophy* 42(4). 317–332.

Iverson, Jana M. & Susan Goldin-Meadow. 2005. Gesture paves the way for language development. *Psychological Science* 16(5). 367–371.

Joshi, Aravind K. 1990. Processing crossed and nested dependencies: an automaton perspective on the psycholinguistic results. *Language and Cognitive Processes* 5(1). 1–27.

Kobele, Gregory M. 2006. *Generating copies: An investigation into structural identity in language and grammar*. Los Angeles, CA: UCLA PhD dissertation.

Kobele, Gregory M., Sabrina Gerth & John T. Hale. 2013. Memory resource allocation in top-down Minimalist parsing. In Glyn Morrill & Mark-Jan Nederhof (eds.), *Formal Grammar: 17th and 18th International Conferences*, 32–51.

Larson, Richard K. & Franc Marušić. 2004. On indefinite pronouns with APs: Reply to Kishimoto. *Linguistic Inquiry* 35(2). 268–287.

Lascarides, Alex & Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26(3). 393–449.

Lee, So Young. 2019. A Minimalist parsing account of attachment ambiguity in English and Korean. *Journal of Cognitive Science* 3(19). 291–329.

Leffel, Timothy. 2014. *The semantics of modification: Adjectives, nouns, and order*. New York, NY: NYU PhD dissertation.

Liu, Lei. 2018. Minimalist parsing of Heavy NP Shift. In *Proceedings of PACLIC 32 (The 32nd Pacific Asia Conference on Language, Information and Computation)*, Hong Kong: Association for Computational Linguistics.

May, Robert. 1977. *The grammar of quantification*. Cambridge, MA: MIT PhD dissertation.

McNeill, David. 2005. *Gesture and Thought*. Chicago: University of Chicago Press.

Morzycki, Marcin. 2008. Nonrestrictive modifiers in non-parenthetical positions. In Louise McNally & Christopher Kennedy (eds.), *Adjectives and Adverbs: Syntax, Semantics, and Discourse*, 101–122. Oxford: Oxford University Press.

Pasternak, Robert. 2019. The projection of co-speech sound effects. ZAS, Ms.

Pasternak, Robert & Thomas Graf. 2021. Cyclic scope and processing difficulty in a Minimalist parser. *Glossa* 6(1): 8. 1–34.

Pasternak, Robert & Lyn Tieu. in revision. Co-linguistic content inferences: From gestures to sound effects and emoji. ZAS and Western Sydney University, Ms.

Pereira, Fernando C.N. & David Warren. 1983. Parsing as deduction. In *21st Annual Meeting of the Association for Computational Linguistics*, 137–144. Cambridge, MA: MIT.

Pierini, Francesco. to appear. Emojis and gestures: a new typology. To appear in *Proceedings of Sinn und Bedeutung 25*.

Rambow, Owen & Aravind K. Joshi. 1994. A processing model for free word order languages. In C. Clifton, L. Frazier & K. Rayner (eds.), *Perspectives on Sentence Processing*, 267–301. Mahwah, NJ: Lawrence Erlbaum Associates.

Schlenker, Philippe. 2018a. Gesture projection and cosuppositions. *Linguistics and Philosophy* 41(3). 295–365.

Schlenker, Philippe. 2018b. Iconic pragmatics. *Natural Language & Linguistic Theory* 36(3). 877–936.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8(3). 333–343.

Stabler, Edward. 1997. Derivational minimalism. In Christian Retoré (ed.), *Logical aspects of computational linguistics* (vol. 1328 of *Lecture Notes in Computer Science*), 68–95. Berlin: Springer.

Syrett, Kristen & Jeffrey Lidz. 2011. Competence, performance, and the locality of quantifier raising: Evidence from 4-year-old children. *Linguistic Inquiry* 42(2). 305–337.

Tieu, Lyn, Robert Pasternak, Philippe Schlenker & Emmanuel Chemla. 2017. Co-speech gesture projection: Evidence from truth-value judgment and picture selection tasks. *Glossa* 2(1): 102. 1–27.

Tieu, Lyn, Robert Pasternak, Philippe Schlenker & Emmanuel Chemla. 2018. Co-speech gesture projection: Evidence from inferential judgments. *Glossa* 3(1): 109. 1–21.

Umbach, Carla. 2006. Non-restrictive modification and backgrounding. In Beáta Gyuris, László Kálmán, Chris Piñón & Károly Varasdi (eds.), *Proceedings of the Ninth Symposium on Logic and Language*, 152–159. Budapest: Hungarian Academy of Sciences.

Wurmbrand, Susi. 2018. The cost of raising quantifiers. *Glossa: a journal of general linguistics* 3(1): 19. 1–39.

Zhang, Chong. 2017. *Stacked relatives: their structure, processing and computation*. Stony Brook, NY: Stony Brook University PhD dissertation.

Zlogar, Christina & Kathryn Davidson. 2018. Effects of linguistic context on the acceptability of co-speech gestures. *Glossa* 3(73). 1–28.