What is Super Semantics?*

Philippe Schlenker**

Revised draft, March 5, 2019
(addition of stable links to audiovisual examples, and of a downloadable folder)

Final version to appear in *Philosophical Perspectives*.

Abstract. Formal semantics was born from an attempt to state explicit rules to predict the truth conditions of natural language sentences. But this goal can be extended beyond language *stricto sensu*, to a variety of representational systems in nature. This was part of the program of semiotics (e.g. Morris 1938), construed as a general theory of signs; but it never got integrated to the research program of formal semantics. We argue for such an integration, on two general grounds. First, the class of objects that interact with or display uncontroversially linguistic properties is larger than was initially thought. Spoken and especially sign language make use of rich iconic resources which interact with logical structure but cannot be captured without a 'formal semantics with iconicity'. In addition, recent comparisons between sign, speech and gestures strongly suggest that language is multi-modal and that gestures are full citizens of the linguistic world: they trigger familiar inferential types (presuppositions or supplements) when they co-occur with or follow words; and when they fully replace words, their informational content gets divided among familiar slots of the inferential typology, and they even follow a 'gestural grammar' that is in part reminiscent of sign language grammar. Second, independently from these language-internal motivations, the proposed extension leads to a far broader typology of meaning operations in nature, one that includes animal meanings, pictorial meanings, musical meanings, and more. This makes it possible to explore new connections among these domains, for instance between vocal and gestural iconicity, between musical inferences and animal signals, or between visual narratives and music.

Keywords: sign language semantics, gesture semantics, picture semantics, visual narratives, music semantics, dance semantics, iconicity, pro-speech gestures, co-speech gestures, post-speech gestures.

Note: Links to audiovisual examples have been included in the text. These examples can also be retrieved in following folder (they are indexed in the text by way of references such as [Bal-4], [MI-15], [DQ]): https://drive.google.com/file/d/1ed5o0jjNgtQEqb_zmmuezIfud7V-mGyc/

1   Introduction

1.1   Goals

Contemporary formal semantics was born from the project to treat natural languages as formal languages, not just in their syntactic component (as generative syntax had done), but also in their meaning component (e.g. Montague 1970a,b). To know the meaning of a sentence is (at least) to know under what conditions it is true, philosophers argued, and thus the methods of logic and especially model theory were deemed appropriate to investigate human language (see Heim and Kratzer 1998 for a textbook account). This project has proven extraordinarily fruitful in the last 50 years, including when it was combined with cross-linguistic and psycholinguistic studies. But it is striking that the project could in principle apply beyond sentences: for any representational form R, one could posit that 'to know the meaning of R is (at least) to know under what conditions it is true'.[1] R could for instance be a visual or an acoustic representation. While the establishment of a formal semantics for pictures and visual narratives was recently advocated in pioneering work by Greenberg (2011, 2013) and Abusch (2013, 2015), formal semantics is still mostly concerned with the meaning of morphemes, words and word complexes. On the other hand, there is another tradition, that of semiotics, which ambitioned to develop a general theory of signs, with a syntactic and especially a semantic and a pragmatic component (Charles W. Morris, 1938). What is striking, however, is the extent to which the two programs have remained distinct: formal semantics has rarely ventured outside of traditional linguistic objects; semiotics has rarely made use of the powerful logical and model-theoretic tools of formal semantics.

We will argue that formal semantics should extend its research program beyond its traditional objects, thus becoming a field of 'Super Semantics' (G. Greenberg uses the term 'formal semiotics' for essentially the same program).[2] We will provide two main arguments for this extension. First, it is necessary on purely linguistic grounds. Spoken and especially sign language make use of rich iconic resources that interact with logical and grammatical structure but cannot be captured without a 'formal semantics with iconicity'. An iconic rule specifies that an expression may only refer to things that resemble aspects of its form, as when *looong* is lengthened to refer to very long durations (and may be further lengthened to refer to extremely long ones). While this mechanism is very different from standard compositional semantics, integrating iconic forms to the research program of formal semantics is mandatory if one is to have a complete theory of meaning in natural language. In addition, recent comparisons between sign, speech and gestures strongly suggest that language can be multi-modal, and that gestures are full citizens of the linguistic world: as we will see, they trigger familiar inferential types (presuppositions or supplements) when they co-occur with or follow words; and when they fully replace words, their informational content gets divided among familiar slots of the inferential typology   (standard entailments, implicatures, presuppositions, supplements, expressives…); and they even follow a 'gestural grammar' that is in part reminiscent of sign language grammar. Strikingly, visual and possibly acoustic animations that are embedded within sentences give rise to a similar inferential typology as well. In other words, *non-standard objects display a characteristically linguistic behavior in terms of their inferential properties*. It is thus natural to integrate them to the research program of formal semantics, and one might even hope that they will bring new light to the cognitive sources of various semantic phenomena.

Second, independently from these language-internal motivations, the proposed extension leads to a far broader typology of semantic operations in nature, one that includes animal meanings, pictorial meanings, musical meanings, and more. This makes it possible to explore new connections among these domains, for instance between vocal and gestural iconicity, between musical inferences and animal signals, or between visual narratives and music. Among these connections are phylogenetic ones: by mapping shared formal properties of human and animal signs to phylogenetic

---

[1] One may be used to a different terminology. For instance, one may think that a picture is 'accurate' or that an animal call is 'applicable' in some situations but not in others. The terminology doesn't matter as long as there is a bipartition between situations that fall under the representation, and ones that don't.

[2] 'Super' is used with its latinate meaning of 'beyond', as in 'supersonic' (super semanticists may thus, without contradiction, be diminutive in stature).

trees (obtained on the basis of DNA data), it is sometimes possible to reconstruct over millions of years the evolutionary history of the form and meaning of some animal and possibly human meaning-bearing forms.

1.2 Structure

The rest of this article is organized as follows. In Section 2, we provide purely linguistic motivations for Super Semantics: the integration of iconic enrichments and gestures to natural language semantics makes the proposed extension mandatory, for sign and spoken language alike; it also leads to the observation that apparently non-linguistic objects, such as visual animations, can display surprisingly linguistic properties. In Section 3, we turn to animal languages. For the most part, they display entirely different syntactic and semantic properties from human language, but the methods of semantics can fruitfully be applied to them, for two reasons: they yield far more precise theories than are commonly found in ethology, and much theoretical action lies in the division of labor between semantics, pragmatics and world knowledge – a staple of semantic expertise. In addition, a comparative approach to animal languages can be combined with genetic data to reconstruct the evolutionary history of some calls and gestures over millions years – and recent work by Hobaiter and colleagues (Kersen et al. 2018) suggests that the reconstruction extends to some human gestures. Finally, we argue for extensions of the research program of formal semantics beyond human or animal languages. The existence of a sophisticated iconic component in language dovetails with recent studies of the semantics of pictures and visual narratives, especially by Greenberg (2011, 2013) and Abusch (2013, 2015). A more abstract version of iconic semantics (a 'source-based semantics') can offer a framework for studies of music semantics. And ideas from iconic, gesture and music semantics play a prominent role in the analysis of dance form and meaning in pioneering formal work by Charnavel (2016) and Patel-Grosz et al. (2018). (Appendix I introduces our notational conventions, Appendix II-III provide complements on music semantics, and Appendix IV provides an example of semantic interaction between dance and music.)

2 Human language

2.1 Linguistic arguments for Super Semantics

In this section, we explain why purely linguistic considerations justify broadening the program of formal semantics. We will discuss five main arguments.

(i) Natural language, and especially sign language, has a productive iconic component that interacts in non-trivial ways with compositional semantics: it is not in general possible to state the truth conditions of a sentence as the conjunction of a standard, compositional component, and of an iconic component. Therefore the very program of formal semantics requires the development of an iconic semantics: without it, no semantic theory can be complete.

(ii) The rich iconic component of sign language has raised a further question: does spoken language have similar means of iconic enrichment when gestures are taken into account? A precise answer requires distinguishing among iconic enrichments: some are at-issue (i.e. are standard entailments), some are presuppositional, some are supplemental (i.e. behave semantically like appositive relative clauses). To determine whether speech with gestures has the same semantic behavior as sign with iconicity, one must develop a precise formal pragmatics for iconic enrichments – and it turns out to rely on, and enrich, standard categories of formal pragmatics.

(iii) Besides iconic enrichments, there are spoken language cases in which a gesture fully replaces a word (such speech-replacing gestures are called 'pro-speech gestures'). While these gestures have an iconic semantics, we can embed them in various logical environments to determine how their informational content gets distributed among the inferential typology of language (at-issue entailments, presuppositions, implicatures, supplements, expressives, etc). The result is that nearly the full inferential typology can be replicated with these pro-speech gestures. This suggest that gestures

are first class citizens of the linguistic world. But because the gestures in question can (thanks to iconicity) be understood with little or no prior exposure, the results also suggest that language has powerful algorithms that make it possible to divide new semantic contents among the inferential typology – and gestures might offer a new tool to determine the nature of these algorithms.

(iv) While one could take these results to just show that language is multi-modal, some further extensions lead to more radical conclusions. The replication of the inferential typology with gestures can be further extended with entirely artificial visual animations; the results are composite utterances of words and animations that are impossible to produce in standard communication and yet elicit robust judgments by naive speakers. Thus whatever algorithms apply to gestures seem to apply more generally to arbitrary iconic contents.

(v) Finally, it is not just semantic properties of language that can be replicated with iconic material. Some non-trivial grammatical properties of sign languages can be guessed 'on the fly' by non-signers if presented with pro-speech gestures that display some grammatical properties of signs. This leads to two possible conclusions. One is that Universal Grammar doesn't just specify some abstract rules, but also some aspects of the form-to-function mapping; for instance, a pointing sign/gesture might be intrinsically specified as being pronominal in nature. An alternative is that some of these form-to-function mappings have deeper cognitive roots – ones that have yet to be fully uncovered.

2.2   *Iconicity and semantics*

2.2.1   *The importance of iconicity*

Iconicity has four roles to play in semantic studies. First, it produces information (i.e. truth conditions) by a completely different mechanism from standard compositional semantics, one that needs to be added to it. Second, iconic elements can enrich normal words in different ways depending on how they are combined with them (e.g. as co-occurring or following them), hence the need for an 'iconic pragmatics' that investigates the relevant typology. Third, iconic rules make it possible to create new word-like elements that have never been seen before and yet are understood on the spot; we can in this way assess whether phenomena that apply to normal words (such as the division between assertion and presupposition) might be more productive than initially thought and thus require the postulation of general algorithms. Fourth, it offers a point of comparison between linguistic and non-linguistic devices, notably visual ones that are at work in pictorial narratives.

In the rest of this section, we explain why standard compositional semantics crucially interacts with iconic semantics. We turn to iconic pragmatics in Section 0, to issues of productivity in gestural semantics in Section 2.4, and then to their extension beyond gestures in Section 2.5 (Section 2.6 discusses the existence of a gestural grammar, possibly involving productive mechanisms as well).

2.2.2   *At-issue contributions of iconic modulations*

As summarized in Schlenker 2018b, c, d iconic modifications of conventional words can modify the truth conditions of sentences. We call such cases 'iconic modulations', as they internally modify the form of an expression which, on its own, already has a conventional meaning (by contrast, co-speech and post-speech gestures provide an external enrichment to an existing form). Now one could imagine that such modulations do not interact with the rest of compositional semantics – not an absurd idea in view of Potts's (2005) proposal that this is precisely how expressives (such as *honkey*) and supplements (such as appositive relative clauses) work in language.[3] But this is definitely not how all iconic modifications function. The point can be made with a spoken word such as *long*, iconically

---

[3] Potts 2005, 2007 took expressives and supplements to make a semantic contribution in an additional dimension of meaning that didn't interact with operators, and hence yielded the impression of a 'wide scope' behavior. While this view has been criticized (e.g. Sauerland 2007, Schlenker, 2007, Schlenker, to appear g), what matters for present purposes is that this is a reasonable model of how some meaning components could be expected to work.
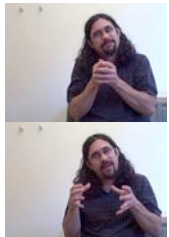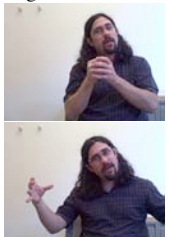
modulated by lengthening the vowel so as to evoke a very long process. The intensification can be interpreted in the scope of the *if*-clause, as is shown in (1).

(1) I am normally rather patient. But if the talk is loooong, I'll leave before the end.
     ≠> if the talk is long, the speaker will leave before the end
     => if the talk is very long, the speaker will leave before the end. (Schlenker 2018 d; see Okrent 2002)

While such iconic modulations might be rare or even anecdotal in speech, things are different in sign language. To take but one example, the sign for *GROW* in American Sign Language (ASL) can easily be modulated along two dimensions, as is illustrated in (2): the broader the end points of the sign, the larger the final size of the group; and the more rapid the movement, the quicker the growth process. While the sign itself is conventional (and would be expressed differently in other sign languages), the modulations are not, as can be seen.

(2) POSS-1 GROUP GROW.
     'My group has been growing.' (ASL, 8, 263; 264) (Schlenker et al. 2013)

(3) **Representation of *GROW***

	Narrow endpoints	Medium endpoints	Broad endpoints
Slow movement	small amount, slowly	medium amount, slowly	large amount, slowly
Fast movement	small amount, quickly	medium amount, quickly	large amount, quickly

In this case as well, the iconic modulation can be interpreted inside an *if*-clause, as illustrated in (4); this suggests that the contribution is at-issue, i.e. that it behaves like a normal entailment.

(4) *Context:* we are discussing the future of the speaker's research group.

IF POSS-1 GROUP _____, JOHN WILL LEAD.
a. [7]GROW_neutral   (ASL, 34, 1942; 2 judgments)
b. [7]GROW_large   (ASL, 34, 1944; 2 judgments)
c. [7]GROW_small   (ASL, 34, 1946; 2 judgments)

'If my group a. (really) grows / b. grows a lot / c. grows a little, John will lead it.' (Schlenker 2018d)

These observations argue for rules such as (5), which highlights that *iconic modifications can be captured by positing that some properties of form are preserved by the interpretation function.* It must be emphasized that these rules are just a 'proof of concept', and only produce information when at least two forms are compared; this is because their general format is that *if Form₁ stands in a certain relation to Form₂, then the meaning of Form₁ stands in a certain relation to the meaning of Form₂*: the conditional will be trivially satisfied if we are dealing with a single iconically modified form.

(5) **Preservation requirements on the interpretation of *GROW***
     Let $GROW_i$ and $GROW_k$ be two realizations of the sign *GROW*, and let e_i and e_k be two events of growth that are in the extension of $GROW_i$ and $GROW_k$ respectively. Then:

     a. Breadth condition
     If the end points of $GROW_i$ are less distant than those of $GROW_k$, then the endpoint of the growth in e_i should be smaller than that of the growth in e_k.

     b. Speed condition
     If $GROW_i$ is realized less fast than $GROW_k$, the growth in e_i should be slower than the growth in e_k.

In this case, iconicity just interacts with lexical meanings, but in other cases it can interact with some grammatical operations. Thus Schlenker and Lamberton, to appear show that repetition-based plurals in ASL can be modulated in rich ways, including: (i) how many iterations are produced (the more iterations, the larger the denoted quantity); (ii) how the iterations are geometrically arranged (their arrangement provides information about the shape of the denoted group). Furthermore, in both cases the iconic enrichment can be seen to make an at-issue contribution.

2.2.3 *Other cases in which iconicity interacts with compositional semantics*

Besides at-issue uses, iconic enrichments can play a more subtle role as well: first, they can contribute presuppositions on the value of pronouns (just like gender features, for instance); second, some new discourse referents can be created on purely iconic grounds, which implies that no theory of anaphora can disregard them.

Pronouns in sign language are typically realized by pointing towards the position of the denoted individuals if they are present in the context, and otherwise by assigning them a position (a 'locus') in signing space. It was argued that these positions are the visible realization of discourse referents, or at least are closely associated with them (Lillo-Martin and Klima 1990, Schlenker 2011, Schlenker 2018b). But there is more: just like the feminine feature of *she* trigger a presupposition that the denoted person is female, ASL and LSF pronouns can have high specifications (realized by pointing upwards) that trigger the presupposition that the denoted person is tall, powerful or important. Thus in (6)a,b, the relevant inference projects out of the scope of negation, as one expects of a presupposition.

(6)   YESTERDAY IX-1 SEE R [= body-anchored proper name[4]]. IX-1 NOT UNDERSTAND IX-a$^{\text{high / normal / low}}$.

a. [7] High locus.   Inference: R is tall, or powerful/important
b. [7] Normal locus.   Inference: nothing special
c. [7] Low locus.   Inference: R is short
'Yesterday I saw R [= body-anchored proper name]. I didn't understand him.' (ASL, 11, 24; Schlenker et al. 2013)

It was also shown that modifications of these cases can display 'iconicity in action' (Liddell 2003, Schlenker et al. 2013, Schlenker 2014, Schlenker 2018b). In a nutshell, sign language loci are simultaneously discourse referents and simplified pictures of their denotations. One typically points towards the part of the representation that corresponds to the head. As a result, when the denoted individuals are rotated in various positions (e.g. because they are training to become astronauts, as in an example discussed in Schlenker 2014), the loci get rotated as well: if one points high for a tall individual in standing position, one will point low when the same individual is in upside down position.

A particularly interesting case of interaction between iconicity and compositional semantics involves iconically inferred discourse referents. Two have been particularly discussed in the literature: one pertains to plural pronouns, the other repetition-based nominal plurals.

Plural loci may be realized by pointing towards semi-circular areas. In special cases, one may sign a plural locus within another one, as is illustrated in (7)-(8): a large locus *ab* denotes the set of all students, while a sublocus *a* denotes the set of students that came to class. As shown in (7)b,c, one obtains different readings depending on whether one points towards the large locus *ab* or towards the sublocus *a*.

---

[4] This proper name is realized in the chin, with the advantage that it cannot introduce a locus on its own. As a result, the following pronoun can freely introduce a high, a low or a normal locus.

(7) POSS-1 STUDENT IX-arc-ab MOST IX-arc-a a-CAME CLASS.
'Most of my students came to class.'
a. [7] IX-arc-b b-STAY HOME
'They [= the students who didn't come] stayed home.'
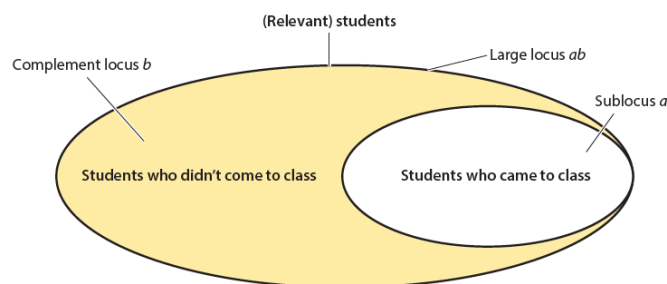b. [7] IX-arc-a a-ASK-1 GOOD QUESTION
'They [= the students who came] asked me good questions.'
c. [7] IX-arc-ab SERIOUS CLASS.
'They [= the students] are a serious class.'
(ASL, 8, 196; Schlenker et al. 2013)

(8)

(figure from Schlenker 2017d)

An interesting phenomenon arises in (7)a: by pointing towards the complement of the sublocus *a* within the large locus *ab* (i.e. by pointing towards the sublocus *b*), one obtains a reading on which the pronoun refers to the students who did not come to class. This reading cannot be obtained in English in the following discourse: *Most students came to class. They stayed home instead*. Nor can it be obtained in ASL when a single default locus is used in a modification of (7): the readings in (7)b,c remain but the 'complement set' locus in (7)a disappears. Thus on its own, the grammar of ASL, just like the grammar of English, does not provide a discourse referent denoting the students that didn't come to class. This is where iconicity kicks in: the mere presence of a large locus *ab* denoting the students and of a sublocus *a* denoting the students that came to class has two consequences. First, a 'complement locus' *b* pops into existence; second, its denotation is specified to preserve the complement relation and thus to denote the students that didn't come to class. In sum, it is because the interpretation function preserves the inclusion and complement relations among loci that this 'complement set' reading can be obtained in the end – thanks to an iconic rule.[5]

Compositional semantics is concerned, among others, with relations of anaphora. As shown in this example, discourse referents can come into existence on purely iconic grounds. But another case is discussed by Schlenker and Lamberton (to appear), who argue that nominal repetition-based plurals lead to the same conclusion. Specifically, an unpunctuated repetition of *TROPHY* on a horizontal line yields the inference that there were trophies horizontally arranged; a triangle-shaped repetition yields the inference that there were trophies arranged as a triangle. But in the first case, one can point towards either edge to refer to the singular trophy found at the left-most or right-most edge. In the triangular case, any of the three tips (vertices) can be anaphorically recovered in this way. Thus the repetition-based plural gives rise to inferred singular discourse referents, apparently on iconic grounds: only objects that can be individuated by their presence at a vertex become available for further anaphoric uptake (exactly why this is remains an open question).

---

[5] Note that the complement set reading can be paraphrased as: *The others stayed home instead*. What is remarkable about (7)a is not that it can express this meaning, but it does so with normal pronouns and without the word *OTHER*.

2.3   Typology of iconic enrichments

2.3.1   Sign with iconicity vs. speech with gestures

While the importance of iconic enrichments in sign language is clear, it raises important questions about the study of Universal Semantics, i.e. of the range of semantic options available to human language. One possible view is that, with respect to iconicity, spoken language is impoverished relative to sign language. If so, along certain dimensions, the full expressive power of the semantics of natural language might be better studied in sign than in speech: both have the same general grammatical and logical properties (Sandler and Lillo-Martin 2006, Schlenker 2018b), but sign language has greater iconic resources. From this perspective, basing a theory of iconicity on English would be like building a theory of case on the sole basis of English data: while English has case distinctions (*I* vs. *me*, *she* vs. *her*, *he* vs. *him*), far richer ones are found in other languages such as Russian, Finnish or Lithuanian. Similarly, there are iconic enrichments of English words such as *loooong*, but there are far richer iconic possibilities in sign language: as we saw above, *GROW* can be modulated both in terms of size and of speed (see Schlenker 2018b for rich iconic modulations of *UNDERSTAND* and *REFLECT* in LSF). There is an alternative view, however, namely that our comparison is unfair to spoken language because it fails to consider the contributions of co-speech gestures: as intimated by Goldin-Meadow and Brentari 2017, sign with iconicity should be compared to speech with gesture rather than to speech alone. In fact, due to strong similarities that have been found between manual gestures, (non-grammatical) facial expressions and onomatopoeias or 'vocal gestures' (e.g. Schlenker 2018d), we will take the challenge to lie in integrating all three enrichment types to speech (and in some cases to sign).

This methodological point is correct, but *even* when gestures are taken into account, systematic differences remain between sign with iconicity and speech with gestures. The reason was already clear in Ebert's pioneering work on co-speech gesture projection (Ebert and Ebert 2014; see also Ebert 2018). While we saw above that iconic modulations of words and signs can make at-issue contributions (though they may also make other types of contributions, as in the case of high pointing signs), Ebert emphasized that co-speech gestures make non-at-issue contributions. For her, this was because they contribute supplements. Later work argued instead that co-speech gestures trigger a species of conditionalized presuppositions, called cosuppositions (for instance, *help* co-occurring with a lifting gesture was taken to trigger a presupposition of the form *if x helps y, lifting will be involved*). Either way, speech with such co-speech gestures does not in general have the same semantic properties as sign with iconic modulations.

The typology of iconic enrichments doesn't just include iconic modulations (illustrated above with *loooong*) and co-speech gestures. There are two further iconic contributions that bear mentioning. *Post-speech gestures* follow the expressions they modify. And *pro-speech gestures* fully replace some words. As we will now see, none of these gesture types have quite the same properties as iconic modulations. In the full typology we will argue for, *co-speech gestures* trigger cosuppositions and are thus not initially at-issue.[6] *Post-speech gestures* trigger supplement, i.e. the same type of meaning as appositive relative clauses (Potts 2005), and thus they too fail to be at-issue. *Pro-speech gestures*, for their part, make an at-issue contribution – but unlike signs (including ones with iconic modulations) they are not conventional words at all, and are thus expressively limited for other reasons.

This typology contributes to the study of 'composite' utterances made of words and iconic depictions, whose importance and diversity was forcefully highlighted by Clark 2016. Clark analyzed these depictions as "physical scenes that people stage for others to use in imagining the scenes they are depicting". This raises two questions. First, how is the semantic content to be formally captured? No general account exists, but the projection-based semantics used for pictures and visual narratives by Greenberg and Abusch (further discussed in Section 4) could prove to be a good model. Second, how are these depictions semantically and grammatically integrated to the sentences they appear in?

---

[6] We say that cosuppositions are not 'initially' at-issue because presuppositions (and thus cosuppositions as well) can, at some cost, be turned into at-issue contributions by the operation of 'local accommodation' (e.g. Heim 1983).

Timing matters, as we will see below: a gesture does not make the same type of semantic contribution if it co-occurs, follows or fully replaces a word. Within gestures that fully replace words, iconic content is by no means unstructured: content is distributed among familiar slots of the inferential typology of language, as we will see in Section 2.4; and there are even traces of a *bona fide* gestural grammar, as we will see in Section 2.6.

2.3.2 Typology[7]

To introduce the typological issue, let us consider the examples in (9). (9)a involves a slapping gesture which co-occurs with the verb punish; it is for this reason called a 'co-speech gesture'. In (9)b, the gesture appears instead after the Verb Phrase it modifies; it is thus called a 'post-speech gesture'. In (9)c, the slapping gesture fully replaces the verb; it is called a 'pro-speech gesture'.[8] In (9)d, a conventional word, *long*, is modified in an iconic fashion by way of an 'iconic modulation' (which by definition is always the modification of a conventional form). The same terminology is extended to sign language by replacing -*speech* with -*sign*.

(9)   a. Co-speech gestures (co-occur with the word they modify [boldfaced])

   I will   **punish** my enemy.
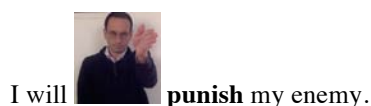
   b. Post-speech gestures (follow the word they modify)

I will punish my enemy   –   .
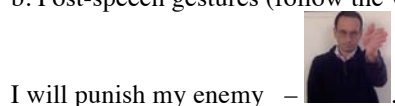
   c. Pro-speech gestures (replace a word)
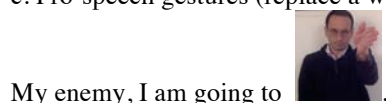
My enemy, I am going to   .
   d. Iconic modulations (modify the form of a conventional word)
   The talk was loooooong.

The main idea behind the typology we will summarize (following Schlenker 2018d) is that different iconic enrichments make different types of contributions depending on whether they are external to the relevant words (and are thus syntactically eliminable), and whether they have a separate time slot. Co-speech/co-sign and post-speech/post-sign gestures are external: they can be eliminated without affecting the integrity of the modified words and the grammaticality of the sentence. Iconic modulations and pro-speech gestures cannot be eliminated in this way. Co-speech/co-sign gestures and iconic modulations do not have a separate time slot; post-speech/post-sign gestures as well as pro-speech gestures do.

The typology is illustrated in (10): in speech and sign alike, iconic modulations can make at-issue contributions, while co-speech gestures and co-sign facial expressions contribute cosuppositions, and post-speech and post-sign gestures and facial expressions contribute supplements, i.e. make the same kind of contributions as appositive relative clauses.

---

[7] This section borrows in part from Schlenker 2018c.
[8] This is the same prefix *pro* that is found in *pronoun* (replacing a noun) and in *proconsul* (someone who acts on behalf of a consul).

(10) **Typology of iconic enrichments** (after Schlenker, to appear d)

	External enrichments (= syntactically eliminable)		Internal enrichments (= syntactically ineliminable)	
	No separate time slot: **Co-speech/co-sign gestures**	Separate time slot: **Post-speech/post-sign gestures**	No separate time slot: **Iconic modulations**	Separate time slot: **Pro-speech/pro-sign gestures**
Speech	I will **punish** my enemy.	I will punish my enemy – .	The talk was loooooong.	My enemy, I am going to .
Sign	IX-arc-b NEVER [SPEND MONEY]	IX-arc-b NEVER SPEND MONEY]ᵦ – .	POSS-1 GROUP GROW_	[currently unclear]
Meaning	**cosuppositions** (= presuppositions of a special sort)	**supplements**	**at-issue or not**, depending on the case	**at-issue**, with an additional non-at-issue component in some cases

Schlenker 2018d proposes that part (but only part) of the typology can be derived from two intuitions. First, elements that are external and thus parasitic on words (in the sense that they can be disregarded without grammatical loss) should not make an at-issue contribution; this might explain why co- and post-speech gestures do not make at-issue contributions, in particular. Second, elements that have their own time slot should not make a trivial (i.e. presuppositional) contribution; this might why pro- and post-speech gestures do not solely make a presuppositional contribution, for instance.

To illustrate this typology, we note that the iconic enrichments in the positive sentences in (9) display radically different behaviors under negation, as seen in (11).

(11) a. I won't **punish** my enemy.
=> if I were to punish my enemy, slapping would be involved

b. #I won' punish my enemy – .

c. My enemy, I am not going to .
d. The talk wasn't loooong.
=>? The talk was long

(i) First, the co-speech gesture in (11)a triggers an inference that projects under negation, to the effect that *if I were to punish my enemy, slapping would be involved*. Other types of embedding studied in the literature (Schlenker 2018a, Tieu et al. 2017, 2018a) suggest that this inference projects like a presupposition; it has received a special name (cosupposition) because the inference is conditionalized on the meaning of the modified expression (here: *punish*).
(ii) Second, the post-speech gesture in (11)b is deviant after a negative statement. Recent literature (Schlenker 2018d) has argued that this is because the post-speech gesture behaves like an appositive relative clause and contributes a supplement (Potts 2005). An alternative would be to take the post-speech gesture to make an at-issue contribution, but to have an anaphoric element that cannot be resolved after a negation, as is illustrated in (12), where the denotation of *this* is hard to interpret:

(12) #I won't punish my enemy, and this will involve slapping him.

But consideration of further examples highlights the similarity with appositive relative clauses, as in the following cases, modified from Schlenker 2018d:

(13)   a. If John punishes his son – SLAP, I might scream.

=> if John punishes his son, slapping will be involved
b. If John punishes his son and this involves slapping him, I might scream.
≠> if John punishes his son, slapping will be involved
c. If John punishes his son, which will/would involve some slapping, I might scream.
=> if John punishes his son, slapping will be involved

(iii) Third, the pro-speech gesture in (11)c makes an at-issue contribution and yields neither a cosupposition nor an implicature.

(iv) Fourth, the iconic modulation in (11)d makes an at-issue contribution and triggers no conditionalized inference akin to cosuppositions. It might yield an implicature, however, to the effect that the talk was long. This is as expected if *loooong* behaves like *very long* in evoking *long* as an alternative; denying the more informative alternative *The talk wasn't long* yields the observed implicature.

Importantly, with the exception of pro-sign gestures (i.e. sign-replacing gestures, whose existence and status is still somewhat unclear), the same typology was argued to hold in speech and in sign (Schlenker 2018d). Disgusted (non-grammaticalized) co-sign facial expressions were used to make the point: in the sign language examples in (10), the disgusted expression can either co-occur with *SPEND MONEY*, just like a co-speech gesture, or it can follow it, just like a post-speech gesture – arguably with the expected semantic results: a cosupposition in the first case, a supplement in the second. The similarity between the iconic modulation of *GROW* and that of *loooong* was also highlighted in our discussion above.

Thus the difference between iconic enrichments in speech and in sign is not one of type: the same abstract typology is found in both modalities. But iconic modulations are arguably common and rich in sign, rare and impoverished in speech. Since they may be at-issue whereas co- and post-speech modifications typically are not, this yields systematic differences between sign with iconicity and speech with gestures.[9]

2.4   *Gestural semantics*

The semantic difference between co-, post- and pro-speech gestures is certainly due to the *manner* in which they are realized, namely as co-occurring, following or replacing a word. The derivation of the typology is thus likely to stem from pragmatics (as speculated in Schlenker 2018d). But in addition, recent research suggests that different pro-speech gestures neatly fall within established categories of the 'inferential typology' of language, which includes not just at-issue entailments and supplements (as discussed above in connection with post-speech gestures), but also implicatures, standard (i.e. non-cosuppositional) presuppositions, expressives, and 'homogeneity inferences' characteristic of definite plurals.[10] This does not falsify our earlier claim that pro-speech gestures make at-issue contributions;

---

[9] As a first approximation, then, if one wants closer speech-based paraphrases of iconic modulations found in sign, one needs to add *like this* after the modified predicates, with appropriate demonstrations to realize the relevant iconic content (but see Schlenker 2018c for a discussion of some limitations of the similarity between *like this* and iconic modulations).

[10] In brief, the characteristic behavior of definite plurals stems from the fact that they behave like quasi-universal quantifiers in positive environments (possibly allowing for some exceptions in the domain of quantification), and as (quasi-)existential quantifiers in negative environments. Thus (i)a entails (*modulo* some exceptions) that Ann found all of her presents, while (ib) entails that she found none. This shows in particular that one cannot just analyze *found her presents* as meaning something like *found all her presents*, as the negated statement would then be too weak. (See Križ 2015, 2016 and Križ and Spector 2017 for a general theory, and Bar-Lev 2018 for an alternative.) Schlenker, to appear f, and Tieu et al. 2018 show that homogeneity inferences can be replicated with manual gestures and also visual animations.

(i)   a. Ann found her presents.
          b. Ann didn't find her presents.

the point is that, depending on their informational content, they may make additional contributions that reflect inferential types (and probably algorithms) that are found in normal words.

These findings hold with gestures that are likely rare, and they were obtained in an experimental setting in Tieu et al. 2018b. But the latter paper goes one step further and replicates part of the typology (pertaining to implicatures, presuppositions, supplements, and homogeneity inferences) in paradigms in which gestures are replaced with visual animations. The resulting composite utterances, made of written words and visual animations, are ones that the subjects could not have seen in a linguistic context before (because these visual animations cannot be produced with gestures – and were in any event non-standard). This suggests that subjects classify 'on the fly' new semantic content within established categories of the inferential typology of language. This, in turn, argues for the existence of productive algorithms that make it possible to do so.

2.4.1 Gestural implicatures

In some cases, the existence of such inferences is expected by current theories. Consider the case of scalar implicatures. In (14), a gesture representing a partial wheel-turning is contrasted with a complete wheel-turning. It can be checked by way of the inferences in the negative sentence in (14)a' that *not TURN-WHEEL* can mean 'not turn the wheel at all' (rather than 'not turn the wheel exactly as depicted', for instance). This suggests that the partial wheel-turning (i.e. *TURN-WHEEL*) can have a weak meaning, akin to 'turn the wheel'. As soon as it evokes a more informative alternative *COMPLETELY-TURN-WHEEL*, standard neo-Gricean theories of implicatures (e.g. Horn 1972) lead one to expect that an implicature should be derived: in the positive case in (14)a, one obtains the inference that the student should turn the wheel, but not completely. Similarly, the negative example in (14)b' (= 'not turn the wheel completely') evokes a stronger alternative meaning *not turn the wheel (at all)*. By negating this stronger alternative, an indirect implicature is triggered to the effect that the student should still turn the wheel.

(14) *A driving instructor to a student:*
     In order to get out, you

     a. should TURN-WHEEL_____.
     => you should turn the wheel a bit but not much

     b. should COMPLETELY-TURN-WHEEL_____.
     => you should completely turn the wheel

     a'. shouldn't TURN-WHEEL_____.
     => you shouldn't turn the wheel at all, OR you shouldn't turn the wheel just a bit.

     b'. you shouldn't COMPLETELY-TURN-WHEEL_____.
     => you shouldn't turn the wheel a lot but you should probably turn it a bit

While the existence of scalar implicatures in the gestural domain is unsurprising, the details raise interesting questions that go beyond standard semantics. First, how does the gesture *TURN-WHEEL* in (14)a' come to have a kind of neutral meaning corresponding to 'turn the wheel'? A simple-minded iconic semantics would lead one to expect that the gesture pictorially represents the denoted wheel-turning, which ought to give rise to a meaning akin to *turn the wheel exactly this much*. While this meaning might well be available, it is not the salient one in this case; why this is needs to be investigated.

Second, there might be a difference between the direct implicature in (14)a and the indirect implicature in (14)b: the former might not be strongly triggered in the absence of a contrast with the contrast between *TURN-WHEEL* and *COMPLETELY-TURN-WHEEL*. But in the negative case,

no such contrast seems to be needed. A possible explanation is that *COMPLETELY-TURN-WHEEL* automatically evokes *TURN-WHEEL* because it contains it as subpart, whereas the converse does not hold. This, in turn, is reminiscent of an asymmetry found with words (Katzir 2007): in the absence of an explicit contrast, *drink* does not evoke the more complex expression *drink a lot*, with the result that (15)a does not trigger the implicature that Robin didn't drink a lot. By contrast, *drink a lot* automatically evokes the simpler expression *drink*, with the result that the sentence in (15)b does trigger the implicature that Robin drank.

(15) a. Robin drank.
      ≉> Robin didn't drink a lot
    b. Robin didn't drink a lot.
      => Robin drank

If a similar contrast is found between (14)a and (14)b, it might call for a more general theory of alternative generation, one in which linguistic or non-linguistic representations alike tend to evoke simpler ones as alternatives, whereas the converse does not hold. In particular, the algorithm of alternative generation developed by Katzir (2007) might need to be extended to the case of iconic representations.

2.4.2   *Gestural presuppositions*

In contrast with scalar implicatures, presuppositions are typically thought to be encoded in the lexical meaning of words (e.g. Heim 1983), although there have been various attempts to propose 'triggering algorithms' that *deduce* the presupposition of an expression from its informational content (see Abrusán 2011 for discussion). Strikingly, pro-speech gestures can trigger presuppositions, as can be illustrated by a modification of our *TURN-WHEEL* examples: the question in (16)a triggers the inference that Mary is behind the wheel, and embedding the same gesture under a *none*-type quantifier arguably gives rise to universal projection of the inference in (16)b; this is significant because such universal projection is sometimes used as a telltale sign of presuppositional behavior (e.g. Chemla 2009; see also Zehr et al. 2015, 2016).

(16) a. Is Mary going to TURN-WHEEL_ ?
      => Mary is currently behind a wheel

    b. In this race, none of your friends is going to TURN-WHEEL_.
      => in this race, each of your friends is behind a wheel
    (Schlenker, to appear f)

The literature has now considered a variety of pro-speech gestures that similarly trigger presuppositions. The next step is to determine how these examples bear on existing 'triggering algorithms', and possibly how they could suggest new ones (see Schlenker 2018c for discussion).

2.4.3   *Further inferential types*

Scalar implicatures and presuppositions are only a beginning: setting aside the supplements triggered by post-speech gestures, pro-speech gestures can trigger expressive inferences characteristic of slurs (Potts 2005, 2007), as well as homogeneity inferences characteristic of definite plurals (Schlenker, to appear f). Experimental results have confirmed the reality of gestural implicatures, presuppositions, supplements, and homogeneity inferences (Tieu et al. 2018b).

2.5   *Extensions of gestural semantics*

There are several important extensions of this program, sketched above with respect to manual gestures.

(i) Do facial expressions participate in the same typologies? With respect to co- and post-speech gestures, the recent literature answers 'yes' (Schlenker 2018d): a disgusted facial expression can trigger a cosupposition when it co-occurs with some words it modifies, and a supplement when it follows them. As hinted above, such (non-grammatical) facial expressions provide a useful bridge between gestures in spoken and in sign language: data from ASL suggest that these non-grammatical facial expressions display the expected semantic behavior in view of the typology in (10). With respect to the typology of inferences triggered by *pro*-speech gestures, by contrast, no comparable results currently exist for facial expressions.

(ii) Do vocal gestures (= onomatopoeias) participate in the same typologies? With respect to post-speech and pro-speech vocal gestures (i.e. onomatopoeias following or replacing words), Schlenker 2018d suggested a positive answer (this justified using the term 'gesture' both in the manual and in the vocal modality). For co-speech gestures, the difficulty is that producing a vocal gesture *while* uttering a spoken word is… non-trivial. With respect to the typology of inferences triggered by pro-speech gestures, ongoing work by Janek Guerrini suggests that most or all results about manual gestures can be replicated with vocal gestures, in the areas of scalar implicatures, presuppositions, supplements, expressives and possibly homogeneity inferences (Guerrini and Schlenker 2019).[11]

(iii) Results on manual, facial and vocal gestures might simply argue for an extension of one's concept of 'language'. It need not come as a surprise that language is multi-modal, nor that non-standard vocal expressions might participate in the same types of semantic typologies as words. But Tieu et al. 2018b argue for a more radical conclusion. They replicate all their results (pertaining to implicatures, presuppositions, supplements and homogeneity inferences) with composite utterances made of written words and of visual animations. Their visual animations are in no way linguistic, and couldn't be reproduced with gestures (for instance because they involve changes of color). Their conclusion is twofold. First, the informational content of entirely novel stimuli (which couldn't have been previously experienced because they cannot be produced by speakers) is productively divided among different parts of the inferential typology. This suggests that this classification is effected by productive algorithms and isn't just memorized as part of the lexicon. Second, these algorithms can apply without difficulty to non-linguistic stimuli, which suggests that they might have a broader cognitive origin.[12]

2.6 *Gestural grammar*[13]

Tieu et al.'s replication of numerous semantic results on gestures with visual animations might suggest that there is nothing grammatical about these generalizations. But in some cases, this is clearly incorrect. The reason is that speakers appear to have some gestural judgments that track some sign language rules that are standardly classified as 'grammatical'. Needless to say, this implies in no way that sign languages are 'merely' gestural: their sophisticated grammars have been described in great detail, and share multiple properties with those of spoken languages (see Sandler and Lillo-Martin 2006 for a survey; there are also shared typological properties among sign languages). Rather, the argument is that despite their expressive limitations, gestures have a proto-grammar reminiscent of sign language.

These properties are particularly striking in the case of pro-speech gestures, for two reasons. First, pro-speech gestures must fulfill some grammatical functions on their own (precisely because they replace rather than accompany words). Second, they might make it possible to test instances of 'zero-shot grammatical learning' because they are arguably extremely uncommon (although their frequency would need to be assessed more rigorously).

---

[11] Rob Pasternak is currently exploring the linguistic behavior of co-speech music, which might trigger cosuppositions as well.

[12] Valentin Richard is currently extending these results to acoustic animations.

[13] This section borrows from Schlenker, to appear h, which summarizes results from Schlenker and Chemla 2018.

Two related examples will make this line of research concrete (we follow Schlenker, to appear h, and also Schlenker and Chemla 2018). Consider first the example in (17), pronounced in English with some specific gestures (see below and Appendix I for our transcription conventions).

(17) Whenever I can hire IX-hand-a **[a mathematician]** or IX-hand-b **[a sociologist]**, I pick IX-a.
        Meaning: whenever I can hire a mathematician or a sociologist, I pick the former.

The first disjunct *a mathematician* is pronounced with an open hand (palm up) on the right (glossed as *IX-hand-a*, and preceding the co-occurring expression, which is boldfaced), while the second disjunct *a sociologist* co-occurs with an open hand on the left (glossed as *IX-hand-b*). Schlenker, to appear h argues that these are gestural counterparts of 'loci', positions in signing space that instantiate discourse referents or variables (Lillo-Martin and Klima 1990). As a result, when the sentence-final object of *pick* is replaced with a pointing gesture towards the right (glossed as *IX-a*), we obtain a sentence that is acceptable, and has a 'donkey' reading on which the gestural 'pronoun' is dependent on the (non-c-commanding) existential quantifier. It is worth noting that in this case *him* or *her* could be ambiguous between the two antecedents, whereas the pointing gesture isn't: it is clear that the gesture is not just a code for a word.

A second example of gestural grammar has the advantage of having been studied with experimental means. In ASL, some 'agreement verbs' (= 'directional verbs') include loci in their realization. For instance, *I give you* could be realized with a movement going from the signer to the addressee, and is for this recent glossed as *1-GIVE-2* ; *I give him* starts from the signer's position and targets a third person locus, for instance *a* - in which case it is glossed as *1-GIVE-a*. These incorporated loci have been argued to display the behavior of agreement markers (Lillo-Martin and Meier 2011), although alternative analyses have been offered as well (e.g. Liddell 2003; see Pfau et al. 2018 and Schembri et al. 2018 for a recent version of the debate). Schlenker and Chemla 2018 argue that agreement verbs have gestural counterparts. They further argue that the gestural construction resembles its sign counterpart in its behavior with respect to ellipsis and focus-sensitive environments involving *only*. To introduce these findings, let us start by considering the ASL paradigm in (18), constructed around the agreement verb *1-GIVE-2* or *1-GIVE-a*.

(18)   a.[7] POSS-2 YOUNG BROTHER MONEY IX-1 1-GIVE-a. IX-2 IX-1 NOT.
        'Your younger brother, I would give money to. You, I wouldn't.'
        b.[4,7] POSS-2 YOUNG BROTHER MONEY IX-1 1-GIVE-a. **IX-2** IX-1 NOT 1-GIVE-**a**.
        c.[7] POSS-2 YOUNG BROTHER MONEY IX-1 1-GIVE-a. IX-2 IX-1 NOT 1-GIVE-2.
        'Your younger brother, I would give money to. You, I wouldn't give money to.'
        (ASL, 34, 1558; 4 judgments)

Here the verb *GIVE* is realized by way of a movement from the first person locus *1* to the third person locus *a* (hence: *1-GIVE-a)* or to the second person locus *2* (*1-GIVE-2*). (18)b,c are controls without ellipsis: they establish, unsurprisingly, that a second person object must trigger second person object agreement. But (18)a shows that under ellipsis things are different: on the assumption that the missing verb is copied from the antecedent clause, its object agreement marker can be disregarded in the elided clause, since otherwise the copied verb *1-GIVE-a* would take a second person object argument.

Related effects are well known in connection with *phi*-features in spoken language. This is illustrated in (19)a, where both the third person features and the feminine features of *her* are ignored under ellipsis.

(19) [Uttered by a male speaker] In my study group,
        a. Mary did her homework, and I did too.
        => available bound variable reading in the second clause
        b. Mary did her homework, and I did her homework too.
        => no bound variable reading in the second clause
        (Schlenker and Chemla 2018)

Strikingly, the ASL data can to some extent be replicated with gestural verbs in English. Things are somewhat complicated by the fact that something like the second person version seems to

do double duty as a neutral form, and hence it is glossed as *(-2)* in parentheses. Still, using a third person form with a second person object yields deviance, as shown by the boldfaced examples in (20)a.[14]

(20)   Your brother, I am going to SLAP-a_____ (/ SLAP(-2)_____),

and then you, I am going to **??PUNCH-a_____** / PUNCH(-2)_____ .
(from Schlenker and Chemla 2018)

Crucially, when the gestural predicate occurs (with a bound variable) under ellipsis-like constructions, third person locus specifications can be ignored, both in VP-ellipsis in the strict sense, as in (21)b, and in the 'stripping' construction in (21)a.

(21)   Your brother, I am going to PUNCH-a / SLAP-a / SHOOT-a, and then
     a. ['stripping'] you, too.
     b. [VP-ellipsis] you, I will as well.

Schlenker and Chemla 2018 conducted an acceptability experiment that confirmed these findings for VP-ellipsis. In three paradigms constructed after (22), acceptability was degraded in the 'mismatch' condition in (22)b relative to the 'match' condition in (22)c, as expected (aggregate results in (23)); but in addition, there was an amelioration of the mismatch under ellipsis, as in (22)a, making this case reminiscent of the behavior of phi-features in (18)a and (19)a.

(22)   a. Ellipsis-Mismatch: Your brother, I am gonna PUNCH-a, then you, too.
     b. Gesture-Mismatch: Your brother, I am gonna SLAP-a, then you, I am gonna PUNCH-a.
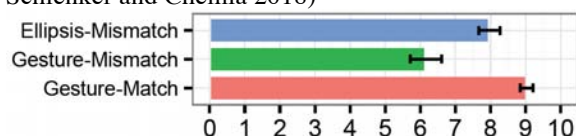   c. Gesture-Match: Your brother, I am gonna SLAP-a, then you, I am gonna PUNCH-2.

(23)   **Mean acceptability responses, averaged over in all conditions**
Error bars represent standard error to the mean after averaging across participants (for details, see Schlenker and Chemla 2018)

Schlenker, to appear h argues that several further properties of sign language grammar related to loci, plurality, telicity, and context shift, can be replicated with pro-speech gestures as well.[15] To mention but two examples, in sign language and in gestures alike, plurals can be realized by the unpunctuated repetition of an expression, and context shift (called 'Role Shift' in sign language) can arguably be represented by a body shift indicating that the signer adopts the perspective of another character.[16]

While these results need to be investigated further, they might have several repercussions. First, gestures in general and pro-speech gestures in particular might be important to understand the historical origins of sign languages. It is noteworthy that homesigners, who grow up without access to sign language, do end up developing gestural languages that share some properties of sign languages,

---

[14] With apologies to the readers, some of our examples refer to objectionable situations.

[15] While several examples involve versions of phi-features, this probably reflects a selection bias in the data that were investigated. Telicity goes beyond this class, and it is often thought that some aspects of sign language prosody, pertaining for instance to raised eyebrows, have counterparts in spoken language facial expressions (Dohen and Loevenbruck 2009, Kuhn and Chemla 2017).

[16] In spoken language, clear cases of context shift have solely been described for reported speech. This corresponds to what is called 'Attitude Role Shift' in sign language (e.g. ASL and LSF). Sign language also has an operation of 'Action Role Shift' in which the signer shifts his or her body to adopt the perspective of a character whose actions rather than words or thoughts are in this way depicted in a particularly vivid fashion. See Davidson 2015 and Schlenker 2017b, c for recent discussions.

but are expressively far more limited (e.g. Abner et al. 2015, Goldin-Meadow 2003). The reason homesigners discover such properties on their own might be that, more generally, non-signers 'know' them.

Second, an important question for future research will be to determine how these instances of 'zero-shot grammatical learning' are possible. One possible view is that Universal Grammar does not just specify the abstract form of grammatical rules, but that in the gestural/signed modality it also specifies part of the mapping between forms and grammatical/semantic content: a pointing sign/gesture might thus be intrinsically endowed with pronominal properties. Another possible view is that some signs/gestures are naturally associated with a fixed grammatical/semantic component for deeper cognitive (and non-specifically linguistic) reasons. This debate is currently open.

2.7   Intermediate conclusion

The extensions of formal semantics advocated in this section proceed 'from the ground up': starting from uncontroversially linguistic properties of human language, we show that they require an extension of the traditional program of formal semantics. Iconic enrichments of various sorts – iconic modulations, as well as co-, pro- and post-speech gestures – clearly affect truth conditions in systematic ways, and in addition trigger inferential types (at-issue, presuppositional/cosuppositional, supplemental) that are familiar from standard semantics. Pro-speech gestures on their own make it possible to replicate a large part of the inferential typology of language with purely iconic means. Finally, the existence of a gestural grammar that shares some (and only some) properties with sign language grammar provides a further argument to expand our view of the formal enterprise.

These findings might suggest that language is makes greater use of multi-modality than was traditionally thought, and that iconic enrichments (some gestural, some not) might be first class citizens of the linguistic world. In addition, the informational content of some gestures seems to be productively divided among established cells of the inferential typology; this suggests that general algorithms are responsible for this division process, including in cases (such as presupposition generation) in which their precise form is still mysterious. This conclusion is further buttressed by the observation that visual animations can replace gestures and yield the same typologies, despite the fact that they couldn't possibly have been observed sentence-internally (as they cannot be produced by speakers). But this observation also suggests a more radical conclusion, namely that the algorithms that productively divide linguistic meaning among various cells of the inferential typology can apply to non-linguistic material as well.

In the next three sections, we turn to further extensions of formal semantics, which are not motivated 'from the ground up', but rather from a desire to obtain a broader typology of meaning operations in nature. Some of these extensions interact in interesting ways with each other, but also with insights obtained from the analysis of human meaning in general and of iconic enrichment in particular.

First, the notion of implicatures and the general issue of the division of labor between literal meaning, pragmatic enrichment and contextual knowledge have a crucial role to play in studies of animal meaning (Section 3). Second, going beyond languages, a study of pictorial semantics can help provide an analysis of iconicity in language, but it can also benefit from linguistic tools (pertaining to anaphoric relations) in the analysis of visual narratives; and since visual animations behave like pro-speech gestures when they are embedded within sentences, we expect that similar results should hold of pictures and picture sequences (Section 4). Third, formal ideas from iconic semantics can be combined with inferential mechanisms from animal signals to yield an explicit (if underspecified) semantics for music; and in turn, ideas from music semantics, gesture semantics and visual narratives are likely needed to provide a semantics for dance (Section 5).

3   Animal languages

3.1   Initial questions

The investigation of meaning in animal languages (e.g. primate gestures and alarm calls) raises several questions. First, can one call such systems languages? Second, in what sense do these calls

and gestures have meaning? Third, what is gained by the application of linguistic methods to them? Fourth, what relation, if any, do they bear to human language? Fifth, should semantic analyses be restricted to intentional communication? The answers given in recent research can be summarized as follows (e.g. Schlenker et al. 2016a, b, c).

(i) One can define 'language' as one wishes, and human language is so unique that one can extract any reasonable subset of its properties to define what 'language' is. Debates on this point might not be illuminating. A more useful attitude is to treat as a language whatever can be analyzed in terms of formal language theory (e.g. Hopcroft and Ullman 1979) with respect to form, and possibly meaning. The requirement is so weak that the interesting question will not be, in salient cases, whether Species X "has language", but rather *what are the formal properties of the language of Species X* (and possibly how they compare to properties of human language).

(ii) A semantics can be defined for a system with a well-defined syntax as soon as it has truth conditions. It is uncontroversial that a given communicative call or gesture is typically applicable in certain situations and not in others. This defines a bipartition between 'true' and 'false' uses of the call (whether the *terms* 'true' and 'false' are used by ethologists is irrelevant to the nature of the theoretical problem). Predicting in detail the truth conditions of animal signals, especially when they are made of sequences of signals, is thus a natural enterprise of animal semantics.[17]

 (iii) Formal methods are useful to make clear and complete predictions. As we will see below, the interaction between various components of analyses in animal linguistics has made such methods useful.

(iv) From this methodological stance, nothing follows about the relation, or lack thereof, between human languages and animal languages. On the other hand, having uniform analytical methods for human and animal languages makes a typological approach far more productive.

(v) Initial work (summarized in Schlenker et al. 2016a, b, c) only analyzed in semantic terms signals that were understood by members of the same species, hence the importance of playback experiments that establish this point. On the other hand, this work did not require that the signals be intentional. While ape gestures are usually defined as being intentional, and some ape calls appear to be intentional as well, monkey calls studied in animal linguistics need not have this property.

All the other semantic systems studied in this piece involve an intentional behavior (this goes without saying for speech and sign, gestures, but also pictorial narratives, music and dance). If one is worried about overextending the subject matter of semantics, one possibility is to restrict it to intentional communication and to animal *signals*, which were defined as follows by Maynard Smith and Harper 2003:

"We define a 'signal' as any act or structure which alters the behavior of other organisms, which evolved because of that effect, and which is effective because the receiver's response has also evolved." (p. 3)

As the authors further explain, "the requirement that a signal evolved *because* of its effect on others distinguishes a signal from a 'cue'", which is "any feature of the world, animate or inanimate, that can be used as a guide to future action."

Still, it is worth noting that there is nothing in the definition of a semantics that involves such a restriction, nor a restriction to intentional communication. This means that precise boundaries of animal semantics (and more generally of Super Semantics) could be the object of further debates in the future.

---

[17] More truth values are typically posited in analysis of human language, for instance to handle presupposition failure (which yields a third value 'neither true nor false'). Similarly, more truth values could be entertained for animal systems as well. We restrict the discussion to the simplest case, however.

3.2   *Main results*

Results of initial studies of primate semantics may be summarized as follows (Schlenker et al. 2016 a,b,c, 2017).

(i) Fruitfulness of a formal approach: Overall, naturalistic observations and field experiments have yielded sufficiently rich and subtle observations on diverse species to make a formal analysis illuminating: informal analyses often fail to make precise predictions, and a modest formal approach can significantly help clarify competing theories. To give but two examples (discussed in greater detail below): the idea that more informative calls are preferred to less informative ones whenever possible lead to new analytical options in the analysis of Campbell's monkey calls; and precise formal analyses helped analyze away the apparent complexity of Titi monkey call sequences, treated in the end by a combination of semantic and pragmatic mechanisms, but no syntax.

(ii) No evidence for a complex syntax in primates: Birdsongs are usually thought to have a sophisticated syntax but no semantics (beyond advertising the caller's quality). In a survey by Berwick et al. 2011, birdsong syntax occupies a part of the 'finite-state' (= 'regular') component of the Chomsky hierarchy of formal languages. Initial studies of primate linguistics have not found a comparable syntax. In fact, all cases that have been analyzed in detail allow for analyses in which every call is comparable to a propositional utterance, and where ordering regularities among calls reflect changes in the calling context rather than genuine syntactic rules (to be clear: this does not mean that there are no syntactic regularities, just that their analysis does not seem to require syntactic rules). One possible exception pertains to the non-predation call *boom* in Campbell's monkeys: it usually appears as a pair at the beginning of sequences (Zuberbühler 2002). But it is also special in other ways: it is produced with air sacs that must be filled before its production, and one might imagine that this plays a role in its syntactic position (for instance because time and energy are needed before the air sacs are filled, which might make it difficult to produce this call sequence-internally).[18]

 (iii) Possible word-internal compositionality: One case plausibly involves word-internal compositionality: in Campbell's monkeys, the calls *krak* and *hok* can be suffixed with *-oo*, yielding further calls *krak-oo* and *hok-oo* with different meanings from the unsuffixed calls; we discuss this case in greater detail below.

(iv) Competition among calls: One common observation in primate linguistics and beyond is that some calls seem to function as general alert calls: they are used in highly diverse situations. Still, if the relevant species has an eagle-related call, one does not normally find the general call at the beginning of an eagle-related sequence. While this could be because the purported general call is in fact specified as being a non-eagle call, this route forces one to posit very unnatural meanings (i.e. ones that do not correspond to what are intuitively natural classes[19]). An alternative is to posit an Informativity Principle, whereby *the most specific call compatible with the caller's state must be used*. This yields a variety of 'primate implicatures': if a general call has been used, one can infer that a more specific call could not be used truly. While the cognitive foundations of the Informativity Principle might be very different from what is found in human language (e.g. it might be entirely automatic, rather than based on a theory of other minds and a cooperativity principle), this has offered a powerful tool in the analysis of call systems (for instance in the case of Campbell's monkey calls, revisited below).

---

[18] These conclusions are solely based on cases that have been studied in detail. Further findings could reverse this conclusion. Demolin et al. 2019 have thus argued that Muriqui monkeys of Brazil display a sophisticated syntax. Further potential cases of non-trivial animal syntax are discussed in Engesser and Townsend 2019 (thanks to E. Chemla for discussion of these points).

[19] The notion of a 'natural class' would need to be investigated with more rigorous means. Work on animal concepts is highly relevant: one might for instance expect that 'non-eagle' is not a natural concept and thus is unlikely to be a natural meaning. See for instance Chemla et al. 2018, to appear for relevant discussion.

(v) Further pragmatic principles: One key ongoing question is whether there are further pragmatic principles at work in primate calls. Schlenker et al. 2016e posited a further 'Urgency Principle' that mandates that calls that convey information about the nature or location of a threat come before those that don't. The goal was to re-analyze data from Arnold and Zuberbühler (2006, 2008, 2012) that suggested that Putty-nosed monkeys have 'idioms', i.e. syntactically complex sequences that have a non-compositional semantics. Schlenker et al. 2016e posited instead that each call had a (weak) propositional meaning, but took the apparently non-compositional sequences to be enriched by the Urgency Principle; whether there is independent evidence for it remains to be seen (but see fn. 20). Less speculatively, it has been argued that sophisticated pragmatic principles are at work in chimpanzees: in some cases, they appear to adapt their calling behavior to the epistemic state of their audience (Crockford et al. 2012), and great ape gesture production is thought to take into account the epistemic state of the audience, and also to be intentional (Byrne et al. 2017).

(vi) Division of labor: More generally, linguists' expertise is particularly useful when it comes to the *division of labor* among rule types such as well-formedness (= morphological and syntactic) principles, call meaning, contextual knowledge, pragmatic principles of enrichment, and rules of meaning combination (if such exist).

Conceptual and analytical issues may be disentangled with linguistic know-how, hence productive collaborations between linguists and primatologists. We hasten to add that in several cases context/world knowledge plays a crucial role in the analyses: regularities that are found in call sequence composition may reflect regularities of the evolution of the environmental context rather than a linguistic phenomenon; an example pertaining to Titi monkeys is given below.

(vi) Biological codes: In addition, there are 'biological codes' that are shared among species: calling speed has been claimed to be an increasing function of the level of urgency of the threat (e.g. Lemasson et al. 2010 on Campbell's monkeys; see also Engesser and Townsend 2019); and the register of a vocal signal provides information about the caller's size: larger sources tend to produce sounds with lower frequencies (e.g. Briefer 2012). In fact, some animals apparently evolved mechanisms – specifically, laryngeal descent – to lower their vocal-tract resonant frequencies so as to exaggerate their perceived body size (Fitch and Reby 2001). Ohala 1994 further argues that human speech makes use of this code as well (he calls it the 'frequency code'). And several acoustic cues, including higher frequency, are associated with greater stress/arousal. Briefer (2012) thus writes about non-human mammals that "with an increase in arousal, vocalizations typically become longer, louder and harsher, with higher and more variable frequencies, and they are produced at faster rates. These changes correspond closely to those described for humans".

(viii) Productivity: one particularly hard question pertains to the productivity of the systems under investigation. Campbell's monkeys appear to have a suffix *-oo* that can be attached to *krak* and to *hok*. But one could posit an alternative (if less interesting) theory on which *krakoo* and *hokoo* are not derived (except possibly in evolutionary times) from *krak* and *hok*. Similarly, general calls seem to be enriched by the negation of more specific calls through the Informativity Principle. But how can we tell that these purported general calls don't have instead a more specific semantics (involving for instance a 'non-eagle' component that explains why they are not used in eagle-related situations)? One would need sophisticated methods to explore semantic and pragmatic rules 'in action'. Two directions have been investigated.

In birds, Suzuki et al. 2017 noted that Japanese tits follow strict ordering rules among two subsequences: 'alert' subsequences regularly came before 'recruitment' subsequences.[20] To test the productivity of this principle, Suzuki et al. created hybrid sequences Japanese: tit 'alert' subsequences were combined with 'recruitment' subsequences used by neighboring species that Japanese tits are known to understand. Although the sequences were entirely novel, Japanese tits understood them in

---

[20] It was suggested in Schlenker et al. 2016b that such ordering principles might stem from a bird version of the Urgency Principle.

the same way as their native sequences, thus providing a rare and fascinating argument for productivity.[21]

In ongoing work, Chemla's team investigates the Informativity Principle 'in action' by studying cases of artificial learning in which (i) a general symbol *S* is learned, after which (ii) a more specific symbol S^+ is learned as a competitor. The Informativity Principle leads one to expect that the meaning of *S* should then be enriched to *S and not S^+*. This is indeed what is found with artificial learning in humans (Buccola et al. 2018). It remains to be seen whether this result can be replicated with animals.

3.3 *Monkey calls*[22]

To make things concrete, we will just discuss two early case studies in primate linguistics. One highlights the analytical fruitfulness of primate implicatures, while the other serves as a reminder that not every regularity found in sequences is linguistic in nature: a sequence may reflect the evolution of the context (in both cases, we closely follow the discussion in Schlenker et al. 2016c, 2017).

3.3.1 *Campbells' monkeys and the Informativity Principle*

We consider first the sophisticated call system of Campbell's monkeys of the Tai Forest, summarized in (24). Male adults have non-predation-related call, *boom*. In addition, they use a call *krak* to raise leopard alerts, and *hok* for raptor alerts. But as was briefly mentioned above, they also have suffixed calls: *krak-oo* is used for unspecific alerts, and *hok-oo* for non-ground disturbances. The initial challenge is thus to assign meanings to *boom*, *krak*, *hok*, and *-oo*.

(24) Campbell's monkey calls
a. Description   b. Analysis   c. Results of call competition

CAMPBELLS MONKEYS

Call	Typical situations
boom boom	non-predation alert
hok	presence of an eagle
krak	*Tai:* presence of a leopard
	Tiwai: unspecific alert
hok-oo	alert from above
krak-oo	unspecfic alert

Literal meanings

boom boom	non-predation alert
hok	non-ground alert
krak	alert
R-oo	weak *R*-alert

Informativity Principle
"Prefer more informative expressions!"

Calls	Competitors	Enriched meanings
hok	hok-oo	serious non-ground alert
krak	krak-oo, hok	*Tai:* alert, serious, ground
		Tiwai: useless enrichment,
		hence literal meaning only

This is not the end of the challenge, as further complexity is added by Campbell's call use on Tiwai Island, where leopards haven't been seen for decades: the Tai calls are used, but *krak* raises unspecific alerts (as does *krak-oo*), rather than leopard alerts. Should we conclude that meaning is subject to a kind of dialectal variation – as it is for *pants* in American English (meaning "trousers") vs. British English (meaning "underpants")?

While possible, the analysis with dialectal variation comes at a cost, not just because it is usually thought that the form and function of primate calls is mostly innate, but also because it yields theoretical difficulties for the analysis of the Tai call system. Theory-neutrally, in Tai *krak* is used to raise ground-related alerts, *hok* is used to raise non-ground alerts; and *hok-oo* is used for broader/weaker non-ground alerts. Thus if *-oo* has the same semantic effect on *krak* as it does on *hok*, one would expect that *krak-oo* is used for braoder/weaker ground alerts. This is entirely incorrect: even in the Tai forest, *krak-oo* is used as a general alert call. Thus in naturalistic data and field experiments alike, *krak-oo* is produced in eagle-related situations. For this reason, an analysis that posits that *krak* has a general meaning on Tiwai island but a ground predator meaning in the Tai forest must still grapple with the fact that in the latter environment the meaning of *krak-oo* seems to be derived from… the general meaning that *krak* has on Tiwai island. In other words, the analysis is

---

[21] This contrasts with work by Zuberbühler 2002 on interspecies communication between Campbell's monkeys and Diana monkeys: Diana monkeys understand Campbell's non-predation *boom* calls but cannot produce them. When a Campbell's leopard-related sequence is prefixed with *boom*, Diana monkeys know that it is not threatening any more. But when one of their own sequences are prefixed with *boom*, they just disregard the 'foreign' component.

[22] This section borrows from Schlenker et al. 2017.

forced to posit an ambiguity within the Tai environment: the *krak* from which *krak-oo* is derived has or can have a general meaning; but unsuffixed *krak* has a leopard/ground predator meaning.

These difficulties motivate the exploration of an alternative analysis. Building on the Informativity Principle in (25), Schlenker et al. 2014 cautiously propose an analysis without dialectal variation.

(25) **Informativity Principle**
 If the speaker uttered a sentence *S* which evokes ('competes with') a sentence *S'*, if *S'* is more informative than *S*, infer that *S'* is false (for if *S'* were true the speaker should have uttered it).

As a first step, they take *krak* to trigger general alerts, and *hok* to trigger non-ground alerts. As a second step, in order to analyze the meaning of the suffix *-oo*, they assume that if *R* is *krak* or *hok, R–oo* indicates a weak alert of the R-type. Thus *hok-oo* indicates a weak *(-oo)* non-ground (*hok*) alert – which is more informative than *hok*.[23]

It is in the third step that one make crucial use of the Informativity Principle, using the informativity relations in (26): *hok* competes with other calls, and because *hok-oo* (pertaining to *weak* non-ground alerts rather than to any non-ground alert) is more specific, the meaning of *hok* is enriched to *hok but not hok-oo*: it only applies to aerial (*hok*) non-weak *(not hok-oo)* alerts – hence the raptor uses. By the same logic, the unspecific alert *krak* competes with *krak-oo,* but also with *hok*. Due to this competition with two more informative calls, in the end *krak* can only be used for *serious (not krak-oo) ground (not hok) disturbances*. This comes very close to the leopard uses observed in Tai.

(26) **Informativity relations among Campbell's calls (a call asymmetrically entails another call that is is linked to and dominates)**
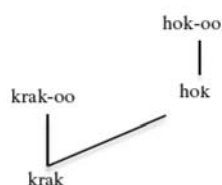
In the fourth and last step, one must account for the different use of *krak* on Tiwai island, where it raises unspecific alerts. Strikingly, this use just corresponds to the basic (unenriched) meaning of *krak*. The question is why this bare meaning fails to be pragmatically enriched on Tiwai. A plausible answer is that this would yield a useless meaning due to the absence of serious ground predators. Without the pragmatic enrichment, one is left with the literal and general meaning of *krak* on Tiwai island.

While more data will be needed to adjudicate between the two main contenders (dialectal variation vs. enrichment by the Informativity Principle), this discussion will hopefully have suggested that the space of possible theories can be considerably enriched and clarified by a formal approach.

3.3.2   *Titi monkeys and the role of contextual knowledge*

As summarized in (27), with two calls (*A* and *B*) re-arranged in various ways, Titi monkeys can provide information about both predator type (cat, raptor) and predator location (on the ground, in the canopy). Simplifying somewhat, and writing X^+ for a series of iterations of call X, B^+ is used for non-predation alerts and for situations involving a cat on the ground, while *A* B^+ is used in situations in involving a cat in the canopy. A raptor on the ground gives rise to

---

[23] Schlenker et al. 2014 discuss problems raised by this analysis, and ways to fix them (in a nutshell, the difficulty is that *krak-oo* appears in sequences that raise serious threats). The general direction is to relativize meanings to the caller's subjective state at the very time at which a call is uttered. The caller may go through phases of lesser or higher alarm even in the presence of an eagle-related situation.

an A^+B^+ sequence, while a raptor in the canopy triggers an A^+ sequence. The main patterns are summarized in (27). Should we conclude that these sequences have a complex syntax/semantics interface? Or should they be treated as very long idioms, with no internal semantics?

(27) **Titi monkey calls**
      a. Description   b. Analysis   c. Results of call competition

TITI MONKEYS

Call	Typical situations	
B^+	non-predation alert	
B^+	cat	on the ground
A B^+	cat	in the canopy
A^+ B^+	raptor	on the ground
A^+	raptor	in the canopy

Literal meanings

A	serious non-ground alert
B	alert

Informativity Principle
"Prefer more informative expressions!"

B^+ correspond to ground or weak alerts
(or alerts that *have become* weak alerts)

(Raptors in the canopy remain serious threats even after having been signaled by As)

      Due to their length and slow time course, it is unlikely that these sequences are interpreted as idioms because hearers would need to wait for too long for the meaning of the message to be effective. A simpler analysis has been explored, in which each call is in effect an independent utterance and thus contributes its meaning independently from the others (Schlenker et al. 2016d). Since the B-call is used in predatory and non-predatory situations alike, one may take it to trigger an unspecific alert. In field experiments, the A-call triggers a 'looking up' behavior, and thus one can posit that it is indicative of *serious non-ground alerts*. These assumptions explain why one finds B^+-sequences (= series of B-calls) in 'cat on the ground' situations, and A^+-sequences in 'raptor in the canopy' situations.

      But why does one find A^+B^+ in 'raptor on the ground' situations? A remark about hunting techniques proved suggestive: raptors on the ground usually attack by *flying*, hence the serious non-ground alerts A^+. Still, being on the ground isn't a typical hunting position, and after a while the alert stops being serious, which only leaves B as a possibility. In 'cat in the canopy' situations, one finds AB^+ sequences, possibly because a serious non-ground danger is indicated, which then transitions to a weaker danger because a cat becomes less dangerous after detection (Zuberbühler et al. 1999).[24]

      On this view, then, the apparent complexity of Titi sequences might reflect the interaction between simple meanings and the evolution of the contextual environment as the sequence unfolds, rather than a complex syntax/semantics interface or very long idioms. This should serve as a sobering reminder that not every regularity that is found in calling sequences is 'linguistic' in nature: sequences reflect an animal's state as it emits the various calls, and this state might well change in regular ways while it produces the sequences, giving the misleading impression of syntactically and semantically complex sequences.

3.3.3 Call evolution

Comparative studies of monkey calls have long been used to reconstruct phylogenies, (i.e. the 'family trees' of monkey species), with results that often converge with DNA methods (e.g. Gautier 1988). But one can turn the problem on its head and start from established phylogenies to reconstruct call evolution.

      Initial results are striking. *Booms* are non-predation-related calls present, not just in Campbell's monkeys, but also in many cousin subspecies of the family cercopithecines. Inspection of their distribution, as in (28), is strongly indicative of their presence in the most recent common ancestor of entire subgroups: *booms* probably existed several million years ago (Schlenker et al. 2016a).

---

[24] More recent fieldwork by Mélissa Berthet (Berthet 2018) fails to replicate the presence of a single A call at the beginning of 'cat in the canopy' situations, where she finds pure B^+ sequences. In addition, she finds two varieties of B-calls, discussed in greater detail in Berthet et al. 2018. Thus the present analyses will have to be revisited as more data become available.

(28) **The evolutionary history of *boom***
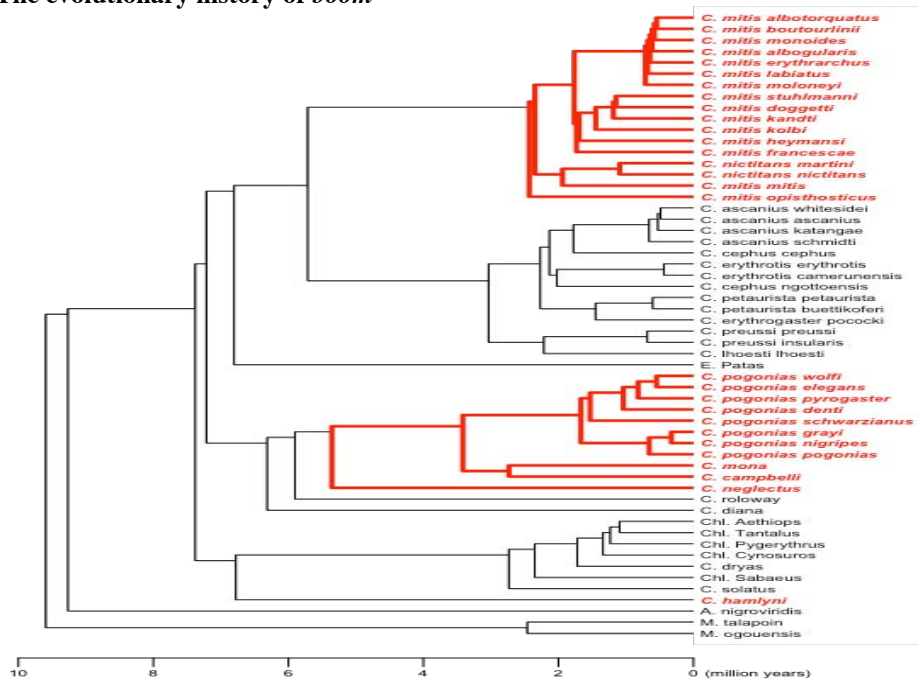
Phylogenetic tree of cercopithecines (redrawn from Schlenker et al. 2016a and Guschanski et al. 2013), with boldfaced names (also in red) for species that have *booms*. It seems very likely that the most common recent ancestor of the top boldfaced (= *mitis*) group (which lived about 2.5 million years ago) had *booms*, since all of its descendants do; and similarly for the most recent common ancestor of the middle blue group (*C. pogonias*, *C. mona*, *C. campbelli*, *C. neglectus*).

In other words, combining DNA-based phylogenetic trees with studies of call distribution across species makes it possible to reconstruct the evolution history of call form and potentially meaning over millions of years. This should be of great interest to evolutionary models of language and meaning.

3.4   Ape gestures and facial expressions

3.4.1   Ape gestures

Ape calls present the same general questions as monkey calls, but in addition their production has been argued (as mentioned above) to yield audience effects (= greater call production when the audience is ignorant) – hence interesting questions for animal pragmatics. The recent study of ape gestures, by contrast, provides qualitatively different research perspectives.

How are gestures defined? In contradistinction to monkey calls, whose intentionality is not always clear, intentionality is taken as a defining criterion of communicative gestures. Hobaiter and Byrne 2017 rely on the following definition:

We defined gestures as discrete, mechanically ineffective physical movements of the body observed during intentional communication (…). Our criterion of intentionality (at least 1st order intentional use) was applied at the level of the gesture instance, not the gesture type: thus, for every instance of gesture analysed, we had evidence that the signaller gestured with the intention of changing the recipient's behaviour, as indicated by one or more of response waiting, audience checking, and/or persistence in communication.

What are the main findings? On the negative side, there is no evidence of syntactic complexity. On the positive side, rich gestural inventories have been found, with dozens of gestures whose meaning was coded by way of 'apparently satisfactory outcomes' (= ASO's), defined by Hobaiter and Byrne 2017 as "an observable change in the recipient that apparently stops the signaller from signalling" and "must conform to some plausible biological function for the signaller". Audience effects are found in this case as well; summarizing research, Byrne et al. 2017 write: "In the wild, we found that chimpanzees were more likely to use a silent visual gesture with an audience who

was actually looking at them, and more likely to use a contact gesture with one who was not attending".

But the most striking result is that the form and to some extent the function of these gestures is preserved over millions of years.

(29) **Shared Great ape gestures (modified from Byrne et al. 2017)**
Figure re-drawn by Lucie Ravaux, adding humans (whose shared gestures with apes have not been fully investigated yet), with divergence dates drawn from Perelman et al. 2011. For clarity, we have replaced *troglodytes* with *chimpanzees*, *paniscus* with *bonobos*, *Pongo* with *orangutans*.
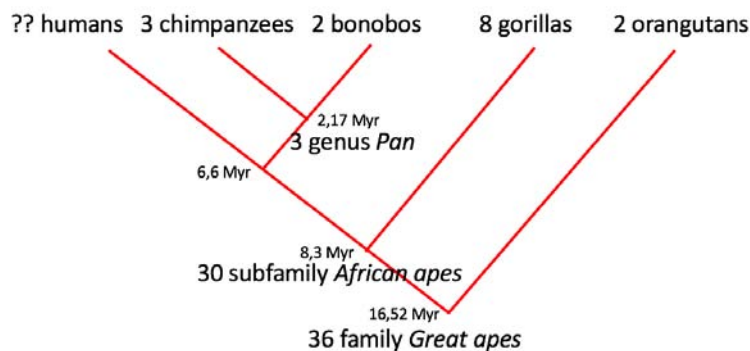
"The distribution of gestures across living great ape species and genera, based on current knowledge: numbers of gestures specific to each clade are shown, revealing extensive overlap at higher taxonomic levels. Where a gesture is found in all of *Pongo* [= orangutan], *Gorilla* and *Pan* [= chimpanzee or bonobo] it has been treated as ape-typical even if it has not yet been recorded in both troglodytes [= chimpanzee] and paniscus [= bonobo]. Note that one gesture, *big loud scratch*, appears to have been lost in the genus Gorilla, although it is of course difficult to be sure of absence."

As shown in (29), Byrne et al. 2017 argue that 36 gestures are shared among the great apes, 30 additional ones are shared among African apes, i.e. chimpanzees, bonobos and gorillas, and 3 additional ones are shared among bonobos and chimpanzees.
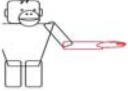
Hobaiter and Byrne 2017 argue on statistical grounds that the similarity of the repertoires is unlikely to be due to limitations of the articulatory possibilities for gestures: they generated more than a thousand morphologically possible gestures, with the result that the observed similarities across repertoires are unlikely to be due to chance and physical limitations. On the meaning side, Byrne et al. 2017 argue that "chimpanzees and bonobos were significantly more similar than expected from this randomization test in how they assigned gestures to ASOs: indeed, not a single pairing of random assignments gave a value as high as the actual similarity between the two species". Two examples of gesture comparisons are provided in (30).

(30) **Comparison of 'arm raise' and 'reach' in bonobos and chimpanzees (Graham et al. 2018)**, with percentages of ASO's for each gesture type.

Gesture Type	Bonobo ASOs	Chimpanzee ASOs
Arm raise	Climb on you 34% Initiate grooming 22% Initiate copulation 20% Initiate GG-rubbing 16% Contact 6% Climb on me 2% *Ambiguous* [9(50): f=3.13, df=12,96 p=0.0009]	Acquire object 48% Move away 19% Move closer 15% Stop behaviour 11% Climb on you 7% *Ambiguous* [x2=65.71, df=14 p<0.0001]
Reach	Climb on me 78% Acquire object 11% Climb on you 11% *Tight* [5(18): f=17.59, df=12,48 p<0.0001]	Acquire object 73% Contact 8% Climb on me 7% Move closer 7% Initiate copulation 2%[1] Climb on you 1% Move away 1% Stop behaviour 1% *Tight* [f=30.24, df=14,336 p<0.001]

The phylogenetic tree in (29) is currently missing information for… humans. Gestures that are shared among bonobos, chimpanzees and gorillas are likely to have been present in their most recent common ancestor, which was also their most recent common ancestor with humans. The next step, taken by Kersken et al. 2018, is to ask whether these shared great ape gestures are in fact found in humans. By observing human infants from Germany and Uganda with the same methods as they observed apes in the wild, they uncover a human infant gestural repertoire, with 50 gestures (96%) shared between children and other apes, and 46 gestures (89%) shared between children and chimpanzees. A further step will be to connect these infant gestures with adult gestures, and/or with

signs. It is too early to tell which connections, if any, could be drawn with the discussion of gestural semantics and gestural grammar outlined in Section 2.

It is worth noting that several ape gestures have a potentially iconic or mimetic component. Genty and Zuberbühler 2014 describe a beckoning bonobo gesture that arguably displays the direction that the subject wishes to go in with his sexual partner. Douglas and Moscovice 2015 describe cases in which female bonobos call attention to their own sexual swellings by way of heel or toe pointing. They also describe a pantomime in which female bonobos' hip movement "pantomimes the rapid lateral movement of the hips that occurs during GG [=Genital-Genital] rubbing", and is given "exclusively when soliciting GG rubbing, and not in other contexts where the action is not relevant, for example when soliciting copulations with males".[25]

3.4.2   Ape facial expressions

Facial expressions convey emotions in humans, but some also play a clear grammatical role in sign language, and possibly in spoken language as well. This is in particular the case of raised eyebrows, which are a marker of focus and topic constructions (among others) in a variety of sign languages (e.g. Wilbur 2012, Sandler 2018). Furthermore, raised eyebrows can help mark focus in spoken language as well (Dohen 2005, Dohen and Loevenbruck 2009). Lowered eyebrows can also have a grammatical function in sign language, for instance to mark *wh-* questions in ASL. Can we trace the evolutionary history of such grammatical markers over millions of years? (See Benitez-Quiroz et al. 2016 for a possible model, which seeks the evolutionary roots of the headshake used to express negation in sign and in gestures.)

While there is an ancient tradition of studying the emotional use and phylogenetic history of facial expressions in the ape lineage, including in comparison with humans (e.g. Parr and Waller 2006), the connection with grammaticalized constructions is rarely made. Strikingly, raised eyebrows seem to be used by baboons as aggression signals (Pellat 1980), and to be related to tension/aggression in some macaque species (Maestripieri 1997, Kanazawa 2016). Could the form and function of raised eyebrows have somehow been preserved over millions of years, and have been inherited from the most recent common ancestor of humans, baboons and macaques, which lived approximately 32 million years ago (Perelman et al. 2011)? An alternative possibility, however, is that raised eyebrows are the product of convergent evolution, in which case they might be a relatively recent innovation in the human lineage. Recent research gives suggestive evidence that this might indeed the case. Godinho et al. 2018 observe that an archaic human, Kabwe 1 (*Homo heidelbergensis*, dated from 125–300 thousand years ago), had a huge brow ridge that made vertical movement of the eyebrows more limited: morphological changes that occurred in modern humans entail that "the eyebrows have the potential to move vertically over a relatively larger area, and to be more readily observed and more mobile". The authors determined by way of modelling that the difference in brow ridge between Kabew 1 and modern humans could *not* be explained by morphological constraints, such as: (i) filling the space where the flat brain cases and eye sockets of archaic hominins met, and (ii) acting to stabilize the skull from the force of chewing).[26] This means that social communication could be a driver of evolution; whether *linguistic* communication played a role as well is another question. But it seems likely that archaic humans didn't have the same eyebrow movement capabilities as contemporary humans.

---

[25] Anecdotally, extraordinarily detailed iconic gestures have been recorded in a zoo chimpanzee who had extensive interaction with humans: https://www.dropbox.com/s/sc61pi5cegp8u8d/Chimp-Iconic-zoo.mp4?dl=0.

[26] The senior author, O'Higgins, summarized the research in these terms: "We used modelling software to shave back Kabwe's huge brow ridge and found that the heavy brow offered no spatial advantage as it could be greatly reduced without causing a problem. Then we simulated the forces of biting on different teeth and found that very little strain was placed on the brow ridge. When we took the ridge away there was no effect on the rest of the face when biting. Since the shape of the brow ridge is not driven by spatial and mechanical requirements alone, and other explanations for brow ridges such as keeping sweat or hair out of eyes have already been discounted, we suggest a plausible contributing explanation can be found in social communication." (ScienceDaily, retrieved on Dec. 22, 2018 at https://www.sciencedaily.com/releases/2018/04/180409112554.htm)

(31) **Model of a modern human skull [left] next to Kabwe 1 [right][27]**

Besides their intrinsic interest, ape gestures might thus take us back to human linguistic capacities, with two main questions: what is the relation between ape and human gestures? and what is the evolutionary history of facial expressions that play a role in sign language and sometimes in gestures?

4   Beyond language I: pictorial semantics

4.1   Motivations

Going beyond language, there are four related reasons to investigate the semantics of visual representations.

(i) First, we saw above that visual iconic contributions are essential in signs as well as in gestures (and possibly in ape gestures as well). We tried to explain above how the content of these iconic contributions is integrated to linguistic representations, including by dividing it among various slots of the inferential typology of language. But how is iconic content derived in the first place? There is currently no completely general theory. In an empirically very rich article, Clark 2016 proposes a "staging theory" according to which "depictions are physical scenes that people stage for others to use in imagining the scenes they are depicting". But this does not aim to be a formal theory. Formal analyses of iconic content mentioned above (notably plural loci, height specifications, iconic modulations of *long* and *GROW*) were entirely *ad hoc*. While it is premature to aim for a general formal theory of iconic gestures (i.e. one that is explicit and explanatory), the semantics of pictures and visual narratives has been studied in detail, notably Greenberg (2011, 2013) and Abusch (2013, 2015). They offer a model of how an iconic semantics could be developed in a more linguistic context in the future (see Giorgolo 2010 for an early formal attempt in the gestural domain).

(ii) Second, understanding in formal detail the semantics of iconic representations is essential to obtain a broader typology of meaning operations in nature. There are numerous visual representations that are not linguistic in nature and yet convey semantic information; several of these (pictures, comics, films, etc.) are also undoubtedly intentional in nature. Understanding how they convey information is essential to the project of a 'Super Semantics'.

(iii) Third, some recent theories of purely visual narratives make use of unmistakably linguistic categories. Thus Abusch (2013, 2015) and Abusch and Rooth (2017) takes visual narratives to establish anaphoric relations among viewpoints, and also to introduce discourse referents familiar from dynamic semantics. If so, there are linguistic structures beyond language, and the appropriate tools to study them involve formal semantics.

(iv) Fourth, one can in principle embed visual representations within sentences, as we briefly mentioned in connection with pro-speech visual animations. This should offer a powerful method to investigate how pictorial information is divided among the inferential typology of language; we will briefly consider such cases (with 'pro-speech pictures') below.

---

[27] Credit: Professor Paul O'Higgins, University of York. Cited in: *ScienceDaily*, retrieved on Dec. 22, 2018 at
https://www.sciencedaily.com/releases/2018/04/180409112554.htm.
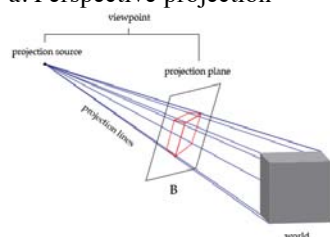
4.2   Picture semantics

4.2.1   Geometric projections

Greenberg 2013, 2018 develops a pioneering approach to picture semantics, crucially based on the notion of a geometric projection. In a nutshell, a picture is true in a world w relative to a viewpoint v if w projects onto the picture relative to v. In Greenberg's words (2018),

a simple type of PERSPECTIVE PROJECTION is illustrated in [(32)a]. Here we begin with a concrete 3-dimensional region of spacetime (possible or actual), which I'll think of as a possible world. In the example below, the world contains only a cube. Next, a PROJECTION SOURCE is located within the space of the world. A projection source is thought of simply as a geometric point in space and time. This in turn defines a system of PROJECTION LINES, which link each point in the world to the source. Finally, a PICTURE PLANE is introduced into this spray of lines, and they are used to map spatially distributed features of the scene back to surface features of the picture plane itself— in this case, the lines of the line drawing. The result of such a projection is displayed at right below.

(32)   **Examples of a projection-based semantics for pictures (Greenberg 2018)**
   a. Perspective projection   b. Parallel projection
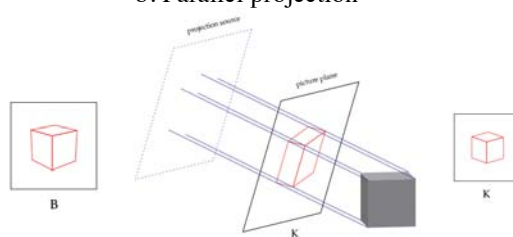
The viewpoint includes both the projection source and the position of the picture. The relevant projection lines are given by the system of projection assumed: a system of "linear projection" is represented in (32)a, and is "characterized by the fact that the projection source is a point, to which all projection lines converge" (Greenberg 2018); other systems can be used, for instance "parallel projection", illustrated in (32)b, whereby "the projection lines, rather than converging on a single point, all run perpendicular to the projection source, hence parallel to one another" (Greenberg 2018).

Simplifying somewhat, this allows Greenberg to define the content of a picture relative to a system of projection.

(33)   **Content of a picture (Greenberg 2018)**
   Relative to system of projection S (such as linear projection, or parallel projection), the **content of a picture P** (notated as $[[P]]_S$) is the set of pairs of the form <w, v> such that:
   w is a world and v is a viewpoint (specifying a projection source and a projection plane)
   and w projects to P from viewpoint v along the system of projection S (notated as: $proj_S(w, v) = P$).
   Formally, $[[P]]_S = \{<w, v>: proj_S(w, v) = P\}$

A further tweak will prove useful later: we can encode time dependency more explicitly by replacing pairs <w, v> (made of a world w and a viewpoint v) with triples <w, t, v> made of a world w, a time t, and a viewpoint v.[28]

Since we have taken truth rather than content to be the primitive notion of semantics and super semantics in this piece, it will be useful to restate things without a notion of content; here we directly adopt the version with time dependency, which will be useful in the analysis of visual animations.

(34)   **Truth of a picture (after Greenberg 2018, adding a time parameter)**
   A picture P is true in world w <u>at time t</u> relative to viewpoint v along the system of projection S iff <u>at time t</u> w projects to P from viewpoint v along S, or in other words:
   $proj_S(w, \underline{t}, v) = P$

---

[28] Greenberg's paragraph cited above makes clear that he takes *w* to refer to a world at a time ("a concrete 3-dimensional region of spacetime"). We henceforth take *w* to refer to a world.

This analysis is based on a causal semantics on which an external event can leave a trace on a perceptual system – for instance, the human retina.[29] This intuition could in principle apply to other senses as well. In an analogous way, the causal semantics of non-linguistic sounds could be specified in terms of the events that caused them, or on their 'causal sources' for short. In pictorial semantics, it is because of the way visual perception works that something like Greenberg's semantics needs to be adopted. Investigating pictures rather than visual perception leads to further possibilities: not only is the retina replaced with the surface of the picture, but diverse projection methods can be used. Still, it is worth keeping in mind for future reference that Greenberg's project can arguably be embedded within a more general program whereby various perceptual stimuli provide information about their causal sources; this will prove crucial in Section 5, when we develop an abstract 'source-based semantics' for music. (Lest one fear that our discussion overextends the category of 'semantics', it should be noted that the phenomena under investigation here – from pictures and comics to films and later to dance and music – are clearly the products of intentional agents; by contrast, the intentionality of monkey calls cannot be taken for granted.)

4.2.2   *Projections vs. ordering preservation*

With the addition of a temporal parameter, Greenberg's semantics can be compared to the iconic rules we posited for English *loooong*, for ASL *GROW*, and for ASL high and plural loci. English *loooong* is not a picture to begin with, but the basic intuition could be that the length of the vowel provides a kind of trace of the length of the denoted event. ASL *GROW* as well as high and plural loci are visual, but they clearly have a conventional component, and thus it is only one aspect of the sign that can be given an iconic semantics. But in any event, the iconic rules we provided for these phenomena are far less ambitious than Greenberg's semantics. First, our rules are entirely *ad hoc*, whereas Greenberg seeks to define a general semantics for pictures. Second, our rules solely pertain to the preservation of certain geometric relations. Take for instance the breadth condition for *GROW,* repeated below:

(35)   If the end points of $GROW_i$ are less distant than those of $GROW_k$, then the endpoint of the growth in e_i should be smaller than that of the growth in e_k.

Strictly speaking, this only constrains the set of possible denotations for multiple realizations of *GROW*. If a single one is produced, anything goes. While our analysis has the advantage of simplicity, it is insufficient: one can infer that a larger than normal realization of *GROW* has to denote a large growth, although this may be because there is always a contextual point of reference for the normal realization of *GROW* as well as for what is taken to be a normal amount of growth in the relevant situation. For high loci, the literature has explored analyses in which the height of the locus should be proportional (*modulo* some contextual parameters) to the height of the head of the denoted individuals (Schlenker et al. 2013). This comes closer to a picture semantics, but again extant rules are *ad hoc*. It would be very interesting to try to combine Greenberg's projection-based semantics with iconic rules for language in a far more systematic fashion, but the issue is non-trivial because of the combination of a conventional and of an iconic component in one and the same expression: it just isn't the case that the conventional form *GROW* is merely a dynamic visual representation of a growth process.

A highly simplified example will clarify the distinction between a weak semantics based on the preservation of some ordering, and a projection-based semantics. Suppose we are given three points A, B, C on an oriented line L, taken as a one-dimensional picture of some other points. We restrict attention to denotations for A, B, C which are themselves on an oriented line. What are the conditions for points A', B', C' to be possible denotations for A, B, C, as in in the picture in (38)?

---

[29] We use the term *trace* in a non-technical sense, with no connection whatsoever to the notion used in syntax.

(36) Ordering preservation vs. projection
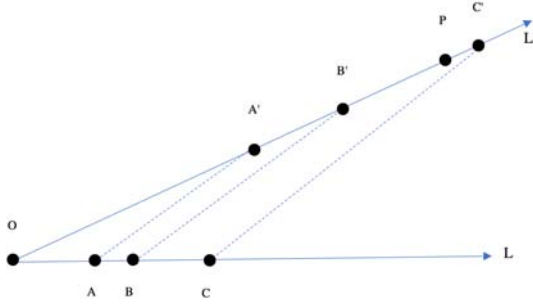
If we only require that the denotations A', B' and C' of A, B, C preserve the ordering of A, B, C, we will get little information about these denotations: the triple <A', B', C'> is a possible denotation of <A, B, C>, but so is <A', P, C'> (although <B', A', C'> and <C', A', P> are not). On the other hand, if we specify that for <A', B', C'> to be a possible denotation for <A, B, C>, there must be a parallel projection of <A', B', C'> onto <A, B, C> (leaving open the precise nature of the parallel projection and hence the position of L), different results are obtained: <A', B', C'> continues to a be possible denotation for <A, B, C>, but <A', P, C'> isn't one any more. The reason is that in this simple case parallel projections preserve proportions among distances, i.e. A'B'/A'C' must be equal to AB/AC. In fact, in this case the converse holds as well: <A', B', C'> is a possible denotation of <A, B, C> just in case A'B'/A'C' = AB/AC.[30] (Here too, consideration of one point alone, say A, leaves possible denotations unconstrained. But if we take the position of A and the direction of projection to be fixed, even A alone comes with non-trivial information about its possible denotations: they must be on the line that meets A along the direction of projection.)

4.3   Visual narratives: comics

4.3.1   Simple cases

The time-sensitive version of the definition truth for pictures in (34) yields a natural notion of semantics for visual narratives. Consider for instance the 2-picture sequence in (37), from Abusch and Rooth 2017, which represents "a short comic of two cubes moving apart".

(37) Picture P1   Picture P2
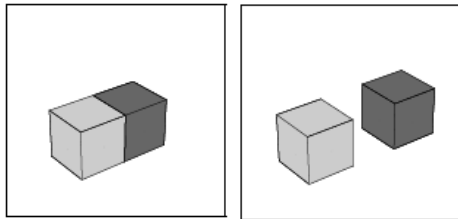
The fact that the two pictures are arranged as a narrative sequence provides information about the world, time and possibly viewpoints. Specifically, we will naturally (i) take the world w to be constant, (ii) take the time t' corresponding to the second picture to follow the time t of the first picture, and possibly (iii) take the viewpoint to be constant. Here and throughout, we will take viewpoints to be spatio-temporal points, which may be associated with a visual perspectival point.

---

[30] Equality of proportions upon parallel projection follows (in this simple case) from the Intercept (= Thalès) theorem of Euclidean geometry, which yields: OB'/OA' = OB/OA, and OC'/OA' = OC/OA. Equality of the other proportions follows, e.g. A'B'/OA' = (OB' - OA')/OA' = (OB'/OA') - 1 = (OB/OA)-1 = (OB-OA)/OA = AB/OA. It then follows as well that A'C'/A'B' = AC/AB. Leaving open the position of L, the converse follows as well: if A'C'/A'B' = AC/AB, there is a position of L which yields a parallel projection of <A', B', C'> onto <A, B, C>. For instance, place L so that A = A', at an arbitrary angle with L', between 0 and 180 degrees. Find C" on L' such that (CC") is parallel to (BB'). The Intercept theorem yields that A'C"/A'B' = AC/AB (still with A = A'), from which it follows that A'C"/A'B' = A'C' / A'B' and hence that C" = C'.

With an eye to future developments, we will also allow two pictures to depict situations that hold at the same time, especially when there is a shift of perspective. This yields the truth conditions in (38).[31]

(38) Assuming that the viewpoint is constant, a picture sequence $P_1, \ldots, P_n$ is true in world w at time t relative to viewpoint v along the system of projection S iff for some times $t_1, \ldots, t_n$ with $t = t_1 \leq \ldots \leq t_n$, $proj_S(w, t_1, v) = P_1$ and …and $proj_S(w, t_n, v) = P_n$.

4.3.2 Viewpoint shift

Still, there are two salient respects in which this analysis is insufficiently fine-grained to account for actual narrative sequences seen in comics. First, viewpoints need not be constant, and may in particular shift to adopt a character's perspective. Second, projections need not be veridical: sometimes they correspond to what a character sees even if this is a hallucination. Both issues are illustrated in (39), again from Abusch and Rooth 2017.

(39) "In Simone Lia's *Fluffy*, the character Michael has lost his rabbit Fluffy on a train. Searching, he looks into a cabin, and hallucinating, sees a girl eating a rabbit in a sandwich. It is subsequently clarified that the girl was eating a kipferl, a kind of pastry." (Abusch and Rooth 2017)
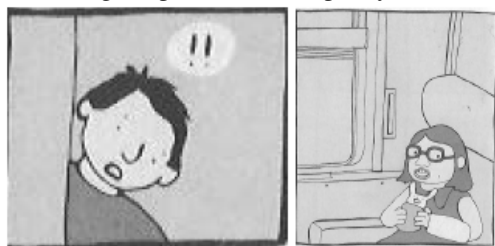
Let us for the moment assume that the entire scene is veridical, and thus that Fluffy is genuinely eaten as a sandwich in the second picture. We must still establish a connection between the perspective of the second picture and that of the first. A natural thought is that the second picture corresponds to a projection onto the visual system of the character in the first picture. As Abusch and Rooth argue, this phenomenon is common in language as well. This is illustrated in (40), where the second sentence is naturally understood to describe a situation viewed from the standpoint of the relevant male character.[32]

(40) He looked at his mother. Her blue eyes were watching the cathedral quietly. (cited in Abusch and Rooth 2017, from Lawrence's *Sons and Lovers*)

What shall we do to address this problem? Departing from Abusch and Rooth 2017 for a moment, the simplest solution would be to relativize the truth of a picture sequence not to a single viewpoint v but rather to a series of viewpoints $v_1, v_2, \ldots$. This would yield a modified version of (38), with changes boldfaced as in (41):

(41) Assuming that **viewpoints can change**, a picture sequence $P_1, \ldots, P_n$ is true in world w at time t **relative to viewpoints $v_1, \ldots, v_n$** along the system of projection S iff for some times $t_1, \ldots, t_n$ with $t = t_1 \leq \ldots \leq t_n$, $proj_S(w, t_1, v_1) = P_1$ and …and $proj_S(w, t_n, v_n) = P_n$.

But this analysis is missing something important about the picture sequence in (39) (interpreted veridically): the viewpoint of the second picture is understood to correspond to the character of the first picture. In other words, the first picture makes a viewpoint (i.e. a character) salient, and the second picture is interpreted as corresponding to that salient viewpoint. Abusch and

---

[31] Further conditions to be added if we do not wish the temporal interval between the scenes depicted by successive pictures to vary in arbitrary ways.

[32] Abusch and Rooth 2017 establish a connection between perspective shift in visual narratives and Free Indirect Discourse in language, but this seems to us to be in error: one key feature of Free Indirect Discourse is that its form contains a mix of direct and indirect discourse (as tenses and pronouns are evaluated as in indirect discourse, while other elements behave as in direct discourse). As a result, the shifted sentence is not just a 'normal' sentence evaluated with respect to a new perspectival point; its very form keeps a trace of the unshifted perspective. See for instance Schlenker 2004, Sharvit 2008, and Eckardt 2014 for discussion.

Rooth propose to introduce anaphoric relations among pictures. Departing from the letter (but hopefully not from the spirit) of their account, we can make the following assumptions:

(i) each picture comes with a viewpoint variable v_i, with v_1 corresponding to the first picture of the sequence;

(ii) pictures that depict a character that can serve as a viewpoint come with a distinguished point (of the picture) that serves as a variable denoting that character's viewpoint;

(iii) only the distinguished variable v_1 may be unintroduced, i.e. may fail to appear in earlier pictures: all further viewpoints must be anaphoric to characters made salient earlier. (This condition may be relaxed if further viewpoints that don't appear in pictures can be made salient.)

To get these rules to work, we need to establish some notations and revise our semantic rules.

• We will write $P(v_1,\ldots,v_n)$ for a picture with viewpoint variables $v_1,\ldots,v_n$ appearing in the picture.

Concretely, the first picture of (39) can be represented with a variable v_k in the picture: ; if the picture is called P, we can notate as $P(v_k)$ the picture with the variable.

• We will write $v_i{}^{\wedge}P(v_1,\ldots,v_n)$ for a picture $P(v_1,\ldots,v_n)$ viewed from a viewpoint denoted by v_i.

• We will need to refine our semantic rules, as in (42).

(42) **Revised semantic rules – with viewpoint variables**

If w is a world, t a time and s an assignment function assigning viewpoints to viewpoint variables:

a. **Atomic pictures**

$v_i{}^{\wedge}P(v_1,\ldots,v_n)$ is true relative to w, t, s iff w, t projects to P from viewpoint $s(v_i)$, and $s(v_1), \ldots, s(v_n)$ are viewpoints in w at t that project to $v_1,\ldots,v_n$ from viewpoint $s(v_i)$ in the picture P.

b. **Picture sequences**

A picture sequence of the form $v_{i_1}{}^{\wedge}P_1, \ldots, v_{i_n}{}^{\wedge}P_n$ (where $P_1, \ldots, P_n$ may contain variables) is true relative to w, t, s iff for some times $t_1,\ldots,t_n$ with $t = t_1 \leq \ldots \leq t_n$, $v_{i_1}{}^{\wedge}P_1$ is true relative to w, t_1, s and … and $v_{i_n}{}^{\wedge}P_n$ is true relative to w, t_n, s.

(42)a specifies that pictures are evaluated with respect to viewpoint variables, and may contain characters that correspond to further viewpoints. (42)b just states that pictures of a sequence are evaluated at times that respect the ordering of the pictures (although the time may not change across contiguous pictures).

In the simplest cases, nothing substantial will change. Thus the sequence $<P_1, P_2>$ depicted in (37) can be evaluated with respect to a single viewpoint variable v_1 as in (43):

(43) **Truth conditions of (37)**

$v_1{}^{\wedge}P_1, v_1{}^{\wedge}P_2$ is true relative to w, t, s iff for some times t_1, t_2 with $t = t_1 \leq t_2$, $v_1{}^{\wedge}P_1$ is true relative to w, t, s and $v_1{}^{\wedge}P_2$ is true relative to w, t_2, s,

iff for some times t_1, t_2 with $t = t_1 \leq t_2$, w, t_1 projects to P_1 from viewpoint $s(v_1)$ and w, t_2 projects to P_2 from viewpoint $s(v_1)$.

In (39) (still on its veridical interpretation), things are more sophisticated: we want the second picture to be interpreted from the viewpoint corresponding to the character in the first picture by way of the viewpoint variable v_k. The rules we introduced suffice to deliver the desired result, as shown in (44).

(44) **Truth conditions of (39) relative to a world, time and assignment function**

$v_1{}^{\wedge}$ $v_k{}^{\wedge}$ is true relative to w, t, s iff for some times t_1, t_2 with $t = t_1 \leq t_2$, $v_1{}^{\wedge}$ is true

relative to w, t_1, s and $v_k{}^{\wedge}$ is true relative to w, t_2, s,

iff for some time t_2 such that $t \leq t_2$, w, t projects to from viewpoint $s(v_1)$ and $s(v_k)$ is a viewpoint in

w at t that projects to v_k, and w, t_2 projects to ⟨picture⟩ from viewpoint $s(v_k)$.

This is almost adequate, as the truth conditions correctly take the second picture to be viewed from the perspective of the boy in the first picture. Still, there is something suboptimal about our result: we do not end up with a notion of truth relative to (a world, a time and) *a viewpoint*, but rather with a notion of truth relative to (a world, a time and) *an assignment of viewpoints to variables*. [33] This may be legitimate when viewpoints change without constraint, but this is not the case in Abusch and Rooth's example in (39), since the viewpoint of the second picture is introduced by the first picture. For such cases, we can make use of the definition in (45), which takes the viewpoint of evaluation to provide the value of the distinguished viewpoint variable v_1, while other viewpoints are constrained to be given by the content of earlier pictures. [34]

(45) **Truth of a picture sequence relative to a world, time and viewpoint**
If in a sequence Σ all pictures are evaluated with respect to viewpoint variables that appear in earlier pictures (or the variable v_1), then Σ is true relative to world w, time t and viewpoint v iff for some assignment s of viewpoints to viewpoint variables, $s(v_1) = v$ and Σ is true relative to world w, time t and s.

When applied to (39), this definition yields the desired result, as shown in (46), where we write assignment functions explicitly (e.g. the assignment function $[v_1 \to v, v_k \to v']$ assigns v to variable v_1 and v' to variable v_k).

(46) $v_1\wedge$ ⟨picture⟩ $v_k\wedge$ ⟨picture⟩ is true relative to world w, time t and viewpoint v

iff for some assignment s of viewpoints to viewpoint variables, $s(v_1) = v$ and $v_1\wedge$ ⟨picture⟩ $v_k\wedge$ ⟨picture⟩ is true relative to w, t, s,

iff for some viewpoint v', $v_1\wedge$ ⟨picture⟩ $v_k\wedge$ ⟨picture⟩ is true relative to world w, time t, $[v_1 \to v, v_k \to v']$,

iff for some viewpoint v', for some time t_2 such that $t \le t_2$, w, t projects to ⟨picture⟩ from v and v' is a viewpoint in w at t that projects to v_k, and w, t_2 projects to ⟨picture⟩ from v'.

An alternative to the analysis we just sketched (following the spirit of Abusch and Rooth 2017) would rely on very weak truth conditions, to the effect that *some* viewpoint is associated with the content of the first picture and *some* viewpoint is associated with the content of the second picture:

(47) **Underspecified (existential) truth conditions**

w, t projects to ⟨picture⟩ from v, and for some time t_2 such that $t \le t_2$, for some viewpoint v', w, t_2 projects to

⟨picture⟩ from v'.

---

[33] With this relativization to an assignment function, we can also interpret sequences of pictures that contain viewpoint variables that are not introduced in earlier pictures: their value could be given by the assignment function. But the definition of truth we turn to in the following lines would need to be modified to take into account such unintroduced viewpoints.

[34] As noted by E. Chemla (p.c.), we could in principle find cases in which this linear constraint is relaxed, and characters appears *after* their viewpoint is introduced.

On this view, the fact that the viewpoint v' is identified with that of the boy in the first picture is due to plausibility reasoning. But taking this reasoning to be outside of the semantics of picture sequences is not very plausible when we embed these sequences in linguistic environments, thus making use of the device of 'pro-speech pictures' (just like we studied pro-speech gestures and pro-speech visual animations in Section 2). Thus in (48), the claim need not be the strong one according to which Michael will require counseling if he is surprised and his bunny is eaten; rather, the more plausible reading is that he will require counseling if he sees that his bunny is being eaten.

(48) If at the end of the story , Michael will require counseling.

In this case, then, existential truth conditions for the second picture seem to be too weak. If indeed plausibility reasoning is at stake, it has to strengthen the existential truth conditions *within* the scope of the conditional. This result is also appropriately achieved if variables connect the viewpoint of the second picture to the character in the first one. If instead we provide existential truth conditions akin to those in (47) (within the scope of the conditional), we get overly strong truth conditions for the entire conditional, along the lines of: if at the end of the story Michael sees something surprising and someone sees Fluffy being eaten as a sandwich, Michael will require counseling.

4.3.3   Intensionality

Abusch and Rooth 2017 note that an extensional analysis is insufficient for cases like (39), which was intended to depict an illusion (a point we have disregarded up to this point). To address the problem, they introduce a covert perspectival operator Π which means something like 'sees', and appears in Logical Forms, as illustrated within our notation in (49).[35]

(49) Abusch and Rooth's covert operator Π (for 'sees')

Here v_k picks out the character in the first picture (we can take the denotation of v_k to be a viewpoint as defined earlier, although this is not what Abusch and Rooth do). Crucially, Π is an intensional operator, and thus there is no requirement that the scene in the second picture should have occurred. This is essential in this case, as "it is subsequently clarified that the girl was eating a kipferl, a kind of pastry", rather than the first character's rabbit.

How strong is the evidence for intensional operators? We could do without them at the cost of developing the semantics differently: instead of providing information about the world, it could provide information about what various real or imagined perceptual systems *see* of the world. Technically, the semantic rule for atomic pictures in (42)a, copied as (50)a, should be modified as in (50)b: truth of a picture is no longer about projection onto a two-dimensional picture, but rather perception by some concrete or abstract perception system.

(50) **Atomic pictures**
a. **Previous statement: projection-based semantics**
$v_i{}^\wedge P(v_1,\ldots,v_n)$ is true relative to w, t, s iff w, t **projects to P from viewpoint $s(v_i)$**, and $s(v_1)$, …, $s(v_n)$ are viewpoints in w at t that project to $v_1,\ldots,v_n$ from viewpoint $s(v_i)$ in the picture P.

b. **Modified statement: subjectivist semantics**
$v_i{}^\wedge P(v_1,\ldots,v_n)$ is true relative to w, t, s iff w, t **is perceived as P by viewpoint $s(v_i)$**, and $s(v_1)$, …, $s(v_n)$

---

[35] Abusch and Rooth 2017 write this intensional operator as *P*. We use Π to avoid confusion symbols referring to pictures.

are viewpoints in w at t that **are perceived as projecting** to $v_1,\ldots,v_n$ from viewpoint $s(v_i)$ in the picture P.[36]

Cutting some corners, the extensional truth conditions we obtained for (46) are modified as in (51), with the changes boldfaced.

(51) $v_1\wedge$ $v_k\wedge$ is true relative to world w, time t and viewpoint v iff for some viewpoint v', for some

time t_2 such that $t \leq t_2$, w, t **is perceived as** **by** v and v' is a viewpoint in w at t that **is perceived**

as projecting to v_k, and w, t_2 **is perceived as** **by** v'.

In order to derive information to the effect that the world is *in fact* one in which the first or the second picture holds true, we need a further assumption, namely that the viewpoint v or v' is veridical, i.e. that the perception is non-hallucinatory (in the intended interpretation, v is veridical while v' is hallucinatory, but this does not follow from the semantics alone). One could then reintroduce a projection-based semantics for this case.

The upshot is that in the subjectivist semantics we sketched, no non-subjective information is provided about the world unless further assumptions are made about the veridicality of certain viewpoints. Thus if one is willing to pay this price, a uniform semantics can be maintained for the veridical and non-veridical cases, without intensional operators.

4.3.4 Coreference

Abusch 2013, 2015 argues that pictures need to be enriched with discourse referents to account for certain intuitive ambiguities, as in (52).

(52) **An ambiguity of coreference in pictures** (Abusch 2015)
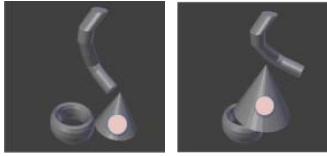
Abusch 2015 writes that on a simple picture semantics, (52) "is consistent with worlds where a single cone moves in front of a torus. It is also consistent with worlds where the cone of the first picture moves out of view, and another cone moves into view. To infer identity between the cones is to eliminate worlds of the second kind. This is done by adding to the discourse representation a syntactic predication of identity between the two indices, serving the same function as co-indexing in linguistic representations".

It is worth asking once again whether a disambiguation device is needed in the representational system, or whether common sense reasoning might suffice. The case for ambiguity would seem to be relatively strong when we embed the picture in a conditional, as in (53) (which replicates the test we performed in (48)).

(53) If what happens next is that , there will be no cone left to close another vase.

Intuitively, this sentence is true on the salient reading on which it is the same cone that appears in the first and in the second image; it is false on the far-fetched reading on which the first cone comes out of view and a second cone appears. The ambiguity is in large part due to the discontinuous nature of

---

[36] In this definition (which should be further refined in the future), $s(v_1), \ldots, s(v_n)$ must *in fact* be viewpoints, although they may be perceived as projecting to things that they are not. But since we take viewpoints to just be spatio-temporal points, the veridical part is rather weak.

the picture sequence: if we were dealing with a video instead, the non-coreferential reading would become entirely implausible.[37]

If one is convinced that there are referential ambiguities, Abusch's basic idea can be implemented by extending to objects (rather than just viewpoints) the semantics for atomic pictures given in (42)a, as in (54). The result is illustrated in (55), using for simplicity the definition of truth in (45), with s an enriched assignment function that assigns viewpoints to viewpoint variables and objects to object variables (changes pertaining to object variables are boldfaced).
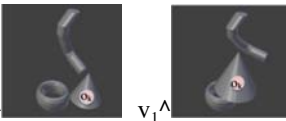
(54) **Revised semantic rules for atomic pictures – with viewpoint and object variables**
Let w be a world, t a time and s an assignment function assigning viewpoints to viewpoint variables of the form v_m, and assigning objects to object variables of the form o_m. Then:

$v_i{}^\wedge P(v_1,\ldots,v_n, o_1,\ldots, o_k)$ is true relative to w, t, s iff w, t projects to P from viewpoint $s(v_i)$, and in w at t $s(v_1), \ldots, s(v_n)$ are viewpoints that project to $v_1,\ldots,v_n$ in P **and $s(o_1), \ldots, s(o_k)$ are objects** that project to $o_1,\ldots, o_n$ in P.

(55) **Truth conditions of (52) with coreference, relative to a world, time and assignment function**

$v_1{}^\wedge$ $v_1{}^\wedge$ is true relative to w, t, s iff for for some assignment s of viewpoints to viewpoint variables and objects to object variables, $s(v_1) = v$ and for some time t_2 with $t \leq t_2$,

$v_1{}^\wedge$ is true relative to w, t, s and $v_1{}^\wedge$ is true relative to w, t_2, s,

iff w, t projects to from viewpoint $s(v_1) = v$ and **$s(o_k)$ is an object in w at t that projects to o_k**

in this picture, and w, t_2 projects to from viewpoint $s(v_1) = v$ and **$s(o_k)$ is an object in w at t_2 that projects to o_k in this picture**.

It is immediate that the boldfaced conditions enforce identity of the cones that appear in the two pictures. Without this condition (e.g. if no indices or different indices appeared), the denoted cones could be different. It will remain to be determined whether variables are necessary to enforce such coreferential readings, or whether common sense reasoning might be sufficient.[38]

4.4 *Visual narratives: film*

Continuous visual narratives are exemplified in film, where a projection-based semantics makes particularly good sense. In film editing, discontinuous shots are pasted together. As a result, the same general questions arise as in comics, pertaining in particular to the need for object and viewpoint variables, as well as potential cases of intensionality/illusion.[39]

The issue of viewpoint shift has received detailed attention in pioneering semantic research by Cumming et al. 2017, which uncovers two additional constraints on viewpoint change in film. Unlike the (highly simplified) cases we discussed in the preceding section, viewpoint shift need not be associated with a salient character, but there are still constraints on possible viewpoint shifts.

---

[37] We were careful to discuss an example in which the viewpoint didn't change. When the viewpoint changes, referential ambiguities will be multiplied *if* we do not resolve viewpoints. But in Abusch's case in (53), a referential ambiguity obtains *even* when the viewpoint is taken to remain constant.

[38] Abusch 2013 makes reference to Pylyshyn's work on indexing in vision (Pylyshyn 2003) to motivate some uses of variables in picture semantics. The connection is definitely relevant and ought to be explored further.

[39] For an introduction, as well as an example of character-anchored viewpoint shift (at 7:26), see Leff 2019. Thanks to E. Chemla for the reference.

One is the X-Constraint, which relies on two notions. The "action line" of a scene is "the most prominent linear relationship in a given scene", such as "the trajectory of a speeding car". The "X-direction" of an action line is "the direction of the action line, as it is projected along the X-axis of the screen. Independent of its upward/downward or forward/backward orientation, an action line pointing screen-rightward has a positive X-direction, while one pointing screen-leftward is negative".[40] Now the X-constraint says that viewpoint shift should not reverse the X direction of the action line, as is stated and illustrated in (56).

(56) If the X-Constraint applies to sequence S_1-S_2, then the X-direction of the action line relative to the viewpoint in S_1 is consistent with its X-direction relative to the viewpoint in S_2.
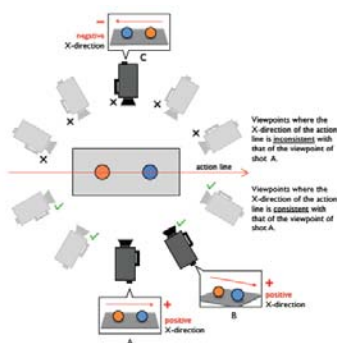
(figure from Cumming et al. 2017)

From the initial viewpoint A, the projection of the action line on the screen is going rightwards. A shift to viewpoint B preserves this property, as does any shift to viewpoints that have checkmarks (and are below the red line). Viewpoints that are crossed (above the checkmark) reverse the X-direction of the action line and shifts to these positions are thus prohibited.

This only scratches the surface of constraints on viewpoint shift, which are investigated in greater detail in Cumming et al. 2017.[41] But it should be clear that film semantics comes with an interesting 'grammar' that ought to be investigated with formal (and experimental) means. Since editing puts together discrete visual sequences, a comparison with constraints at work in comics should be conducted. In addition, specific questions can be raised about viewpoint manipulations: Can constraints on viewpoint shift be derived from elementary principles? What relation, if any, do they bear to linguistic phenomena?[42]

4.5   Relation to the inferential typology of language

We saw in Sections 2.4 and 2.5 that information provided by iconic gestures and visual animations is divided 'on the fly' among familiar slots of the inferential typology of language. We thus expect that the same should be true of visual narratives, which as a first approximation are discontinuous or continuous visual animations. The visual animation part of Tieu et al.'s (2018) experiments involved videos embedded in written text. Only minimal modifications are needed to turn them into (admittedly minimalist!) discontinuous visual narratives embedded in text, as shown in (57).

(57) **Pictures from Tieu et al.'s videos testing presuppositions generated by visual animations**
       (here: a change of state animation pertaining to an alien's antenna turning from green to blue)

---

[40] Nothing hinges on the choice rightward = positive, leftward = negative: the goal is just to assess identity of direction along the X-axis in order to state the relevant constraint.

[41] Cumming et al. 2017 also motivate a "T-Constraint", which "requires that screen angle, not just direction, be maintained in the transition between shots in a sequence".

[42] For Cumming et al. 2017, "viewpoint constraints are most nearly parallel not to lexical conventions, nor to the compositional rules of subsentential semantics, but instead to those inter-sentential semantic relations which seem to organize all forms of linguistic discourse", i.e. to coherence relations in discourse.

(green)   (green+blue)   blue

So we certainly expect that presuppositions should be triggered by various visual narratives, although proving that this is so might require considering mixes of text and visual narratives, just as was the case of our analysis of pro-speech gestures in visual animations in Sections 2.4 and 2.5[43].

Could presuppositions be triggered by a single picture? On views of presupposition on which a 'triggering algorithm' applies to diverse information-bearing units, this is to be expected. As an example, a running theme of Gosciny and Uderzo's *Asterix* is that the eponymous hero acquires superhuman force after drinking a magic potion prepared by druid Getafix (Panoramix in the French version)". With this background in mind, the questions in (58) arguably trigger presuppositions pertaining to the preconditions of the relevant actions: in one case, that Getafix in fact prepared the potion; in the second, that Asterix has some magic potion with him.[44]

(58)   a. Will Getafix ?[45]

naturally understood as: will Getafix give some of the potion to Asterix?
=> Getafix prepared some magic potion

b. Will Asterix ?[46]

naturally understood as: will Asterix drink some (of his) magic potion?
=> Asterix has some magic potion with him

Besides these potential cases of presupposition triggering on the basis of the informational content of the pictures, there might be further cases that are specifically due to the anaphoric component unearthed by Abusch and Rooth. In (53), repeated below as (59), we arguably get an inference that if the two situations depicted take place, *they will involve the same cone*, which means that whatever supplies the connection (the anaphoric component, or possibly general reasoning) is not at-issue.

(59)   If what happens next is that , there will be no cone left to close another vase.

More generally, a systematic investigation of the inferential status of inferences triggered by static and dynamic iconic representations might prove interesting both for our understanding of iconicity and of the inferential typology of language; this work is just in its infancy.

---

[43] Abusch and Rooth 2017 discuss presuppositions introduced by their covert attitudinal operator P (notated as Π in the present piece). They also discuss 'preconditions' of various scenes, but without technically treating them as presuppositions.

[44] In addition, we believe that (58)b triggers an inference to the effect that *if Asterix drinks the magic potion, he will do so in the way depicted*. This can be described as a cosupposition (a conditionalized presupposition) triggered by purely iconic means. See Schlenker 2018h for potentially related examples involving pro-speech gestures and ASL classifier predicates.

[45] Picture retrieved online on January 5, 2019 at https://culturebox.francetvinfo.fr/livres/une-grande-exposition-dopee-a-la-potion-magique-consacre-asterix-a-la-bnf-137293.

[46] Picture retrieved online on January 5, 2019 at http://free4web.free.fr/BD/Resume/Asterix/persos.htm.

5 Beyond language II: music and dance semantics

5.1 Auditory scenes and music semantics

5.1.1 Visual vs. auditory iconicity

As we mentioned in Section 4.2, Greenberg's semantics can be seen as a particular incarnation of a more general 'source-based semantics', whereby the meaning of certain percepts is given by the information they provide on their causal sources. This more general notion can be applied to other senses, including audition.[47] Now a sound analogue of pictures would be an attempt to depict a scene by way of near-instantaneous (non-linguistic) sounds: instead of mentioning the projection of a scene onto a picture, we would need to consider the auditory 'trace' of a situation on an audio recording, for instance. Similarly, a sound animation could do the same thing with a temporally indexed modification of Greenberg's semantics, with the notion of a 'projection' replaced with that of a 'sound trace', as sketched in (60).

(60) **Truth of a sound animation (modified from Greenberg's picture semantics)**
A sound representation P is true in world w at time t relative to perspectival point v along the system of sound perception S iff at t w produces the sound trace corresponding to P along S at perspectival point v, or in other words: sound-trace$_S$(w, t, v) = P.

Unfortunately, we know of no Greenbergian semantics for sound animations, possibly because these are less common and prominent than pictures or dynamic visual narratives. But a more complicated case has been discussed in recent research: it pertains to music semantics. Instead of Greenbergian projections, this line of research constructs toy models based on preservation conditions similar to those we posited for English *loooong* and ASL *GROW*. This is partly for convenience, as the resulting models are particularly simple (if overly weak). But this is also because the conceptual problem addressed is the same, though for different reasons: due to the conventional character of *GROW* and *long,* a theory that treats these as simple visual or auditory animations is a complete non-starter. Similarly, there is little interest to a semantics that posits that music is a standard auditory animation: on such a view, music could just convey information about its actual causal sources, e.g. the violinist, cellist, oboist, and possibly the conductor. One may respect musicians and still think that this isn't what the meaning of music is about.

Following in part insights in Bregman 1994, the meaning of music has been argued instead to lie in inferences about 'virtual' sources of the music (Schlenker 2017a, 2018g): not the musicians, but virtual objects that satisfy certain inferences triggered by the music. As we will see, in recent proposals these virtual sources need not be sound-producing, which allows music to evoke extremely diverse scenes and objects (albeit in a highly abstract, underspecified fashion).

5.1.2 Motivations for music semantics

But first, is there really any empirical motivation for a music semantics? Does music genuinely trigger inferences about a music-external reality?

Old and recent work have unearthed a variety of inferential effects. They are of two kinds: some are lifted from normal auditory cognition; others derive from specifically musical properties of tonal pitch space. An illustration is given in Schlenker 2018g:

Both kinds of inferences can be used to signal the end of a piece. One common way to signal the end is to gradually decrease the loudness and/or the speed. While this device could be taken to be conventional, it is plausible that it is in fact derived from normal auditory cognition: a source that produces softer and softer sounds, and/or produces them more and more slowly, may be losing energy. But on the tonal side, it is also standard to mark the end of a piece by a sequence of chords that gradually reach maximal repose, ending on a tonic. Plausibly, an inference is drawn to the effect that a virtual source that manifests itself by a tonic is in the most stable physical position, with no tendency to move any further. Thus these two types of inference combined conspire to signal the end of a piece.

---

[47] As before, one may elect to restrict the term 'semantics' to apply to artificial and intentional representations (rather than to naturalistic percepts).

Further examples are listed and illustrated (by way of links to sound examples) in Appendix II. For instance, lower pitch is associated with larger sources; it is put to use, to comical effect, in Saint Saëns's Carnival of the Animals: a waltz is played by a double-bass in order to evoke an elephant. When the source is fixed, lower pitch is associated with a less excited or energetic object. Lower loudness can be associated with a less energetic source, or with a source moving away. Lower speed in the music is associated with a slower source, while silence is associated with the interruption of an action. Harmonic dissonances can be associated with states of physical disequilibrium, or with emotional tension.

Importantly, these inferences need not even pertain to sound-producing sources: if anything, Saint-Saëns's elephant is evoked as dancing rather than as trumpeting. And music has been used to evoke all sorts of silent scenes, from Strauss's sunrise (in Zarathustra) to Saint-Saëns's Aquarium (in his Carnival) to Debussy's Prelude to the Afternoon of a Faun.

5.1.3   Models for music semantics

How should music semantics be modelled? The formal difficulty lies in ensuring that some but not all properties of the music are semantically interpreted, and to aggregate rather heterogeneous inferences. As mentioned at the outset, recent (toy) models of music semantics are based on weak preservation conditions reminiscent of our analysis of iconic effects in English *loooong* and in ASL *GROW*. They have the advantage of delivering the kind of abstract inferences that are needed: diverse objects and situations will satisfy the inferences triggered, including objects that produce no sound at all.

In our analysis of the iconic component of ASL *GROW* in (5), we stated two preservation requirements: the larger the sign, the greater the growth; the faster the sign, the quicker the growth. Schlenker 2017a, 2018g defines a (weak) music semantics based on formally similar preservation requirements. A toy example is discussed in which two properties are taken into account: loudness, with the requirement that lower loudness is interpreted in terms of lower energy of the source, or greater distance of the source from the perspectival point; and harmonic stability, with the requirement that less harmonically stable chords denote less physically stable events.

Concretely, if we consider a series of three chords as in (61), it will denote three real-world events, one corresponding to each chord. The tonic chord I is more stable (in classical music theory) than the dominant chord V, and as a result the first and third denoted events (corresponding on the initial and final I) should be more stable than the second one. The three-chord sequence features a crescendo, with loudness going from 70db, to 75db, to 80db; correspondingly, the three events should either correspond to a source that gains energy, or one that approaches the perspectival point.

(61) M = $<<I, 70db>, <V, 75db>, <I, 80db>>$

While the entire semantics could be developed in terms of events, the basic intuition of the framework is that musical voices are associated with virtual sources that are *objects*, and participate in certain events. Correspondingly, a voice involving n musical events will be taken to denote a pair of an event and of n real world events, as is stated in (62):

(62)   Let M be a voice, with $M = <M_1, \ldots, M_n>$. A possible denotation for M is a pair $<O, <e_1, \ldots, e_n>>$ of an object and a series of n events, with the requirement that O be a participant in each of $e_1, \ldots, e_n$.

Starting from the piece in (61) and the specification of possible denotations in (62), we will say that the musical piece $M = <M_1, \ldots, M_n>$ is true of the pair of an object and events it undergoes, $<O, <e_1, \ldots, e_n>>$, just in case $<O, <e_1, \ldots, e_n>>$ is a possible denotation for M, and in addition the mapping from $<M_1, \ldots, M_n>$ to $<e_1, \ldots, e_n>$ preserves certain requirements, listed in (63). Informally: the denoted events should preserve the temporal ordering of the musical events, as well as the loudness and stability ordering among them.

(63) **Defining 'true of' in music**

Let $M = <M_1, \ldots, M_n>$ be a voice, and let $<O, <e_1, \ldots, e_n>>$ be a possible denotation for M. **M is true of $<O, <e_1, \ldots, e_n>>$** if it obeys the following requirements.

a. Time
The temporal ordering of $<M_1, \ldots, M_n>$ should be preserved, i.e. we should have $e_1 < \ldots < e_n$, where $<$ is ordering in time.

b. Loudness
If M_i is less loud than M_k, then either:
(i) O has less energy in e_i than in e_k; or
(ii) O is further from the perceiver in e_i than in e_k.

c. Harmonic stability
If M_i is less harmonically stable than M_k, then O is in a less stable position in e_i than it is in e_k.

Schlenker 2017a, 2018g then shows that these rules make it possible to take the sequence M in (61) to be true of a sunrise involving three subevents: minimal luminosity, rising luminosity, maximal luminosity. The apparent energy of the source rises, as mandated by the Loudness condition; and the first and third subevents are more stable than the second one, as mandated by the Harmonic stability condition (this is on the assumption that events of 'minimal luminosity' and 'maximal luminosity' involve little or no change, whereas 'rising luminosity' involves a faster change). By contrast, a sunset would fail the Loudness condition, as the apparent level of energy of the source does not rise. Similarly, interpreting the Loudness condition in terms of proximity rather than in terms of level of energy, the same sequence could be satisfied by a boat approaching, with three subevents: maximal distance, movement towards the source, minimal distance (here too, with the assumption that the first and last event are more stable that the second). By contrast, a boat *departing* could not satisfy the Loudness condition.

Importantly, these preservation conditions are abstract enough that they can be satisfied by real world events that need not be sound-producing, and may be very diverse in nature. This is as desired: the larger the set of event sequences that satisfy the music, the more abstract the corresponding meaning will be. And musical meaning is in general *very* abstract.

5.1.4   *Source-based semantics vs. iconic semantics*[48]

In this piece, we have taken the source-based semantics developed for music to be a generalization of the iconic semantics developed for gestures, pictures and visual animations. Peirce famously established a distinction between icons, which involve a 'likeness' between a signal and its denotation, and indices, which involve a causal connection between a signal and its source (e.g. the smoke is an index of the fire; see Atkin 2010, Peirce 1868, and Koelsch 2011, 2012). We argued above that projection-based semantics is based on a notion of source and thus of index, since a perspectival projection makes it possible to draw in inferences on the causal source of a visual percept.

Could we also take our music semantics to involve icons? After all, the technical theory we sketched in Section 5.1.3 is based on certain preservation conditions that could qualify as 'iconic'. But it all depends on how iconicity is defined. If it involves a kind of intuitive *resemblance* between the signal and its denotation, our semantics need not be iconic. For instance, the three-chord sequence in (61) can be true, among others, of a sunrise. But a sunrise is a silent event that doesn't much resemble a musical piece. On the other hand, if the notion of iconicity is made more abstract, the preservation principles we introduced do qualify as iconic: a sunrise could be denoted by (61) because the mapping between the relevant series of notes and the relevant series of subevents satisfies pre-determined preservation principles. There is thus a terminological point that might require further conceptual elaboration.

5.1.5   *Connections*

While this framework only offers the bare bones of a music semantics, it immediately establishes connections with other domains of Super Semantics.

---

[48] This discussion follows a similar one in Schlenker 2018g.

First, it is striking that a lot of inferential devices at work in music are lifted from biological systems. We noted in Section 3.2 that lower frequency is associated with larger body size, an inferential device of music semantics as well. Similarly, higher frequency and faster production rates are associated with greater stress/arousal, and these inferential effects too play a role in music. And Blumstein et al. 2012 further argue that distortion noise (nonlinearities) are signals of alarm in many vertebrates, and can be added to music to induce in listeners "increased arousal (i.e. perceived emotional stimulation) and negative valence (i.e. perceived degree of negativity or sadness)". More generally, several triggers of emotions in music have been traced to animal signals (e.g. Juslin and Laukka 2003).

Second, formally, a semantics based on the preservation of certain orderings is common to music semantics and to the iconic semantics we proposed for ASL *GROW* and English *loooong*. But it is also too weak. Strikingly, this semantics has nothing to say about the semantic associations of a single sound (because any ordering will be trivial and thus trivially preserved). But a low-frequency sound is probably associated with a large source even independently of other sounds. Stronger semantic rules should be explored in the future.

Third, the simplified nature of recent models of music semantics should not obscure potential connections with the semantics of visual representations. Auditory inferences are always relativized to a perspectival point; this was implicitly the case in our discussion of the Loudness condition: louder musical sounds could be associated to a source with more energy, or to a source closer *to the perspectival point*. It remains to be seen whether the issue of perspectival shift that played a role in our discussion of comic and film semantics can be observed in 'real' music as well.

Fourth, we can go through the exercise of comparing music semantics to picture semantics. The formats used in (38) (picture sequences) and in (63) (music semantics) are different along several dimensions, but they may be brought closer by making some adjustments. First, we replace truth at a world and time in the definition of pictures with truth of a sequence of events, as in (64). Second, we drop reference to an object in the definition of our music semantics, and just keep a requirement that the denoted events be temporally ordered in the same way as musical events, and that the Loudness and Harmonic stability conditions should be preserved. On the other hand, we add an explicit reference to an auditory (perspectival) point, which in any event played an implicit role in our definition of the Loudness condition.

(64) **Modified definition (for comparison): picture sequences true of tuples of events**
A picture sequence $< P_1, \ldots, P_n>$ is true of events $< e_1, \ldots, e_n>$ relative to viewpoint v along the system of projection S iff
(1) temporally, $e_1 < \ldots < e_n$;
(2) $\text{proj}_S(e_1, v) = P_1$ and $\ldots$ and $\text{proj}_S(e_n, v) = P_n$.

(65) **Modified definition (for comparison): musical sequences true of tuples of events**
A musical sequence $< P_1, \ldots, P_n>$ is true of events $<e_1, \ldots, e_n>$ relative to auditory point v iff
(1) temporally, $e_1 < \ldots < e_n$;
(2) the Loudness and Harmonic stability conditions are satisfied, i.e.:
a. If M_i is less loud than M_k, then either:
(i) O has less energy in e_i than in e_k; or
(ii) O is further from the auditory point v in e_i than in e_k.
b. Harmonic stability
If M_i is less harmonically stable than M_k, then O is in a less stable position in e_i than it is in e_k.

It is now clear what these two definitions have in common: a pictorial or musical sequence functions as the visual or auditory trace of some events. Furthermore, meaning is produced by a way of requirement that some real world/physical properties of the events should be preserved on the visual or musical surface. But picture sequences produce meaning by reference to a specific mode of projection, with the result that a single picture can provide information about the world. By contrast, in the very weak music semantics we provided, a single musical sound cannot produce information about the world (because preservation rules pertain to the relation among sounds, and thus for a unique sound preservation principles will trivially be satisfied).[49]

---

[49] For a discussion of the difference between preservation-based and projection-based views of pictorial representation, see Greenberg 2013, 2018. As E. Chemla (p.c.) notes, we could explore stronger music

Finally, if music can produce semantic effects, one might expect that it is possible to use music as a gesture that can enrich or even replace some words. And one might further expect that music, just like visual and possibly auditory animations, could trigger inferences that fill various slots of the inferential typology of language. The investigation of these issues is in its infancy but could yield interesting results in the coming years.[50]

5.1.6   Musical vs. visual narratives

Abusch's work on visual narratives suggests related questions about musical narratives. They are undoubtedly harder in the musical case because even supposedly referential music is by its very nature abstract; but this should not detract us from exploring some of the conceptual similarities between musical and visual narratives.

To make the point concrete, consider Richard Strauss's tone poem *Don Quixote*. In his *Young People's Concerts* (Bernstein 2005), Leonard Bernstein focused on Variation II, and argued that it could fit perfectly well with completely different story lines (despite Strauss's stated intentions). One, consistent with Strauss's goals, has Don Quixote departing to conquer the world, encountering a flock of sheep that he takes for an enemy army, charging at them, and ending up proud of his 'knightly deed'. Bernstein's alternative version features Superman seeking to free an innocent prisoner, approaching the prison with the prisoners snoring at night, charging into the prison, and taking his innocent friend back to freedom. There are striking structural correspondences between Bernstein's two stories, and they are very much in the spirit of what a source-based semantics would lead one to expect, as discussed in Appendix III: since music semantics is abstract, diverse stories can be made consistent with it, but from this it doesn't follow that anything goes, nor that music doesn't have a semantics at all.[51]

But when we consider the details, Abusch's question about the need for variables in pictorial representations arguably arises in music as well. A key element of the Strauss piece, and of Bernstein's interpretations, is that the same source plays a role at the beginning of the piece (Don Quixote departing to conquer the world, Superman departing to free the innocent prisoner), and at the end (Don Quixote feeling proud of his knightly deed, Superman and his friend reaching freedom). But what guarantees this cross-identification? A source-based semantics on its own may or may not be sufficient. This is a different version of the problem of cross-identification across pictures that we saw in Abusch's discussion. The question was whether the similarity of shapes we see in (59) is sufficient to license a reading with cross-reference, or whether a special device (variables) is needed to obtain this result. In Strauss's Variation II, the same tune played by the cellos at the beginning and at the end (in   (66)) is indicative of Don Quixote's (or Superman's!) presence. The question is whether this similarity is enough to ensure coreference; if not, musical counterparts of Abusch's variables might be needed.[52]

(66)

Abusch's question about the need for perspectival shift and intensional operators can also be raised in a musical context, albeit in a less transparent fashion. Here too, Strauss's piece offers a nice

---

semantics, e.g. ones that not just certain orderings but also certain proportions (as was the case of the projection-based semantics in the simple figure in (36)).

[50] Current work along these lines is being developed by Robert Pasternak (co-speech music) and Janek Guerrini and Léo Migotti (pro- and post-speech music).

[51] Thanks to Paul Egré (p.c.) for highlighting the relevance of Bernstein's discussion to the present analysis.

[52] To see a case in which there is a discrepancy in the cross-identification of the sources, we can note in Bernstein's retelling, Superman takes over the role of both Don Quixote and Sancho Panza. Specifically, Superman takes the same kind of heroic actions as Don Quixote, but also whistles a tune (to alert his prisoner friend of his presence). The latter action corresponds to Sancho Panza's chuckling (see Appendix III for details). This means that the two interpretations do not quite posit the same coreference relations among sources.

illustration. In Bernstein's Don Quixote interpretation of Variation II, some inferences are veridical: the presence of the sheep is taken to be real. Others are not, or not clearly: Don Quixote's heroic departure and later triumph have an element of delusion, and are likely interpreted from Don Quixote's perspective: the perceiver might not share it. Bernstein's Superman interpretation need not come with this perspectival shift: the sense of triumph expressed at the end may be Superman's, his friend's, or the perceiver's.

5.1.7 *Interfaces: syntax/semantics and semantics/pragmatics*

Once a (highly simplified) music semantics is in place, two further questions naturally arise, pertaining to the syntax/semantics interface, and to the semantics/pragmatics interface.

First, how does music semantics interact with music syntax? While music semantics is in its infancy, music syntax is not: besides a long history in musicology, it led to important formal developments in Lerdahl and Jackendoff's pioneering work (1983) (followed by several others, including Rohrmeier 2011, Granroth-Wilding and Steedman 2014 and Pesetsky and Katz 2009). Schlenker 2017a, 2018g tentatively suggests that some of Lerdahl and Jackendoff's syntactic structures could be reinterpreted from a semantic perspective. Specifically, Lerdahl and Jackendoff start from a notion of 'grouping structure', which they take to derive from Gestalt principles of perception. They further argue that musical groups are 'headed': at each level, each group contains a musical event that is more important than the others and thus counts as its head. Importance is defined in terms of a mix of metrical prominence and harmonic stability. The result is a hierarchical structure with heads at every level; just keeping the heads yields a reduction of the most important elements of a musical passage. A semantic reinterpretation would go like this: grouping structure might stem from an attempt to recover the structure of the denoted events, as is likely the case in normal auditory perception; and heads might correspond to subevents that are relatively more important or stable. It remains to be seen whether this idea is correct, and can be extended to further aspects of music syntax.[53]

Second, music is construed as an intentional activity, and thus besides whatever semantic effects it may give rise to, its form and meaning can trigger further inferences about the intentions of musical narrator, i.e. the intentional agent that is seen as the author of the music. This naturally gives rise to questions about a music pragmatics. This area too is in its infancy (see Schlenker 2018g for some speculations); but it is clear that the same conceptual issues should arise across (intentional) semantic systems, including visual representations.

5.2 *Dance semantics*

In view of the traditional connections between music and dance, one might ask whether syntactic and semantic investigations can prove enlightening for dance as well. From a syntactic perspective, Charnavel 2016 proposes to extend to dance Lerdahl and Jackendoff's notion of 'grouping structure'; and like Lerdahl and Jackendoff, she takes the notion of a group to derive from Gestalt principles of perception. Charnavel even proposes that dance groups, just like musical groups, are headed.[54] She sketches further potential analogies with aspects of music syntax.

5.2.1 *General questions*

From the present perspective, two questions naturally arise. First, can a semantics for (abstract) dance be defined, possibly with an analogue of the source-based semantics we outlined for music? Second, when dance is accompanied by music, what is the relation between dance semantics and music semantics?

---

[53] One could further ask whether similar issues – pertaining to hierarchical organization and its cognitive source – arises in visual perception and in the 'syntax' of pictures or narrative sequences in particular.

[54] In her words: "we can define the head as the most important element of a rhythmic unit, which remains in its reduced version (…). Moreover, I assume that just like in music, both stability and rhythmic criteria determine headedness in dance."

The first question is speculative at this point: no semantics for abstract dance has, to our knowledge, been defined. But this would be a project worth pursuing, starting from the observation that ballet, for instance, may be used to evoke inferences that are not just about the dancers nor what they literally resemble. It would make much sense to treat each dancer or group of dancers as an equivalent of the sources we posited for music semantics, and to take their movement to trigger abstract inferences about the events experienced by these sources. How to develop this semantics remains to be seen.[55]

The second question, pertaining to the interaction between music and dance semantics, is equally speculative, but also conceptually subtle. When two semantic systems interact, there is no reason to assume that they will provide the same semantic information. If anything, this might be something to *avoid* for fear that the result would be redundant or boring. In film music, the term 'Mickey Mousing' is used to refer to a music that attempts to literally depict the action that appears on screen; the term need not be laudatory. Still, the fact that such effects are *possible* is in itself telling, as it suggests that despite the vast difference in medium and expressive possibilities, some clear correspondences – possibly semantic ones – can be established.

The same issue arises with respect to the dance-music interface. Some connections might be syntactic in nature. For instance, Charnavel 2016 hypothesizes that in a dance-music event, grouping and metrical structures coincide on the music and on the dance side. While grouping might in the end be a semantic notion, we should still ask whether *clearly* semantic notions can help coordinate music and dance. One would expect that each system has its own semantics, that the abstract messages may be in part different (so as to avoid Mickey Mousing), but that they should also have certain points of contact – i.e. time points in which music and dance denote the same events so as to allow for a clear coordination between the two mediums.[56] There could in principle be cases in which this coordination gives rise to disambiguation of one medium by the other on semantics grounds; a possible example is discussed in Appendix IV.

5.2.2   *Referential dance*

Particularly interesting and challenging questions are raised by referential dance. In pioneering work, Patel-Grosz et al. 2018 investigate Bharatanatyam, a classical South Indian dance that is figurative and typically serves to tell a story. While this dance has an inventory of conventionalized and meaningful gestures (around 31 types of one-handed gestures and 27 types of two-handed gestures), the authors investigate a more abstract issue, the realization of coreference vs. disjoint reference. In a production experiment (with marker-based motion capture), they seek to understand how disjoint reference is marked in the dance sequence in (68), explicitly produced by the dancer to 'translate' the story in (67).

(67)   The artist sees a strong man [P21 sitting on the ground]. Then she sees that [P22+P23+P24 *another man*] [P25 is holding a spear].

(68)   5 stills from a motion capture video for the dance realization of the story in (67), corresponding to the components [P21] to [P25], as indicated on the 2nd line. Body orientation is represented by way of arrows ↙ and ↘ on the 3rd line. Grouping is represented by ---------|--------- on the 4th line (the first group ends

---

[55] Léo Migotti is currently working along these lines.

[56] Analyses might start from ballets that remain particularly close to the music, as is the case for Balanchine's Symphonie Concertante (on Mozart's music). In the words of the Balanchine Trust, "the two principal ballerina roles correspond to the solo instruments; one suggesting the violin part and the other, the viola." (http://balanchine.com/symphonie-concertante/, retrieved online on January 23, 2019)

after [P23])

[P21]	[P22]	[P23]	[P24]	[P25]
↙	↙	↙	↘	↘
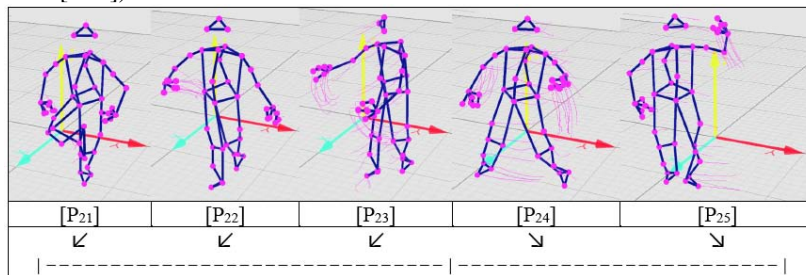
A control task involving coreference involves, as one might expect, a smooth transition from one action to the next. The question is how disjoint reference is marked in the present case. It involves at least four devices.

(i) First, "the dancer uses a designated mudra (hand-and-arm gesture) that symbolizes 'another, a different', as visible in [P22]". But the authors note that two naïve subjects "found this mudra difficult to track" even after being informed about its function (to mean 'another').

(ii) Second, there was a grouping boundary after [P23] to indicate that a new character appeared in [P24]. It is marked by a change of direction or orientation, as indicated by the arrows ↙ and ↘ in (68). We noted above that grouping in music might be analyzed in semantic terms, as deriving from the representation of boundaries between the denoted events. Patel-Grosz et al. extend this analysis to dance semantics: they propose "that grouping in dance serves as a way to organize (sub-)events". Thus "the introduction of larger-level group boundaries serves to signal discontinuity. Such a signal can have different functions; in other words, it is not necessarily the case that every single grouping boundary indicates a change of character; yet, it is quite plausible that every change of character requires a grouping boundary to be placed."

(iii) Third, disjoint reference involves the creation of a new position for the new character, which Patel-Grosz et al. analogize to sign language and gestural loci. As they note, the analogy is strengthened by the fact that in sign language (and possibly gestures), a locus may simultaneously function as a discourse referent and as a simplified picture of its denotation.

(iv) Fourth, for the dancer at least, it appears that the change of body orientation is in itself important to signal the change of character. Since in this case the dancer embodies the character's actions, it is tempting to relate this change of orientation to the operation of Role Shift in sign language (see Patel-Grosz et al. 2018 for a brief discussion of this connection).[57]

One key question for the future is how to integrate insights about visual narratives, loci, and groups in dance semantics. Patel-Grosz et al. 2018 propose to extend Abusch's (2013) analysis of coreference in visual narratives, while incorporating constraints due to grouping and loci, and possibly making the analysis more abstract by adopting a dance analogue of the 'source-based semantics' we outlined above for music.

Let us add that it would be interesting to investigate ways in which semantic properties of dance can be investigated by creating composite utterances made of words of dance snippets, just as we did above with gestures and visual animations.

6   Conclusion

We have developed two lines of argument in favor of an extension of the domain of formal semantics. The first line starts from unmistakably linguistic properties of non-standard objects. Natural language semantics cannot coherently ignore these objects. And they raise foundational questions: is language a broader natural class than was initially thought? do some semantic rules originate in other cognitive

---

[57] Since the dancer depicts actions rather than words or thoughts, the relevant point of comparison is Action rather than Attitude Role Shift (see Davidson 2015, Schlenker 2017a,b).

systems, which might explain the ease with which they apply to apparently non-linguistic objects? or are semantic rules just easy to recycle for non-linguistic objects?[58] The second line starts from a broader typology of meaning operations in nature, and yields unexpected connections among semantic domains. Natural language semantics may refrain from the latter extensions, but at a cost: it will fail to understand the place of human meaning among semantic systems in nature, and it will miss out on fruitful connections unearthed by this comparative approach.

Let us take stock. The first line of argument started from human language, and sought to establish the following points:

(i) In speech and sign alike, iconic enrichments interact in non-trivial ways with logical operators, and thus a semantic theory must include an iconic component.

(ii) Greenberg and Abusch's semantics for pictures and visual narratives might provide helpful formalisms in this connection, but they cannot be applied without modification to iconic modulations. The reason is that in iconic modulations include a given expression simultaneously has a conventional and an iconic component, and that a projection-based semantics would wrongly predict that the conventional component is also interpreted iconically. Extant accounts posit weaker preservation rules, but they will have to be refined in the future.

(iii) Different types of iconic enrichments make different types of semantics contributions (at-issue, cosuppositional, supplemental), depending on whether they co-occur with, modulate, follow or replace words. But the typology has yet to be fully derived on theoretical grounds.

(iv) The informational content of pro-speech gestures appears to be divided among familiar slots of the inferential typology of language, possibly by way of the same algorithms.

(v) This division can be effected 'on the fly' for stimuli that one has never seen before (uncommon gestures, but also visual animations). This finding has or might be extended to further types of stimuli, such as vocal gestures and acoustic animations.

(vi) Gestures don't just have a semantics, they also have a grammar, which can be investigated with greater ease in the case of pro-speech gestures. These arguably obey some non-trivial properties of sign language grammar (which in no way implies that signs are just gestures, of course).

The second line of argument is different: by extending the methods of formal semantics beyond language, we obtain a broader and more interesting typology of meaning phenomena in nature. But it turns out that there are fruitful connections among linguistic and non-linguistic phenomena. We discussed the following:

(vii) Monkey calls have a completely different syntax and semantics than human language, but there is potential evidence for word-internal compositionality (*-oo* suffix in Campbell's monkeys), and rules of competition among calls (by way of the Informativity Principle) might be crucial to understand why general calls fail to be used when more specific calls are available. If correct, the Informativity Principle offers a rich analytical tool in the analysis of animal meanings. While ape calls are not understood well yet, work on ape gestures has unearthed rich repertoires of communicative gestures.

(viii) Monkey calls and ape gestures alike can give rise to a phylogenetic study of the evolution of form and meaning. Ape gestures are arguably continuous with gestures found in human infants, which raises questions about their possible connection with some adult gestures.

(ix) Besides its relevance to iconic enrichment in human language and possibly to some ape gestures, picture semantics (especially Greenberg's) raises interesting issues of its own. Abusch and Rooth's

---

[58] Thanks to E. Chemla for helpful comments on these points.

work convincingly argues for the existence of anaphoric relations among viewpoint in pictures, and possibly also for the existence of discourse referents (variables) associated with objects depicted in pictures. Viewpoint shift might be productively compared to Role Shift in sign language. In addition, viewpoint changes are controlled by sophisticated rules in film, whose grammar is beginning to be studied with formal means.

(x) A generalization of iconic semantics, based on virtual sources, can form the backbone of semantics for music. Importantly, this semantics collects some but not all inferences that can be drawn on the sources of the music; this, in turn, is essential to deliver abstract inferences that diverse objects and events can satisfy, including ones that have nothing to do with sound production. Several inferential mechanisms seem to be borrowed from biological systems, including human voice and animal calls.

(xi) While dance semantics is in its infancy, it already seems clear that some referential dances make use of devices seen in gestures and signs, such as loci and possibly some analogues of Role Shift. A semantics for abstract dance has yet to be developed, possibly by analogy with the source-based semantics for music.

Appendix I: Notational conventions

We follow the notational conventions of Schlenker, to appear h, which are summarized below.

❑ *Sign language transcription conventions*

In the following, sign language sentences are glossed in capital letters, as is standard. A suffixed locus, as in *WORD–i*, indicates that the realization of *WORD* points towards locus *i*. Locus names are assigned from right to left from the signer's perspective; thus when loci *a, b, c* are mentioned, *a* appears on the signer's right, *c* on the left, and *b* somewhere in between. *IX* (for 'index') is a pointing sign towards a locus, while *POSS* is possessive; they are glossed as *IX-i* and *POSS-i* if they point towards (or 'index') locus *i*; the numbers *1* and *2* correspond to the position of the signer and addressee respectively. Agreement verbs include loci in their realization – for instance the verb *a-ASK-1* starts out from the locus *a* and targets the first person locus *1;* it means that the third person individual denoted by *a* asks something to the signer. *IX-arc-i* refers to a plural pronoun indexing locus *i,* as it involves an arc motion towards *i* rather than a simple pointing sign.

Acceptability scores (on a 7-point scale, with 7 =best) on sign language data appear as subscripts at the beginning of examples.

❑ *Spoken language transcription conventions*

Glossing conventions for gestures were chosen to be reminiscent of sign language: here too, we used capital letters to gloss elements that are produced manually. (This choice should definitely not suggest that signs are gestures or conversely.)

For legibility, we use a non-standard font to transcribe gestures. A gesture that co-occurs with a spoken word (= a co-speech gesture) is written in capital letters or as a picture (or both) *preceding* the expression it modifies (in some cases, we have added a link to a video to illustrate some gestures). The modified spoken expression will be boldfaced, and enclosed in square brackets if it contains several words.

Examples: John SLAP **punished** his enemy.

John SLAP_ **punished** his enemy.

John **punished** his enemy.

A gesture that follows a spoken word (= a post-speech gesture) is written in capital letters or as a picture *following* the expression it modifies, and preceded by a dash: – .

John punished his enemy – SLAP.

John punished his enemy – SLAP_.

John punished his enemy – .

A gesture that replaces a spoken word (i.e. a 'pro-speech gesture') is written in capital letters:

My enemy, I will SLAP.

My enemy, I will SLAP_.

My enemy, I will .

As in sign language, pointing gestures are alphabetized from right to left from the speaker's perspective. *IX-a* encodes index pointing towards position *a*, while *IX-hand-a* encodes pointing with an open hand, palm up, towards position *a*. A gestural verb involving slapping is glossed as *SLAP-2* if it is realized towards the addressee, and as *SLAP-a* if it is realized towards a third person position. Refining the notation, we write *SLAP(-2)* if we think that this form is both a second person and a neutral form, usable in all persons.

Appendix II: Examples of musical inferences[59]

In the following, we list examples of musical effects triggered by music. We list musical attributes, their inferential effects, reasons for the inferences, real-world examples outside of music (if applicable), and musical examples in which the effect is found, as well as some modified examples in which the effect has been removed or amplified. Modified versions are due to Arthur Bonetto.

(69) **Attribute:** lower pitch.
Inferential effect (1) (for different sound sources): larger sound source.
Reason: lower pitches are emitted by sound sources with larger resonance chambers (larger animals, larger instruments).
Real-world example: elephants; double-basses (compared to violins, for instance).
Musical example: Saint Saëns, Carnival of the Animals, The Elephant.
Normal version: https://bit.ly/2mea8pQ **[MI-01]**
Modified version (removing the effect): if the pitch is raised (by three octaves), this removes the impression of a large source: https://bit.ly/2CI6Xhk **[MI-02]**

(70) **Attribute:** lower pitch.
Inferential effect (2) (for a given sound source): less excited or less energetic sound source.
Reason: lower-frequency sounds are emitted when a sound-producing movement slows down (technically, pitch is given by the number of vibrations per time unit).
Real-world examples: a tape recorder or vinyl record player that is slowing down because its battery is down will start producing lower-pitch sounds; a fast-moving jumping rope may produce sound, but the frequency goes down when the rope is slowed down.
Musical example: Chopin's Nocturne Op. 9/2, last two measures.
Normal version: the original version ends with two identical chords, the second one 2 octaves below the first one: https://bit.ly/2CKX0zH **[MI-03]**
Modified version (removing the effect): if instead the second chord is raised by 3 octaves and thus ends up being 1 octave above the first one, the effect is arguably less conclusive: https://bit.ly/2Eprsjq **[MI-04]**

(71) **Attribute:** lower loudness.
Inferential effect (1): less energetic sound source.
Reason: loudness is related to sound pressure, and lower-energy sources can be expected to produce less sound pressure.
Real-world example: a whistle will be heard less well if one blows air in it with less energy.
Musical example: last bars of Chopin's Prelude 15 ('Raindrop').
Normal version: the last two bars feature a diminished speed (ritenuto) and loudness (diminuendo): https://bit.ly/2qHPSmj **[MI-05]**
Modified version (increasing the effect): in an exaggerated version of the diminuendo in the normal version, realized with a ritenuto, the source seems to gradually lose energy, becoming slower and softer: https://bit.ly/2CJWHVJ **[MI-06]**

(72) **Attribute:** lower loudness.
Inferential effect (2): sound source which is further away.
Reason: loudness is related to sound pressure, and less pressure will reach the perceiver when the source is further away.
Real-world example: a car moving away is heard with dimishing loudness.
Musical example: Mahler's Frère Jacques (First Symphony, 3rd movement).
Normal version: the beginning features an increasing loudness (crescendo), which can (but need not) be interpreted as a procession approaching the perceiver: https://bit.ly/2ma7rFW **[MI-07]**
Modified version (increasing the effect): by artificially modifying the sound level so that the loudness greatly increases, this can yield the impression that a procession is approaching: https://bit.ly/2m9WnIS **[MI-08]**

---

[59] This list was originally prepared by the author for an interview incorporated in Keats 2018.
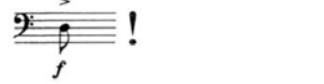
(73) **Attribute:** lower speed.
Inferential effect: the source is becoming slower.
Reason: sounds are indicative of what the source does.
Real-world example: a carpenter hammering nails will produce slower sounds as his actions slow down.
Musical example: Saint-Saëns's, Carnival of the Animals, Tortoises (a radically slowed down version of the Can-Can dance): https://bit.ly/2EpPQ4p [**MI-09**]
Offenbach's original Can-Can (from Orpheus In The Underworld - Overture, Can Can Section - Selection)[60]: https://bit.ly/2T38GtX [**MI-10**]

(74) **Attribute:** silence.
Inferential effect: an event is interrupted.
Reason: sounds are indicative of what the source does (or in this case doesn't do).
Real-world example: a carpenter hammering nails will produce no sounds when he takes a break.
Musical example: Saint-Saëns, Carnival of the Animals, Kangaroos, beginning: when the first piano enters, it plays a series of short notes separated by short silences. This evokes a succession of brief events separated by interruptions. In the context of Saint-Saëns's piece, these sequences evoke kangaroo jumps: for each jump, the ground is hit, hence a brief note, and then the kangaroo rebounds, hence a brief silence. https://bit.ly/2m98kPd [**MI-11**]

(75) **Attribute:** greater dissonance.
Inferential effect (1): the sound source is a less stable *physical* position.
Reason: in tonal music, dissonances are unstable tonal position.
Real-world example: [not applicable - this is a tonal inference, not an inference from normal auditory cognition]
Musical example: Saint Saëns, Carnival of the Animals, Tortoises, measures 10-13.
Normal version: In the original version, there is a dissonance in the first half of measure 12 [because a chord F A C is played with an G# added], as can be heard by focusing only on the violin and piano parts: https://bit.ly/2ECNWNJ [**MI-12**]
Modified version (removing the effect): The dissonance can be removed [by turning the G#'s into A'] and the impression that tortoises disappears (as can be heard by focusing only on the violin and piano part): https://bit.ly/2CWFVCT [**MI-13**]

(76) **Attribute:** greater dissonance.
Inferential effect (2): the sound source is less stable *emotional* position.
Reason: in tonal music, dissonances are unstable tonal position.
Real-world example: [not applicable - this is a tonal inference, not an inference from normal auditory cognition]
Musical example: Herrmann, music for Hitchcock's Psycho, piano reduction (around 13:11): http://bit.ly/2mAjZGL [**MI-14**]
Normal version: dissonances are indicative of anguish: https://bit.ly/2D2NlEK [**MI-15**]
Modified versions: by removing the dissonances, some of the anguished character of the original disappears.
Modified version 1 (removing the effect): https://bit.ly/2EH4iFt [**MI-16**]
Modified version 2 (closer to the original harmony; removing the effect): https://bit.ly/2mtDXmL [**MI-17**]

(77) **Attribute:** change of key [= modulation].
Inferential effect: the sound source is moving to a new kind of environment.
Reason: a change of key corresponds to a transition to a different part of tonal pitch space.
Real-world example: [not applicable - this is a tonal inference, not an inference from normal auditory cognition]
Musical example: Saint Saëns, Carnival of the Animals, The Swan.
Normal version: there is a change of key [= modulation] in measures 7-10, suggesting the swan is moving to a different kind of environment: https://bit.ly/2D6TcNq [**MI-18**]
Modified versions (removing the effect): when the change of key is eliminated, the impression of a different surrounding disappears.
Modified version 1: https://bit.ly/2DqCC80 [**MI-19**]
Modified version 2: https://bit.ly/2ED4yVY [**MI-20**]

---

[60]Excerpt at 32s from https://commons.wikimedia.org/w/index.php?title=File%3AOffenbach_-_Orpheus_in_the_Underworld_-_Overture%2C_Can_Can_section.ogg, retrieved on February 28, 2019.

Appendix III: Bernstein's two interpretations of Strauss's Don Quixote[61]

In a session of his *Young People's Concerts* devoted to the meaning of music (Bernstein 2005), Leonard Bernstein sought to convince his young audience that music doesn't have any referential meaning, and that its true meaning is "the way it makes you feel when you hear it". One of his key arguments was that one can tell the *wrong* story and still have something that fits as 'descriptive' a music as a symphonic poem. To get his point across, Bernstein had his orchestra play Variation II of Strauss's Don Quixote to illustrate a story about Superman; then he had it play the very same music, but now to illustrate an episode of Don Quixote, in accordance with Strauss's intentions. He argued that both versions fit the music equally well. A simplified version of the two stories is given in (78).

(78) **Simplified structure of Bernstein's Don Quixote and Superman interpretations of Strauss's Variation II of *Don Quixote*** (Kriegerisch. "Der siegreiche Kampf gegen das Heer des großen Kaisers Alifanfaron" ("The victorious struggle against the army of the great emperor Alifanfaron") [actually a flock of sheep])   Entire discussion: https://youtu.be/XFZ7wORtj2A **[DQ]**

Don Quixote interpretation	Superman interpretation	Salient musical passage
Context: Don Quixote is a foolish old man who has read too many books about knighthood and decides he is a marvelous knight himself. Sancho Panza is his devoted servant. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=5m17s	**Context:** An innocent man can't sleep in a prison where he was put unjustly. He spends his night playing the kazoo while other prisoners snore. But his friend Superman is coming to rescue him. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=28s	
Don Quixote departs on his horse to conquer the world. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=5m36s	Superman comes charging along through the alley on his motorcycle. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=1m8s	
We hear Sancho chuckling to himself[62]. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=5m45s	Superman whistles his secret whistle (in the woodwinds) so the prisoner will know he's coming. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=1m20s	
They see a flock of sheep in the field going *baa-baa*. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=6m3s	Superman hears all the prisoners snoring away peacefully in the dead silence of night. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=1m28s	
A shepherd is playing on his pipe. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=6m16s	Over this snoring, Superman hears his imprisoned friend playing his kazoo over the snoring, which gets louder as he gets nearer. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=1m50s	
Don Quixote charges at the sheep, taking them to be an army. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=6m27s	Superman charges into the prison yard and bops the guard over the head, done in the orchestra with a loud bang in the percussion. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=2m14s	loud bang in the percussion:
The sheep run off in all directions baaing wildly. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=6m40s	The kazoo stops playing, and with all the snoring still going on, Superman grabs his friend and carries him away on his motorcycle. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=2m22s The snoring gets farther and farther away, until we don't hear it any more. https://www.youtube.com/watch?v=XFZ7wORtj2A&t=2m37s	
Don Quixote is convinced he has done a truly knightly deed, and is he proud! https://www.youtube.com/watch?v=XFZ7wORtj2A&t=6m45s	Our hero at last reaches freedom! https://www.youtube.com/watch?v=XFZ7wORtj2A&t=2m50s	

Strikingly, some key structural elements remain constant across the two interpretations, with small modifications. The main correspondences are listed in (79): Don Quixote corresponds to Superman, with a twist: a melody attributed to Sancho Panza chuckling is assigned to Superman whistling; thus two virtual sources in the Don Quixote interpretation are merged in the Superman

---

[61] Thanks to Arthur Bonetto for discussion.
[62] The text has "chuckling to himself", Bernstein's live performance has: "laughing at Don Quixote" (there are several small differences between the live and the printed version).

interpretation. The sheep of the Don Quixote interpretation become prisoners snoring in the Superman interpretation. The shepherd becomes an innocent prisoner playing his kazoo. And the triumphant ending represents in one case Don Quixote's pride, in the other Superman and his friend's obtaining freedom.[63]

(79) **Correspondence in terms of sources between Bernstein's Don Quixote and Superman interpretations**

Don Quixote interpretation	Superman interpretation
Don Quixote on his horse + Sancho Panza chuckling	Superman on his motorcycle + Superman whistling
Sheep going ba ba	Prisoners snoring
Shepherd playing on his pipe, with the sound becoming louder as Don Quixote comer nearer	Innocent prisoner playing his kazoo, with the sound becoming louder as Don Quixote comer nearer
Don Quixote charges the sheep	Superman charges into the prison
The sheep run off in all directions baaing wildly.	Superman grabs his friend and carries him away on his motorcycle.
Don Quixote is proud of his knightly deed	Superman and his friend reach freedom

In the end, Bernstein's example doesn't at all show that music has no meaning, or that its true meaning is "the way it makes you feel when you hear it". Rather, it suggests that music has an *abstract* meaning, which allows highly diverse but structurally analogous sets of events to be denoted by it.

Appendix IV: An example of semantic interaction between dance and music[64]

Charnavel 2016 writes that "the correspondence between music and dance can (...) resolve cases of ambiguity: if the musical structure is ambiguous between different interpretations, the position of the movement phrases on the music can contribute to disambiguating it." It is clear that Charnavel has in mind cases of syntactic ambiguity. Can we find more semantic cases of disambiguation?

At the beginning of Balanchine's ballet Symphony in C, on Bizet's music, the alternation of dancers imposes on the music (we think) a reading on which parallel musical groups are attributed to different sources. This is striking because on its own the music forces no such interpretation, as there are no relevant changes (in particular of instruments) across the groups.

The 1st violin part appears in (80). Starting in bar 57 (i.e. [2] in the score) alternation, which we indicated by 'right' and 'left', between musical groups that correspond to actions by the right-hand ballerina and the left-hand ballerina (from the viewer's perspective). We have boxed five passages: motif A corresponds to a movement by the right ballerina ('A-right'), it is repeated with modification with a reply by the left ballerina ('A-left'); motif B corresponds to a movement by the right ballerina ('B-right') followed again by a reply by the left ballerina ('B-left'), before the two act in concert ('together').

---

[63] Note that the musical chaos corresponding to the sheep's baaing wildly is not easily interpreted in the Superman story.

[64] Thanks to Isabelle Charnavel and Léo Migotti for very helpful discussion (I am also indebted to Migotti for showing an excerpt of the beginning of Symphony in C). Thanks to Arthur Bonetto for discussion of the score.

(80) Association of the main action with the right vs. left ballerina in Balanchine's Symphony in C, 1st violins, bars 57 sqq.

Three realizations of the relevant part of this ballet are linked to (81):

(81) Three realizations of Balanchine's Symphony in C, starting around the ff in bar 49.
a. New York City Ballet 1973 https://youtu.be/HKG6v4a2DvA [Bal-1]

b. Bolshoi Ballet 2008 (Marianna Ryzhkina, Chinara Alizade, Anna Tikhomirova & Karim Abdullin) https://youtu.be/p3r0dDe43aw [Bal-2]
c. Opéra National de Paris, Musical direction Philippe Jordan https://youtu.be/x_FSbnMGvnM [Bal-3]

Our impression is that the ballet imposes on the music a structure with two different sources in a kind of dialogue (or at least interaction). To check that the music itself does not force a reading with two sources between bars 57-61 (= A-right) and 61-65 (=A-left), one can listen to the music alone. It can also help to consult a visualization of the score corresponding to A-right and A-left, from Stephen Malinowski's Music Animation Machine (in (82)). The part that corresponds to A-right and A-left in (80) appears in (83). As can be heard and seen, there is no obvious asymmetry in the music that would force one to posit different abstract sources or a dialogue; if it is brought out, this interpretation is due to the dance (although it is compatible with the music).[65]

(82) Visualization with Malinowski's Music Animation Machine, starting around the ff in bar 49
https://youtu.be/u5WEAvJatWg [Bal-4]

(83) Visualization of the orchestral score in A-right and A-left in (80)

---
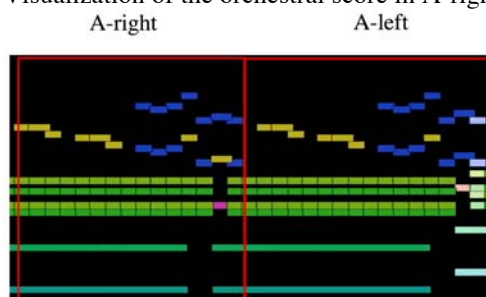
[65] Several remarks should be added (thanks to Arthur Bonetto for suggestions). (i) First, the score has a gradual crescendo starting at the end of A-right. If it is indeed realized in a gradual fashion (as in Haitink's concert performance with the Royal Concertgebouw Orchestra https://youtu.be/RpzpO7u5QZM [Bal-5]), the dialogue-based interpretation becomes less salient from the music alone. Interestingly, some of the ballet interpretations, such as that in (81)c, do not realize this gradual crescendo in a salient fashion. (ii) Second, Balanchine's interpretation is certainly permitted (but not forced) by the music, for harmonic reasons: A-right resembles B-right in being harmonically stable (staying in A^7 and D^7 respectively); A-left resembles B-left in moving from one harmony to another (from A^7 to D^7, and from D^7 to G^7). (iii) If one wanted to force a dialogical interpretation on the music, one could probably use timbre or even location, with A-right and B-right played by one instrument/from one location, and A-left and B-left played by/from another.

Importantly, the inference obtained doesn't pertain to the grouping structure per se, or at least not just to the grouping structure: the music makes clear that A-right and A-left are different group, but what the dance contributes is the impression that they are due to different virtual sources – arguably a semantic notion.

References

Abner N., Cooperrider K. and Goldwin-Meadow S.: 2015, Gesture for Linguists: A Handy Primer, *Lang. Linguist. Compass*, 9/11, 437–449

Abrusán, Márta. 2011. Predicting the presuppositions of soft triggers. *Linguistics and Philosophy*, **34**(6), 491–535.

Abusch, Dorit: 2013, Applying discourse semantics and pragmatics to co-reference in picture sequences. In *Proceedings of Sinn und Bedeutung* 17: 19–25.

Abusch, Dorit: 2015, Possible Worlds Semantics for Pictures. Manuscript, Cornell University.

Abusch, Dorit and Rooth, Mats: 2017, The formal semantics of free perception in pictorial narratives. In Alexandre Cremers, Thom van Gessel & Floris Roelofsen (eds), *Proceedings of the 21st Amsterdam Colloquium*. Available online at https://semanticsarchive.net/Archive/jZiM2FhZ/AC2017-Proceedings.pdf.

Arnold Kate, Zuberbühler Klaus: 2006, Semantic combinations in primate calls. *Nature* 441: 303.

Arnold, K. and Zuberbühler K.: 2008, Meaningful call combinations in a non-human primate. *Current Biology* 18(5): R202–R203

Arnold Kate, Zuberbühler Klaus, 2012. Call combinations in monkeys: Compositional or idiomatic expressions? *Brain and Language* 120 (3): 303-309.

Atkin, Albert: 2013, Peirce's Theory of Signs. *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/>.

Bar-Lev, Moshe E.: 2018, Free Choice, Homogeneity, and Innocent Inclusion. Doctoral dissertation, Hebrew University, Jerusalem.

Benitez-Quiroz C. F., Wilbur R. B., Martinez A. M.: 2016, The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150, 77–84

Bernstein, Leonard: 2005, *Young People's Concerts*. Foreword by Michael Tilson Thomas. Amadeus Press, Pompton Plains, New Jersey.

Berthet, Mélissa, *Semantic content in Titi monkey alarm call sequences*. Doctoral dissertation, Université de Neuchâtel, 2018.

Berthet, Mélissa; Neumann, Christof; Mesbahi, Geoffrey; Cäsar Damas, Cristiane; Zuberbuhler, Klaus: 2018, Contextual encoding in titi monkey alarm call sequences. *Behav Ecol Sociobiol* (2018) 72: 8. https://doi.org/10.1007/s00265-017-2424-z

Berwick Robert C., Okanoya Kazuo, Beckers Gabriel J.L. and Bolhuis Johan J., 2011. Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences* 15 (3): 113-121.

Blumstein, Daniel T., Bryant, Gregory A. and Kaye, Peter: 2012, The sound of arousal in music is context-dependent. *Biol. Lett.* 8, 744-747

Bregman, Albert S.: 1994, *Auditory Scene Analysis*. MIT Press.

Briefer, E. F.: 2012,Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, 288(1), 1-20

Buccola, B.; Dautriche, I.; and Chemla, E.: 2018, Competition and symmetry in an artificial word learning task. *Frontiers in Psychology*.

Byrne RW, Cartmill E, Genty E, Graham KE, Hobaiter C, Tanner J. Great ape gestures: 2017, Intentional communication with a rich set of innate signals. Animal Cognition 20(4):755–69.

Charnavel, Isabelle: 2016, Steps Towards a Theory of Dance Cognition. Manuscript, Harvard University.

Chemla, Emmanuel: 2009, Presuppositions of quantified sentences: experimental data. *Natural Language Semantics*,17(4):299-340.

Chemla, E.; Dautriche, I.; Buccola, B.; and Fagot, J.: 2018, Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio). Ms. CNRS, University of Edinburgh, Aix-Marseille University.

Chemla, E.; Buccola, B.; and Dautriche, I.: to appear, Connecting content and logical words. *Journal of Semantics*.

Clark, H. H.: 2016, Depicting as a method of communication. *Psychological Review*, 124, 324-347.

Clarke E, Reichard UH, Zuberbühler K (2006) The Syntax and Meaning of Wild Gibbon Songs. *PLoS ONE* 1(1): e73. doi:10.1371/ journal.pone.0000073

Crockford, Catherine, Wittig, Roman M., Mundry, Roger, & Zuberbühler, Klaus: 2012, Wild chimpanzees inform ignorant group members of danger. *Current Biology*, *22*(2), 142-146.

Cumming, Samuel, Greenberg, Gabriel & Kelly, Rory: 2017, Conventions of Viewpoint Coherence in Film. *Philosophers Imprint* 17, 1.

Davidson, Kathryn: 2015, Quotation, Demonstration, and Iconicity. *Linguistics & Philosophy,* 38: 477-520.

Demolin, D., Ades, C. & Mendes, F. D., 2019: Context-sensitive grammars in Muriqui vocalizations. Manuscript.

Dohen, Marion. 2005. Deixis prosodique multisensorielle: Production et perception audiovisuelle de la Focalisation contrastive en français. Doctoral dissertation, Institut National Polytechnique de Grenoble.

Dohen, Marion, and Hélène Loevenbruck. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language & Speech* 52(2-3): 177-206.

Douglas, P.H., and L.R. Moscovice. 2015. Pointing and pantomime in wild apes? Female bonobos use referential and iconic gestures to request genito-genital rubbing. Sci Rep DOI: 10.1038/srep13999

Ebert, Cornelia: 2018, A comparison of sign language with speech plus gesture. *Theoretical Linguistics*.

Ebert, Cornelia and Ebert, Christian: 2014, Gestures, Demonstratives, and the Attributive/Referential Distinction. Handout of a talk given at Semantics and Philosophy in Europe (SPE 7), Berlin, June 28, 2014.

Eckardt, Regine. 2014. *The Semantics of Free Indirect Discourse. How Texts Allow to Mind-read and Eavesdrop*. Leiden: Brill.

Engesser, Sabrina and Townsend, Simon: 2019, Combinatoriality in the vocal systems of nonhuman animals. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1493. DOI: 10.1002/wcs.1493

Fitch, Tecumseh W., Reby, D.: 2001, The descended larynx is not uniquely human. *Proceedings of the Royal Society of London*. Series B, 268, 1669-1675.

Gautier, J-P. (1988) Interspecific affinities among guenons as deduced from vocalizations. In Gautier-Hion, A., Bourlière, F., Gautier, J.P. & Kingdon (Eds.), *A Primate Radiation - Evolutionary Radiation of the African Guenons* (pp. 194-226). Cambridge University Press.

Genty, Emilie and Zuberbühler, Klaus: 2014, Spatial Reference in a Bonobo Gesture, *Current Biology*, http:// dx.doi.org/10.1016/j.cub.2014.05.065

Giorgolo, Gianluca: 2010, *Space and Time in Our Hands*, Uil-OTS, Universiteit Utrecht.

Godinho, Ricardo Miguel; Spikins, Penny; O'Higgins, Paul: 2018, Supraorbital morphology and social dynamics in human evolution. *Nature Ecology and Evolution*, 2018 DOI: 10.1038/s41559-018-0528-0

Goldin-Meadow, S.: 2003, *The Resilience of Language*. Taylor & Francis.

Goldin-Meadow, Susan and Brentari, Diane: 2017, Gesture, sign and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, doi:10.1017/S0140525X15001247

Graham KE, Hobaiter C, Ounsley J, Furuichi T, Byrne RW: 2018, Bonobo and chimpanzee gestures overlap extensively in meaning. *PLoS Biol* 16(2):e2004825

Granroth-Wilding, Mark and Steedman, Mark: 2014, A robust parser-interpreter for jazz chord sequences. *Journal of New Music Research* 43 (4), 355-374.

Greenberg, Gabriel: 2011, *The Semiotic Spectrum*. PhD dissertation, Rutgers.

Greenberg, Gabriel: 2013. Beyond Resemblance. *Philosophical Review* 122:2, 2013

Greenberg, Gabriel: 2018, The Geometry of Pictorial Representation. Manuscript, UCLA.

Guerrini, Janek and Schlenker, Philippe: 2019, Linguistic inferences without words: the case for pro-speech vocal gestures. Poster to be given at GLOW 42, May 8-10, 2019.

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., ... & Lenglet, G. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic biology*, *62*(4), 539-554.

Heim, Irene, 1983, "On the projection problem for presuppositions", in Barlow, M. and Flickinger, D. and Westcoat, M. (eds.), Second Annual West Coast Conference on Formal Linguistics, Stanford University, 114–126.

Heim, Irene and Kratzer, Angelika: 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.

Hobaiter C, Byrne RW: 2017, What is a gesture? A meaning-based approach to defining gestural repertoires. *Neuro Biobehav Rev* 82:3–12

Hopcroft, John and Ullman, Jeffrey: 1979, *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Horn, Laurence R. 1972. On the semantic properties of the logical operators in English. Ph.D. thesis, University of California at Los Angeles.

Juslin P. and Laukka P.: 2003, Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*; 129(5):770–814.

Kanazawa, S.: 2016, Recognition of facial expressions in a japanese monkey (Macaca fuscata) and humans (Homo sapiens). Primates, 37 (1): 25-38.

Katzir, Roni: 2007, Structurally-defined alternatives. *Linguistics and Philosophy*, **30**(6), 669–690.

Keats, Jonathan: 2018, Science of Music: Listen up! *Discover Magazine*, special issue on *Everything worth knowing,* August 2018. http://discovermagazine.com/2018/jul-aug/science-of-music

Kersken, V., Gómez, JC., Liszkowski, U., Soldati, A., Hobaiter, C.: 2018, A gestural repertoire of 1- to 2-year-old human children: in search of the ape gestures. *Animal Cognition*. https://doi.org/10.1007/s10071-018-1213-z

Koelsch S.: 2011, Towards a neural basis of processing musical semantics. *Physics of Life Reviews* 8(2):89–105

Koelsch, S.: 2012, 'Musical Semantics'. Chapter from *Brain and Music*, Wiley-Blackwell.

Križ, Manuel, 2015. *Aspects of homogeneity in the semantics of natural language* (Doctoral dissertation, PhD thesis, University of Vienna).

Križ, Manuel: 2016. Homogeneity, Non-maximality, and all. *Journal of Semantics* 33/3.

Križ, Manuel and Spector, Benjamin: 2017. Interpreting Plural Predication: Homogeneity and Non-Maximality. Manuscript, Institut Jean-Nicod.

Kuhn, Jeremy and Chemla, Emmanuel: 2017, Facial expressions and speech acts in non-signers. Poster, 6th Meeting of Signed and Spoken Language Linguistics (SSLL 6), Japan.

Leff, Koby: 2019, Spatial Coherence in Film. Video, accessed at https://www.youtube.com/watch?v=SjKcGSzfnn0 on January 18, 2019.

Lemasson Alban, Ouattara Karim, Bouchet Hélène and Zuberbühler Klaus, 2010. Speed of call delivery is related to context and caller identity in Campbell's monkey males. *Naturwissenschaften* 97 (11): 1023-1027.

Lerdahl, Fred and Ray Jackendoff: 1983, *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Liddell, Scott K.: 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

Lillo-Martin, Diane & Klima, Edward S.: 1990, Pointing out Differences: ASL Pronouns in Syntactic Theory. In Susan D. Fischer & Patricia Siple (Eds.), *Theoretical Issues in Sign Language Research*, Volume 1: Linguistics, 191-210. Chicago: The University of Chicago Press.

Lillo-Martin, Diane & Richard P. Meier. 2011. On the linguistic status of 'agreement' in sign languages. *Theoretical Linguistics* 37(3/4). 95–141. DOI: https://doi.org/10.1515/thli.2011.009 Maestripieri D.: 1997, Gestural communication in macaques: Usage and meaning of nonvocal signals. *Evolution of communication* 1: 193–222.

Maynard Smith, John and Harper, David: 2003, *Animal Signals*. Oxford Series in Ecology and Evolution, Oxford University Press.

Montague, Richard, 1970a. English as a Formal Language. In Visentini, B. et al. (eds.) Linguaggi nella società e nella tecnica. Milan: Edizioni di Comunità. 189-224.

Montague, Richard, 1970b. Universal grammar. *Theoria* 36: 373-398.

Morris, Charles W.: 1938, Foundations of the Theory of Signs. In Otto Neurath (ed.), *International Encyclopedia of Unified Science*, vol. 1 no. 2. Chicago: University of Chicago Press.

Ohala, J. J.: 1994, The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols & J. J. Ohala (Eds.), Sound Symbolism, 325- 347. Cambridge: Cambridge University Press.

Okrent, Arika: 2002, A modality-free notion of gesture and how it can help us with the morpheme vs. gesture question in sigh language linguistics, or at least give us some criteria to work with. In R.P. Meier, D.G. Quinto-Pozos, & K.A. Cormier (eds). *Modality and structure in signed and spoken*

languages (pp. 175-198). Cambridge: Cambridge University Press.

Parr LA, Waller BM: 2006, Understanding chimpanzee facial expression: insights into the evolution of communication. *Soc Cogn Affect Neurosci*. 2006; 1: 221–8. doi: 10.1093/scan/nsl031

Patel-Grosz, Pritty; Grosz, Patrick Georg; Kelkar, Tejaswinee & Jensenius, Alexander Refsum (2018). Coreference and disjoint reference in the semantics of narrative dance, In Uli Sauerland & Stephanie Solt (ed.), Proceedings of Sinn und Bedeutung 22, vol. 2, ZASPiL 61. Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS). Chapter in Vol. 2. s 199 - 216

Peirce, Charles S.: 1868, On a New List of Categories. *Proceedings of the American Academy of Arts and Sciences* 7, 287-298.

Pellat, A.: 1980, Facial expressions of Papio Ursinus and some other higher primates. S. Afr. J. ScL, 76, 413-418.

Perelman, Polina, Warren E. Johnson, Christian Roos, Hector N. Seuánez, Julie E. Horvath, Miguel A. M. Moreira, Bailey Kessing, Joan Ponitus, Melody Roelke, Yves Rumpler, Maria Paula C. Schneider, Artur Silva, Stephen J. O'Brien and Jill Pecon-Slattery, 2011. A molecular phylogeny of living primates. *PLoS genetics* 7(3): e1001342.

Pesetsky, David and Katz, Jonah. 2009. The Identity Thesis for Music and Language. Manuscript, MIT.

Pfau, R., Salzmann, M., & Steinbach, M. (2018). The syntax of sign language agreement: Common ingredients, but unusual recipe. *Glossa: a journal of general linguistics*, 3(1): 107. 1–46, DOI: https://doi.org/10.5334/gjgl.511

Potts, Christopher. 2005. *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics, Oxford: Oxford University Press.

Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2): 165–197. Published with commentaries by several researchers, and replies by Potts.

Pylyshyn, Zenon: 2003, *Seeing and Visualizing: It's Not What You Think*. MIT Press.

Rohrmeier, Martin: 2011, Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music* 5 (1), 35–53.

Sandler, Wendy: 2018, The body as evidence for the nature of language. *Frontiers in Psychology*. doi.org/10.3389/fpsyg

Sandler, Wendy and Lillo-Martin, Diane: 2006, *Sign Language and Linguistic Universals*. Cambridge University Press.

Sauerland, Uli. 2007. Beyond unpluggability. *Theoretical Linguistics*, 33:231–236.

Schembri, Adam, et al. 2018. Indicating verbs as typologically unique constructions: Reconsidering verb 'agreement' in sign languages. *Glossa: a journal of general linguistics* 3(1): 89. 1–40, DOI: https://doi. org/10.5334/gjgl.468

Schlenker, Philippe: 2004, Context of Thought and Context of Utterance (A Note on Free Indirect Discourse and the Historical Present), *Mind & Language* 19: 3, 279-304

Schlenker, Philippe: 2007, Expressive Presuppositions. *Theoretical Linguistics* 33 (2) : 237-246

Schlenker, Philippe: 2011. Donkey Anaphora: the View from Sign Language (ASL and LSF). *Linguistics and Philosophy* **34**(4): 341-395.

Schlenker, Philippe: 2014. Iconic Features. *Natural Language Semantics* **22**(4): 299-356.

Schlenker, Philippe: 2017a, Outline of Music Semantics. *Music Perception: An Interdisciplinary Journal* 35, 1: 3-37 DOI: 10.1525/mp.2017.35.1.

Schlenker, Philippe: 2017b, Super Monsters I: Attitude and Action Role Shift in Sign Language. *Semantics & Pragmatics,* Volume 10, 2017.

Schlenker, Philippe: 2017c, Super Monsters II: Role Shift, Iconicity and Quotation in Sign Language. *Semantics & Pragmatics*. Volume 10, 2017.

Schlenker, Philippe: 2017d, Sign Language and the Foundations of Anaphora. *Annual Review of Linguistics*, 3:149–77

Schlenker, Philippe: 2018a, Gesture Projection and Cosuppositions. *Linguistics & Philosophy* 41, 3:295–365.

Schlenker, Philippe: 2018b, Visible Meaning: Sign Language and the Foundations of Semantics. *Theoretical Linguistics*.

Schlenker, Philippe: 2018c, Sign Language Semantics: Problems and Prospects. *Theoretical Linguistics* 44(3-4): 295–353.

Schlenker, Philippe: 2018d, Iconic Pragmatics. *Natural Language & Linguistic Theory* 36, 3:877–936

Schlenker, Philippe: 2018g, Prolegomena to Music Semantics. *Review of Philosophy & Psychology*. https://doi.org/10.1007/s13164-018-0384-5

Schlenker, Philippe: 2018h, Iconic Presuppositions. Manuscript, Institut Jean-Nicod and New York University.

Schlenker, Philippe: to appear, d. Iconic Pragmatics. To appear in *Natural Language & Linguistic Theory*.

Schlenker, Philippe: to appear, f. Gestural Semantics: Replicating the typology of linguistic inferences with pro- and post-speech gestures. *Natural Language & Linguistic Theory*.

Schlenker, Philippe: to appear g, The Semantics and Pragmatics of Appositives. To appear in Lisa Matthewson, Cécile Meier, Hotze Rullmann, Thomas Ede Zimmermann (eds), *Companion to Semantics*, Wiley.

Schlenker, Philippe: to appear h, Gestural Grammar. Accepted with minor revisions, *Natural Language & Linguistic Theory*.

Schlenker, Philippe, and Chemla, Emmanuel: 2018, Gestural Agreement. *Natural Language & Linguistic Theory*. 36, 2: 87–625587. https://doi.org/10.1007/s11049-017-9378-8

Schlenker, Philippe, Lamberton, Jonathan & Santoro, Mirko: 2013. Iconic Variables. *Linguistics & Philosophy* **36**(2): 91-149.

Schlenker, Philippe, Chemla, Emmanuel, Arnold, Kate, Lemasson, Alban, Ouattara, Karim , Keenan, Sumir , Stephan, Claudia, Ryder, Robin, Zuberbühler, Klaus: 2014, Monkey Semantics: Two 'Dialects' of Campbell's Monkey Alarm Calls. *Linguistics & Philosophy* 37, 6: 439-501. *DOI 10.1007/s10988-014-9155-7*

Schlenker, Philippe; Chemla, Emmanuel; Schel, Anne; Fuller, James; Gautier, Jean-Pierre; Kuhn, Jeremy; Veselinovic, Dunja; Arnold, Kate; Cäsar, Cristiane; Keenan, Sumir; Lemasson, Alban; Ouattara, Karim; Ryder, Robin; Zuberbühler, Klaus: 2016a, Formal Monkey Linguistics. *Theoretical Linguistics* 42,1-2:1–90, DOI: 10.1515/tl-2016-0001

Schlenker, Philippe; Chemla, Emmanuel; Schel, Anne; Fuller, James; Gautier, Jean-Pierre; Kuhn, Jeremy; Veselinovic, Dunja; Arnold, Kate; Cäsar, Cristiane; Keenan, Sumir; Lemasson, Alban; Ouattara, Karim; Ryder, Robin; Zuberbühler, Klaus: 2016b, Formal Monkey Linguistics: the Debate. (Replies to commentaries). *Theoretical Linguistics* 42, 1-2:173–201, DOI: 10.1515/tl-2016-0010

Schlenker, Philippe; Chemla, Emmanuel; Zuberbühler, Klaus: 2016c, What do Monkey Calls Mean? *Trends in Cognitive Sciences,* 20, 12, 894–904.

Schlenker, Philippe; Chemla, Emmanuel; Cäsar, Cristiane; Ryder, Robin; Zuberbühler, Klaus: 2016d, Titi Semantics: Context and Meaning in Titi Monkey Call Sequences. *Natural Language & Linguistic Theory* doi:10.1007/s11049-016-9337-9

Schlenker, Philippe; Chemla, Emmanuel; Arnold, Kate; Zuberbühler, Klaus: 2016e, Pyow-Hack Revisited: Two Analyses of Putty-nosed Monkey Alarm Calls. *Lingua* 171:1-23

Schlenker, Philippe; Chemla, Emmanuel; Zuberbühler, Klaus: 2017. Semantics and Pragmatics of Monkey Communication. *Oxford Research Encyclopedia of Linguistics*. http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-220

Schlenker, Philippe, and Lamberton, Jonathan, to appear. Iconic Plurality. *Linguistics & Philosophy*.

Sharvit, Yael: 2008. The puzzle of free indirect discourse. *Linguistics and Philosophy* **31**: 353 - 395.

Suzuki, Toshitaka N; Wheatcroft, David; Griesser, Michael (2017). Wild Birds Use an Ordering Rule to Decode Novel Call Sequences. *Current Biology*, 27(15):2331-2336.e3. DOI: https://doi.org/10.1016/j.cub.2017.06.031

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

Tieu, Lyn; Pasternak, Robert; Schlenker, Philippe; Chemla, Emmanuel: 2018a, Co-speech gesture projection: Evidence from inferential judgments. *Glossa* 3(1), 109. DOI: http://doi.org/10.5334/gjgl.580

Tieu, Lyn; Schlenker, Philippe; & Chemla, Emmanuel: 2018b, Linguistic inferences without words: Replicating the inferential typology with gestures. Manuscript.

Wilbur Ronnie B.. 2012. Information structure. In *Sign language. An international handbook* (*HSK - Handbooks of linguistics and communication science*), ed by Pfau, Roland, Markus Steinbach & Bencie Woll, 462-489. Berlin: Mouton de Gruyter.

Zehr, Jérémy, Cory Bill, Lyn Tieu, Jacopo Romoli & Florian Schwarz. 2015. Existential presupposition projection from none: An experimental investigation. In Thomas Brochhagen, Floris Roelofsen & Nadine Theiler (eds.), Proceedings of the 20th Amsterdam Colloquium, 448–457.

Zehr, Jérémy, Cory Bill, Lyn Tieu, Jacopo Romoli & Florian Schwarz. 2016. Presupposition projection from the scope of None: Universal, existential, or both? In Mary Moroney, Carol-Rose Little, Jacob Collard & Dan Burgdorf (eds.), Proceedings of the 26th Semantics and Linguistic Theory Conference, 754–774. DOI: https://doi.org/10.3765/salt.v26i0.3837

Zuberbühler Klaus: 2002, A syntactic rule in forest monkey communication. *Animal Behaviour* 63 (2): 293-299.

Zuberbühler, Klaus, Jenny David, and Bshary Redouan. 1999. The predator deterrence function of primate alarm calls. *Ethology* 105(6): 477–490. doi:10.1046/j.1439-0310.1999.00396.x.