TRANSPARENCY OF TRANSITIVITY IN PANTOMIME, SIGN LANGUAGE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Chuck Bradley

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Ronnie B Wilbur, Chair

Department of Speech, Language, Hearing Sciences and School of Interdisciplinary Studies

Dr. Jeffrey Mark Siskind

School of Electrical and Computer Engineering

Dr. Elaine Francis

Department of English and School of Interdisciplinary Studies

Dr. Evguenia Malaia

Department of Communicative Disorders, University of Alabama

Approved by:

Dr. Ronnie B Wilbur

Department of Speech, Language, Hearing Sciences and School of Interdisciplinary Studies

To Chuck!

ACKNOWLEDGMENTS

Thank you to my committee, Drs. Ronnie Wilbur, Jeff Siskind, Evie Malaia and Elaine Francis, for their substantial contributions to this effort. At the very least, I hope I've learned half as much as you've taught me.

In no particular order I would like to thank the following folks for their support, whether they were aware of it or not. Thank you to my friend, lab mate, and colleague Kadir Gököz for inspiring a strong curiosity and creative spirit in me. Thank you to my friend, lab mate, and colleague Charles Lam for teaching me how to apply a critical eye to them. To this end, I am also very grateful to the members of the Lafayette Brain Trust, in its many instantiations.

For many fruitful discussions of parts and wholes of this dissertation, I am indebted to Ashley Kentner, Serpil Karabüklü, Guilherme Lourenço, and Ruth Crabtree. In particular, I am grateful to Ashley and Guilherme for their deft displays of athleticism in helping me with stimulus creation. Thank you to Robin Shay, Rita Mowl, Sandra Wood, Rana Abu Zhaya, Jess Rohr, Josh Perry, Stacey Enneking, Ryan van Nood, Ingrid Parra, Jake Desmond, John Diamond, Niall Peach, Heather Offerman, Christopher Munt, and Carolyn Szuter for the raw material used in this dissertation, and to Tori Glass and Lexi Rozycki for shaping it into an analyzable form. Thanks, too, to the scores of Turkers whose brains I picked on a measly grad student's budget. For their companionship along the way, I would like to thank Ben Van Zee, Carol Zheng, Dillon Mills, Caleb Trujillo, my roommate Boomer (and his pet, Peter Perreault), Brian McManus, the IO side of the Psychology Department, and scores others (some of whom have already been thanked for other things). For administrative assists, I would like to thank Elsa Schirmer.

And, for their love and support, I would like to thank Daryl and Laurie in their dual roles of doting parents and CEOs of the first national bank of mom and dad.

For the idea to start the PhD program in the first place, I owe a considerable debt of gratitude to Dr. Lee Etta Powell.

The greatest thanks goes to Ronnie Wilbur, a strong intellect, adviser, advocate, and friend, who has seen me through eight or more years, from the first day of my master's to well beyond when I should have graduated with my PhD.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Bradley, Chuck Ph.D, Purdue University, May 2019. Transparency of Transitivity in Pantomime, Sign Language. Major Professor: Ronnie B. Wilbur.

This dissertation investigates whether transitivity distinctions are manifest in the phonetics of linguistic and paralinguistic manual actions, namely lexical verbs and classifier constructions in American Sign Language (ASL) and gestures produced by hearing non-signers without speech (i.e., pantomime). A positive result would indicate that grammatical features are (a) transparent and (b) may thus arise from non-linguistic sources, here the visual-praxic domain. Given previous literature, we predict that transitivity is transparent in pantomime and classifier constructions, but opaque in lexical verbs.

We first collected judgments from hearing non-signers who classed pantomimes, classifier constructions, and ASL lexical verbs as *unergative*, *unaccusative*, *transitive*, or *ditransitive*. We found that non-signers consistently judged items across all three stimulus types, suggesting that there is transitivity-related information in the signed signal.

We then asked whether non-signers' judging ability has its roots in a top-down or bottom-up strategy. A top-down strategy might entail guessing the meaning of the sign or pantomime and then using the guessed meaning to assess/guess its transitivity. A bottom-up strategy entails using one or more meaningful phonetic features available in the formation of the signal to judge an item. We predicted that both strategies would be available in classing pantomimes and classifier constructions, but that transitivity information would only be available top-down in lexical verbs, given that the former are argued to be more imagistic generally than lexical verbs. Further, each strategy makes a different prediction with respect to the internal representation

of signs and pantomimes. The top-down strategy would suggest signs and pantomimes are unanalyzable wholes, whereas the bottom-up strategy would suggest the same are compositional.

For the top-down analysis, we correlated lexical iconicity score and a measure of the degree to which non-signers 'agreed' on the transitivity of an item. We found that lexical iconicity only weakly predicts non-signer judgments of transitivity, on average explaining 10-20% of the variance for each stimulus class. However, we note that this is the only strategy available for lexical verbs.

For the bottom-up analysis, we annotate our stimuli for phonetic and phonological features known to be relevant to transitivity and/ or event semantics in sign languages. We then apply a text classification model to try to predict transitivity from these features. As expected, our classifiers achieved stably high accuracy for pantomimes and classifier constructions, but only chance accuracy for lexical verbs.

Taken together, the top-down and bottom-up analyses were able to predict non-signer transitivity judgments for the pantomimes and classifier constructions, with the bottom-up analysis providing a stronger, more convincing result. For lexical verbs, only the top-down analysis was relevant and it performed weakly, providing little explanatory power. When interpreting these results, we look to the semantics of the stimuli to explain the observed differences between classes: pantomimes and classifier constructions both encode events of motion and manipulation (by human hands), the transitivity of which may be encoded using a limited set of strategies. By contrast, lexical verbs denote a multitude of event types, with properties of those events (and not necessarily their transitivity) being preferentially encoded compared to the encoding of transitivity. That is, the resolution of transitivity is a much more difficult problem when looking at lexical verbs.

This dissertation contributes to the growing body of literature that appreciates how linguistic and paralinguistic forms may be both (para)linguistic and iconic at the same time. It further helps disentangle at least two different types of iconicities (lexical vs. structural), which may be selectively active in some signs or constructions but not others. We also argue from our results that pantomimes are not holistic units, but instead combine elements of form and meaning in an *analogous* way to classifier constructions. Finally, this work also contributes to the discussion of how Language could have evolved in the species from a gesture-first perspective: The ‘understanding’ of others’ object-directed (i.e. transitive) manual actions becomes communicative.

1. INTRODUCTION

1.1 Scope of study

What underlies our ability to understand non-manual actions? Those of us old enough will remember how to signal rolling down a car window. We'll know how to order a shot in a noisy bar, or wave someone through a four-way stop. We know to trace our thumbs across our necks at people who are thinking about doing something we'd very much like them not to. You might remember how to mock shoot someone and then put your 'gun' in its holster. Outside of these stereotyped situations, we may also have the intuition that you cannot possibly want someone to pass you the ball if you gesture using an extending ring finger, peace-sign or thumbs-up.

Iconicity, or a motivated link between how a form (linguistic or not) looks and what we intend to communicate, may play an important role in understanding these phenomena. Gesturers recruit certain handshapes and movements, but not others, to represent their intentions, which–in the case of the *rolling-down-window* gesture–may become codified in a particular culture, with the form-meaning correspondence becoming completely uncoupled over time (e.g., we have power windows now).

In some theories of Language evolution or emergence (e.g., Arbib, 2005), iconicity bootstraps an eventual linguistic system. However, iconicity is free to remain in (e.g., Lepic & Padden, 2017) and constrain (e.g. Meir, Padden, Aronoff, & Sandler, 2013) language, be processed as such (e.g., Thompson, 2011; Thompson et al., 2009), accessed subconsciously, or completely ignored (Emmorey et al., 2004). Through conventionalization and then perhaps lexicalization or grammaticalization, iconicity may then be completely erased (Napoli, 2017). Of course, this is not to say that iconicity is affected the same way across the linguistic system (including the lexicon, morphology, syntax, etc.; e.g., Lepic & Padden, 2017). This project takes the point of view that a language is free to exploit non-linguistic visual and temporal resources for grammatical purposes, from the linear ordering of constituents, and lengthening and reduplication processes in spoken languages, to the spatial arrangement of event participants and the co-opting of object affordances for argument realization in sign languages. It is this last example that is in focus here. The empirical questions we hope to contribute to are:

1. *Whether and to what degree argument structure can be linguistic and iconic at the same time*, and
2. *Could the understanding of others' actions and object affordances have led to a vision-based argument structure in the nascence of Language?*

The idea sprouts from neurolinguistic work on the association of language with action, particularly praxic (manual) action. For instance, Arbib (2005, 2012) argues that the origins of Language lie in the manual modality: Among other innovations in the species, human beings learned to gainfully use 'action comprehension' for communicative purposes, with object affordances helping to identify intended referents. That is, the shape of the hand tells the language-ready proto-signer that it's a fruit, and not a stick, that's in focus; the referent object is 'read off' of the shape of the hand.

One way to explore this connection is through the inferences hearing non-signers make on (arguably) paralinguistic communicative strategies, like pantomime (a subset of gesture; McNeill, 2005).[1] To bridge to language, we might additionally explore non-signer inferences on the argument structure of classifier constructions, a subset of highly iconic constructions in sign languages (Supalla, 1983).

These constructions, while imagistic (e.g., Liddell, 2003; Cogill-Koez, 2000) and not universally accepted as true classifiers (on par with those found in spoken language; e.g., Schembri, 2003), are nevertheless linguistically structured, with argument

[1] *Pantomime* is also called *silent gesture* (and *pantomimers*, *silent gesturers*), especially in the field of sign language x gesture studies (e.g., Goldin-Meadow, 2015).

structure encoded by handshape and handpart morphemes (Benedicto & Brentari, 2004; Benedicto et al., 2007; Gökgöz, 2013). Further still, we can probe non-signer inferences on the argument structure of lexical verbs in sign languages, which form a cline with respect to their lexical iconicity (e.g., Caselli et al., 2016), but may not be iconic with respect to their argument structure. We might expect the ability to accurately guess the argument structure of a form is predicated on that form's imagistic properties, with pantomimes and classifier constructions being more imagistic than lexical verbs, showing (synchronically) a transition from (a) the encoding of argument structure features in a motivated way to ensure communicative success (pantomimic stage; paralinguistic, iconic), to (b) the conventionalization of these iconic strategies into linguistic form (classifier construction stage; linguistic, iconic), to the lexicalization of argument structure and possible erasure of iconicity (lexical verb stage; linguistic, not iconic).

We address these questions in a series of experiments, each looking at transitivity marking on *verbs*, specifically. We first test whether non-signers perceive transitivity distinctions in everyday actions and encode them into language. We then test whether non-signers have intuitions on how pantomimes should be performed. Finally, we test whether non-signers have intuitions about the argument structure of pantomimes, classifier constructions from American Sign Language (ASL), and ASL lexical verbs. We also assess the degree to which knowledge of a sign's meaning guides the analysis of its parts, or whether analysis proceeds bottom-up from the assemblage of phonetic and phonological features.

1.2 Research Questions, predictions

For the present study, we asked the following questions:

1. Are non-signers consistent in their transitivity classing? If yes, we have evidence to suggest that there is something in the signal that guides this classing behavior.

2. Are non-signers accurate in their transitivity scoring with respect to the actual transitivity of the ASL lexical verbs and classifier constructions, and the 'ground truth' transitivity of the pantomimes? If yes, we have evidence to suggest that production and perception models of transitivity are similar and make similar use of iconicity.

3. What effect does the class of stimulus–lexical verb, classifier construction, or pantomime–have on questions (1) and (2)? We predict that lexical verbs will be classed (a) less consistently and (b) less accurately than classifier constructions and pantomimes given that there is no attested, consistent object marking strategy in lexical signs (though see §2.2.2). Further, should there be one, it may be opaque on par with verbs in spoken languages, such that on the basis of visual analysis, non-signers should not be able to detect it. The object marking strategy in classifier constructions, by contrast, is more consistent and may be transparent. And, while the actual realization of object-marking may be variable in pantomimes, the strategy (e.g., using object affordances) is consistent and transparent.

4. Could any differences between stimulus classes observed in question (3) be attributable to lexical iconicity, or, how easily an item's lexical meaning could be guessed from its form? While we predict that lexical verbs will be generally rated as being less iconic than classifier constructions and pantomimes, this may not explain the differences found in question (3). Instead, we predict that transitivity information is only available in lexical verbs via access to its lexical meaning.

5. Alternatively, could any differences between stimulus classes observed in question (3) be attributable to the phonetic make-up of these items? We predict that the phonetic form of the relatively more iconic stimuli, pantomimes and classifier constructions, may inform transitivity coding and perception. However, lexical items, having integrated into the linguistic system and having become

more opaque with respect to their lexical iconicity over time, may be or have become opaque with respect to their argument structure as well.

1.3 Organization of the dissertation

This dissertation is organized as follows: We first provide a basic account of how argument structure is expressed in sign languages (§2.2.2) and how subject-object relations may be marked in the gestures produced by hearing non-signers (§13). We provide definitions for these and related phenomena as we go.

Next, we set the stage for our discussion by reviewing how certain grammatical features have their roots in iconicity. We illustrate first with some data from spoken languages (§2.3.1), while conceding that the oral modality does not generally lend itself well to iconic representations, especially concerning visual information. We then turn our attention to a case study, regarding the iconic underpinnings of the formal feature, telicity, in sign languages (§2.3.2). We then discuss a few studies that examined the emergence of formal or formal-looking properties in new sign languages and in gestures created *de novo* by hearing non-signers (§2.4). We then pose the question, *what types of iconicity are there, which type(s) is or are responsible for encoding of grammatical features in sign languages, and whether this iconicity is holistic or componentially structured?* (§2.5). Our proposal, the iconic structuring of argument structure in sign and pantomime, closes out Chapter 2 (§2.7).

Our contribution begins with an analysis of classifier constructions and pantomimes in Chapter 3. We discuss first how we elicited both (§3.2), before describing our two main experiments (§3.3, 3.4). We provide a top-down and bottom-up analysis of the results in the following two sections (§3.5.1 and 3.5.2, respectively), and weigh the strength of the results of both analyses against each other (§3.6). We then walk through the same analysis, only using ASL lexical verbs, in Chapter 4.

We synthesize the results obtained from all analyses in Chapter 5, and discuss them in light of our research questions. A few suggestions for future studies are also offered (§5.3), before finally concluding this dissertation in Chapter 5.4.

More detailed information concerning the results of each study can be found in the Appendices. Such information includes survey material, raw survey results, discussion of different participant populations (where appropriate), and full results of the machine learning analyses. We also provide a few ancillary analyses that help us clarify and support our discussion.

2. BACKGROUND

2.1 Brief primer on lexical verbs, classifier constructions, and pantomime:

More information regarding the nature of, and differences/ similarities between pantomime, classifier constructions, and lexical verbs will be elucidated as we go. However, we provide a few very quick, cursory notes on each system to get us started.

According to Brentari and Padden (2001) the lexicon of ASL can be divided into two parts: the core lexicon ('lexical verbs') and spatial lexicon ('classifier constructions'), ignoring non-native vocabulary (e.g., fingerspelled loan signs, etc.). Lexical verbs are monomorphemic (depending on how you view compounds), obey phonological constraints (e.g., Battison, 1978; Brentari, 1998), convey relational–as opposed to spatial–information (e.g., Bradley, 2013), have more or less fixed phonological forms, and can (mostly) be described using the same linguistic theories built around spoken languages. Or, roughly hewn, lexical verbs are those that could be listed in a dictionary.

Classifier constructions, however, are polymorphemic, consisting minimally of a handshape morpheme (which by some analyses may itself be decomposed), a movement root, and location features (Supalla, 1986). The handshape morpheme restricts (or classifies) a coreferential noun.[1] Handshape morphemes are deformable to indicate, e.g., the size of referents (Emmorey & Herzig, 2003), but are otherwise generally fixed in form. Further, individual sign languages will have their own inventory of

[1]For example, the 3-handshape co-occurs with vehicle referents (but not human referents) while the reverse is true of the 1-handshape. A given referent may be referred to using different classifier handshapes, depending on what is in focus. For instance, human (and other bipedal) referents may also be referred to using an inverted V-handshape (symbolizing legs). To that end, a given classifier handshape is not uniquely paired to a particular semantic class (such as *animate*, *human*, etc.) and can be coreferent with entities that share certain visual characteristics. The 1-handshape, e.g., not only occurs with human referents, but any referent whose dominant dimensions are long, skinny, and/ or vertically oriented.

classifier handshapes (Brentari & Eccarius, 2010). The movement root and location features, by contrast, have variable exponence and do not vary greatly across sign languages (Schembri et al., 2005).

Pantomimes are a unique subset of non-signer manual behavior. They occur in the absence of speech and are therefore bear the full brunt of the communicative load (McNeill, 1992, 2005). Given that there is no attested, rigid system governing their use, pantomime production is highly variable. Nevertheless, pantomimes are noted to share some (para-)linguistic qualities, some of which we explore in more detail later.

2.2 Argument structure

2.2.1 Brief primer on argument structure

Canonically, in an utterance the verb names the event and specifies the event participants (Levin & Hovav, 2005). That is, it is structured. Event participants are referred to as the verb's arguments, which may be subjects, objects, indirect objects, or obliques. In which case the verb only specifies a subject, the verb is said to be intransitive (1a). If the verb requires a subject and an object, it is transitive (1b). And, if a verb requires a subject, object and indirect object, its ditransitive (1c). Many verbs are variable in this regard. For instance, *read* optionally takes an object: *I read a book* and *I read all morning* are both acceptable. Further, the verb will assign meaning, encapsulated in theta-roles, to its arguments. For instance, while *run* assigns an agentive interpretation to its subject (e.g., *Jake ran on purpose*), a similarly intransitive verb *die* assigns an undergoer (or theme) interpretation (e.g., *?Jake died on purpose*).

Argument structure not only refers to how many arguments a verb may take and what interpretations they receive, but also to the syntactic categories of arguments a verb may take (its *subcategorization*). For instance, the verb *know* may take a nominal or sentential complements (e.g., *I know Paul* vs. *I know Paul smokes*). Further, argument structure can be expanded to include larger frames involving the types of adjuncts licensed and their effects on the meaning of the predicate (Goldberg, 1995). For instance, the sense of *kick* changes between the sentences *Susan kicked the ball into the goal* and *Susan kicked Paul into submission*. At present we are chiefly focused on the first sense of *argument structure*, namely how many arguments a verb may take. As such, *argument structure* and *transitivity* will be used interchangeably throughout.

(1) a. I slept (*the bed)
    b. I have *(a brother)
    c. I gave *(myself) *(a headache)

2.2.2 Argument structure in the visual modality

Research into the argument structure of sign languages is in its infancy, with a few notable exceptions. Early discussions of argument structure in ASL comes from Fischer and Gough (1978) and Kegl (1990), who give a broad overview of different argument relations and verb types (e.g., *transitives*, *unergatives*, *psych verbs*, etc.), as well as argument structure alternations and some selectional restrictions. Interest in the topic has been intermittent (e.g., Benedicto & Brentari, 2004; Benedicto et al., 2007; Grose et al., 2007; Kimmelman, 2016, 2018, *inter alia.*)

However, argument structure has been examined obliquely, especially in the discussion of *agreement* and agreement verbs, wherein one class of (transitive) verbs agrees with a subject and an (indirect) object and one class of (intransitive) verbs agree with one or more locations (i.e., the source and goal of the movement). We explore this in the following discussion (§2.2.2). Here, we also discuss other coding strategies relevant to argument structure, including handshape, the role of the non-dominant hand, certain non-manuals, and viewpoint perspective.

With respect to the argument structure of pantomime, very little has been done, with most work concentrating on handshape production and viewpoint perspective. However, there has been some work on the perceptual categorization of handshape.

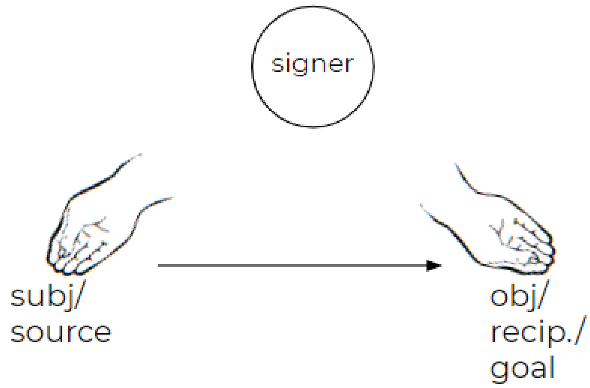


Figure 2.1. Diagram of an agreement verb. The verb starts at the locus assigned to the subject or source and moves towards the locus of the (indirect) object, goal, etc. In many cases, it is the choice of handshape that distinguishes between different directional verbs. Here, the handshape is consistent with GIVE, but an S-handshape would be consistent with HIT. To note, not all directional verbs use path movement. Verbs like HATE instead use the facing of the palm.

We review what has been done in the section after next. To note, we ignore transitivity via word order, as NP V sequences vs. NP V NP sequences are by and large intransitive and transitive, respectively (excluding verb chains or pragmatic transitivity). And, while it still may need to be resolved whether the second NP in an NP V NP string is an argument, oblique or adjunct, we discuss other phenomena in sign language that encounter this same problem.

Lexical Verbs & Classifier Constructions

The two most consistent (and well-studied) markers of argument structure in sign languages are directionality and handshape (e.g., Gökgöz, 2013; Börstell, 2017), though there are considerable nuances to these (as we will explore in some detail below). The former strategy, Börstell (2017) argues, generally encodes Patient and Recipient roles, while the latter strategy encodes Theme and Patient roles. Throughout, we make note of if/ how these cues are iconic.

Directionality: Directionality refers to a property shared by some verbs in sign languages, called *agreement verbs*,[2] that are articulated such that their movement begins at the locus of the source of the action (often, though not uniquely coinciding with the grammatical subject) and ends at the locus of the goal of the action (VP-internal element; i.e., object, indirect object, or VP-adjunct), depending on the semantics of the verb.

As Kimmelman (2018) notes, if a verb happens to be an agreement verb denoting transfer, it is likely because it is transitive. But this is a unidirectional observation as not all transitive signs denote transfer or are directions (e.g., the case of LOVE, which is transitive despite being body-anchored, and not denoting transfer). A further restriction is that both arguments of an agreement verb must be animate (at least in ASL; Rathmann & Mathur, 2011), as exemplified in 2. The variant in (a) is articulated with a short movement in the signing space in front of the signer's face. The variant in (b) leaves this location and moves towards the location of the person being taught.

(2) a. IX1 TEACH MATH[3]
‘I teach math’
b. IX1 $_1$TEACH$_2$ IX2
‘I teach you’

As mentioned parenthetically above, the loci to which agreement verbs move are not always assigned a unique function. In agreement verbs, again, the verb optionally originates at the locus associated with the subject of the sentence and (obligatorily)

[2]These verbs may also be referred to as *directional verbs*, *inflecting verbs*, *indicating verbs*, and so on, depending on your (a)theoretical starting point. While I use the term *agreement verb*, I am not making a claim on whether directionality is truly a case of agreement.

[3]By convention, signs are reported in all capital letters. Indexicals, which are visually similar to points but are formal elements, are indicated by IX. The numbers 1, 2, and 3 indicate whether the indexical refers to 1st, 2nd, or 3rd person referents, respectively. In some cases, these numbers may be in subscript, as in $_1$VERB$_2$. This indicates that the verb is articulated such that it moves from the location of the first subscripted number/ referent to the location of the second subscripted number/ referent.

terminates at locus of the object (3a) or oblique (indirect object 3b). Verbs of this type are transitive and ditransitive, respectively. In spatial verbs, the source and goal are marked as in the ditransitive PUT (3c). Here, the locus refers to the adjunct, not argument. However, intransitive verbs may also make use of points in space to indicate the source and goal of a motion event (3d). Here again, the loci are not interpreted as subject and object. As Quer (2011, 2017) points out, the interpretation of a locus in space is ambiguous, and a given point may take on more than one role at a time (e.g., a location y and a referent located at location y). Here, except for the number of overt NPs, you cannot infer argument structure, specifically *transitivity*, from the directionality of the verb.

(3) a. IX1 $_a$PITY$_b$ IX2

'I pity you' [transitive]

b. $\overline{BOOK}^{\text{br}}$ JOHN$_a$ $_a$GIVE$_b$ MARY$_b$

'John gave Mary the book' [ditransitive (dative)]

c. $\overline{BOX_a}^{\text{br}}$ BOOK PUT$_a$

'Put the book in the box' [ditransitive (locative)]

d. IX1 ARRIVE$_a$ SCHOOL$_a$

'I arrived at school' [intransitive]

In addition to directionality, some verbs may also be specified for *facing* (Meir, 1998), or, the direction the palm of the hand is oriented towards. Unlike directionality, facing seems to target syntactic subjects and objects, specifically: the palm is oriented towards the syntactic object, while the back of the hand is oriented (as always optionally) towards the subject. In some cases, directionality and facing are coupled (e.g., ASK, where *source* and *subject*, and *goal* and *object* are aligned) and sometimes not. Verbs that do not denote transfer may not be directional but still exhibit facing (e.g., HATE, where there is no path movement). Verbs that denote transfer may not also exhibit facing (e.g., GIVE, where the palm faces upwards). In the case of 'backwards verbs,' the verb's semantics call for the verb to be articulated source-to-goal, though the source may be the syntactic object and goal syntactic subject (e.g., STEAL: I [*goal*/ subj.] steal from you [*source*/ obj.]). If present in a sign, facing, then, may be a more reliable of a transitivity cue than directionality.

One final note concerns the nature and consequent prevalence of agreement in sign languages. Since Padden's (1988) original tripartite verb class system (in modern terms: agreeing, plain, and spatial), there have been several attempts at clarifying or reorganizing it. Some have argued for collapsing agreeing and spatial verbs into one category to the exclusion of plain verbs (e.g. de Quadros & Quer, 2008), but one recent proposal argues for agreement to be underlyingly more pervasive than once thought: Lourenço (2018) argues that location (and not movement, or movement + location) is the sole exponence of agreement (which thus also explains so-called Single Argument Agreement; Meir, 1998; Zwitserlood & Gijn, 2006; Costello, 2016), placing agreement, spatial and non-body-anchored plain verbs in a single category. This explanation, we argue, further uncouples argument structure from the agreement system and further highlights the difficulties of resolving transitivity information from agreement information.[4]

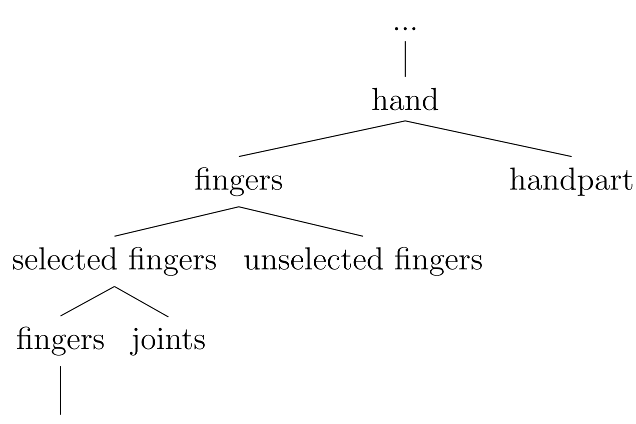
As a non-signer, then, one must resolve whether facing is relevant (e.g., it is in HATE, but not in TEACH) and whether the movement of a verb is transitional (in the case of Single Argument Agreement) or not. And, if not, whether the movement is directed towards an object, a recipient, or a location. Other cues may be informative. For instance, we might predict that signs like HIT will give rise to a transitive parse, it being two handed (with each hand having a different handshape). We might expect verbs like GIVE to be interpreted as directed towards recipients, as the handshape again potentially yields an interpretation of holding something– the patient/ theme is already accounted for. Finally, for the spatial parse, a handshape that's incompatible

[4]That is, though the agreement process may be iconic, the mapping from agreement to transitivity may not be.

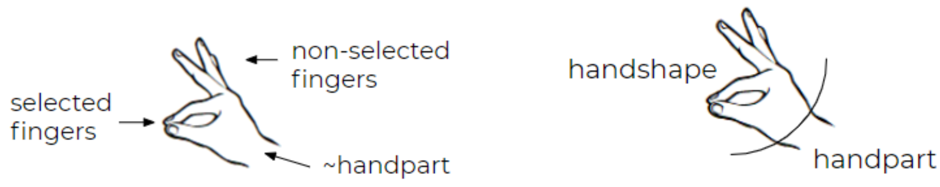
with a handling interpretation might suffice. To that end, we discuss the role of handshape in argument encoding next.

Handshape Handshape, quite naturally, refers to the shape of the hand in both lexical signs and classifier constructions. According to Brentari's (1998) model of sign language phonology, *handshape* is decomposable into two components: *fingers*, which is itself further decomposable, and *handpart* (see feature geometry tree in 4). *Fingers* is comprised of *selected* and *unselected fingers*. Historically, the field refers to particular configurations of selected and unselected fingers (i.e., the *finger* node) as *handshapes* without also invoking handpart, a convention we respect here. For example, for the F-handshape (see Fig. 2.2), the index finger and thumb are the selected finger and the middle-, ring- and pinky-fingers are all unselected. Though it is referred to as the F-handshape, handpart is not generally not included. So, going forward, I will refer to selected- and unselected-fingers as *handshape* and handpart as *handpart*.

(4) Feature geometry of hand node



For lexical verbs, both handshape and handpart features are fixed (i.e., they're lexical). However, for classifier constructions, handshape is morphological, changing to reflect perceptual or ontological properties of its referent NP (Benedicto & Brentari, 2004; hereafter B&B). Handling and whole entity classifier constructions

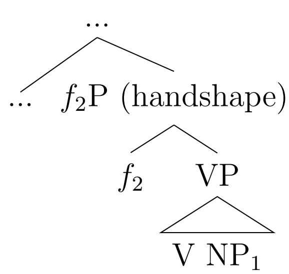


(a) (b)

Figure 2.2. Diagrams showing regions of the hand. For a more detailed diagram (pertaining to argument structure), see Benedicto and Brentari (2004).

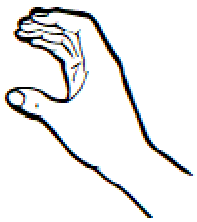
differ with respect to the morphological or phonological status of handpart. In whole entity classifier constructions, handpart is phonological, and in the handling classifier constructions, morphological. When morphological, handpart marks the presence of an agent, thus making handling classifiers transitive (cf. 5 and 6).[5,6]

(5) Structure for intransitive, whole-entity classifiers



[5]Two quick notes: (1) Handpart morphemes have variable exponence. (2) It may be functionally equivalent to posit either a handpart morpheme (which is defined with respect to the postures that the hand takes alone) or as the signer's entire body (e.g., the 'body as subject' analysis of Meir, Padden, Aronoff, & Sandler, 2007), which is consistent with descriptions of character viewpoint and classifier constructions (e.g., Perniss, 2007). See below.

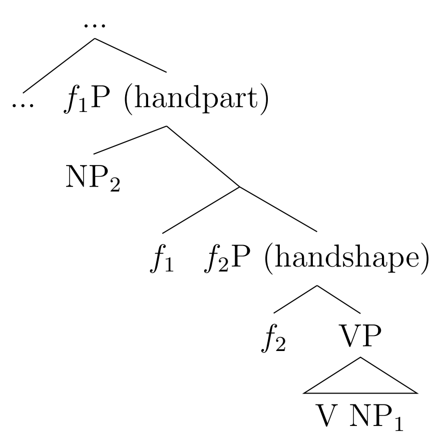
[6]Explicitly, in the trees, NP_1 refers to the internal argument of the verb and NP_2 refers to the external argument. The functional projection f_2P is responsible, in a way, for the determination of the selected fingers node based on the referent of NP_1. F_2P is comparable to little-v, assuming a VP-shell analysis, and corresponds with the handpart morpheme. Unlike the handshape morpheme, the handpart morpheme is not supplied by its NP (NP_2).



Sample signs	
CCs	CL-MOVE$_x$ (trans), CL-$_x$MOVE (intrans)
verbs	LOOK^FOR, GUESS, DRINK
nouns	PICTURE, CUP, CHARACTER
adjectives	STRANGE, HUNGRY*

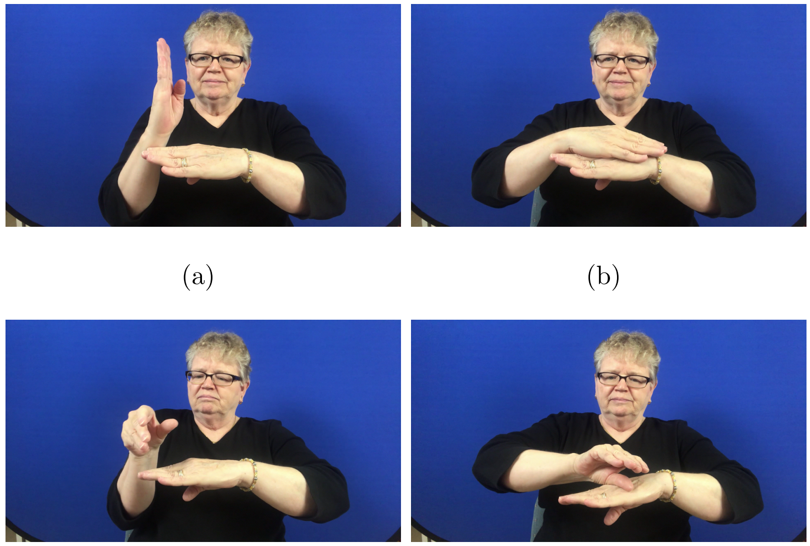
Figure 2.3. The C-handshape and a table of ASL signs where it occurs. There is not a universal mapping between handshape and transitivity or even lexical class. If transitivity distinctions are manifest in the form of a sign, other visual elements–we predict–must conspire. *May actually be a verb.

(6) Structure for transitive, handling classifiers



Benedicto and Brentari (2004) provide a battery of tests, calibrated on ASL lexical signs, that support the existence and characteristics of the handshape and handpart morphemes. Their tests for internal arguments include the distribution of a distributive morpheme, [+distr], which does not affix to external arguments, demonstrated in 7 for lexical verbs and 8 for classifier constructions.[7]

[7]Throughout this dissertation, we report the examples from other sources as given, and do not attempt to 'translate' them into our glossing conventions.



(a) (b) (c) (d)

Figure 2.4. Handshape contrasts in transitive and intransitive classifier constructions. (a,b) are two stills from the intransitive classifier construction that corresponds to 8a, without the distributive morpheme. (c,d) are two stills from the transitive classifier construction that corresponds to 8b, again, without the distributive morpheme.

(7) a. ICE-CREAM MELT$_{[distr]}$

'Each of those ice creams melted' [B&B (ibid.), 15b]

b. *WOMAN LAUGH$_{[distr]}$

Intended: 'Each woman laughed' [B&B (ibid.), 16b]

c. GIRL SHIRT BUY$_{[distr]}$

'The girl bought each one of the shirts'

'Each of the girls bought a shirt' [B&B (ibid.), 17a,b]

(8) a. BOOK w/e-CL:MOVE$_{[distr]}$

'Each of the books feel on its side' [B&B (ibid.), 38a]

b. BOOK handl-CL:MOVE$_{[distr]}$

'S/he put down each book (on its side)'

Each of them put down the book (on its side)' [B&B (ibid.), 38a]

The analysis is really neat and tidy, and explains gross differences between intransitive and transitive classifier constructions. However, there are numerous complications once other data are considered, most of which are related to the polysemy of the classifier handshape itself, or the exponence of the handpart morpheme.

As an example (based on Zwitserlood, 2003, 128), if we look at the C-handshape, we can see that the handshape morpheme is common among intransitive (9a) and transitive classifier constructions (9b). What distinguishes (a) from (b) is the movement root. (a) is articulated with the BE-AT ('be located at') root, which is articulated as a short, terse movement downwards and occurs only in intransitive contexts. (b), by contrast, is articulated with a MOVE root, which has variable exponence and occurs in both intransitive (e.g., *something moves*) and transitive contexts (e.g., *someone moves something*).[8]

(9) a. CUP C-BE-AT$_a$
'A/ the cup is there'
b. CUP C-MOVE$_a$
'Move the cup (there)'

Another example is the bent-V handshape, which may be used intransitively to represent the movement of an animal (10a) or the dropping of one's jaw in awe (10b).[9]

(10) a. CAT V-MOVE$_a$
'A/ the cat moved (there)'
b. IX1 V-JAW-DROP
'My jaw dropped (in awe)'

[8](b) is also articulated with a short, terse movement that codes imperatives in ASL, thus unambiguously denoting a transitive action. There may be other modulations of the MOVE root that distinguish between its transitive and intransitive uses.

[9]While Benedicto & Brentari argue that bodypart classifiers (BPCLs) like the one in (10b) are intransitive, Grose et al. (2007) argue that they are in fact transitive, with the internal argument being the body part of the external argument. We assume following Grose et al. that BPCLs are transitive.

The addition of the handpart morpheme, which has its exponence in the orientation of the hand with respect to some plane, is what separates the *a, b* examples above.[10] Although the handpart morpheme in (9a) is homophonous with the handpart specification of (9b), it need not be. One can sign in such a way as to convey manually inverting a glass, but one cannot sign that a glass is located somewhere, but upside down. Similarly, the classifier handshape coreferent with CAT in (10a) has the same handpart restrictions as the classifier handshape in (9a), while the two hands involved in JAW-DROP are even oriented differently from each other.

Despite the fact that the C-handshape and V-handshapes discussed here cannot be oriented differently when used in intransitive classifier constructions, there seem to be other cases where handpart specification is not so rigid. Benedicto & Brentari claim that in intransitive classifier constructions, when co-referential with a human referent, the base of 1-handshape may only be articulated perpendicular to the horizontal plane. However, there is no problem with the 1-handshape being articulated along either the horizontal or vertical planes if it is co-referential with a pen. Further still, *water dripping* is also articulated with the 1-handshape, but the fingertip is now oriented down. It would seem that the restrictions on handpart are particular to specific classifier-referent pairs, and not the classifier handshape itself (specifically, its [under-]specification for handpart).

Further, others have noted that–for other SLs, too–intransitive classifier constructions are possible in transitive contexts and *vice versa*. For instance, Kimmelman, Pfau, and Aboh (2016) demonstrates that entity (intransitive) and handling (transitive) handshape occur in both transitive and intransitive contexts for Russian Sign Language. Simper-Allen (2016) (via Börstell, 2017) shows the same for Swedish Sign Language. More recently, He and Tang (2018) demonstrated that entity classifier constructions are possible in imperative contexts in Tianjin Sign Language, demon-

[10] To be more precise, lexical signs and intransitive classifier constructions do have a specification for handpart, except that it is phonological in this case. That is, a change in handpart for these signs results in an illicit form or an unrelated concept. As an example using lexical signs, the signs TRAIN and SHORT minimally contrast in their handpart specifications, where TRAIN is articulated with the palms facing downwards and SHORT is articulated with the palms facing each other.

strating their transitive use. In our own data, too, our signer produces a transitive handshape in an intransitive context, but is corrected by another signer. In her first production, she produces a flat-C-handshape in response to the stimulus *The book fell on its side*, which is consistent with a handling (transitive) strategy. She then produces the intransitive handshape variant, the B-handshape. Finally, to note, ASL may use the signs DOOR and WINDOW, which may have their roots in entity classifier constructions, transitively or intransitively.

(11) a. MAN DOOR OPEN
'The man opened the door'
b. DOOR OPEN
'Open the door!'
c. DOOR OPEN
'The door opened'

For non-signers, then, handshape may not be a strictly reliable cue. Even in best case scenarios, the same handshape may be used in transitive and intransitive contexts, varying only–perhaps–by the form of the movement root or exponence of the hand-part morpheme. That is, the manipulation of all of these variables seems important to arrive at a consistent, accurate transitivity parse. However, there are some perceptual biases non-signers (or human beings in general) may have that will help them settle on a (correct) parse. We contend, though, that these biases must have probabilistic outcomes at first, given that the transitivity-coding system in established sign languages is not yet completely understood.

Nevertheless, as we go on, we will continue to use *transitive*, *handling (strategy)*, and *transitive handshape*) interchangeably, and so too for *intransitive*, *(whole) entity (strategy)*, and *intransitive handshape*. Given how robust these associations are (see next section), we argue that the presence of handling handshapes in intransitive contexts and *vice versa* to be exceptions, not the rule, and offer (without proof) that ultimately handshape selection is due to a number of constraints competing against each other.

Finally, though, we report that there are regularities in classifier handshape choice and that these have been revealed experimentally. As part of a larger study on handshape use in Nicaraguan signers (on which more below), Goldin-Meadow, Brentari, Coppola, Horton, and Senghas (2015) additionally had native ASL signers sign responses to vignettes that featured actions with and without an agent. The video clips contained one of several small, familiar items occurring in events of location or falling (agentless) and events of placement (agent). The authors report that ASL signers use object handshapes in agentless contexts a a vast majority of the time. While these participants used handling handshapes more than entity handshapes in clips with an agent, an almost equal number of entity handshapes were used. As such, the double-dissociation doesn't quite hold for ASL, at least in this study, but a difference in the coding of transitive and intransitive is nevertheless manifest.

Handshape complexity Handshape complexity (Eccarius, 2008; Brentari & Eccarius, 2010) is defined with respect to (a) the number of nodes in (4) needed to describe the handshape, (b) the pervasiveness of the handshape cross-linguistically (where simpler handshapes are more pervasive), and (c) the age at which children acquire them (simpler handshapes are learned earlier). Handshape complexity can be decomposed into two different types of complexity: joint complexity (how many and joints are active) and finger complexity (how many and which fingers are selected). The measures dissociate, such that a handshape with high finger complexity does not necessarily have high joint complexity. Some examples are given in Fig. 2.5.

Eccarius (ibid.) demonstrates that whole entity and handling classifier handshapes differ with respect to finger- and joint-complexity, with entity handshapes on average having higher finger complexity and handling handshapes on average having higher joint complexity. It is not the case, though, that *every* entity and handling handshape will fit this rule, but the contrast is nevertheless robust when considering entity

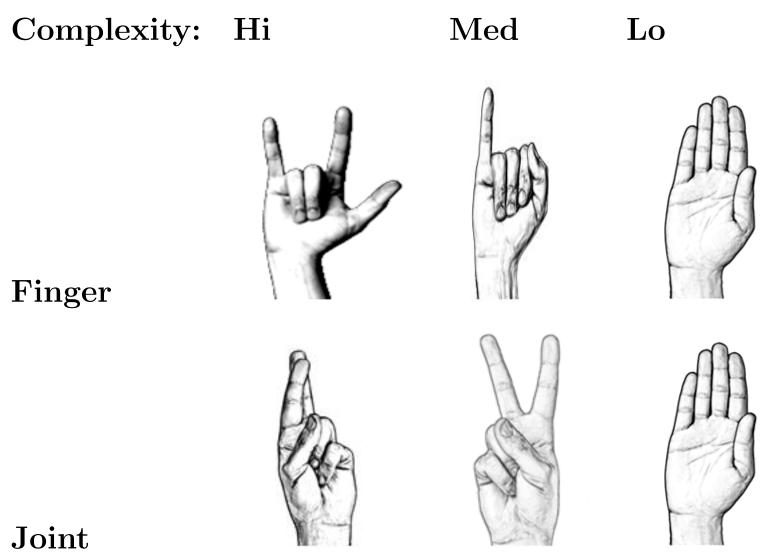


Figure 2.5. Example handshapes exhibiting high, medium, and low joint complexity. Images retrieved from `lifeprint.com`

classifiers as a group and handling handshapes as a group. Her findings for ASL, Hong Kong Sign Language (HKSL) and Swiss German Sign Language (DSGS) have then been experimentally demonstrated for ASL, Italian Sign Language (LIS), Chinese Sign Language (CSL) and Nicaraguan Sign Language (NSL, or sometimes ISN).

In two experiments, Brentari et al. (2012, 2017) probe whether signers make transitivity distinctions (i.e., distinctions between entity and handling handshapes) using handshape complexity (both studies) and joint complexity (only the latter study). The authors had signers from each country produce classifier constructions in response to short video clips depicting the placement (transitive), or location or movement (intransitive) of 11 different items. The items were chosen to elicit a number of different handshapes, both for handling and entity strategies.

Results from both experiments showed that signers consistently manipulated finger complexity in response to vignette type: higher finger complexity for intransitive events, and lower for transitive. Results from the second study replicated the findings for finger complexity, but also reported that joint complexity was, by contrast, higher in handling handshapes over entity handshapes. There were very few between-language effects, suggesting to us the effects of iconicity on transitivity distinctions.[11] Finally, we'll mention briefly here that the use of both joint and finger complexity to code transitivity distinctions is redundant (cf. the exclusive use of entity handshapes in intransitive events, yet the non-exclusive use of handling handshapes in transitive events). The information can be coded in just one of these measures. We discuss these last two points later when we discuss what non-signers and new sign languages do.

Scale, perspective taking, and two-handed productions: Two nuances conditioning where 'transitive' and 'intransitive' handshapes may appear are the scale of the event (by which we mean *perspective taking* and *agent focusing*) and the manipulability of the internal argument.

[11] It is also conceivable that modality effects are at work, such that there's something about using a sign language that would naturally give rise to this divide.

(a)

(b)

(c)

Figure 2.6. Stills from the classifier construction elicitation task that demonstrate observer (a), character (b), and mixed observer/character (c) viewpoint. Images of the referents were added to the stills to illustrate scale.

With respect to scaling, Perniss (2007) (see also Cormier et al., 2012) observes that transitive classifier constructions occur in scenes in which we are to understand that referents are life size; what she calls *character perspective*. That is, in BOOK C-MOVE ('s/she moved the book'), the C-handshape takes on the real-life size characteristics of the referent book and the agent is represented by the signer's body. We observe the action from within the scene. By contrast, intransitive constructions occur in scenes where we are understood to be watching events unfold in miniature; what Perniss calls *observer perspective*. For instance, in w/e-CL:GO ('someone went [like this]'), the 1-handshape represents a person, understood to be many times larger than the finger. Here, we observe the movements of the finger outside of the scene.[12]

Three examples are illustrated in Fig. 2.6. In (a), the signer depicts the events *The toy car passed the block tower*. The referents *toy car* and *block tower* are not life-size. (In this case, the hands are larger than the referents, but referents can just as easily be many times larger than the hands.) In (b), the signer depicts the transitive event *Someone broke a stick*. The referent *stick*, by contrast, *is* life-size; character viewpoint is observed. In (c), the signer depicts the event *Someone hit the bottle with a ball*. The bottle is represented from an observer viewpoint, while the ball is from a character viewpoint. The entire construction is transitive, despite the presence of observer viewpoint. We talk about this example a bit more below.

However, as referents get larger and, thus, less manipulable, the manipulation of handshape for transitivity coding becomes less and less viable. As Börstell (2017) notes, the referent HOUSE is more likely to be referred to using directionality than handshape. This is in spite of the fact that directionality is usually only available with human (or animate) referents. By contrast, even though babies are human, and thus candidates for directionality, Börstell claims that handshape may be used to mark them as objects due to their manipulability. Similar results were obtained by Ortega and Özyürek (2016), who observed that object manipulability and scale

[12]We lump the discussion of hand-as-object and hand-as-hand iconicity (used by Padden et al., 2013; Brentari, Renzo, Keane, & Volterra, 2015; *inter alia*) in with the discussion of perspective and scale.

often predicts gesturers' representational strategy: gesturers trace the outline of large, non-manipulable referents, but use handling-like handshapes for smaller, manipulable referents.

Agent-focusing or defocusing also plays a role in handshape selection, and thus transitivity coding. This has been experimentally verified by Rissman et al. (2016), who found that how much of the agent is present in a scene may condition the use of intransitive and transitive handshapes in NSL signers. For the same event, in cases where just the agent's hand was visible in a vignette, signers were more likely to produce intransitive classifier constructions, while they were more likely to produce transitive classifier constructions when more of the agent's body was visible.

However, the two perspectives can be mixed in at least two ways. In one, one hand articulates a sign or classifier construction embedded in one perspective, and the other hand is embedded within the other. As we began discussing above, our signer conveys the meaning *Someone knocked over the water bottle with a ball* (Fig. 2.6(c)). She does this by first establishing the water bottle with her left hand, using a 1-handshape whole entity classifier (observer perspective). She then uses a Claw-handshape handling classifier to show an agent hitting the water bottle using a ball. We see this type of mixed perspective not only in classifier constructions, but in certain lexical verbs in ASL. For instance, the verbs HIT, FLATTER, and ARREST are articulated such that the dominant hand interacts with (strikes, fans, and grabs, respectively) the non-dominant hand, in each case the 1-handshape (possibly derived from the homophonous classifier entity handshape used for human referents).

In the second way, both hands exist in both spaces at once. Kimmelman (2016), for instance, cites a case (12) where the signer produces a complex event: both manipulation (the cat's holding of the bird) and the cat's movement are conveyed simultaneously. Perhaps a more ordinary case is the addition of path to certain bodypart classifiers (BPCLs). BPCLs are demonstrably transitive (Grose et al., 2007) yet occur in motion events (e.g., BE-AT, MOVE [intrans.]), which may have been a motivating factor for their original classification as unergative predicates (Benedicto

& Brentari, 2004). For BPCLs, then, the rule is that transitive handshapes occur in observer perspective, in contrast to other types of classifier handshapes.

(12) CL(hl){canary}-MOVE.DOWN
'The cat falls while holding the canary.'

(13) IX3 bpcl-WALK$_a$
'S/he walks (there)'

The importance for the present studies is this: as part of our experiment, we elicit pantomimes and classifier constructions using a set of video clips. We want to capture transitivity marking (or demonstrate a lack thereof) as modulated by properties of the event itself, and not modulated by taking a particular perspective. As we describe in our pantomime/ classifier construction elicitation methods (§3.2.2), we force possible character and observer viewpoints into the same space. We additionally restrict ourselves to the use of objects that can be manipulated. Many of the objects are the same size relative to each other, too, in events with potentially more than one figure.[13]

Finally, we return to the point we made a moment ago about the hands' ability to occupy two different spaces. We note here that the asymmetry[14] of the hands is also a potential transitivity cue. Lepic, Börstell, Belsitzman, and Sandler (2016) demonstrate that certain concepts are encoded using two hands across Al-Sayyid Bedouin Sign Language (ABSL), ASL, Israeli Sign Language (ISL), and Swedish Sign Language (SSL). One such concept is *interaction*, such that the hands may move one against the other (e.g., HIT), together (e.g., ACCOMPANY), in opposition (e.g., ARGUE, MEET), away from each other (e.g., NEAR), and so on to show how two entities interact with each other. All of these signs are transitive. Other

[13] Objects denoting grounds were larger than figure objects.

[14] In some cases we truly mean 'asymmetry' in that the hands have different handshapes. In other cases, we mean that one hand moves and the other doesn't or that the hands more in opposition to each other (mirror symmetry). Yet in other cases, the hands have an identical handshape and movement, though seem clearly to refer to more than one entity. An all-encompassing term escapes us at the moment.

types of interactions can also be encoded, though, like EXIT (two-handed variant), JOIN, ENTER, ZOOM^OFF, and so on that show the relationship between a figure and a ground. These verbs are also ostensibly intransitive.[15] While a more detailed explanation and a corpus analysis are surely necessary, our cursory description of the distinction between transitive and intransitive interactional signs has to do with orientation and handshape. In at least the signs we've thought of (and listed above) transitive signs either have the same handshape, have opposing orientations or both. Intransitive signs, however, may have mismatching handshapes and the hands may have varied relative orientations.

Non-manuals: Eye-gaze and head tilt have been identified (Bahan, 1997; Neidle, Shepard-Kegl, MacLaughlin, Robert, & Bahan, 2000) as potential, though perhaps probabilistic[16] (Thompson, Emmorey, & Kluender, 2006, p. 587) markers of agreement, where the eyes typically focus on the locus of the object, and the head tilts towards the locus of the subject. Although, again, these non-manuals have been identified with respect to agreement, there is at least a partial overlap with argument structure.

Nevertheless, Thompson et al. (2006) show that eye gaze mostly co-occurs with agreement verbs, and less frequently with plain verbs (i.e., verbs that do not exhibit any kind of directional modification), with the gaze directed at the locus of the direct object. Interestingly, in cases where there is a choice to look at either the grammatical object or the semantic goal (they're separable in the case of 'backwards verbs'; see above), signers track the grammatical object.

The also note that gaze is directed at locative arguments in the case of transitive spatial verbs (e.g., PUT, MOVE [trans.]) much more so than at direct objects. For intransitive spatial verbs (e.g., STAND), gaze is predominantly directed at the subject. However, we might be curious as to gaze direction in other types of intransitive

[15]We make an argument for treating locative arguments in, e.g. EXIT ROOM, as adjuncts. Thus, verbs like EXIT are intransitive. That argument can be found in Appendix E.

[16]There has to our knowledge been no one theory that predicts with complete certainty when eye-gaze will be used.

events (unergatives and unaccusatives), and in other verbs types (e.g., ACT, which is intransitive and plain). For instance, would eye gaze be directed at surface subjects or underlying objects in the case of unaccusatives?

In any event, the results obtained by Thompson et al. (2006) corroborate the issues in transitivity marking via directionality/ agreement generally: the gaze may be directed at VP-internal constituents, direct objects, indirect objects or locations. As such, eye gaze is ambiguous with respect to argument structure, and does not distinguish between ditransitives, transitives, and locatives (intransitives). The additional complexity here is that in intransitives (and only in intransitives with a single location it seems) may eye gaze be directed at the subject. We'll mention here (instead of in the discussion of pantomimes to follow) that Thompson (2006) reports that non-signers look at the addressee while signing over 90% of the time. Further, appropriate eye gaze takes time for L2 learners of ASL to acquire, with novice signers in Thompson's study gazing at the addressee, direct object, and so on nearly equally. Even proficient L2 signers overgeneralize, e.g., looking at the locus of the object while signing plain verbs (native signers look at the addressee or elsewhere). As such, eye gaze does not appear to be an iconic strategy at first blush.

Pantomime

To be sure, gesture isn't widely considered to be a part of language proper, although many argue for its special relationship and interaction with language (e.g., Goldin-Meadow, So, Özyürek, & Mylander, 2008; Wilbur & Malaia, 2008; Özyürek et al., 2008; Kita & Özyürek, 2003). To ask, then, whether transitive and intransitive distinctions are made in pantomime might assume that there is argument structure in gesture. At the level of granularity at which we approach the main focus of this work, we neither argue for or against (linguistic) structuring of transitivity in gesture. However, we–and others–have observed that non-signers adopt different strategies when gesturing about object-directed and non-object-directed events. Some of these strategies bear at least superficial resemblance to those employed in sign languages (see below). Also note that we say *gesture* here instead of *pantomime*, as the studies we detail below report on either pantomime or co-speech gesture.[17]

We survey three different strategies used by non-signers to convey transitive and intransitive events: handshape, handshape complexity, and character and viewer perspective. We also describe a two studies on how non-signers perceive handshape with respect to transitive and intransitive events.

Handshape, Handshape complexity Brentari et al. (2015) aim to discover, among other things, whether the robust object-handling dichotomy observed in sign language classifier constructions is also present in the pantomimes of hearing non-signers, and–if so–whether there are cultural differences in the robustness of this contrast. The experiment proceeded as follows: 12 adults, six from the United States and six from Italy, were filmed pantomiming the movement, location (intransitive) and manipulation (transitive) of 11 everyday items (e.g., a lollipop, a book, a coin, etc.) Their productions were coded for the use of object handshapes and handling handshapes–though no clear definition of either was provided.[18] Productions were also coded for 'other' handshapes, which mostly consisted of points.

The particular scoring method employed here was probability of a match, where a 'match' is the use of an object handshape in an agentless context or a handling handshape in an agent context. Mismatches were the inverse cases *and* cases where participants produced a non-target handshape (i.e., 'other' cases). The found that Italian and American gesturers were not significantly likely to produce handling handshapes to describe vignettes including an agent. Conversely, gesturers were not significantly

[17] *Co-speech gesture* is a form of manual action that accompanies speech (McNeill, 1992). It is temporally tied with the speech it 'modifies' (in a loose, non-linguistic sense until proven otherwise) and often adds spatial information. Sometimes the information is redundant with what is spoken, but not necessarily. It does not bear the communicative load that pantomime does, as this is carried by the accompanying speech. See Kendon (2004).

[18] It may be evident from the stimulus objects themselves which strategy–object or handling–was employed.

likely to produce object handshapes when describing agentless vignettes. A portion of Fig. C in their appendix is shown in the table below.

	Matches	Mismatches	Other
USA	0.53,	0.14	0.33
IT	0.70	0.18	0.13

Despite not getting significant results–which depend on which statistical tests they run and how–we can still appreciate that a disproportionately large set of productions did match. However, we might argue that what counts should only be the *relevant* handshapes, i.e., just the object and handling handshapes. That is, instead of comparing 0.53 against (0.14 + 0.33 =) 0.47, we might just compare matches vs. true mismatches, whatever the new proportions would turn out to be. In the case where the 'other' production was a full body gesture (0.08 of 'other' productions for American gesturers, 0.18 of 'other' for Italian gesturers), an appropriate transitive or intransitive strategy (that may not rely on handshape distinctions) may be employed instead– the information is still there and is still consistent with the context. Again, the vast majority of 'other' gestures were points, which do not bear transitivity information one way or the other (until proven otherwise). While it may be a 'mismatch,' it is not untrue or inappropriate, strictly speaking. It should be noted, too, that other work from Brentari (Brentari et al., 2017) does exclude extraneous pointing gestures from analysis.

We would also argue that, given that gesturers vary considerably in their gesture production, it may not be appropriate to class the handshapes they use as either object or handling. Further, given that handshape complexity is used differently between signers and non-signers, it may be the case that what was identified as an object handshape by the researcher counts as a handling (or some other type of) handshape for the gesturer. This question is addressed in this thesis by obtaining gesture users' judgments without influence from expectations derived from work on sign language classifier constructions (see §3.2.2).

Switching now to handshape complexity, transitivity distinctions are also manifest in the handshape complexity measures, joint complexity and finger (group) complexity. In a series of experiments, Brentari and colleagues (Brentari et al., 2012, 2017) task to find out whether handshape contrasts with respect to finger and joint complexity are morphemic or phonological in sign languages and pantomime. In the first experiment, Brentari et al. (2012) examine finger complexity specifically, noting that this feature reliably distinguishes intransitive and transitive classifier constructions in sign languages (Eccarius, 2008; Brentari & Eccarius, 2010).

In the most recent iteration, Brentari et al. (2017) elicit pantomimes from hearing non-signers from the United States, China, Nicaragua and Italy, to see whether the reported differences in handshape complexity between whole entity and handling classifiers for sign languages would be manifest in their gestures. This adds a more universal flavor to their findings, should there be any patterns. Participants were again shown video clips of 11 different objects participating in transitive (events of *placing*) and intransitive (events of *being*, *moving*, or *falling*) contexts. Productions of the 'verbs' only were annotated for finger and joint complexity.

Mean joint complexity differentiated transitive from intransitive productions, with transitive productions being more complex. This was true for non-signers from each country. However, finger complexity *did not* distinguish transitive from intransitive productions among most non-signers, with the exception being for Italian gesturers. Here, Italian gesturers exhibited more finger complexity for transitive productions than intransitive productions. However, finger complexity was low generally across the board, suggesting to us that it is not an iconic strategy or it is not an iconic strategy that non-signers 'know' to use.[19] Our conjecture here, then, is that joint complexity measures might be communicative enough in expressing and perceiving events that differ in their transitivity.

[19]We note in passing that the situation is a little more complicated, in that hearing children often make weak transitivity distinctions using finger complexity with transitive productions having high complexity than intransitive ones in the US and Italy, and the reverse pattern in Nicaragua. (The authors did not collect data from Chinese children.) However, we ignore this.

Character vs. Observer perspective: The manipulation of perspectives has also been observed in co-speech gesture (McNeill, 1992) and compared with similar representations in sign languages (e.g., Cormier et al., 2012; Perniss, Özyürek, & Morgan, 2015). The discussion is mostly similar to the one above about the same phenomenon in sign languages: For instance, Parrill (2010) found that gesturers tend to take on the role of a character and use handling handshapes to represent events of manipulation. This is also true of cases where the gesturer wishes to demonstrate affect of the referent (e.g., smiling, terror, reluctance, etc.), or to perform whole-body reenactments. By contrast, to show movement of location of referents, an observer perspective is often assumed. Again, mixed events, where the referent is moving while manipulating another referent, a mixed strategy may be used. In a follow up study, Quinto-Pozos and Parrill (2015) found that the use of handling classifier constructions in signers and character viewpoint in gesturers, and the use of entity classifier constructions in signers and observer viewpoint in gesturers overlapped significantly, though some differences emerged (expectedly) in the groups' use of affect and body posture.

The similarities between perspective use between signers and gesturers has intrigued these (and other) authors about the cognitive processes that may underlie them (e.g., Perniss et al., 2015). However, again, this particular co-incidence of grammatical form (handling vs. entity handshape) with perspective is not our focus here. As such, that we construct our to-be-depicted action videos using a single vantage point seems necessary in order to avoid entangling our final discussion of transitivity marking with one of perspective.

Production, Perception of handshape: In our discussion of handshape, and especially the one-to-many mapping between a particular handshape and a transitive or intransitive parse (§2.2.2, *handshape*), we concluded that a given handshape is not informative enough on its own to be given a consistent parse, or at least, this problem arises for several specific handshapes. However, there may be subtle variations of a particular handshape that coincide with a particular parse at least probabilistically.

Hassemer and Winter (2016, 2018), for instance, considers handshapes like those in Fig. 2.7. The (a,b) handshapes are potentially ambiguous between a shape parse and a size parse, depending on how the fingers and the space between them are interpreted. For instance, in both (a,b), the object being referred to can take the shape of the fingers (i.e., round, likely flat). Likewise, the object could have any shape, so long as it can be held between the index finger and thumb (i.e., only size is communicated). The handshapes in (c,d), while still consistent with a 'size' interpretation, arguably becomes infelicitous with a (round) shape interpretation. The position of the non-selected fingers (i.e., the middle, ring, and pinky fingers) may also bias observers towards one parse or another.

The authors test these claims in both a series of production and perception experiments. In the production experiments, the authors simply asked participants to indicate the size or shape of an object, using just their thumb and index finger. For the shape interpretation, the non-selected fingers were raised most consistently, the opposite being true for the size interpretation, thus demonstrating that–all else being equal–shape and size information can be conveyed in the handshapes of hearing non-signers using a consistent marker.

In the comprehension experiments, participants were shown one of 54 computer generated images of a hand, which varied with respect to the curvature of the index finger and thumb, and non-selected fingers (cf. Fig. 2.7a,c, for index finger curvature; a,b for non-selected finger curvature). The increase curvature of the index finger correlated with an increase proportion of shape responses, while the increase curvature of the non-selected fingers correlated with an increase in the proportion of size responses. The specifics, however, are a little more complicated and, what is more, there was a general, substantial bias towards shape interpretations.

For our purposes, although the authors cast their hypotheses in terms of size- and shape parses, crucially, how they define these terms is consistent with a handling or

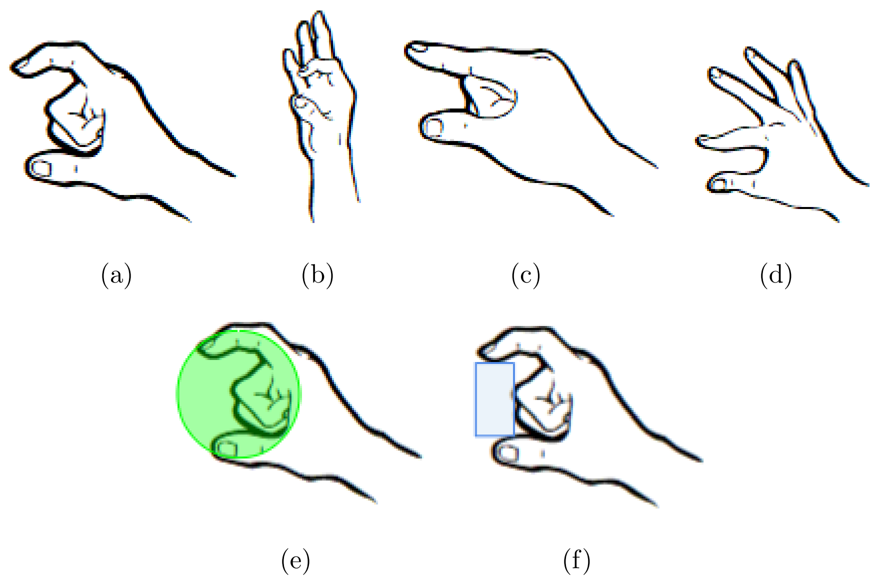


Figure 2.7. A series of handshapes demonstrating contrasts explored in Hassemer & Winter's (2016, 2018) work. The index finger and thumb create a round shape in the handshapes in (a,b), whereas they are flat in (c,d). The non-selected fingers (i.e., middle, ring, and pinky fingers) are closed in (a,c), but open in (b,d). The handshape in (a) is repeated in (e,f), but with outlines of two possible interpretations. In (e), the handshape itself represents the referent. In (f), the space between the first finger and thumb is interpreted as the referent.

entity handshape strategy, respectively. That is, their 'size' condition is predicated on holding an object between thumb and index finger, and their 'shape' condition is predicated on the fingers representing the shape of the object. As such, we are able to make inferences about how non-signers produce and perceive transitive and intransitive handshapes from the results of their study.

For one, the authors demonstrate that the production and perception of transitive and intransitive distinctions is not categorical in non-signers, but that strong tendencies emerge on both ends. On the production end, 87% of participants in Hassemer and Winter (2016) produced handshapes according to the hypothesis that furled non-selected fingers would result in a size parse and unfurled non-selected fingers would results in a shape parse. As previously mentioned, on the perception end, there was a demonstrable bias for one interpretation over the other, but this bias was nevertheless modulated by handshape features. This also points to the fact that producers and perceivers are beholden to different biases or strategies when en- or de-coding manual actions. And, although the results of (Brentari et al., 2012, 2017) do not speak to perception, we would like to argue that here, too, do non-signers come up with a consistent, though not categorical, strategy for encoding transitivity. We extend this argument to the work presented here, in that we are not expecting categorical responses from non-signers in their judgments, but only expect above-chance consistency.

Finally, we note that others have done perception studies on handshapes, though unfortunately those have focused solely on categorical perception within handshape types (e.g., Emmorey & Herzig, 2003; Emmorey, McCullough, & Brentari, 2003; Baker, Idsardi, Golinkoff, & Petitto, 2005.) That is, even though Hassemer's focus is on a type of categorical perception, their task allowed us to make inferences on a transitive and intransitive parse, while other studies of categorical perception have not.

2.3 Iconicity in formal domains

The function, pervasiveness and utility of iconicity in Language has been coming into focus since the mid 1990's and has been explored in earnest only in the most recent decade. This is due in part to looking past purely vocal phenomena (illustrated in 14a) and recasting certain morphological and syntactic phenomena in a new, iconic light. For instance, some accounts posit that reduplication in the nominal domain is iconic, in that the operation generally denotes summation (e.g., Inkelas, 2014; illustrated with the distributive in 14b).[20] with respect to syntax, the observation that a great majority of the world's languages are SOV or SVO (e.g., Dryer, 2011b, 2007) can be analyzed as the grammaticalization of causal chains (Levin & Hovav, 2005; Croft, 1998) or force dynamics (Talmy, 1988). That is, the unfolding of an event in real life is mirrored by the ordering of constituents in most of the world's languages (including sign languages; e.g., Napoli, Spence, & de Quadros, 2017).

(14) a. That meeting was sooooo loooong

b. go2 go2 sai3lo6 dou1 hour2 lek1
CL CL child distr very smart
'Every child is very smart' [Cantonese; Lam, 2013]

The expansion of work on iconicity is also due in part to the recognition that users of spoken languages do not limit their communicative strategies to just the oral modality, but also incorporate affective facial cues and body language (Argyle, 1975; Sendra, Kaland, Swerts, & Prieto, 2013), and gesture (McNeill, 1992, 2005; Kendon, 2004; Kita & Özyürek, 2003). Concomitant is the relaxation of Hockett's (1960) design

[20]To note, reduplication in some languages denotes *diminution* instead or as well as summation. For instance, diminutive reduplication also occurs in Cantonese: hung4 ('red') → hung4 hung2 dei2 ('reddish.') One albeit speculative explanation for this opposite pattern is that, once grammaticalized, iconic properties may take on new and sometimes unpredictable meanings. A nearly analogous example is this: ASL FALL is derived from a body-part classifier (BPCL) wherein the first and second fingers are extended as if to represent human legs. The fingertips are first oriented towards the ground and then invert to 'show' a fall. As an iconic form, FALL could only refer to human beings (or perhaps other bipeds). After lexicalization, FALL can now be predicative of non-bipeds, including animals and even collapsing structures (our data; see also Aronoff, Meir, Padden, & Sandler, 2003).

features of Language: with the obvious admission that sign languages are natural languages, we already must abandon the tenet that Language is only delivered in the auditory-vocal channel. A further relaxation is the admission that Language is not entirely arbitrary, even excepting a small set of motivated words (such as onomatopoeias and ideophones).

The converse is the redefining of *iconicity* itself. In one sense, iconicity is a motivated link between meaning and form. Because the search domain has historically been in the lexicon, iconicity was deemed scarce in spoken languages (i.e., *Why do we have different labels for the same object cross-linguistically?*). The expansion of the search domain has revealed iconicity elsewhere, and not only in the interface between meaning and form (e.g., see again the motivated link between meaning and structure in 14b).

In what follows, we briefly summarize the iconicity of formal features in formal domains in spoken languages (§2.3.1). We then present one, more well-examined cases of formal iconicity–namely, telicity (§2.3.2)– in sign languages to motivate our exploration of the iconicity of argument structure.

2.3.1 Iconicity in spoken languages

Interest in iconic elements in speech has been decidedly dominated by the exploration and discussion of lexical iconicity. Two best studied phenomena are, of course, onomatopoeias and ideophones. The former refers to acoustic mimicry of a source, such as *coockle-doodle-doo* (English), *kikiriki* (Italian/ Romance), etc. for the crowing of a rooster, or *meow* (English), *miao* (Mandarin), etc. for the meowing of a cat. Ideophones go beyond onomatopoeias by encapsulating more complex events, even denoting those that are not acoustic (e.g., *kpebebee* 'rigid posture of a muscular person'; Dingemanse, 2011). Even though these phenomena are more common than once thought, especially outside of Western languages, they are not iconic in the way that we intend here: what we're after specifically is the iconicity of grammatical devices (the recovery of grammatical meaning/ structure from form) in language, and not–we'll say–lexical iconicity (or the recovery of lexical meaning from form).

However, we'll explore one example of iconic vocal phenomena to help illustrate the connection between biology and language we are after. For instance, there is a robust observation that vowel height corresponds with the size of a referent, with low frequencies correlating with large referents and high frequencies with small. The correlation has been documented in the lexicon of several genetically unrelated languages (e.g., English, Japanese; Kammu, Svantesson, 2017 *inter alia*) and is illustrated in (15) below:

(15) 'fire'

a. sñtɨ̀ɨñ (of a bonfire)
b. sñtɔ̀ɔñ (of a torch)
c. sñtʌ̀ʌñ (of a candle)
d. sñtɛ̀ɛñ (of a match) [Kammu; Svantesson, 2017]

Several explanations for the phenomenon are proposed, each invoking some motivated link between form and perception. In one explanation, the size of the articulator (i.e., mouth) is mapped onto the size of the referent: low vowels require the jaw to lower, with the resultant mouth shape being relatively larger than mouth shapes used to produce high vowels (Liberman & Mattingly, 1985). Another line of reasoning comes from the observation that larger animals tend to make lower frequency sounds. Once this connection was established, the interpretation of lower frequencies could meander. That is, lower frequencies have been experimentally demonstrated to correlate with judgments of strength, power, and social dominance; a pattern repeated cross-linguistically (see discussion in Auracher, 2017). This observation, by the by, is not a human-specific one: frogs, birds, and other animal species use frequency cues to pursue mates that are larger than competitors (e.g. Ohala, 1984, 1994; Morton, 1977).

With respect to iconicity in grammar, we already briefly mentioned two examples: reduplication[21] and constituent order. Beyond these examples, we could not readily find other instances of iconicity in this domain. Most of the work we could find, it seems, stems from the 1980's by linguists, Joan Bybee, John Haiman, T. Givòn, Dan Slobin, Joseph Greenberg, and others. We discuss two examples here: one that has a universal flavor, but lacks an embodied explanation, and the other that has a strictly-local flavor, but can be readily given an embodied origin.

The first is an isomorphism between the semantic and morphological domains with respect to the prevalence and relative ordering of different classes of morphemes around the verb stem, so called 'diagrammatic iconicity,' as argued in Bybee (1985). Bybee argues that the more relevant to the stem an affix is, (a) the more likely it is to be bound to the verb, (b) the more likely it is to occur closer to the verb stem, and (c) the more likely it is fuse phonologically with the verb stem. Relevance is defined in turn as whether the morpheme in question modifies the event denoted by the verb in some way, with less relevant morphemes additionally or exclusively modifying other elements (e.g., gender marking coreferencing event participants). Using a small set of verb-related concepts (e.g., valency, mood, gender), Bybee predicts a cline of relevance, with more relevant morphemes being more prevalent as affixes on verbs across languages. Bybee then compiles a database of 50 genetically and areally unrelated languages to test this claim. The hypothesis was borne out with a few minor exceptions. As such, this morphological-semantic isomorphism, modulated by a simple notion of relevance, correctly predicts the prevalence and distribution of verbal affixes. This natural ordering is taken to be iconic, in a way, in its intuitiveness.

Some of these concepts are active at the syntactic level, too. As a simple example, Givón (1980) argues that the degree of (in)dependence of a complement clause has evidential-like import: while 16a is true only of events of direct perception, 16b is additionally true if the speaker hears or reasons that the event occurred.

[21]Insofar as reduplication is a syntactic and not a [morpho-]lexical process.

(16) a. I see you pee funny colors

b. I see that you pee funny colors

However, putting like things together and unlike things apart seems simply intuitive at face value. In Bybee's work, there was no real link presented or hypothesized between the intuitiveness of diagrammatic 'iconicity' and some embodied experience human beings have. Further, by our requirement that embodied experience must be perceptual in some way and not couched in, say, the human reasoning faculty alone, this example is not particularly relevant.

The second example of the iconicity of syntax we discuss is word ordering beyond simple gross constituent order (i.e., relative order of subjects, verbs, and objects) in Mandarin Chinese. H-Y. Tai (1985) proposes the Principle of Temporal Sequence, which amounts to the observation that the ordering of constituents in Mandarin seems to follow sequence logic. That is, for example, *cong Zhongguo* ('from China') represents the starting point of the action *lai* ('come') and therefore must precede it in an utterance, cf. 17a and 17b. In other cases, word order can be switched, but the interpretation of the clause also changes. Consequently, *dao* ('arrive') in 17c conveys what Tai calls a 'projected goal,' but a 'reached goal' in 17d. He demonstrates this phenomenon in a range of different constructions (including the *ba*-constructions, where the object counterintuitively[22] precedes the verb).

(17) a. tā cóng Zhōngguó lái
3s. from China come
'He came from China'

b. *tā lái cóng Zhōngguó
3s. come from China
'He came from China'

c. tā zuòtiān dào Měiguó lái
3s. yesterday arrive USA come

[22]This is, of course, only counterintuitive according to the Principle of Temporal Sequence.

‘He left for the United States yesterday’

d. tā zuòtiān lái dào Měiguó
3s. yesterday come arrive USA
‘He arrived in the United States yesterday’

However, as he himself admits, the Principle of Temporal Sequence only seems to apply to Mandarin Chinese. He offers a few theories as to why this may be so, including a lack of morphology in the language (and, hence, a reliance of strict word ordering) and word-order-interrupting strategies, like focusing and topicalization.[23] While we are generally sympathetic to this type of argument (an iconic universal, trumped by language-specific innovations, or different constraints put on the signal, etc.)–as we will argue for later anyway–the analysis does seem strictly local: iconic, maybe, but indicative of a human universal with its basis in human cognition, perhaps not.

From this very cursory overview of specific iconic phenomena in spoken languages, the impression is that there’s not much there. The finger can be pointed squarely at a few different obstacles: (a) our own cherry-picked literature, (b) an actual paucity of literature on the iconicity of syntactic phenomena, (c) a focus on Western languages, which may not make use of iconicity in grammatical domains, and so on.

To note also is that none of the examples above make use of visual or imagistic iconicity. This is, as Dingemanse et al. (2015) point out, a modality effect: while

[23] Incidentally, the analysis mentions a certain type of serial verb construction, and can accommodate it. However, it doesn’t seem to cover all types. For instance, in (i) (from Zheng, 2012), you cannot reorder the verbs within the motion SVC to indicate whether the *crossing* event or the deictic event (‘come’) happened first, second, or concurrently. While the example is from a Sinitic language spoken in Southern China, we assume that similar examples exist in Mandarin.

(i) a. ziah^4-gao^2 guê^3 lai^5 bhê^{2}lou^7 zio^3 boin^5
Cl-dog cross come road this side
‘The dog crossed toward this side of the road’

b. *ziah^4-gao^2 lai^5 guê^3 bhê^{2}lou^7 zio^3 boin^5
Cl-dog come cross road this side
‘The dog crossed toward this side of the road’ [Suan^{1}tao^{5}Uê^7]

denoting sequences is easy for spoken languages to do, as the speech stream itself is linearly ordered, expressing visual concepts is relatively difficult.

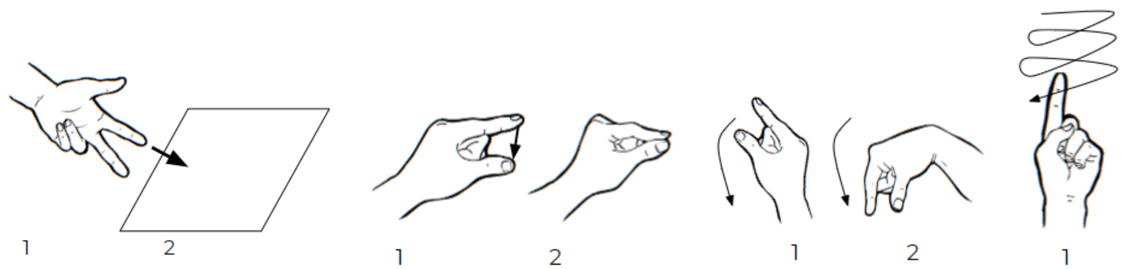
2.3.2 Case: Iconicity of telicity in sign languages

Perhaps the most well examined link between cognition, iconicity and grammar has been demonstrated with telicity marking. As reviewed below, researchers have demonstrated that non-signers perceive event boundaries in everyday actions using kinematic cues; that telicity marking is present in the sign language signal and processed as such; that non-signers perceive telicity distinctions in sign languages; and that, tentatively, non-signers produce aspectual differences in their gestures.

Telicity refers to whether the event expressed by a predicate possesses a natural endpoint (it's telic) or not (it's atelic). The classic test for the telicity of a predicate is the *in an hour/ for an hour* test, with the former adverbial occurring felicitously with telic predicates and the latter occurring with atelic events. The phenomenon is illustrated in 18.

(18) a. Brian twirled his hair for an hour/ *in an hour (atelic)
b. Brian permed his hair *for an hour/ in an hour (telic)

Telicity may be overtly marked in spoken languages (e.g., Japanese, Fujimori, 2012; Slavic, Svenonius, 2005), but note that the marking is itself arbitrary with respect to its sound-meaning correspondence. However, Wilbur (2003) identifies phonetic markings on ASL lexical signs that correspond to telic verbs. These are deceleration towards a point/ plane (POSTPONE), contact with the second hand or body (HIT), and/ or orientation change (DIE), and handshape aperture chance (SEND). Atelic predicates, by contrast, do not have such markings (e.g. PLAY) or are characterized by the addition of one of several reduplicative morphemes (cf. SICK 'become sick' vs. SICK$_{[+cont.]}$, 'Be continually sick'). This correspondence between semantics and phonology she calls the Event Visibility Hypothesis (EVH). These strategies are il-



(a) (b) (c) (d)

Figure 2.8. Telicity marking, illustrated with possible signs. (a) Contact with a plane, marked by, e.g., a sharp deceleration towards it. Contact with the second hand (not shown) is a similar cue. (b) Aperture change. (c) Orientation change. (d) Atelic sign, exhibiting tracing. In all, the numbers '1' and '2' refer to the number of subevents comprise the sign: two for telics, one for atelics. Note that these do *not* refer to timing units á la Brentari (1998), which Wilbur (2003, 2008) uses in her formulation of the EVH.

lustrated in Fig. 2.8. Further work by Grose et al. (2007) demonstrates that this cue is also available in classifier constructions.

This strategy is not only active in ASL, but in Croatian Sign Language (HZJ, Dukić et al., 2010; Milković & Malaia., 2010; Malaia et al., 2013) and Austrian Sign Language (ÖGS, Schalber, 2006), suggesting generally that sign languages may use the same visual/ kinematic resources to encode similar concepts. And, although these telicity markings are appreciable to the naked eye, Wilbur and colleagues quantify them in a series of experiments (Malaia & Wilbur, 2012a, 2012b; Malaia et al., 2013). Using motion capture, they identified that deceleration distinguishes telic from atelic signs in ASL and HZJ.

We also want to note that there is another, separate layer of iconicity here, besides the visual characteristics of end-marking, heavily implied by Wilbur's description of the phenomena but perhaps never explicitly stated: The counting of subevents identifies atelic vs. telic signs. Atelic events are homogenous. They are characterized by some unchanging state or process. They are singular, in a sense (a single S[tate] or a P[rocess] in Putstejovskian terms (Pustejovsky, 1991). Telics, by contrast are characterized by two subevents, one leading to the other (S $\rightarrow$ S or P $\rightarrow$ S). They are plural, in this sense.[24]

Telicity information is available in the production of signs, and it is argued to be iconic, but what about perception? Is telicity marking is perceptible and iconic? To that end, Strickland et al. (2015) ask whether telicity cues are appreciable to non-signers, who by definition do not have access to the structure of any sign language. The authors test American non-signers using data from three historically and geographically unrelated sign languages (LIS; Turkish Sign Language, TİD; and Sign Language of the Netherlands, NGT), and one artificially created set of signs. For all categories, participants were given a video displaying a verb sign and two possible single-word labels that could describe its meaning. Neither of the labels were the correct label,[25] discouraging non-signers from using any potential lexical iconicity to guide their decisions in a top-down fashion.[26] For all languages (and the set of invented signs), participants classified telic vs. atelic verbs correctly significantly above chance levels.

To force the issue of whether it was the presence/ absence of a visual boundary that informed non-signer judgments, the authors conducted additional experiments in which they explicitly asked participants to rate on a scale from 1 to 7 whether they believe they detected the presence of a boundary (corresponding to telic signs) or whether they detected repeated movement (corresponding to atelic signs). Participants were more likely to detect a gestural boundary in telic signs over atelic signs as predicted.

[24]Recall that we've mentioned this sort of counting in our discussion of directionality, which is a partial cue for argument structure, in that a perceiver can count the number of locations a verb is articular at and deduce transitivity from that: one location for intransitives, and two for transitives (including ditransitives), except in the case of motion events.

[25]This discounts experiments 1 and 2.

[26]The authors additionally controlled for the possibility of a top-down parse by having the sign depicted in the stimulus be from a different conceptual domain as the two provided labels. For instance, if the sign were THINK (cognition), the two labels might be DIE and RUN (physical).

On the perception end, Strickland et al. (2015) demonstrate that non-signers perceive event boundaries in sign language verbs, and they have communicative effect. To complete the picture, to see whether this is a universal but perception-only effect or a universal encoding-decoding phenomena, it behooves us to ask whether non-signers produce such contrasts. There is some indication that non-signers do produce this contrast. Duncan (2002) notes for Mandarin and English speakers, gestures co-occurring with unbounded predicates were longer than those occurring with bounded ones. While the results do not weigh in on telicity *per se*, these results demonstrate that gesturers from two geographically and genetically different languages produce co-speech gestures that vary with respect to a linguistic contrast (outer aspect) using similar kinematics (duration).

But what might be the origins for this ability? How is it that non-signers putatively use kinematic cues to make decisions about the telicity of verbs they would have never seen before in languages they don't know? As, e.g., Malaia (2014) points out, research on the perception of real-life event boundaries relies on some of the same kinetic cues that are employed in telicity marking in sign languages.

For instance, Zacks et al. (2009) had non-signers segment a series of everyday actions into subevents. These actions were quantified using motion capture, allowing the authors to correlate non-signer boundary judgments with kinematic cues. The authors found that participants generally agreed on where event boundaries occurred. Further, they demonstrated that the acceleration and velocity of individual limbs largely predicted participant-identified event boundaries.

In sum, then, telicity distinctions are manifest in the form of sign language verbs. The same kinematics that describe telicity in sign languages also describe event boundaries in the perception of everyday actions. Non-signers can accurately guess the telicity of signs, using these boundary cues. And, non-signers (tentatively) produce verbs in accordance with the kinematics when gesturing. This has been demonstrated over speakers and signers of different (sign) languages, pointing to a universal mapping between a certain set of kinematic features, event perception, and event semantics.

Lastly, despite the fact that telic and atelic verbs are differentiated in sign languages using visual/ kinematic cues, like peak velocity, there remain some language-specific nuances in (a) the specifics of the cues themselves and (b) how these cues integrate into the linguistic system. For instance, Malaia et al. (2013) demonstrate that, while both ASL and HZJ both use velocity (and/or its derivative) to encode telicity distinctions, (a) ASL uses deceleration and slope of deceleration, but HZJ uses peak velocity, and (b) the kinematics of telics are affected by prosodic factors (such as phrase-final lengthening) in different ways. Further, telicity marking is a productive process in a majority of HZJ verbs (e.g., akin to Croatian *gledati* 'look at' vs. *ugledati* 'spot'), while it is inherent in only the telic subset of ASL verbs.

In the same vein, one final caveat here is that although non-signers can reliably distinguish telic from atelic verbs in unfamiliar sign languages, using kinematic features that also help them parse real-life actions into subevents, there are some salient differences in the visual abilities of signers and non-signers, evidently due to the former group's lifetime experience using a visual language: an fMRI study by Malaia et al. (2012) demonstrates that telicity distinctions are not processed in the same way across signers and non-signers.

In both cases–differences in the adoption of kinematic features into the linguistic system between sign languages and differences in the neural processing of (telicity distinctions in) sign language–demonstrate the influence of the linguistic system atop of pure perceptual phenomena. With respect to the origins and iconicity of argument structure, then, our ultimate hypothesis is that key concepts from the visual-praxic domain (e.g., object affordances, force dynamics) are recruited for the purpose of argument structure marking.

2.4 Emergence of grammatical features in visual communication systems

In what follows, we discuss the emergence of two grammatical phenomena in visual communication systems. We use the general term 'visual communication system' to include non-signer manual behavior (co-speech gesture, pantomime, etc.), homesign, young sign languages, and established sign languages. We should also clarify what we intend my 'emergence.' Here, we are talking about the roots of grammatical phenomena in the communication systems of modern humans, and not the emergence of these devices in the species. We further assume that the range of human non-manual communicative behaviors forms a cline with respect to formalization, with non-signer gestural behavior constituting the lower extreme and established sign languages being the upper extreme, though additional elaboration seems necessary (we ignore it here).

2.4.1 Case: Argument structure & directionality

In previous sections, we've explored how argument structure is manifest in both established sign languages and pantomime, representing two poles on the continuum: full fledged vs. *de novo*. But what of newly budding sign languages? Here we briefly review what has been discovered with respect to argument structure in young sign languages.

Handshape, handshape complexity: One of the most well studied young sign languages is Nicaraguan Sign Language (e.g., work done by Senghas originally, and the Brentari group at the University of Chicago currently. See below for sample references). Briefly, NSL was born after an influx of deaf children (and adults) from various regions of Nicaragua into Managua. The linguistic situation there now sees signers stratifying into a few distinct groups: those who first arrived were homesigners–users of a family-specific gestural system. The first wave of homesigners, then, was an instance of language contact. The second wave, or cohort, of children to arrive at the deaf school learned the system created by their older peers, and the third cohort learned from the second. As a result, NSL has been observed to take on more expected linguistic qualities of older, established sign languages and of Language more generally. Because members of each cohort are still alive, it is possible to track diachronic developments in the language synchronically.

In one study, Goldin-Meadow et al. (2015) measured the proportion of handling and entity handshapes used in response to agent and agentless vignettes, using a similar methodology to the studies reported above. Groups included in this study were Nicaraguan homesigners, members of Cohort 1, members of Cohort 2, and native ASL signers. Perhaps surprisingly, all groups behaved the same way: all participants used more entity handshapes in response to agentless vignettes than handling handshapes; there was no significant difference in performance between groups. Further, all groups used similar proportions of entity and handling handshapes in response to vignettes with an agent; again, there was no difference between groups. There are a couple of things to note: All groups used roughly the same proportion of entity and handling handshapes in response to vignettes with agents. This is consistent with the results reported above for ASL. Second, although mean responses for homesigners matched mean responses from the other groups, the authors note that within-subject variability in this group was predictably higher, given the heterogeneous nature of homesign systems. Nevertheless, this variability decreases as a function of cohort, suggesting to the authors that communicative and linguistic pressure reinforces tendencies in object-handling handshape use.

Similar, but messier results, were obtained in initial research on Central Taurus Sign Language (CTSL), a village sign language of central Turkey (Ergin & Brentari, 2017). CTSL is reportedly distinct from Turkish Sign Language (TİD). In a small study of seven signers (which stratify into three cohorts) using the same paradigm as Brentari et al. (2012, and subsequent), a small preference for handling handshapes in response to vignettes with agents was observed. The opposite pattern, a preference for object handshapes when describing agentless vignettes, was also observed.[27]

Finally, with respect to handshape complexity, Brentari et al. (2012) also report on the homesigning group in Nicaragua. They find that, unlike Italian and American gesturers (missing are Nicaraguan gestuerers), homesigners exhibit more finger complexity in their entity handshapes (used in agentless contexts) than their handling handshapes (per above, used in both agent- and agentless contexts). This pattern matches what adult signers of established sign languages, here ASL and LIS, though the specifics vary a little.

Taken together, the data from young sign languages and homesigners match more closely to established sign languages than they do to pantomimes. However, transitivity distinctions are made in each communication type using the same phonetic resources (here, handshape and handshape complexity). This suggests an iconic beginning to transitivity distinctions (everyone manipulates [aspects of] handshapes) but that the presence of a linguistic system (young sign languages, established sign languages) and/ or the persistence of use (homesign) changes the raw material into a new shape over time.

Directionality: All sign languages studied to date exhibit directionality (Aronoff, Meir, Padden, & Sandler, 2005), no matter if they are genetically related or not, or areally related or not. Directionality, in some form, is even found in new, developing sign languages, such as ISL, ABSL, and NSL. What is more, there is some evidence of directionality in gesture, suggesting that the eventual encoding of syntactic and/ or semantic roles may stem from iconic origins.

Padden, Meir, Aronoff, and Sandler (2010) tracked the development of the use of space for directionality across three generations of ISL signers ('older group': 65–90; 'younger group': 45–65; 'youngest group': 30–40) and find differences in the propor-

[27] There was in general a lot of variation between cohorts with cohort 1 (a single signer) expressing an opposite pattern to subsequent cohorts. Per-vignette data also reports that the entity-handling preference was only really observed for some object types, but not others.

tion of Z- and X-axis use between them. The Z-axis here is the one that extends straight out from the signer's chest. This is the axis used in many sign languages for 2nd person reference. The X-axis lies along the plane that bisects the body horizontally (i.e., it is parallel to the ground). This is the axis used to show the relationship between two 3rd person referents in many sign languages. The combination of axes, the X/Z-axis, is the one that extends from the body out to the left or right of the signer. This is the axis used to show the relationship between 1st person and 3rd person referents in many sign languages.

In their study, the authors asked ISL signers to sign sentences in response to simple transitive and intransitive videotaped actions. All of the video clips contained, naturally, 3rd person referents. The authors then counted the number of times particular axes were used to express the events, whether or not event participants were localized separately, and whether the verbs were directed at those event participants across two functions of space: space used for spatial reference (e.g., in motion events) and space used for personal reference.

The authors also found that there is a difference in prevalence of directionality between generations of signers: Older signers were far less likely than younger signers to inflect verbs agreement, as defined. However, they do see an increase in the use of single agreement, or the inflection of the verb towards its (indirect) object from the oldest group to the second oldest group. The youngest group is the only group that consistently used double agreement, or the modulation of the verb to agree with its subject and (indirect) object.

Further, the form of the 'agreement' is generally different. Among older signers, a referent tends to be localized directly in front of the signer and the verb moves from the chest of the signer outwards towards that locus. Meir (2016) notes that the form of the verb is the same in this newly inflected form as it is in the uninflected form, only the referent is now localized. In younger signers, referents are freed from this axis, and are able to be localized to the left or the right of the signer. As a result, the verb end may move towards the locus of the referent to inflect for what Meir calls 'single argument agreement.' A further innovation was the detachment of the initial part of the sign from the body and the its orientation towards the locus of the subject (should the subject not be the signer).

The authors also note that the older groups contained more variability than the younger group. For instance, around half of the older participants used single agreement the most frequently, while the other half did not use agreement at all. The picture, then, is that the use of agreement becomes more prevalent and more consistent in younger and younger generations of ISL signers.

In their analysis of ABSL, Padden et al. (ibid.) show that generations of these signers also differ in their use of directionality. In this case, only two generations of signers are included (the second and third generations). As with the ISL signers, younger signers tended to use the X and X/Z axes more so than older signers. The younger signers also showed a difference in signing verbs denoting transfer (e.g., the relationship between participants) and those denoting space (e.g., how participants move in space): Younger signers used the X axis more in events denoting space, but the X/Z axis more in events denoting transfer.

In very few cases did ABSL signers use space to for agreement purposes. Of the 65 verbs of transfer they collected (i.e., verbs eligible for person agreement), only 16 of them they consider to exhibit single- or double-agreement, where single agreement refers to the movement of the verb towards its object and double agreement refers to the movement of the verb from its subject to its (indirect) object. What is striking here, is that a similar move away from the Z-axis and towards the Z+X- or X-axis was observed in ISL for younger and younger signers. However, in the ISL case, agreement was much more common. It seems, then, that the movement off the Z-axis is a prerequisite, but not an immediately determining factor in the emergence of agreement, as defined.[28]

[28] That is, unless my math is off. It seems that the count of off-Z-axis productions was higher than the count of single- and double-agreement.

One explanation Padden et al. give for the paucity of agreement forms in ABSL is the notion of 'body as subject,' wherein the subject is preferentially mapped to the body, leaving the tail end of the verb free to move to another locus in space (or not). The specifics do not concern us here, as the point is just that other (iconic) strategies may interfere with the emergence of an agreement system like we see in ASL, among other sign languages.

One thing of note from an earlier investigation of agreement in ABSL (Aronoff et al., 2005), though, is that there were five occurrences of 'directionality,' only that the form of the production appeared to be more mimetic: a grasping gesture was used in lieu of the sign GIVE and it moved from the locus of the source to the locus of the goal. This is particularly interesting for our present purposes in that the gesture incorporates what we'll call 'non-linguistic directionality' while the lexical sign does not. This either indicates that the concept 'directionality' initially becomes far less iconic as it is coopted by the linguistic system and then resurfaces in a constrained way, or that lexical verbs and 'non-linguistic directionality' remain separate. Of course, the notion of 'body-as-subject' could also be at play here, in that the signer takes on the body of the 'grabber.'

In any case, candidate directional verbs always denote some sort of transfer (e.g., GIVE, a transfer of an object) before the strategy extends to other verbs (e.g., HIT, a transfer of force; TELL, a transfer of a story/ words, etc.). As Meir (2012) puts it, directionality emerges as a semantic phenomenon, but eventually progresses towards encoding grammatical categories (e.g., subject and indirect object).

There is somewhat less information about the emergence of directionality in NSL. Senghas (1997) had signers from Cohort 1 and Cohort 2 watch video clips of simple transitive actions and then sign what they saw. She then looked at verbs produced with 'spatial modulation,' which can mean the addition of aspectual, plural or other information to the verb stem. It can also mean movement towards the locus of a referent established in the signing space. Here, though, her focus here is on the rotated or unrotated versions of this directionality (i.e., what perspective the signer takes), but it is immediately evident from this account that directionality is used by both cohorts with some regularity. In work elsewhere, Senghas (1995); Senghas and Coppola (2001) calculate the number of modulations per verb in the productions of Cohort 1 and Cohort 2 signers, finding that Cohort 2 signers–and specifically those who entered the Deaf school at a younger age–produced significantly more modulations per verb than Cohort 1 signers. Further, signers from Cohort 1 did not produce modulations related to agreement.

Returning to the discussion of character and signer views, Senghas (2003) demonstrates that the production and interpretation of these views differ in Cohort 1 and Cohort 2 signers. Cohort 2 signers use character view in their productions and interpret signing from character view in perception. Cohort 1 by and large does not–the resolution of who is the object or recipient of the utterance is ambiguous. As such, although both cohorts modulate verbs, only Cohort 2 assigns this modulation a consistent function and interpretation. That is, the seeds of directionality were already present in Cohort 1, but it took a generation (loosely speaking) for a consistent, grammatical use of directionality to emerge.

As for non-signers, sensitivity to directionality has been demonstrated by Cassell, McNeill, and McCullough (1999) using co-speech gesture. In their experiment, the authors have a confederate narrate a short story about two cartoon characters. The confederate locates the characters in space early on in the narrative. In parts of the narrative, when one character acts on the other, the confederate performs one hand acting on the other. In matched cases, the hand corresponding to the actor's locus in space moves towards the hand associated with the patient's locus in space. In the mismatch condition, the reverse is true. The authors then had participants retell the narrative to a third person and counted errors in the participants' retellings. Errors abounded when the gesture mismatched speech, indicating to the authors that participants attended to co-speech gesture. We go further to suggest that implicit here is that person or thematic information is established in co-speech gesture in a functionally similar way to directionality in sign languages. We say *functionally* here, as it is impossible to make a formal linguistic inference about the status of loci as variables (Kuhn, 2015), referential loci (Lillo-Martin & Klima, 1990), etc. without formal tests.

Similar results were obtained by Schlenker and Chemla (2018), who asked hearing non-signers to judge the appropriateness of pro-speech gestures, or, pantomimes that are embedded in speech.[29] Specifically, the authors were after whether person information was available in the directionality of the co-speech gestures. In some cases, the direction of the gestures matched what what in speech, as in *Your brother, I am gonna PUNCH*$_3$*, then you, I am gonna SHOOT*$_2$, where the subscripts refer to a third person and second person locus, respectively. In other cases, they were mismatched. The match cases were rated as more acceptable than the mismatch cases, suggesting that non-signers have intuitions about the use of space for encoding (2nd and 3rd) person reference. This further suggests that the mechanism used in sign language agreement systems is already available to hearing non-signers.

Again, the above suggests a functional similarity between sign and co-speech gesture, but the authors further provide evidence for a *formal* similarity between both. In the same experiment, some items contained elided material, where the resolution of who was to be slapped, punched, etc. was non-overt. For instance, in *Your brother, I am gonna SHOOT*$_3$*, then you, too*, the interpretation is that you are going to shot after the speaker shoots your brother, indicating in non-theoretical terms that SHOOT$_3$ is understood as SHOOT$_2$ when elided. Judgment data reveals that this 'mismatch' under ellipsis is significantly better than when SHOOT$_3$ is overt in both clauses (i.e., the mismatch case). The same pattern exists in spoken languages, here English: in *Ulyssa loves her job and Rollie does, too.* it is more felicitous in this case to posit that Rollie loves his (own) job, changing *her job* to *his job* under ellipsis, though the other interpretation–that Rollie loves Ulyssa's job–is certainly available.

[29]More specifically, pro-speech gestures are silent gestures that occur embedded in speech, as in *I wanna [traces thumb across neck] him.* Here, the tracing gesture does not occur with speech, but appears in the middle of the utterance. In pantomime, by contrast, there is no embedding: the gestures are performed totally in the absence of speech and take on the full propositional load.

It should be noted, though, that the reports on ABSL, ISL and NSL have all centered around *production*, while the reports on gesture have centered around *perception*. It could be the case that non-signers are also subject to competing iconicities (like 'body-as-subject') when producing 'agreement' forms, and that ABSL and ISL signers are perceptive to directionality even though they do not produce it. A more complete picture, linking production with perception/ judgment, would indicate that the system is iconic (and thus likely informative for, e.g., communication).

In sum, directionality is found in many established sign languages, at least those that have even a basic description. These sign languages are not necessarily genetically or areally related (Sandler & Lillo-Martin, 2006). The pervasiveness of directionality across many sign languages, its development in the young sign languages surveyed here, and in the judgments of non-signers suggests its iconic origins.

2.4.2 Case: Constituent order

Constituent (or word) order is defined with respect to three main components of a transitive sentence, namely the subject (S), verb (V), and object (O; though in some cases it may be more appropriate to talk about these constituents by their semantic roles: agent [Ag], action [A], and patient/ theme [P]). While languages may demonstrate a number of different word orders internally, there is usually one 'default' order, which is deemed the basic word order of the language. According to Dryer (2007), a basic word order is determined along the following criteria: (1) *Frequency*, or how often a particular word order is used with respect to others; (2) *(Un)Markedness*, or the absence of any special phonological, morphological or syntactic marking; and, (3) *Pragmatic neutrality*, or the word order does not convey any particular pragmatic meaning, i.e., it is a simple active declarative.

Logically, then, there are six combinations of these three major constituents: SVO, SOV, VSO, VOS, OSV, and OVS. And, logically, we might expect an equal distribution of these word orders across the world's languages, all else being equal. However, this is not the case. In a survey of 1,377 languages, Dryer (2011a) counts that 1188 (86%) have a basic word order (two exceptions, e.g., are Russian and Walpiri). Of these languages, 565 (48%) have SOV as their basic word order, 488 (41%) have SVO, 95 (8%) have VSO, 25 (2%) have VOS, 11 (1%) have OVS, and 4 (0.5%) have OSV. That is, the lion's share of languages exhibit just two of six possible basic word orders.

The distribution is skewed, too, when considering established sign languages alone. In Kimmelman's (2012) survey of 24 sign languages, 21 (88%) have dominant SOV or SVO word orders. In subsequent work by Napoli and Sutton-Spence (2014), all sign languages in their sample have either SOV or SVO as their basic word order. Nevertheless, all still permit SOV order. We return to some of their findings below when we discuss constraints placed on word order by the linguistic system.

In our brief survey of the emergence of word order, we'll discuss this issue with respect to pantomime (instantaneous, new 'languages'), young sign languages, and established sign languages (which themselves are younger than most spoken languages). We cast this in the light of (a) extra-linguistic sources of linguistic (here grammatical) phenomena and (b) the pressures of a linguistic system on these sources. We further cast this in light of a recent argument for SOV as the evolutionarily basic word order in human language/ communication.

From descriptive studies of homesign systems–systems that emerge in the homes of deaf children without access to a language model–it appears that SV and OV, orders consistent with an overall SOV order, emerge in geographically, culturally, and linguistic diverse climates (e.g., for American and Chinese homesigners, Goldin-Meadow & Mylander, 1998; for American, Chinese and Turkish homesigners, Goldin-Meadow, Özyürek, et al., 2008). Further, studies of young sign languages have demonstrated that basic SOV word order develops relatively quickly (e.g., Sandler, Meir, Padden, & Aronoff, 2005 for ABSL; Senghas, Newport, & Supalla, 1997 for NSL). Some important caveats are discussed below.

Experimentally, it has been shown that constituent order in elicited pantomime is fairly stably SOV, at least in non-reversible events (i.e., events of an agent acting on an inanimate object), irrespective of the pantomimer's native language. For instance, Goldin-Meadow, So, et al. (2008) tested English-, Turkish-, Mandarin-, and Spanish speakers in a pantomime elicitation task. Despite the languages having different basic word orders (English/ Mandarin/ Spanish, SVO; Turkish, SOV) all participants produced primarily SOV strings. These findings were replicated in subsequent studies (Langus & Nespor, 2010; Gibson et al., 2013). Meir et al. (2017), though, do find appreciable effects of language contact (i.e., a pantomimer's L1, a signer's contact with spoken languages, among other possibilities), but otherwise find comparable results for Hebrew, Turkish and Arabic speakers, and signers of ISL, ABSL, and Kafr Qasem Sign Language.

The authors all purport their own versions of a few explanatory themes: one order is less ambiguous than another (e.g., SVO is less ambiguous than SOV in reversible events), one order allows for information to be lost without damaging the entire message, one order avoids embodying both subject and object in reversible events, and so on. These are all more or less functionalist accounts that take communication as the predominant driving factor of SOV and SVO alternations. Newmeyer (2000) additionally adds linguistic motivations: SOV respects theta positions (e.g., SOV languages exhibit fewer movement operations than SVO), but needs Case marking for role identification. Similar trade-offs are discussed for SVO.

While we do want to illustrate that motivations both within and external to the linguistic system proper influence decisions languages make with respect to word order, we've lost touch our iconic grounding. And, although we panned Tai's (1985) discussion if Mandarin word order patterns in §2.3.1, there may be some corroborating evidence in elicited pantomime: Christensen, Fusaroli, and Tylén (2016) demonstrated that the unfolding of the event in time can also be predictive of the gross constituent order in pantomime. For instance, in events of manipulation (e.g., *putting a book down*) reliably elicited an SOV strategy, as the action and object are coextensive in time. However, in creation verbs, where the object exists only after some action has taken place (e.g., *building a sandcastle*), a consistent SVO order emerged.

In a series of three experiments the authors demonstrate what they claim are three extra-linguistic, not-even-cognitive motivating factors in constituent ordering: structural iconicity (of the kind invoked but not named in H-Y. Tai, 1985), interactive alignment and distributional frequency.

However, I think the most neutral, yet perhaps the most exploratory explanation for the prevalence of SOV and SVO word orders comes from Kemmerer (2012). Kemmerer invokes two independently motivated explanations: subject salience and verb-object contiguity. At their base, both can in turn be explained by causal chains and temporal sequence. These serve as iconic motivating factors in the SOV and SVO orders, such that the subject lies at the head of the causal chain and is linearized first, and the verb and object, proceeding from the doings of the subject form a tight unit and come last in the sequence. This explanation remains neutral with respect to which order–SOV or SVO–ultimately surfaces in a language, but as such is consistent with the constituent order phenomena surveyed above. Kemmerer's account is further consistent with neural systems underlying sequential and hierarchical processing, both language internally and externally, suggesting an albeit speculative iconic, isomorphic link between neural processing, causal chains, and linguistic constituent ordering.

In this brief review, we hoped to have shown that linguistic phenomena, such as basic (and derived) constituent order, can have its roots outside of the linguistic system. Cross-modal preferences for SOV were discovered in non-linguistic picture sequencing (Goldin-Meadow et al.), para-linguistic pantomime elicitation, homesign systems and new sign languages, and–of course–in the world's developed (signed and spoken) languages. The shift from SVO to other constituent orders (SVO in particular) has been argued to be due to agent first biases, communication strategies (e.g., avoiding role-conflict or ambiguity), among other pressures (none of them chiefly linguistic). However, it was also reviewed that these functional and cognitive biases do have effects on the shape of linguistic systems, and not only on the change from one constituent order to another. As Hall et al. point out, some systems evolve Case marking, agreement (/directionality), or similar devices–all linguistically controlled–to solve these functional problems.

We use the discussion here to scaffold our argument that consistency found in pantomime argues for the recruitment of non-linguistic sources for a communicative task, and our further argument that this consistency can become linguistic over time (with community, etc.). The source and target domains we examine–visual-praxic, argument structure–are different, but we mean to say that the proof of concept already exists.

2.5 Nature of coding, emergence of grammatical features: Holistic or compositional?

Before Stokoe (1960) (and even for some time after) scientific and lay communities conceived of signs as being unanalyzable, or holistic, and conflated signs with gestures and pantomimes produced by hearing communities. Stokoe showed, however, that signs could be decomposed into meaningless parts, *cheremes*, including *handshape*, *location*, and *(palm) orientation*. The details have changed over time, but the spirit hasn't: signs are analyzable, and for a number of different meaningful or meaningless parts across multiple domains of inquiry. See Goldin-Meadow and Brentari (2017) for a recent review of the (pre-)history of sign language research and variable attitudes towards iconicity in the field.

We suggest, as with Wilbur and Malaia (2008), that gesture (generally) and pantomime (specifically) may be subject to linguistic analysis on par with signs and classifier constructions. Supposing that pantomimes are holistic, at any rate, prohibits discovery of subtle form-meaning correspondences, regularities within like pantomimes, dissimilarities between pantomimes conveying different information (or even propositions), *inter alia* (Martell, 2005).

Ahead of time, though, we would like to stress that there are many top-down and bottom-up processes in language, often working in tandem (e.g., word processing:

matching phones to stored phonologically representations; predicting words based on discourse, etc. etc.). We further note that top-down and bottom-up strategies are likely both available even when it is less intuitive that top-down processes should already be in place (e.g., first language acquisition, language emergence). With this, we note that top-down and bottom-up processes operating within a given (cognitive) system do not need to have originated in that system (e.g., reaching, joint attention, etc. aid in [object] word learning). This relates in particular to our question of where grammatical devices in (sign) languages come from: reanalysis of extant holistic forms into meaningful pieces, and/ or the construction of new forms from already available pieces, perhaps originating from other cognitive modules. We demonstrate with a couple examples below and a recapitulation of two grammatical processes surveyed above that both top-down (reanalysis) and bottom-up (composition) processes are at work in ASL lexical signs, classifier constructions, and their affixes.

A useful starting point to our discussion of pantomimes being holistic or derived from meaningful parts is to discuss what is known about lexical signs and classifier constructions in sign languages. Both are compositional, but in different ways. As for classifier constructions, we mentioned above that they are analyzed as being composed of a handshape, movement root, and location affixes. Changing the handshape changes the possible referent of the action. Changing the movement changes the action the referent performs, and so on. These constructions are decomposable at the syntactic level.

As for lexical signs, again, Stokoe was the first to analyze them with respect to their component phonological parameters (*cheremes*, what we might call segments or phonemes in spoken languages). These seemingly meaningless parts combine to form a meaningful whole, thus illustrating duality of patterning. Lexical signs are thus decomposable into component, sublexical parts. However, while this phenomenon is exhibited in sign languages, it is underexpressed with respect to spoken languages: there are generally very few minimal pairs, at least in ASL, due in part to the pervasiveness of iconicity (Brentari & Eccarius, 2010).

Instead, many signs fall into families, with changes in sign certain sign parameters (but not others) in sometimes iconic ways deriving each family member. For instance, Fernald and Napoli (2000) map out the 'nuclear' and 'extended' family of kinship signs. Signs like FATHER and MOTHER are related by handshape and movement, but differ in place of articulation. MOTHER and GIRL, and FATHER and BOY are related by place of articulation and movement, but differ with respect to handshape. The signs GRANDFATHER and GRANDMOTHER are related to both FATHER and MOTHER (handshape and location) and the sign FUTURE (movement), iconically incorporating notions of continuing time (i.e., age) with mother- and fatherhood. In general, too, the forehead (place of articulation) is the locus of many male-denoting signs (FATHER, BOY, NEPHEW, BASTARD) while the chin is the locus of many female-denoting signs (MOTHER, GIRL, NIECE, BITCH). Our albeit speculative conjecture here is that even though the parameters alone or together do not add up to an iconic sign (FATHER, e.g., is still arbitrary: 1.259/ 7 mean iconicity [`asl-lex.org`]), that these signs express related concepts might not be.[30]

Lepic and Padden (2017) argue that the ASL lexicon is organized in this way– around sometimes iconic, sometimes arbitrary 'family resemblances.' And, what is more, these resemblance paradigms do not amount to sublexical structure: iconic aspects of ASL lexical signs are identified in a top-down fashion. That is, they argue that the identification of a sign's parts as iconic follows identifying the sign as a whole as iconic.

> Crucially, the perception of iconicity arises as a consequence of the fact that signs are conventional pairings of a potentially complex form and potentially complex meaning, and not from a compositional analysis of

[30]Similar work has been done on spoken languages, using large corpora to discover whether there are systematic phonetic/ phonological correlations between words and their lexical categories (Monaghan et al., 2005, 2007. The authors found reliable within- and across-language cues to word category membership, especially in corpora of child directed speech, indicating that these cues may be used in language acquisition. The authors note, however, that these are *regularities* that exist in languages, which are not necessarily iconic. In the FATHER-MOTHER-etc. example we give above, it could be argued that the manipulation of these parameters or the (conjectured) ability to class related signs reflects a sensitivity to systematicity and not iconicity.

> the sign's parts. The meaning of the whole facilitates the (re)analysis of its parts, rather than the other way around. Lepic and Padden (2017, p. 497)

They cite the sign TIME as one example. The sign is articulated by tapping the back of the wrist of the non-dominant hand with the index finger. Historically, this represented the clapper or hammer striking a bell (like in a clock tower), though today it is iconic of a wrist-watch. The authors note that the sign was invented long before the invention of the wrist-watch, and therefore argue that the sign is only iconic today through the reanalysis of its parts. They use Taub's (2001) analog-building model of linguistic iconicity, wherein elements of a symbol are mapped in a one-to-one fashion with experience outside of the symbol, to build their case, though we skimp on the details here.

On the other hand, Emmorey (2014) argues for a compositional treatment of the same phenomena, citing structure-mapping theory and also Taub's (2001) analog-building model.

> Perceptual symbols are not holistic representations of experience; rather, they are componential and structured representations that schematize multi-modal aspects of experience with entities or events in the world. (pp. 4)

She gives as an example the sign for BIRD in ASL. In ASL, the beak of a bird is iconically represented– that is, there is an alignment between the source (perceptual) domain, represented by the visual features of a bird('s beak), and the target domain, the articulation of the sign. She notes that other languages may pick out some other salient characteristic from the source domain to represent iconically. For instance, the sign for BIRD in Turkish Sign Language is articulated to show a bird's wings flapping. Emmorey draws support from Thompson et al. (2009) (see also Thompson, 2011), who show that iconic parts of signs speed picture-sign matching when the iconic features of the sign were prominent in the picture. She also supports her argument with an observation by Meir (2010) that iconicity in signs blocks metaphorical extension: the sign EAT in ASL, for instance, is articulated with a closed hand making contact with the signer's lips, iconically showing the hand bringing food to the mouth. Because of this, EAT cannot extend to non-animate entities, as it does in English, so a different verb would have to be used in sentences like *The acid ate through the metal.* The reader is referred to Emmorey (2014) for other supporting evidence.

It appears that Emmorey (2014) and Lepic and Padden (2017) have come to different conclusions concerning whether signs merit a holistic or compositional analysis, while invoking the same framework (i.e., Taub, 2001). Part of this is attributable to the domains they treat: Emmorey explored both functional morphology and sublexical phenomena, while Lepic & Padden focus on the latter. Neither admits (Lepic/ Padden by argument, Emmorey incidentally) that iconic sublexical components are morphological (as Zwitserlood, 2008 does), but both do admit that they exist. The source Emmorey draws upon (i.e., Thompson et al., 2009), we argue, is consistent with both perspectives: speeded reaction times in response to iconic elements in signs being present in pictures could argue for the analysis of components from a whole or the summation of iconic components. Perhaps the unsatisfying, yet necessary conclusion is that both theories may be true of some parts of the sign language lexicon, and even perhaps sometimes they overlap.

We are sympathetic to both accounts and do not know *a priori* which might be (more) correct. In part, this dissertation impinges on the question by probing whether the argument structure of ASL lexical verbs is accessible from a bottom-up perspective. One important point we want to make, though, is that due to ASL's strong tendency to conspire towards monosyllabicity, many visually 'singular' (holistic) signs may in fact be multimorphemic (i.e., compositional). Some examples include compounds (e.g., THINK^SELF, 'think for yourself'), affixation (SEE^ZERO, 'haven't seen'), inflection for aspect (GO-TO++, '(to) frequent'), telicity marking (cf. ARRIVE and WALK), and inflection for agreement ($_1$GIVE$_2$, 'I give you'), to which we turn now.

Discussion of directionality: In our discussion of the emergence of directionality (verb agreement) in young sign languages, we glossed over the perspective from the respective researchers that the emergence of this feature was from a reanalysis of extant verb forms. In the case of ISL, Meir (2012) argues that the particular trajectory of the emergence of directionality came from reanalyzing the 'ends' of the verbs as possible sites for agreement morphology. For NSL, it appears that younger signers took the modulation of verbs that show location to actually encode thematic (as opposed to spatial) relations.[31] Both cases are slightly different: the former carves new morphemes out of an arguably monomorphemic sign, while the latter reinterprets spatial markers as relational ones. As we only have a small comment to make on the latter, we'll start there first: this appears to be a very simple case of reanalysis, along the lines of Lepic & Padden's (2017) analysis above. One complexity is that these spatial modulations apparently did not come with a set function or form-meaning correspondence, leaving them wide open for reinterpretation.

With regard to the reanalysis of the endpoints of the verb as being slots for agreement markers, this is ostensibly a case of taking a simplex, potentially complex form and assuming it is complex. Meir (2012) cites the sign GIVE as an example of this phenomenon. In its non-inflected form, the sign is not directed at a locus, but moves along the Z-axis (i.e., from the signer's body straight outward). In its inflected form, the verb is ostensibly produced the same way, only the intended referent is established in the space to which the verb moves. The last step is the articular of the verbs towards a point not on the z-axis (e.g., to the side of the signer) to refer to 3rd person entities. Here, we might say that the raw material to produce an agreeing form were always present: the verb likely involves path movement, as GIVE denotes transfer and the path movement, a line, has two endpoints. While we might argue that the creation of this sign might have been iconically motivated (e.g., GIVE

[31] There's no description of the emergence of spatial modulation available (to our knowledge), so it remains an open question as to whether that arose from reanalysis or not.

isn't body anchored; GIVE's timing slots have two different specifications), the older cohorts do not use (or extend) this iconicity in production.

Telicity marking: As might have been clear from our discussion of telicity marking in sign languages, we take this phenomenon to be indicative of historical (or in the case of non-signers on-line) bottom-up process *par excellence*. Here, the 'pieces' do not come from the linguistic system proper, but are borrowed in from the visual domain. Again, we have gathered evidence from non-signer perception of non-linguistic events, non-signer perception of linguistic events (signs), signer and non-signer neurological behavior, linguistic analysis of signs from different sign languages, and kinematic data all pointing in this direction. Further, Strickland et al. (2015) show that the identity of the sign is not needed to be able to surmise different grammatical properties of those signs. For instance, non-signers classified signs matching in telicity but not in meaning over signs that mismatched in both, despite not having access to the (true) global meaning of the sign.

We make one final point on this matter, though: these features are recruited by the linguistic system in potentially varying ways, such that, e.g., different aspects of velocity are borrowed or, as we discuss more below, borrowed into different linguistic modules. In ASL, telicity marking is lexically specified, and so makes up the phonological component of telic verbs. On the other hand, in HZJ, end-marking seems to behave morphemically, deriving telic predicates from atelic ones. We want to stress that both are cases of compositionality, lest one confuse 'morphologial decomposition' with 'decomposition' generally.

With respect to pantomime: Circling back to pantomimes and the current research at hand, one motivation for this series of experiments and their analyses is to better understand the notion of non-decompositionality of pantomime. The argument, put forth by McNeill (2000, i.a.), is that the determination of the meaning of a pantomime is global, or *top down*. That is, the understanding of the individual parts of the pantomime are derived from the interpretation of the pantomime as a whole. Very few examples are given here or elsewhere as to why pantomimes are non-decompositional, but plenty are given of co-speech gestures, which–while ontologically very different (see Kendon, 2004 for a detailed overview)–may sometimes convey information in a strikingly similar way to pantomime. Wilbur and Malaia (2008) discuss one of McNeil's most referenced examples–the pantomimed dropping of a bowling ball down a drainpipe–and how a sign linguistic might tackle it.

We provide an example of our own: For instance, Goldin-Meadow and Brentari (2017), citing McNeill (1992) and Goldin-Meadow et al. (1995), argue that co-speech gestures are not decompositional. They offer as an example the use of finger wiggle to communicate the event *running*. Because finger-wiggle is used in other contexts (e.g., in their dataset, to offer someone two options) and is used inconsistently to mean *running*, they conclude that it cannot be used morphemically. While I agree that giving finger-wiggle morphemic status in co-speech gesture (and also pantomime) is unfounded, I nevertheless suggest that their conclusion biases us from discovering more elusive usage patterns.

Addressing the first issue, the non-specificity of finger-wiggle to *running* events, this ignores the wide-spread observation that a particular form may map onto several distinct meanings or have different uses. Brow-raise in ASL is an example, in that it occurs over topics, conditionals, yes-no questions, and so forth. These functions seem disparate enough to perhaps posit that brow-raise is not morphemic (by just this first criterion). However, Wilbur and Patschke (1999) provide a uniform analysis of brow-raise wherein brow raise occurs over the restriction of a [-wh] operator. The situation with finger-wiggle may be similar.

With respect to consistency, it is again not surprising that non-signers use different strategies to describe the same event, even internally, as different aspects of the event may be more (or less) discourse or perceptually salient each time. In action-naming tasks, we would not expect all participants to provide the same answer (as below; §3.2.1, so placing the burden on co-speech gesture or pantomime is unfounded. Just the same, we might look for consistency elsewhere, or note that non-signers consistently use a certain range of strategies and/ or consistently *do not* use some other range. For instance, in our gesture dataset, finger wiggle was used fairly consistently (within and across non-signers) to represent (a) internal movement and (b) atelic events (e.g., *fire flickering*, *plants blowing*, and *typing*), noting that precisely what the finger-wiggling was denoting differed considerably (e.g., flames, leaves, and fingers). Incidentally, this was also true of our signer's productions.[32]

For their part, Mark Aronoff, Irit Meir, Carol A. Padden and Wendy Sandler argue that holistic pantomime may be interspersed between signs in ABSL. They note:

> The individual signs contrast with pantomimic expressions in several ways: they are conventionalized, much shorter, confined largely to the hands (rather than involving the entire body), and express concepts that are members of individual lexical categories (e.g. noun, verb, modifier) and distributed accordingly in the syntax... In established sign languages, the individual signs are not holistic, but are instead each made up of a specific hand configuration, location, and movement, which pattern like the phonemes of spoken languages.

To which we say: we would argue, too, that an expressly syntactic, morphological or phonemic treatment is inappropriate for pantomimes such as they describe, and it is true that this contrasts with analyses available for 'real' sign. However, in pantomiming one has to select certain features (and not others) to portray, a mapping schema, and so forth. These considerations, we argue, come from somewhere.

To reiterate, I do not believe, nor do I have evidence to suggest that any feature of pantomime is used morphemically (or phonologically; Brentari et al., 2012). However, my argument is that some features may be used with enough regularity to guide production, perception, and even intuition (acceptability judgments). To take an extreme example, we might expect that no one would pantomime the consumption of an apple with just a pinky finger, that no one would perceive it as such, and

[32]One might also consider the lexical signs WAIT and ONCE-IN-A-LONG-TIME, or the addition of wiggling, aka USET (Unchanging State in Elapsing Time; Wilbur, 2003, 2008), to a sign like RUN-OUT to mean RUN-OUT-SLOWLY. Further, USET is used to fill pauses (think *um* in English) and encode the delayed completive.

further that it is intuitively 'wrong' to express this event in this way. A holophrastic/ reanalysis account puts the emergence of subparts/ morphemes squarely on perceivers, and does not necessarily take into consideration what strategies producers do (which we address in our study).

However, the biggest point that this line of thinking misses is that we have seen language evolve out of gesture (e.g., the NSL case; Goldin-Meadow et al., 2015; Brentari et al., 2012; Brentari & Coppola, 2013; Brentari et al., 2017). The process of going from a noncompositional whole to a sign with full morphophonological specifications is left partially unexplained. There must be some guidance to the producer (coiner) to choose some forms over others. There must be something in the gesture signal that can be analyzed by the receiver as compositional, such that this cue can be grammaticalized. As Wilbur and Malaia (2008) point out:

> Indeed, if one is to take seriously the argument that gestures are the forerunners to language, with mediation through sign languages, it is practically an imperative that some gestures are analytical in order to permit the development of sign language with its clear phonological and morphological structure.

We test this claim experimentally in Experiment 2. Experiment 2a asks nonsigners to guess whether a given pantomime is transitive, ditransitive, or intransitive unergative, or intransitive unaccusative. In Experiment 2b, participants are shown the 'meaning' of the pantomime and asked on a scale of 1-to-7 whether that meaning is captured by the form on the pantomime. Here, it is possible that participants correctly guess the transitivity of a pantomime while indicating that its meaning is opaque. In this case, thus, the (global) meaning of the pantomime cannot feed the interpretation of its parts, specifically those features that guide transitivity judgments.

A few last comments: We are excusing ourselves from a couple considerations here. Nowhere in our discussion have we mentioned, e.g., *headedness*, *stems*, or other terms a morphologist might look for. With respect to verbs, the head of a classifier construction is the movement root, and the head of a lexical verb is the verb itself. We do not assume that such terms are immediately amenable to the discussion of compositionality in pantomime, though they could be.

2.6 Summary of background

In the above, we hoped to have conveyed a few arguments. We repeat them here:

1. We review some of the visual cues to transitivity (distinctions) in sign languages and pantomime (where there was available evidence), noting that they are often ambiguous. What evidence was available from non-signer productions of argument-marking strategies argued for (a) pre-linguistic tendencies that may carry over into the development of a full language, and (b) pre-linguistic tendencies that morph as they edge closer (in)to a linguistic system. Regarding (a), we reviewed arguments that analyses (not just descriptions) of sign language agreement seem to extend to co/pro-speech gesture; and, regarding (b), we reviewed literature that showed that handshape and handshape complexity undergo a transformation from what non-signers produce to what homesigners produce, and finally to what signers of established sign languages produce. This comes with the caveat that, to our knowledge, no demonstrable link between these phenomena and the cognitive processes that may underlie them has been specifically, experimentally explored.

2. We demonstrated two cases where preexisting cognitive biases, abilities, etc. can be co-opted into formal linguistic systems in an iconic way: telicity marking in sign languages, and the emergence of word order.

3. We likewise showed that these iconic underpinnings can be profitably explored in non-signers, by examining their judgments on tasks in the source (cognitive) and target (linguistic) domains. The same can be done by examining their manual productions. For telicity marking, we reviewed literature that explored

boundary detection and kinematics in a non-linguistic judgment task, the theoretical and kinematic underpinnings of telicity marking in sign languages, and non-signer judgments about telicity marking in sign languages. With respect to word order, we reviewed literature on how word order may be a reflection of causal chains and event perception, generally, and how these chains might have become grammaticalized into languages (new and old; signed and spoken), specifically.

4. We further demonstrated how languages can move beyond these sources, due to pressures from a linguistic system. While telicity and event perception can both be described in similar kinematic terms, sign languages choose which kinematic sources to incorporate into the linguistic system. That is, (a) generally, the source material is sampled from and not taken wholesale, and (b) different languages may choose different elements. Further, while there are strong word order biases in language, generally, and sign language and pantomime, specifically, languages in the course of their evolution may choose to order constituents differently (or evolve free word order, among other options).

5. We also reviewed cases where a top-down explanation seems most convincing. The emergence of grammatical devices through the reanalysis of extant linguistic forms was demonstrated with recourse to the emergence and development of directionality in young sign languages. Sublexical access to iconic elements, at least with respect to lexical iconicity (e.g., recognizing the sign BIRD from the beak-like form of the sign), remains equivocal.

2.7 Proposal: Iconicity of argument structure and its emergence

We propose that argument structure–though it is variably coded in sign languages (across and within both lexical verbs and classifier constructions)–is nevertheless accessible through iconicity. We propose that it is accessible through both top-down and bottom-up processes, though to different strengths among categories–lexical verbs, classifier constructions, and pantomimes. We propose to test these claims by first asking non-signers to identify the argument structure of all three categories. If there is consistency, there may be iconicity. We probe this consistency by correlating it with lexical iconicity scores to gauge whether the identity of the sign aided in the identity of its parts (here, its argument structure). To assess the degree to which individual perceptual features add up to an iconic representation of argument structure, we annotate verbs for the phonetic features we identified in this chapter as being relevant to transitivity coding. We use a text classification algorithm to see whether one or an assemblage of features predicts non-signer transitivity judgments.

3. TRANSPARENCY OF TRANSITIVITY: CLASSIFIER CONSTRUCTIONS AND PANTOMIMES

3.1 Statement of problem, hypotheses, some (more) relevant background

How do grammatical features arise in new sign languages? To address this question, we explore how transitivity distinctions are or are not mapped to perceptual features in pantomimes (non-linguistic, iconic) and ASL classifier constructions (linguistic, iconic). We predict that argument structure is iconic and thus inferable from form in both, noting–of course–that pantomimes have not yet been shown to have formal syntactic features.

Further, we predict that the visual features that code transitivity iconically will be the same generally, though perhaps not specifically. That is, for example, it is possible that particular handshapes may differ between transitive pantomimes and transitive classifier constructions (as has been demonstrated in production tasks; e.g., Schembri et al., 2005; Emmorey & Herzig, 2003; Brentari et al., 2012), but the use of handshape to code transitivity will not. To that end, we are interested in whether gesturing non-signers do code and perceive transitivity distinctions and, if so, what explains this ability. Concerning this last point, we entertain two hypotheses: one in which transitivity information is accessed top-down, that is, in which the transitivity of a sign or pantomime is only accessed via its meaning as a whole[1]; the other in which transitivity information is available in the individual pieces of the pantomime or classifier construction, such that transitivity classing proceeds bottom-up. We note that classifier constructions, by analysis, are compositional and that pantomimes, by

[1]In two ways it seems inappropriate to talk about 'lexical iconicity' with regard to classifier constructions and pantomimes. First, we might argue that there *is no lexicon* for pantomimes, since they are–by definition–produced on the fly. Second, classifier constructions are multimorphemic, and are thus not lexical items. Going forward, when we say *lexical iconicity*, what we really mean is *the meaning of the whole form.*

hypothesis, are holistic. However, we do not know how non-signers perceive both. We assume for convenience of exposition that non-signers view both classes as singular words with or without internal structure.

The first hypothesis is championed by McNeil (1992 and subsequent) for co-speech gestures, and for the beginnings of sign languages (homesign system, pidgin sign, and young sign languages) by Senghas, Ozyurek, and Goldin-Meadow (2013). Lepic and Padden (2017) claim that certain components of meaning within ASL lexical signs are only accessible by analysis (i.e., top-down) and do not participate in sign meaning morphologically. The second hypothesis is championed by Emmorey (2014) and tested empirically by Strickland et al. (2015): For instance, Strickland et al. (2015) show that non-signers are able to make correct inferences about the telicity of signs in several unrelated (and one nonce) sign languages, irrespective of the meaning of those signs. Here, the identification of grammatical features is due to perceptual biases already present in the visual system, that are co-opted for linguistic categorization. The narrative with transitivity would be similar, only the source would be the visual-praxic domain, where objects, object affordances (via vision) and manual action all coalesce (Arbib, 2005).

First, though, given that there are no reported limitations on what a pantomime can be, it is important to survey the diversity of coding strategies and ask which strategies are most accepted by other non-signers. That is, we attempt to establish a 'ground truth' with respect to the form-meaning correspondence of pantomimes to achieve equal footing when we compare these forms to classifier constructions (and then to ASL lexical verbs in §5.1). This is the focus of Experiment 1.

To answer (a), we conduct a transitivity classing experiment, in which non-signers are asked whether pantomimes and classifier constructions are transitive, ditransitive, intransitive unergative or intransitive unaccusative. We find that non-signers class these forms, for which they have no (linguistic) experience, at a rate significantly greater than chance, suggesting that a model of transitivity can be built around visual features.

To choose between our phonetically-grounded and holistic hypotheses, we perform two analyses on our data from (a): To decide whether non-signers use a top-down strategy, we additionally collect iconicity ratings for both classifier constructions and pantomimes (following Vinson et al., 2008; Caselli et al., 2016) and correlate these scores with the consistency of transitivity classing. We also correlate accuracy scores (whether non-signers accurately classed pantomimes and classifier constructions), which tells us whether the production and perception models of visually-grounded argument structure are the same or different. We contend that a high correlation between consistency and iconicity score indicate top-down access to transitivity, with a high correlation between consistency and accuracy indicating further that the model of transitivity non-signers build approximates the production strategy of (a) other non-signing pantomimers and (b) the actual argument structure of classifier constructions.

To decide whether non-signers use a bottom-up strategy, we annotate our corpus of classifier constructions and pantomimes for phonetic features thought to be relevant to transitivity distinctions in sign languages. We then run these features through a machine learning algorithm, using both non-signer derived and ground-truth labels, to see whether any feature or set of features reliably codes transitivity information.

We make the following predictions: we predict that due to the lack of a linguistic system coercing forms to become more consistent and thus more arbitrary, pantomimes will be classed more consistently than classifier constructions. For these same reasons, we anticipate that (a) pantomimes will be classed more accurately than classifier constructions, (b) pantomimes will be regarded as more iconic than classifier constructions, (c) we will find higher correlations between non-signer agreement (consistency) and/ or accuracy and iconicity scores in pantomimes over classifier constructions, and (d) phonetic features will more reliably code transitivity distinctions in pantomimes than in classifier constructions.

Again, in what follows, we first outline our two material preparation experiments: we discuss our pantomime and classifier construction elicitation methods and discuss our 'best pantomime' study in §3.2.2. The transitivity classing study is reported in (§3.3) and the iconicity rating study in (§3.4). Our main analyses, the explanation of non-signer classing behavior via top-down or bottom-up strategies, are reported in §3.5.1 and §3.5.2, respectively. We discuss our results in §3.5.2.

3.2 Stimuli Creation

3.2.1 Experiment 1a: Action Naming

The goal of Experiment 1a is to establish that consistent event-labeling is achievable using our action video stimuli. Importantly, this provides us with a defensible ground truth, in that the classification of actions does not fall on the experimenter. Here, participants watched videos of action clips and then provided sentences (referred to as 'labels') that describe those actions.

Materials: One hundred eighty-five short action videos were produced, of which 80 were chosen for inclusion in the study. Forty of these videos were intended to be transitive and 40 intransitive. Of the intransitives, the majority (37/40) were unaccusative. Videos depicted actions recorded in a laboratory setting, performed with various objects. The performer appeared in each video, and was the agent in the transitive videos (i.e., used, and or acted on objects), and witness (i.e., watched actions unfold) or sole participant in the intransitive videos. To ensure that the proportion of 'transitive' responses to transitive videos would be robust, the agent's full body (or minimally, the torso) was present in the frame (Rissman et al., 2016; Horton et al., 2017; see also discussion of view-point perspective in Perniss, 2007 and Cormier et al., 2012). Stills from a transitive (a) and intransitive (b) video are shown in Fig. 3.1.

Note also that we wanted to avoid events involving two potential agents, events with abstract objects, and so on, per our discussion in §2.4.2 of the variability and subsequent difficulty in interpretation these elements bring. As a point of fact, though,

(a)

(b)

Figure 3.1. Two stills from the action video dataset. (a) depicts a sample from the transitive dataset, 'adjust picture,' and (b) depicts a sample from the intransitive dataset, 'race car drive into box.' Note that the agent (and his torso) are present in both transitive and intransitive videos, as it has been demonstrated that perspective–which may vary as a function of how much of an agent is present in a scene–may determine whether a handling or entity strategy is elicited in classifier constructions.

we do not know how these elements may affect verb exponence, as we are not chiefly interested in constituent ordering (or elements outside of the verb) here.

On practical grounds, videos where the object or action was difficult to see were excluded. Videos that were too long (>5 s) were also excluded since it is assumed that increased length of experiments risks an increase in inattention or attrition (both for the labeling experiment and pantomime elicitation task). Computational processing loads associated with loading many longer (and therefore larger) videos in a browser were also considered.[2]

[2]There were additional theoretical selection criteria. As these actions would be pantomimed in the subsequent experiment, actions for filming were selected according to a few criteria and *a priori* hypotheses about coding transitivity in pantomime. We first predicted that, like in ASL, handshape could be gainfully employed for transitivity coding. To this end, we varied the types of objects used in an effort to elicit different object and handling handshapes. However, just as handshape is not enough to reliably code transitivity (Kimmelman et al., 2016; He & Tang, 2018), we predicted as much for pantomime. As such, we additionally hypothesized that contact with or passage through a (projected) plane would be used to code transitive actions. Pantomimes depicting actions occurring along a plane, or stationary actions–we hypothesized–would be more likely classed as intransitive. Further, we predicted that the role of the second hand would be informative (Lepic et al., 2016), so we chose actions where the second hand could contribute a ground (e.g., *going in a box*; intransitive action), could be holding an object that the dominant hand acts on (e.g., *unscrewing a lid*; transitive

Participants: Sixty participants were recruited on Amazon Mechanical Turk (AMT) and compensated $1.00 for completing the survey. Participants' IP addresses were limited to the United States to increase the likelihood that they had some English proficiency. Proficiency was sought so that participants would be more likely to understand the instructions and provide desirable English labels to the stimuli. All participants indicated that they were fluent in English and had normal or corrected-to-normal vision.

Design: Eighty action videos were included in the experiment, 40 intransitive and 40 transitive. To ease workload on participants, the videos were divided into two sets of 40 videos, with 20 transitive and 20 intransitive videos each. Videos were randomized before being assigned to a set. Thirty participants saw one set of 40 videos, and 30 the other set. Participants saw one of two orders, where Order B was the reverse of Order A. This was done to assess the effect of video presentation order, should we have reason to believe that this factor was skewing our results.[3,4] The designs of all of the experiments in this dissertation were prepared using Turktools (`turktools.net`), which is a suite of python scripts designed to prepare studies in AMT-readable formats. To note, the *lister* function breaks a mother list of stimulus items into smaller lists, randomizing and percolating items into counterbalanced Latin Square designs.

Procedure: Participants first read the instructions. These included a list of criteria for sentence labeling and two example items, each with a set of acceptable and unacceptable sentences.[5] Afterwards, participants immediately started the task. For each

action), and so on. We chose actions that might elicit such strategies. Videos using redundant strategies were not selected. A full list of these actions can be found in Appendix A.1.

[3]For clarity, the total number of stimuli lists was four: a given participant saw one of two possible orders of one of two possible video lists.

[4]To note, we did not have any reason to suspect presentation order had an effect.

[5]Specifically, the instructions asked participants to use one sentence to describe the action they saw in the video. Participants were instructed to use terms like 'Object A' or 'An object' for items they could not see well enough to name, or otherwise didn't know the name of. A few criteria for acceptable sentences are given here below.

i. Please use complete sentences.
ii. Please do not use more than one sentence.

item, participants watched an action video and then provided a sentence describing what they saw. Videos could be replayed. All items were presented on a single page. Participants indicated their consent to take the survey by clicking appropriately after being presented with a consent form at the end of the survey.

Coding & Results In total, 2,400 labels (i.e., sentences) were produced, of which 2,256 were kept and classed as intransitive or transitive. Three participants from one set were disqualified for providing sub-par labels,[6] but all other participants' responses were kept ($(27 \times 40) + (30 \times 40) = 2{,}280$ labels).

Individual responses were classified as either transitive, intransitive, or other. Transitive sentences all contained a transitive verb and a direct object (plus or minus adjuncts). Ditransitive sentences were coded as transitive sentences. Intransitive sentences contain just an intransitive verb plus or minus adjuncts. Sentences coded as *other* used an ambitransitive verb without an object (e.g., *The man ate [an apple]*), or misidentified the event (e.g., misclassing an eating event for a drinking event) even if the transitivity of the verb used was appropriate. Sentences classed as *other* were excluded from the analysis (n=24, or 1.06% of the dataset).[7]

Here, we simply wanted to test whether a given video received a given label (transitive, intransitive or other) significantly more than chance. As such, we used a one-tailed, one-sample test of proportion against the hypothesized chance mean (here,

iii. Please do not use compound sentences (i.e., two sentences joined by the word *and*, *or*, or *but.*).
iv. Please keep your sentences as simple as possible.
v. Please make your sentences true to the video.
vi. Please do not write irrelevant sentences.

Participants were told that failure to comply with these criteria on three or more videos would disqualify their answers and they would not be paid. Although many participants broke rules (i), (ii), and/ or (iii) consistently, their answers were kept and they were paid so long as their sentences were relevant to the video and the transitivity of their sentences was clear. In rare cases where more than one clause was provided, the transitivity of the first clause was taken to be representative.

[6]Here, the labels were either too vague (e.g., *An object moved*) or were completely absent.

[7]In some cases, participants responded to intransitive actions with indefinite subjects (e.g., *someone or something did X*) or passive sentences (e.g., *Y was X'ed*). Both these cases were counted as transitive. As transitive videos were equally likely to elicit passive sentences, we could not justify classing passives as intransitives.

50%). We used $\alpha = 0.05$ as our cut off. Seventy-five videos met this threshold; five did not.

Paring down results: Again, the goal was to determine which of 80 videos were highly transitive and highly intransitive for eventual use in Experiment 1b, where 'highly transitive' and 'highly intransitive' mean 'videos that were significantly likely to be labeled with transitive or intransitive verbs.' Seventy-five met this criteria. As 75 is an odd number, we pared the list down to 72 actions, 36 transitive and 36 intransitive.[8] For the most part, this was done by choosing the videos with the highest participant agreement. However, we chose to include some videos that did not meet this requirement (2), while excluding others that did on (5) the following grounds: some actions were converses of other selected actions (e.g., the video for bubbles fizzing up was excluded because a video of bubbles fizzing down was selected); some were thought to elicit ASL lexical verbs instead of classifier constructions (e.g., writing on a whiteboard); and so one.

Discussion: For the most part, videos were labeled in agreement with their intended transitivity. This is so despite the fact that the actor in the transitive videos was also present in the intransitive ones, indicating that the presence of a potential (but not actual) agent does not greatly interfere with the labeling of intransitive actions. There were two videos, however, that participants rated as intransitive where a transitive label was expected. In both cases, the desired verb was *approach* (*The coat rack approached the man* and *The man approached the coat rack*), but the most common strategies were to (a) completely ignore the second event participant or (b) relegate that participant to a satellite. As such, these action videos were reclassified as *intransitive.* There were no other anomalies in the results.

As there was no specific question for this transitivity-verification study, no further analyses were conducted. Again, the result here was 72 action videos that were

[8]Seventy-four is also divisible by two, but the quotient, 37, is not further divisible. We envisioned that this would result in surveys with odd numbers of transitive or intransitive items down the road. Seventy-two and its factors are further divisible.

consistently labeled transitive (36) or intransitive (36). We have grounds to assume, then, that these transitivity distinctions will be salient to subjects participating in the manual coding of these actions.

3.2.2 Experiment 1b: Best Pantomime

The goal of Experiment 1b was to derive a set of ‘best pantomimes’ for use in Experiment 2, using the action videos that were consistently classed as transitive or intransitive from Experiment 1a. Given reported variability in pantomime production (e.g., Schembri et al., 2005; Goldin-Meadow et al., 2015; Brentari et al., 2012), we wanted to elicit a number of pantomimes, and from that diversity select only those that non-signers decide are most representative of the events they describe. This is akin to a grammaticality judgment experiment, but for paralinguistic stimuli.

Pantomime elicitation: Pantomimes were elicited from six sign-naïve participants with no significant acting experience (3 female; 27-35 years old, mean 31.66). Five gesturers were graduate students at Purdue University; one was a friend from the community. Five were born and raised in the US, one was born and raised in Wales but had been living in the US for a few years.[9] All consented to being filmed and to have their videos distributed online as part of Studies 2a,b. Participation was unpaid. The elicitation and subsequent use of participants’ videos were cleared by the Internal Review Board of Purdue University, West Lafayette.

Participants were seated in front of a blue backdrop, with a laptop computer situated at their left. Participants were instructed to hit a key on the laptop to view each video clip (played one at a time), signal to the experimenter that they were ready to perform, and produce silent gestures of the actions they saw. Specifically, participants were told that the experimenter was familiar with each of the videos they’d be watching and that they should communicate to him which video it was. In this way, participants were encouraged to gesture in such a way as to convey

[9]We do not anticipate that these or any other demographic measures had an effect on productions.

(a) (b)

Figure 3.2. Two stills from the pantomime dataset. (a) depicts a sample from the transitive dataset, ‘adjust picture,’ and (b) depicts a sample from the intransitive dataset, ‘race car drive into box.’

meaningful information, rather than to (a) provide too much extraneous information or (b) just mimic the actions in the videos. Otherwise, participants were free to pick out any aspect of the action in the video to portray.[10] Participants were additionally asked to make their productions below the top of the head, above the waist, and within one foot of each shoulder (i.e., roughly the signing space employed by native ASL users). Participants could practice or re-shoot any of their pantomimes. Filming took between 15 and 30 minutes for each subject.

An additional elicitation session was conducted with a native signer (currently in her 60's; learned ASL at birth from Deaf parents and went to Deaf school) to elicit semantically matched classifier constructions. The signer was filmed in her home by another signer. The signer was asked to produce full sentences in response to the action videos, pause, and then reproduce the verb she used in the sentence in isolation. The signer was also explicitly asked to try to use classifier constructions.[11] Elicitation was otherwise identical to that of the non-signers. However, the signer provided two options for one of the action videos, which resulted in a total 73 classifier constructions. To note, these classifier constructions were not included in Experiment 1, but were used in the second study, detailed in §3.3.

The result of the elicitation sessions was 432 pantomimes and 73 classifier constructions. Since we wanted to find the 72 best pantomimes for inclusion in Studies 2a and 2b, we percolated the pantomimes into another study (described below), which had sign-naïve participants rank the pantomimes according to their faithfulness to a sentence that described the action that generated them.

'Best' production study design: There were in total 72 unique items. Each item consisted of a sentence printed at the top of the screen, with six pantomime videos below. To ease workload, and mitigate problems with holding $72 \times 6 = 432$ videos in

[10]This was to maximize the diversity of responses. A diversity of responses may increase the likelihood that an underlying preferred transitivity-coding strategy emerges and may also increase the number of features potentially relevant to transitivity classification in the machine learning analysis (§3.5.2).

[11]The signer is an ASL instructor and has access to linguistic terms like classifier construction.

(a)

(b)

Figure 3.3. Two stills from the classifier construction dataset. (a) depicts a sample from the transitive dataset, ‘adjust picture,’ and (b) depicts a sample from the intransitive dataset, ‘race car drive into box.’

memory (virtual and cognitive) the survey was split into eight smaller surveys. Nine items of each survey were unique, but two items were shared with two other surveys for inter-survey reliability. In total, then, each survey had 11 questions. Thirty participants were assigned to each survey, for a total of 240 participants.Items were randomly percolated into each survey. We assume that participant rankings were independent across items.

Procedure: Participants read the instructions[12] and were presented with an example. Afterwards, participants immediately started the task. For each item, participants first read an intransitive or (di-)transitive active declarative sentence. Each sentence was derived from sentences provided in Experiment 1a. Specifically, for a given item, the verb that appeared most (i.e., simple majority) was selected. Other event participants were selected identically. Adjuncts (like *into the box* in *The car drove into the box*) were included even if they did not appear in the majority of sentences from Study 1a, as this information was conveyed in many of the pantomimes (e.g., pantomimes often included goal information).

Then, participants watched six pantomime videos. Each pantomime video within an item was produced in response to the same action video (e.g., all videos were pantomimes of *put down book* or *book fall*, etc.). Participants used a series of drop-down menus to indicate their ranking of the six pantomimes, with '6' being the best, and '1' being the worst. Participants were instructed to only use each number once.

Several measures were put into effect to ensure that participants viewed all six videos per item. To be sure that participants were not simply choosing a favorite pantomimer, a pink poster was placed in front of each video, which disappeared only after a video's play button was hit. Further, for each item, video order was randomized, such that, e.g., Pantomimer 1 was the first video in Item 1, but the 4th video in Item 2, and so on. An attention check item was also included to prevent

[12] The instructions included an abbreviated description of the task, and a few rules. Specifically, participants were told that they were not to pick the best overall pantomimer, but to pick the best pantomimer for each item. Participants were also told to ignore pantomimers' race, ethnicity, etc. and technical aspects of the video (e.g., lighting, resolution, etc.).

against random rank assignment. The attention check video was a 4 s long video with white text against a plain pink background that read 'please rank this video a 3.' Lastly, the participants had to provide some basic justification for choosing their ranking, including what made good pantomimes good and bad pantomimes bad.

All items were presented on a single page. As before, participants indicated their consent at the end of the survey. Consent was logged as participants clicking the 'submit' button to submit their responses.

Participants: Participants were recruited on AMT and were compensated $1.00 for completing the survey. Participants' IP addresses were limited to the United States to increase the likelihood that they had some English proficiency. This was done purely to increase the likelihood that the instructions were understood. The experiment was run twice, the first time with 80 participants, the second with 160. This was due to a high rate of subject and response rejection. See below.

Data preprocessing: Before analysis, the data were first scrubbed of undesirable items and participants. As the scrubbing left few participants/ responses for analysis, additional data were collected. If a participant took the survey twice (n = 3 subjects), data from the first attempt was discarded. Then, data from participants admitting to knowledge of a sign language beyond the manual alphabet and a few signs were completely discarded (n = 24 subjects). Second, a participant's response to a specific item was discarded if (a) there were two or more blank responses[13] (n = 1 item) or (b) if two ranks were identical (e.g., assigning two videos a rank of '1'; n = 21 items). Participants providing two or more undesirable responses were excluded entirely (n = 3). Further, data from participants failing the foil trial (i.e., failing to assign the foil video a '3' rank) were excluded (n = 12 subjects; replacement subjects were recruited in this case). 2,217 rankings remained for analysis.

[13]Responses were kept if only one rank was missing as its value could be deduced from process of elimination.

Analysis: We asked whether there is consistency across non-signers as to which pantomime best represents an action. That is, we were after what could be considered an 'acceptability judgment' for pantomimes. For this, we summed up rankings for each video for each item. Next, we identified the lowest scoring video and compared its response vector to the hypothetical chance mean. This mean was calculated as $((1+2+3+4+5+6)/6 =)$ 3.5.

Results & Discussion: Thirty-four transitive videos and all 36 intransitive videos met our statistical threshold. The top results of this analysis are presented in Tabs. 3.1 and 3.2. The remaining results can be found in Appendix A.2.

Here we simply asked whether non-signers have intuitions about what constitutes an appropriate gestural representation of a linguistic description, where that description varied with respect to transitivity. We did achieve consistency, but it remains to be seen whether that consistency is a result (even in part) of faithfulness of the gesture to the transitivity of its label.

There are a few more justifications to make. For one, the one-sample t-test of proportion is likely the weakest test we could apply. Again, we could instead compare the means of the two most frequently selected options per item, keeping only those items where the most frequently selected option was selected significantly more frequently than the second most frequently selected. However, one conceptual problem with this analysis is that it potentially pits two pantomimes that employ roughly the same strategy against each other.

One could, then, avoid the problem by collapsing individual pantomimes into classes according to whatever depictive strategy was being used, and then make comparisons between strategies. Then, the pantomime ranked highest within the category would be selected. For example, if three pantomimes used a handling handshape strategy and three used an entity handshape strategy, we sum the responses for each groups and make a comparison between groups as opposed to individual productions. If the handling strategy group won, for example, we would then choose the production that was lowest ranked.

Table 3.1.
Top 20 intransitive pantomimes. Our exclusion criteria and the repetition of certain items across surveys resulted in items with unequal numbers of participants (here, ranging from eight to 53). As such, notice that the degrees of freedom vary considerably.

Pubj	Item	t(df)	p (1-tailed)
IP	The man bowed at the waist.	t(7) = inf	0.0000
NP	The coat rack moved toward the man.	t(21) = -20.3516	0.0000
HO	The paper airplane landed on the table.	t(34) = -10.6122	0.0000
HO	The box moved across the table.	t(23) = -11.9345	0.0000
NP	The hanger swung [...] on a coat hook.	t(27) = -10.4003	0.0000
IP	The bread spun.	t(52) = -7.6433	0.0000
HO	The box slid across the table.	t(22) = -9.625	0.0000
NP	The cards scattered everywhere.	t(49) = -7.2427	0.0000
HO	The jar of bottle caps spilled over.	t(23) = -8.9842	0.0000
HO	The toy crawled up the incline.	t(20) = -8.7881	0.0000
CM	The man walked backwards.	t(25) = -7.5222	0.0000
HO	The fan oscillated.	t(25) = -7.4386	0.0000
HO	The poster rolled up.	t(24) = -7.5056	0.0000
NP	The ball bounced.	t(25) = -6.6966	0.0000
IP	The bowl broke.	t(24) = -5.5807	0.0000
RVN	The toy skittered on the table.	t(25) = -4.5455	0.0001
IP	The microwave door closed.	t(22) = -4.6126	0.0001
HO	The light turned on.	t(24) = -4.3339	0.0001
NP	The drink bubbled down.	t(25) = -4.3028	0.0001
HO	The book fell over.	t(28) = -4.152	0.0001

Table 3.2.
Top 20 transitive pantomimes. Our exclusion criteria and repetition of certain items across surveys resulted in items with unequal numbers of participants (here, ranging from eight to 53). As such, notice that the degrees of freedom vary considerably.

Item	t(df)			p (1-tailed)
The man took the keys out of the box.	t(26)	=	-10.1114	0.0000
The man took the lid off of a jar.	t(29)	=	-9.3467	0.0000
The man dipped his finger into the jar.	t(51)	=	-7.9089	0.0000
The man tightened the string.	t(25)	=	-7.8335	0.0000
The man hammered the nail.	t(56)	=	-7.085	0.0000
The man plugged in the charger.	t(27)	=	-6.6938	0.0000
The man uncorked the wine bottle.	t(24)	=	-6.6842	0.0000
The man turned the fan.	t(27)	=	-6.1325	0.0000
The man pulled out the measuring tape.	t(29)	=	-6.0937	0.0000
The man rolled the tube back and forth.	t(27)	=	-5.5794	0.0000
The man hit the bottle with a ball.	t(28)	=	-5.4306	0.0000
The man shook the shaker.	t(20)	=	-5.2796	0.0000
The man pushed a button on the microwave.	t(25)	=	-5.1657	0.0000
The man put the cup on a coaster.	t(25)	=	-5.0883	0.0000
The man spun the bread.	t(27)	=	-4.8364	0.0000
The man closed the microwave door.	t(23)	=	-4.7718	0.0000
The man bounced a ball.	t(17)	=	-4.6048	0.0001
The man broke the stick.	t(26)	=	-4.3239	0.0001
The man lit a candle.	t(50)	=	-3.9822	0.0001
The man cut the bread in half.	t(21)	=	-3.8537	0.0005

However, there were at least two challenges with this approach. The first is that this assumes that we know what the strategies are *a priori*. We could, in keeping with the example above, build a category around handshape strategy (e.g., handing *vs.* entity). But, we argue that this imposition of linguistic theory does not necessarily reflect how non-signers rank. For instance, three handshape strategies occurred in the pantomiming of the event *adjust picture*: one involving two claw handshapes, one involving two L-handshapes, and one involving a B-handshape. The first strategy represents how one would manipulate a picture (handling strategy); the second represents the shape of the frame of the picture as it moves (size and shape strategy, potentially); and the third represents the picture in its entirety and its movement (entity strategy). However, the productions were also made in different areas of the signing space, involved different movements, and so on. While the handshape strategy is salient to sign linguists, other strategies may be more important to participants. Further, while collapsing handshape strategies might work in this example, it may not for other pantomimes: we would need to devise and justify buckets for upwards of 72 pantomimes. Rather, we see that in future studies, the features that come to light in our machine learning analysis, if any, could then guide bucketing.

Second, we are looking for phonetic correlates of transitivity distinction, i.e., predictors at lower levels (e.g., perception). Representational strategies, by contrast, are higher-order. While much of what non-signers produce can be usefully described in higher-order terms, it is not known whether non-signers perceive pantomimes in these terms. To demonstrate our points we provide a few representative criteria participants listed as relevant to their selections in (1).

(1) a. "The person's mood also was a factor. Those who just had this "whatever" look on their face were not as good as those who got into the action."

b. "I looked for videos that showcased the persons appropriate facial expression, hand gestures that related to the charade, and mouth movement."

c. "If the description said that "a man" did something, it was more likely that I would select a man, as it fit the description better."[14]

d. "I looked for clear motions of depiction. I looked for appearances of being involved and putting effort into the motions."

The winning pantomimes were then percolated into Experiment 2a, which tests the availability of linguistic constructions (here transitivity) in non-linguistic (pantomime) and linguistic (classifier construction) stimuli.

3.3 Experiment 2a: Transitivity Classing

The result of the pre-studies was a set of pantomimes that non-signers agree are the best available pairings of form and meaning. This puts these pantomimes on roughly equal footing with ASL lexical verbs and classifier constructions, which have linguistically codified form-meaning correspondences.

The goal of Experiment 2a is to see whether non-signers have consistent intuitions about perceived transitivity of these linguistic and paralinguistic forms (§3.3). We then ask where these intuitions stem from: top-down processing, via the lexical iconicity of the forms (§3.5.1), or bottom-up processing, via their phonetic characteristics (§3.5.2). We collect our measure of lexical iconicity for pantomimes and classifier constructions in Experiment 2b (§3.4).

3.3.1 Methods

Participants: As with Experiment 1, participants were recruited on AMT. They were compensated $1.00 for completing the survey. Again, participants' IP addresses were restricted to the United States to increase the likelihood that responders had some English proficiency in order to understand the instructions.

[14]Note that the instructions specifically asked participants not to rank based on (perceived) gender identity.

Materials: All classifier constructions (73) and 71 of the 72 winning pantomimes were included for a total of 144 experimental items.[15] We split the survey into four subsurveys, each with a complementary set of 36 items (37 for one of the surveys). In addition, two to three comprehension videos and one foil video were included per survey to bring the total number of items to 40. The comprehension videos were all live-action videos (either filmed by the experimenter or downloaded from YouTube) depicting a tower collapsing (intended response: *intransitive unaccusative*), a man hammering a nail into a box (*transitive*) and two women exchanging business card (*ditransitive*). The foil item was a video displaying text that read 'Please select answer (b).' Again, posters covered the videos so participants could not discover which items were comprehension or foil items at a glance. The lengths of these videos fell within the range of the other stimuli (between 2 and 4 s), so participants would not be able to identify these trials this way.[16] While not airtight, this helped ensure that at least all videos were played, if not also attended to.

Survey design & implementation: In this experiment, we indirectly ask sign-naïve participants to guess whether a given manual action (pantomime or classifier construction) is transitive, ditransitive, intransitive unaccusative, or intransitive unergative, using the descriptions below. In each case, a single English sentence with several verb options was provided in order to calibrate participants to how we intended the labels to be used.

(2) a. Someone/ something is acting on someone/something else

Ex. *The person is grabbing/ picking up/ hitting/ squeezing/ kicking/ dropping the ball*

b. An object changes possession or is placed somewhere

Ex. *The person is giving/ taking/ passing/ borrowing/ stealing the ball*

[15]One winning pantomime, *approach coat rack*, was not included in the stimuli. It was not excluded for any particular reason; it was simply forgotten.

[16]This is additionally relevant as these comprehension/ foil items were shared with previous AMT studies and pilots. Participants participating in earlier studies would theoretically be able to identify comprehension/ foil videos by length if meaningfully different.

to/ from the other person

Ex. *The person is placing/ setting/ putting the ball on the table*

c. Something changes shape or location by itself

Ex.: *The ball is moving/ falling/ rolling/ deflating/ exploding/ wilting*

d. Someone is performing an action without an object

Ex.: *The person is walking/ running/ singing/ dancing/ whistling/ sneezing*

An example was also given: A video depicting a live action video of a man crushing a soda can was presented with the option *(a) Someone/ something is acting on someone/ something else* pre-selected. An explanation (*Here, we chose option (a) since someone (the man) is acting on (crushing) something (a soda can).*) was provided, as well as a reminder that the remaining videos would all depict pantomimes *without* accompanying objects. To avoid any unanticipated bias in responses, we simply told participants that all videos were pantomimes (charades), rather than explaining that some were actually classifier constructions/ ASL signs.

To note, although there are four transitivity class options, the classifier constructions and pantomimes were largely transitive and intransitive unaccusative, with very few intransitive unergatives and zero ditransitives. Participants were advised, thus, that not all options were equally likely.

Finally, all 40 items were presented together on a single page. Participants were confronted with a consent statement at the end of the survey. Consent was considered granted when participants hit 'submit' to submit their responses.

Data preprocessing: Data were downloaded from AMT and preprocessed through a suite of python routines. As before, data from signers were discarded (n = 17). Data from participants failing the foil trial (n = 1) were also discarded, but were replaced by recruiting an additional subject. Next, data from participants who did not perform as intended on more than one comprehension trial were excluded (n = 10). In total, then, data from 27/96 participants were unusable, leaving data from just

69 participants. Fortunately, excluded subjects were equally likely in each subsurvey: seven were excluded from Subsurvey 1 (leaving 17), eight from Subsurvey 2 (leaving 16), five from Subsurvey 3 (leaving 19) and seven from Subsurvey 4 (leaving 17).

Data from remaining items and subjects were further analyzed. We use a one-sample t-test of proportion against a chance (=25% of each transitive, ditransitive, etc. responses) distribution as a measure of consistency. Specifically, we first identify the top scoring response (using a simple tally), assign that response a value of 1, and zero-out all remaining responses. We compare this vector of 1's and 0's against the chance vector, which is matched in length but has precisely 25% 1's and 75% 0's. This generates a t-value for each item. This analysis addresses the question *Do non-signers have intuitions on the transitivity of lexical signs?*

To note, we consider chance to be 0.25 for the following reasons: if there is no transitivity information inferable from a given verb, participants are equally likely to choose any of the four labels. We acknowledge that informing participants that there may not be an equal distribution of stimuli of each type may have biased them away from a proportional assignment of labels. However, we do not assume that participants were biased towards choosing mostly transitive and intransitive unergative labels and not others.

Finally, it was discovered after the survey that one of the classifier videos, *shake shaker*, was incorrectly clipped and instead displayed our signer at rest between two items. This item was excluded from analysis, leaving 143 total items for analysis.

3.3.2 Data preprocessing

We discovered a very strong intransitive bias in our results and originally entertained an 'intransitive until proven otherwise' classing strategy. However, we note that this is inconsistent with the results obtained for the ASL-LEX study, presented in the next chapter (§4.3.2). The participants in that study likely did not employ such a strategy, despite lexical verbs being even more opaque with respect to their lexical meaning and (probably) argument structure (see the next section; §3.4).

Instead, we entertained the possibility that some participants or block of participants (i.e., and not all participants, generally) responded to the survey in a way that was different from what was expected. To assess this, we calculated each participant's mode response. We then divided the frequency of this response against the participant's entire response distribution, to obtain the what proportion that mode response constituted.

We found that 24 of the 98 participants[17] had mode responses that represented more than 85% of all of their responses (including responses to foil and comprehension trials). Of these 24 participants, 23 (95%) of them chose answer '4' (*Someone is performing an action without an object*) the most. Only 1 out of the 24 (5%) chose '1' as their default answer. While we assume that it is because the wording of option 4 was the most general of all of the options that most of these participants defaulted to it, we have no solid explanation for the one participant who defaulted to option 1.

As such, we decided to run a post-hoc analysis of the two populations we observed: the 'target group'–or those who responded to the survey as we intended–and the 'bias group,' who answered with a single response for most items. We refer to the union of the target and bias groups as the 'whole' (or sometimes 'full') group. In what follows, figures, tables, and discussion refer to performance by the target group. Information concerning the bias group will be mentioned in less detail and in footnotes.

To note, a different number of non-conforming participants were excluded from each subsurvey: two subjects were excluded from subsurvey 1, nine from subsurvey 2, seven from subsurvey 3, and six from subsurvey 4, leaving 22, 15, 17, and 18 subjects in each subsurvey, respectively. After removing comprehension- and foil-failing subjects, some items had as few as nine ratings, while most had between 13 and 15. Perhaps surprisingly, none of the 'bias group' participants could be excluded for fail-

[17]For clarity, these 98 participants represent those who were not excluded by the criteria outlined in §3.3.

Table 3.3.
Post-hoc: Tallies of consistently classed classifier constructions and pantomimes, where consistency is defined as maximum votes that were chosen significantly above chance (at $\alpha = 0.05$). Both classes of stimuli had well over chance (=25%) rates of consistent responses, indicating that participants had some model of transitivity.

	Classifier Constructions	Pantomimes
Transitive	25	27
Ditransitive	1	4
Intransitive (E)	5	2
Intransitive (A)	13	10
Total	**44**/72	**43**/71
% dataset	61.11%	60.56%

ing comprehension or foil trials. For the analysis on data from all study participants, see Appendix B.

3.3.3 Results & Discussion

Of the 144 pantomimes and classifier constructions, 87 were classifiable as transitive, ditransitive, intransitive unergative or intransitive unaccusative according to our criteria. Of the 87, 44 were classifier constructions and 43 were pantomimes. As such, significantly more items were classified than chance ($p < 0.0001$ for each).[18] This suggests that classifier constructions and pantomimes are iconic with respect to their transitivity. We explore how this may be the case in §3.5.1 (top-down explanation) and §3.5.2 (bottom-up explanation).

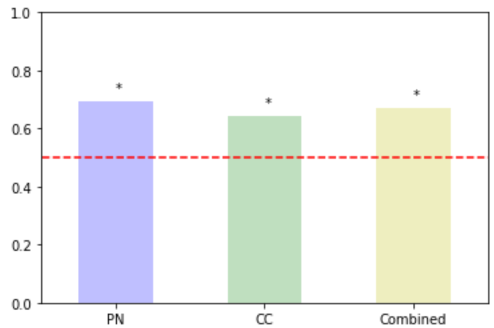
[18] Here we used a binomial test with probability $p = 0.25$.

The breakdown of classifier constructions and pantomimes into transitive, ditransitive, etc. classes is presented in Tab. 3.3. For both classes of stimuli, there was a bimodal distribution, with most stimuli being classed as transitive or unaccusative. The composition of the stimuli themselves may explain this: As most of the input videos (i.e., the action videos) were either transitive or intransitive unaccusative, with few intransitive unergatives and zero ditransitive videos, it is unsurprising that the classifier constructions and pantomimes depicting these actions were generally classed as transitive and unaccusative.

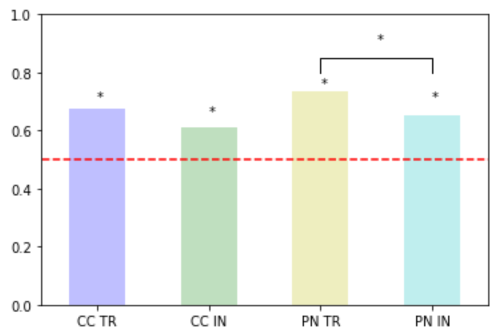
Target group, Accuracy: To assess accuracy, we decided to bin responses into 'transitive' and 'intransitive' categories, where *transitive* meant an action involving one or more objects (transitive and ditransitive stimuli) and *intransitive* meant an action that does not involve an object (intransitive unergative and intransitive unaccusative stimuli). We did this due to the low incidence of unergative stimuli in the action video dataset (3) and the zero incidence of ditransitive stimuli. Further, given that participants consistently chose unergative labels for unaccusative stimuli, perhaps due to an agency bias, accuracy might be artificially low.

Accuracy was measured in two ways. In the first method, we assessed accuracy by consensus: items were categorized as 'transitive' or 'intransitive' (a) by simple tally (i.e., an item gets more 'transitive' than 'intransitive' labels or *vice versa*), and (b) by our measure of consistency (i.e., an item gets significantly more 'transitive' than 'intransitive' labels or *vice versa*). This second method allows us to assess accuracy at the group level, should individual performance be low. This analysis also allows us to examine biases in non-signer classing at a manageable granularity, as creating confusion matrices (which plots non-signer guesses against ground truth) for individual items would be too time and space consuming. Further, assessing bias on a per-item basis might not yield interesting or generalizable information.

In the second, individual responses for an item were classed as 'hit' (1) or 'miss' (0), then the responses were averaged to get a percent correct figure. This allows us



Comparison	t(*df*)	p
PN vs. chance	t(70) = 8.3913	< 0.0001
CC vs. chance	t(72) = 5.7154	< 0.0001
Combined vs. chance	t(142) = 9.8155	< 0.0001
PN vs. CC	t(142) = 1.4823	0.0702



Comparison	t(*df*)	p
CC TR vs. chance	t(35) = 5.0888	< 0.0001
CC IN vs. chance	t(35) = 3.0417	0.0022
PN TR vs. chance	t(35) = 8.4623	< 0.0001
PN IN vs. chance	t(34) = 4.1941	< 0.0001
CC TR vs. CC IN	t(70) = 1.3745	0.0868
PN TR vs. PN IN	t(69) = 1.8114	0.0372

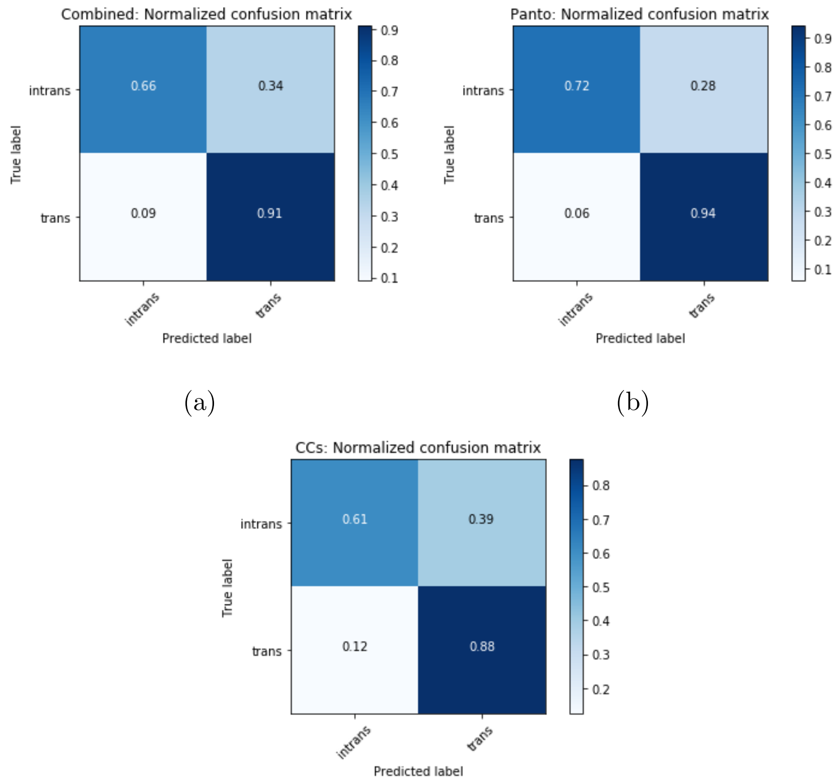
Figure 3.4. (Target Group) Accuracy of pantomimes and classifier constructions, all items, 'individual-level' accuracy. (Top) Average accuracy on classifier constructions juxtaposed with average accuracy on pantomimes. Both are significantly greater than chance, but neither is significantly greater than the other. (Bottom) Same as (top) but with accuracy on intransitive and transitive items displayed separately. In contrast to the results presented in Fig. B.1, only one comparison came back significant: transitive pantomimes were classed significantly more accurately that intransitive pantomimes. 'CC' = *classifier construction*; 'PN' = *pantomime*; 'TR' = *transitive*; 'IN' = *intransitive*.

to assess accuracy on a subject-by-subject basis. This measure is also continuous, and can thus figure into the correlations we run in §3.5.1, e.g., to quantify the relationship between accuracy and consistency of classing.

By the first method, the consensus method, accuracy was much higher in the 'target group' versus the whole group. Among pantomimes and classifier constructions that were consistently classed, accuracy was 87.35%, up from 75.29% (Pantomimes: 91.11% from 76.74%, Classifier constructions: 83.33% from 73.81%). As before, we calculated the Matthew's Correlation Coefficient to complement our accuracy measure. The MCC for classifier constructions and pantomimes combined is 0.7455 (up from 0.5088), for just classifier constructions 0.6661 (up from .4571), and for just pantomimes 0.8231 (up from 0.5574).

Taken all together, the 'target group' behaved as expected, under the hypothesis that the transitivity of iconic forms (like pantomimes and classifier constructions) would be transparent. These participants were highly accurate in their classing, at least among those forms that were consistently classed. (That is, if they were certain about the transitivity of a form, they were also likely accurate.) This accuracy, though, is likely not due to some (simplistic) bias, as the one that plagued the analysis of the whole group. This was revealed in the increased MCCs, and further elucidated in the confusion matrices in Fig. 3.5.

By the second method–or individual level analysis–again, participants accurately classed transitive and intransitive pantomimes and classifier constructions significantly above chance levels. Pooled, participants were 66.84% accurate across all items. Of just pantomimes, accuracy reached 69.41% and of just classifier constructions, accuracy was 64.35%. Separating out transitive and intransitive items, participants were 73.74% accurate in classing transitive pantomimes, 65.23% accurate in classing intransitive pantomimes, 67.73% accurate in classing transitive classifier constructions, and 60.87% accurate in classing intransitive classifier constructions. All accuracies were significant compared to chance, but only transitive pantomimes were accurately classed significantly more than intransitive pantomimes. See Fig.3.4.



(c)

Figure 3.5. Confusion matrices illustrating the performance of the ‘target group.’ Visualization is of the ‘consensus’ level analysis. These plots show a slight bias towards transitive responses.

In general, then, participants accurately guessed the argument structure of both pantomime and classifier constructions, suggesting that it is iconic in both of these systems. Further, while participants were descriptively more accurate in classing pantomimes over classifier constructions, this comparison was not significant (consensus method: Welch's $t(85) = 1.0757$, $p = 0.1427$; individual-level method: Welch's $t(142) = 1.4823$, $p = 0.0702$). While it may be tempting to conclude that the insignificant difference here is attributable to similarities in the signals of both systems (and descriptively, they are similar), we don't strictly speaking know why non-signers behaved this way. That is, it is feasible that what makes a classifier construction's argument structure guessable is not the same as what makes a pantomime's guessable. Nevertheless, our tentative claim will be that the similarities in the coding of argument structure in both systems modulates similarities in the perception of argument structure in both systems.

3.4 Experiment 2b: Iconicity Rating

The goal of Experiment 2b is to assess the degree to which non-signers' 'knowledge' of the meaning of a manual action correlates with the consistency at which they class it. To that end, we collect iconicity scores, or, an average rating of how well a form matches or describes its intended meaning.

3.4.1 Methods

Materials, participants, etc. The materials used in Experiment 2b were identical to those in Experiment 2a. Participant demographics were also identical, noting that none of the participants from Experiment 2a participated in Experiment 2b.

Survey design & implementation: The implementation of Experiment 2b was identical to that of Experiment 2a, including the number of items per survey, number of surveys, number of participants, demographic questions, etc. Specific videos may have been shuffled into different surveys. The only other difference was the task itself.

Each item contained the following: a video of a pantomime or classifier construction, a sentence that described that pantomime or classifier construction (sentences were based on those collected in Study 1a), and a 1 - 7 Likert scale. Participants were instructed to read the sentence and then watch the video. Participants were then instructed to rate how well the performer captured the meaning of the *action* of the sentence, where a rating of 1 meant *The person did not capture the meaning of the sentence at all*, a 7 rating meant *The person did a really good job of capturing the meaning of the sentence*, and a 4 rating meant *The person kind of captured the meaning of the sentence.*[19] Participants were encouraged to use the full scale, and not just 1's, 4's and 7's. One example was given, using the verb ACCIDENT from ASL-LEX. An example rating and justification were provided to calibrate participants to how we wanted them to think about their ratings. At the end of the survey, participants had the opportunity to read a consent statement. We considered submission of the survey as consent.

3.4.2 Results

Both classifier constructions and pantomimes tended to skew towards being iconic ($M_{CC} = 4.4556$, $SD_{CC} = 1.1377$; $M_{Panto} = 4.9293$, $SD_{Panto} = 0.9879$). However, the results of a D'Agostino-Pearson normality test indicate that both distributions of iconicity scores do not differ from a normal distribution: $s^2_{CC} + k^2_{CC} = 0.8652$, $p_{CC} = 0.6488$; $s^2_{Panto} + k^2_{Panto} = 4.178$, $p_{Panto} = 0.1238$. To note, the s statistic is a measure of *skewness* and the k a measure of *kurtosis*. The top five least and most iconic classifier constructions and pantomimes are listed in Tab. 3.4. A histogram of iconicity scores of both stimulus types is shown in Fig. 3.6.

[19] We realize that asking participants to attend to the *action* of a sentence, but then rate how well pantomimers represented *the sentence* was an oversight. If the instructions here were too unclear, we might have expected iconicity ratings to be much lower than they were.

Table 3.4.
Top five least and most iconic classifier constructions and pantomimes from Experiment 2b.

	Lo Iconicity		Hi Iconicity	
	Item	Iconicity (M)	Item	Iconicity (M)
CC	stick break	2.1429	hammer nail	6.6471
	cards scatter	2.6	take lid off	6.55
	toy crawl	2.6	tear paper	6.3529
	approach coat rack	2.65	pull out meas. tape	6.3333
	tower fall	3	paper drop	6.2857
PN	coat rack approach	2.65	tear paper	6.6471
	measure book	3.0476	take off lid	6.5882
	shaving cream spray	3.1765	break stick	6.5714
	bread spin	3.2353	person bend	6.5714
	lid blow off	3.5882	shake shaker	6.55

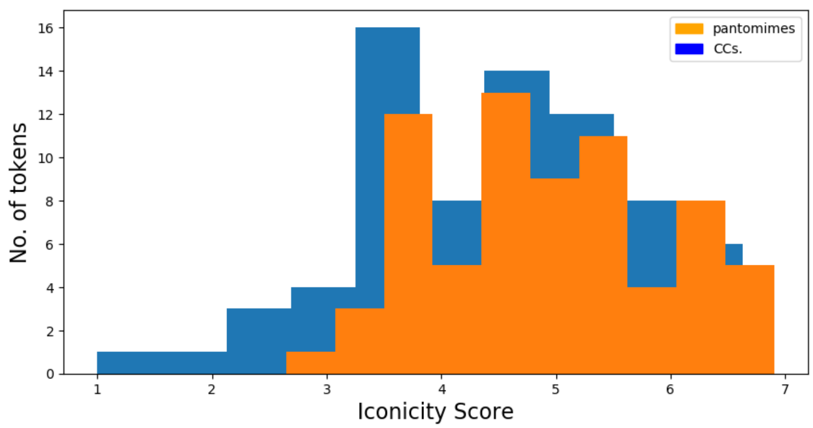
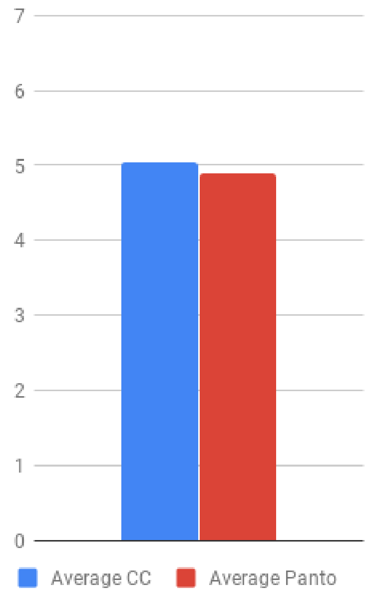


Figure 3.6. Histogram showing the distribution of iconicity scores across pantomimes (orange) and classifier constructions (blue). Both classifier constructions and pantomimes tended to skew towards being iconic (M_{CC} = 4.4556, SD_{CC} = 1.1377; M_{Panto} = 4.9293, SD_{Panto} = 0.9879). However, the results of a D'Agostino-Pearson normality test indicate that both distributions of iconicity scores do not differ from a normal distribution: $s^2_{CC}+k^2_{CC} = 0.8652$, $p_{CC} = 0.6488$; $s^2_{Panto}+k^2_{Panto} = 4.178$, $p_{Panto} = 0.1238$.



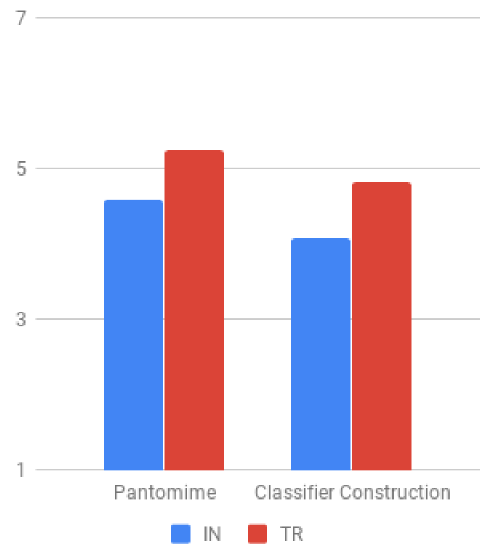
Type	Mean	SD
Pantomimes	4.9293	0.9879
CCs	4.4556	1.1377

Comparison	t	p (uncorr.)
Panto x CC	t(142) = 2.7675	p = 0.0064

Figure 3.7. Mean iconicity of pantomimes and classifier constructions from Experiment 2b (§3.4).

Our results show that pantomimes, classifier constructions, and lexical verbs (from ASL-LEX) form a cline with respect to their lexical iconicity. On the whole, pantomimes were rated as significantly more iconic than classifier constructions, which were in turn rated as significantly more iconic than lexical verbs. Specifically, the mean iconicity score for pantomimes was significantly greater than the mean iconicity for classifier constructions ($t(142) = 2.7675$, $p = 0.0064$). Further, the mean iconicity score for classifier constructions was significantly greater than mean iconicity score for lexical verbs ($t(267) = 8.3728$, $p < 0.0001$). This illustrates, as reported in the literature (Frishberg, 1975; Napoli et al., 2017)that iconicity is lost as one moves from a paralinguistic system to a core lexicon (Brentari & Padden, 2001). Finally, mean iconicity of pantomimes was significantly greater than lexical verbs ($t(266) = 8.3728$, $p < 0.0001$). The results are plotted and summarized in Fig. 3.7.

Among pantomimes and classifier constructions, transitive forms were rated as significantly more iconic than intransitive forms. Specifically, mean iconicity for transitive pantomimes was 0.6638 greater than intransitive pantomimes ($t(69) = 2.978$, p



Type	Mean	SD
IN_{Panto}	4.5818	0.8551
TR_{Panto}	5.2456	1.0038
IN_{CC}	4.0829	0.9298
TR_{CC}	4.8182	1.2025

Comparison	t	1-tailed p
IN_{PN} x TR_{PN}	$t(69) = 2.978$	0.002
IN_{CC} x TR_{CC}	$t(71) = 2.7338$	0.004

Figure 3.8. Graph illustrating that, among both pantomime and classifier construction datasets, transitive items were rated as significantly more iconic than intransitive items. Iconicity score (1-7) plotted on the y-axis. Summary statistics and two-sample t-tests comparing mean iconicity scores are presented in the tables at right.

= 0.004). Mean iconicity for transitive classifier constructions was 0.7353 greater than mean iconicity of intransitive classifier constructions ($t(71) = 2.7338$, $p = 0.0079$).

Discussion:

Overall, both pantomimes and classifier constructions were rated as being fairly iconic with respect to their lexical iconicity,[20] with pantomimes being rated as more iconic than classifier constructions. We say ‘fairly iconic’ here in that the distributions of iconicity scores from both the classifier construction and pantomime datasets were not significantly different from a Normal distribution. This suggests to us that pantomimes and classifier constructions are not as iconic as is often suggested, at least with respect to their lexical iconicity. Note, too, that we do not mean *transparent*, as all of our videos were presented alongside their meanings.

[20]Though this is hard to appreciate only looking at pantomimes and classifier constructions, the iconicity ratings of ASL lexical verbs is much lower. See §4.2.2.

With regard to transitivity and iconicity, it was found that transitives were more likely to be rated as more iconic than intransitives. From Tab. 3.4, we can see that the top five most iconicity classifier constructions were all transitive, and the top five least iconic classifier constructions were all intransitive. The same is essentially true of the top five most and least iconic pantomimes. We suggest that transitives were rated as more iconic than intransitives given that the lexical items presented in text could be reliably mapped onto the handshapes in the videos. In the case of transitive items, the handshapes were those that everyday people use to move or manipulate objects. For the intransitive stimuli, however, the object described in text has to be mapped to the hand (and in some cases forearm)– that is, the hand represents the object. Intuitively, there are limits to how well complex shapes can be mapped onto the hand conceptually. By contrast, there are only a few categories of grasps that could be used to show the shape, size, and texture, etc. of an object (MacKenzie & Iberall, 1994).

One thing we note for future work is that the iconicity scores may be influenced by the specific objects involved in the event.Some of the objects–particularly in the intransitive events–would not necessarily be easily accessible to raters. For instance, the toy in *toy-skitter* looked somewhat like a robotic spider and not, say, a plush bear, doll, or other prototypical toys. Future studies might control for this by using only objects typical to particular events, or even place a picture of the referents in the task. That is, our expectation is that we could derive even higher iconicity scores when taking object or verb-object familiarity into account.

3.5 Main analyses

We are now prepared to move on to our main analyses. Given that a significant proportion of classifier constructions and pantomimes were classed, it behooves us to ask what underlies this ability. A bottom-up explanation would argue for the presence of a certain set of privative features that additively encode transitivity distinctions.

We explore this possibility in §3.5.2. On the contrary, a top-down explanation would argue that knowledge of what a form means (or might mean) guides the analysis of its parts. To this end, we next (§3.5.1) correlate participants' consistency in classing pantomimes with the iconicity ratings obtained from Experiment 2b. We discuss the results in light of perceiver accuracy in identifying the argument structure encoded by our pantomimers and signer. To note, the below analyses are on the *target group only*. We include the analyses on the whole group in the Appendix (see Appendix B.3 for the top-down analysis of the whole group; Appendix B.4 for the bottom-up analysis).

3.5.1 Top-down: Classing behavior explained by previous linguistic/ conceptual experience

If the top-down approach is on the right track, we might expect to see consistency (i.e., agreement on a label)[21] and/ or accuracy increase with lexical iconicity score. We say *and/ or*, in that both measures make slightly different assumptions about what participants are thinking. A correlation with accuracy would indicate not only that top-down processing might be occurring, but also that the production model of transitivity coding matches the perception model. A correlation with consistency, however, does not make this particular assumption about a production-perception correspondence. Instead, non-signers may be responding to the lexical meaning of the sign or pantomime, which may then be fitted into some argument frame or another.

That is, in our own pantomime/ classifier construction dataset, sometimes the signer or non-signers introduced an agent or causer in response to intransitive stimuli. For instance, several pantomimers imitated dropping a ball in response to the video *ball drop*, perhaps intending to clarify the lexical meaning *drop*. We might assume that the same is possible on the perceiving end. Incidentally, as Napoli (2017) points

[21]In what follows, *consistency* corresponds to the magnitude of the t-value associated with an item. This t-value was derived, again, by comparing the frequency of the most selected option against chance. The t-values we used were for all four labels (*transitive*, *ditransitive*, *intransitive unergative*, etc.).

out, this is true in lexical signs, too, and even in those where argument structure is irrelevant. For instance, the sign BABY in ASL is articulated by rocking the arms back and forth, as one would rock a baby. However, the sign does not mean 'to rock a baby.' Work by Padden et al. (2013) illustrate other cases across different sign languages where nouns implicated in an action are denoted by visual analogs of those actions (e.g., in ASL, the verb CL:HAMMER and its derived noun HAMMER are visually very similar, only the latter is reduplicated).

Finally, a correlation between accuracy and consistency would be unsurprising, since accuracy and consistency are partially overlapping concepts as we define them. As accuracy increases, so too does consistency. As accuracy reaches zero, however, consistency again increases (e.g., if participants are 100% wrong, they are nevertheless 100% consistent): The relationship between accuracy and consistency is not linear, but polynomial. That is, while we report statistics concerning the relationship between consistency and accuracy, we do not take these statistics to be particularly meaningful with respect to our main objective.

Given that the mean iconicity between pantomimes and classifier constructions was off by less than half a point, and given that a comparable number of classifier constructions and pantomimes were consistently classed, we might expect that we might obtain similar figures (and if not, patterns) for both.

Analysis

We started with the following hypotheses: If non-signers are using linguistic knowledge or knowledge of conceptual structure, then we should expect them to to be more be more accurate as iconicity ratings increase. But, just as an event can be lexicalized or conceptualized in a number of ways (e.g., *break the stick* vs. *the stick breaks*), then it is possible that only consistency increases as iconicity scores do.

We calculated correlations using SciPy's `pearsonr` function. The function returns the Pearson correlation coefficient, r, and a two-tailed p-value, which can roughly be interpreted as the probability that an uncorrelated system would produce an r-value as great. The statistic becomes more reliable as datasets get larger, but since our dataset is quite small, we also performed correlations on randomly generated data. We also report an R^2 statistic.

Data were randomly generated as follows: values from one of the two measures under comparison were selected. The minimum and maximum values from that measure were determined and a random set of values was generated between those maximum and minimum values. For the comparison *Accuracy x. Iconicity*, iconicity values were randomly generated. For the comparisons *Accuracy x. T-values* and *Iconicity x. T-values*, the t-values were randomly generated.

Results

As shown in Tab. 3.5, all three measures are significantly correlated with each other in turn. First, again, it is unsurprising that the strength of the correlation between accuracy and consistency is relatively strong. We report the linear coefficient, r, in the table, but we additionally fit the data with a second order polynomial (as shown in Fig. 3.9). The R^2 value of this analysis is 0.456

For both Pantomimes and Classifier Constructions, a small correlation exists between accuracy and iconicity (PN: $r = 0.3199$, $R^2 = 0.1023$; CC: $r = 0.4011$, $R^2 = 0.1609$). A relatively larger correlation exists between iconicity and consistency (PN: $r = 0.4073$, $R^2 = 0.1659$; CC: $r = 0.4904$, $R^2 = 0.2405$). As expected, the additional analysis we performed using random data did not produce a strong (or significant) correlation. Both comparisons are represented graphically in Figs. 3.11 & 3.10.

Taken together, the results seem to suggest that if non-signers were consistent in classing an item, that item tended to have a higher iconicity score. This includes cases where non-signers were accurate (thus consistent). We take this to mean that, if non-signers are able to figure out what a classifier construction or pantomime means, they are able to use their linguistic or conceptual knowledge to make inferences about whether it involves action with or without an object.

The picture is slightly more complicated when transitive and intransitive items are analyzed separately, as seen in Tab. 3.6. Among pantomimes, correlations across all three measures are relatively more robust for intransitive stimuli and relatively less so for transitive stimuli. Even so, the comparison *Accuracy vs. Iconicity Score* did not return a significant result for either transitives or intransitives. However, the correlation between iconicity scores and t-values was still as robust, suggesting on the one hand that if participants were able to identify an event, they were more likely to class it one way, and on the other hand that this ability does not wholly depend on any putative transitivity coding strategy actually available in the stimuli.

An opposite pattern emerges for Classifier Constructions. Here, transitive stimuli showed higher correlations across the board when compared to intransitive stimuli, as might be expected given that iconicity scores were generally higher for transitive stimuli and more transitive stimuli were consistently classed. Still, though, it remains true here, too, that the comparison *Iconicity x. Consistency* is stronger than *Accuracy x. Iconicity.*

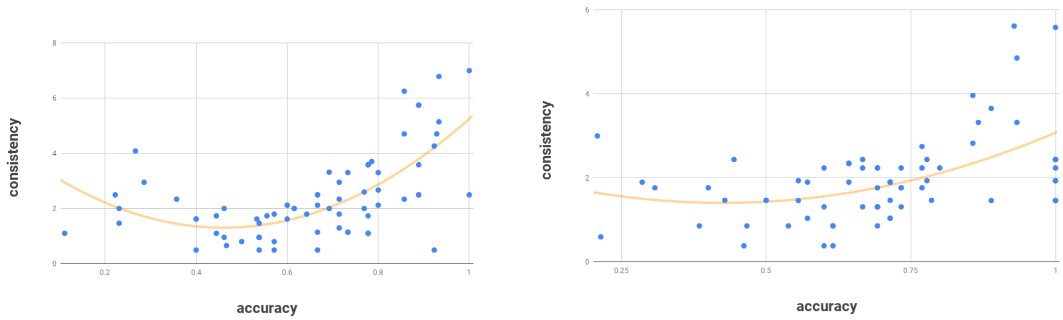
Discussion

In the above, we correlated *iconicity ratings* obtained from Experiment 2b (§3.4) with t-values–our measure of *consistency*–from Experiment 2a (§3.3) and percent correct transitivity guesses, or *accuracy.* The general pattern that emerged was that, of the two measures, iconicity scores correlated more strongly with consistency. This suggests that participants were more likely to consistently class an item if it was more transparent with respect to its meaning. We suggest that this means that participants, in some small way, are accessing prior linguistic or conceptual experience in order to help inform their transitivity classing.

One surprising result was that there was a stronger correlation in all three comparisons for intransitive pantomimes than transitive pantomimes, given that transitive stimuli (across the board) (a) were rated as more iconic than intransitive stimuli and (b) were more consistently classed. There were no obvious outliers in the data which could have a pronounced effect on the r/ R^2 values, so we're left with somewhat of a puzzle. However, we do note that even though more transitive pantomimes were consistently classed, the mean t-value for transitive items was actually a little bit lower (TR: 1.93, IN: 1.99), the maximum t-value for transitive items was also lower (TR: 4.8536 IN: 5.6412), while the minimum value was identical (TR = IN: 0.3873). We take this to mean that even though fewer intransitive items were ultimately consistently classed, there was considerable agreement on those that were when compared to consistently classed transitive items. However, the expected pattern, that correlations between measures should be higher among transitive than intransitive stimuli, was seen among classifier constructions. This all being said, the R^2 values were very low, so we may be organizing the wastebasket here.

However, do note a few more caveats: We mention again, though, that our measure of lexical iconicity only goes so far. This measure was collected from a different pool of non-signers who were presented with a pantomime or classifier construction and also its meaning. In Experiment 2a, where participants assigned transitivity labels to the same, they did not have direct, printed access to their meanings. Although we attempt to make claims about the transparency of argument structure (by way of lexical iconicity), we can only achieve an approximation here.

All along, our slogan has been *If someone can figure out what a production means...* and the rest, but we have not been precisely careful with what we intend by *mean*. We cannot be sure from our measurements what participants were thinking or what strategies they might have been using while taking the experiments. For one participant, an item might recall a specific word, with a single–or at least most unmarked–argument structure. For another a different word, or the same word but with an



(a) Classifier Constructions (b) Pantomimes

Figure 3.9. **Top-down analysis: accuracy and consistency:** Plots showing the expected polynomial relationship between t-values, as a proxy of non-signer agreement on a label for a particular item (consistency), and accuracy, measured in percent of participants guessing the correct label for an item.

alternate argument structure. We leave it to future studies to suss out this possibility.

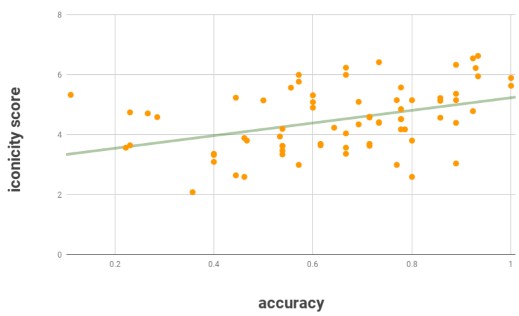
All being said, the results of the top-down analysis are fairly weak, though our methods were certainly not exhaustive. Just the same, these results argue again a holistic explanation of non-signer classing behavior. As we mentioned before, although we could not measure or detect reanalysis using our method, our results nevertheless suggest that this would not be the most fruitful enterprise.

We now turn to our second analysis, which uses the phonetic features of our stimuli, paired with non-signer judgments about their transitivity class, to see whether argument structure can be assessed bottom-up.

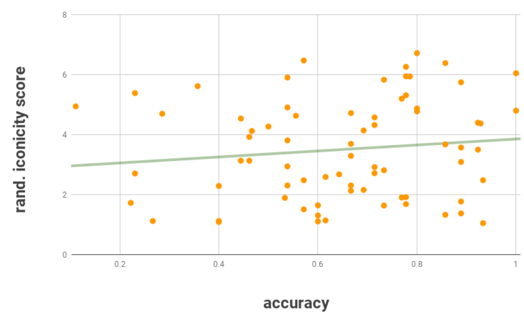
3.5.2 Bottom-up: Classing behavior explained by perceptual features of the stimuli

If the bottom-up approach is on the right track, we predict that individual phonetic features or visual cues–alone or in concert–(additively) predict transitivity class.

Classifier Constructions:

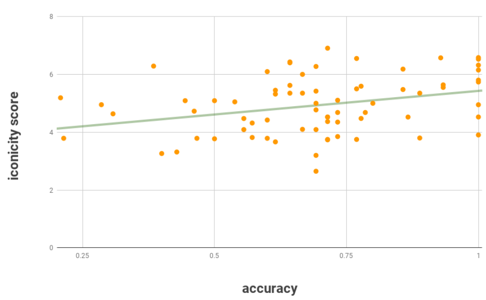


(a)

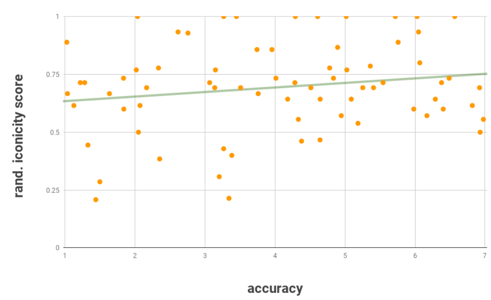


(b)

Pantomimes:



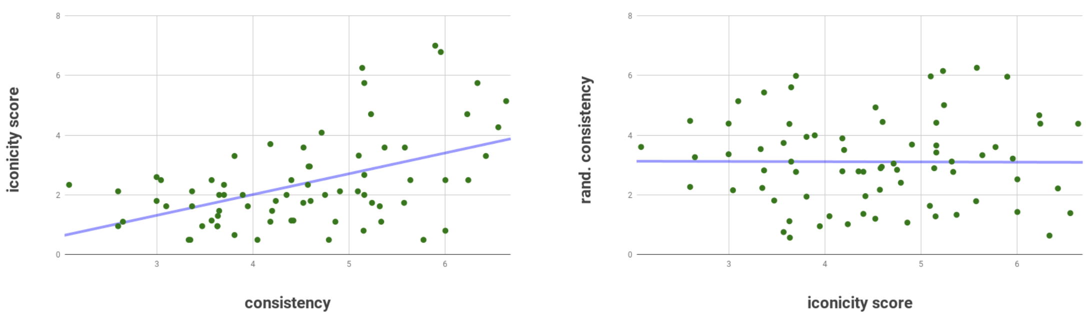
(c)



(d)

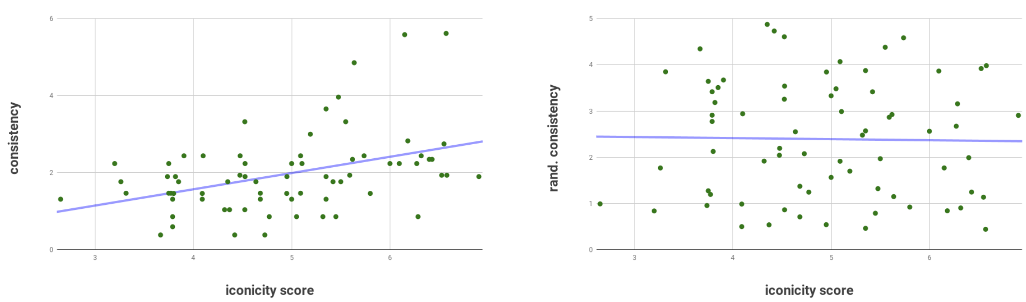
Figure 3.10. **Top-down analysis: accuracy and iconicity scores:** Plots (a,c) showing a medium positive linear relationship between iconicity scores (derived from Experiment 2b) and accuracy, measured in percent of participants guessing the correct label for an item. The (b,d) plots are the same, only iconicity scores were randomly generated. All iconicity score were between the minimum and maximum limits of actually observed iconicity scores.

Classifier Constructions:



(a) (b)

Pantomimes:



(c) (d)

Figure 3.11. **Top-down analysis: iconicity and consistency:** Plots (a,c) showing a medium positive linear relationship between iconicity scores (derived from Experiment 2b) and accuracy, measured in percent of participants guessing the correct label for an item. The (b,d) plots are the same, only iconicity scores were randomly generated. All iconicity score were between the minimum and maximum limits of actually observed iconicity scores.

Table 3.5.
Table of correlations: Correspondences between top-down measures, accuracy (Acc.), consistency (measure by magnitude of t-values, Tval.), and iconicity scores (Icon.). Medium sized correlations hold between all three measures for both pantomimes and classifier constructions, suggesting that non-signers were both consistent and accurate in their transitivity classing if they had a sense as to what a given pantomime or classifier construction meant. 'ran.' = random.

		Pantomimes				Classifier Constructions			
		r	p	r ran.	p ran.	r	p	r ran.	p ran.
Acc.	Icon.	0.3199	0.0065	0.1776	0.1383	0.4011	0.0005	0.1237	0.3005
Acc.	Tval.	0.4539	0.0001	-0.1051	0.3832	0.5094	0.0001	0.0512	0.6695
Icon.	Tval.	0.4073	0.0004	-0.0937	0.4370	0.4904	0.0001	0.0565	0.6371

Table 3.6.
Table of correlations (IN vs. TR): Same as Tab. 3.5 but separating out transitive and intransitive items. Results show that correlation between top-down features and consistency are stronger among intransitive items in pantomimes, but among transitive items in classifier constructions.

		Pantomimes				Classifier Constructions			
		trans		intrans		trans		intrans	
		r	p	r	p	r	p	r	p
Acc.	Icon.	0.2449	0.1500	0.3071	0.0684	0.3829	0.0212	0.3316	0.0482
Acc.	Tval.	0.3807	0.0220	0.5283	0.0009	0.6806	0.0001	0.2413	0.1563
Icon.	Tval.	0.4153	0.0118	0.4877	0.0026	0.4898	0.0024	0.3886	0.0192

Due to the flexible nature of machine learning analyses, we probe our data in the following ways. First, to establish whether transitivity information is present in the signal (i.e., encoded by our signer and non-signers) we run an analysis on the entirety of the signer and non-signer datasets (in turn) using 'ground truth' labels. Then, to establish whether non-signers use these cues to class the stimuli, we run an analysis on just those items that non-signers consistently classed. Finally, to assess what features are important to senders and receivers, we run an analysis on just those items that were consistently and accurately classed.

Given the similarities between classifier constructions and pantomimes that we have seen so far (comparable number classed, §3.3; significantly different yet near equal mean iconicity scores, §3.4), we predict that the phonetic features that describe both stimuli types will be equally informative. As such, we should see comparable classifier accuracies for both stimulus types across all analyses. Further, we predict that all analyses will provide significant results as the phonetic features we code for have (a) been identified as contributing to transitivity encoding in sign languages and (b) been argued to be iconic (§2.2.2).

In what follows, we describe the phonetic features we code for and what their hypothesized relationship to transitivity marking is (§3.5.2). We then outline our classifier and (hyper)parameter selection, and processing pipeline (§3.5.2). We report our results in §3.5.1 and discuss them in §3.5.2.

Phonetic Features

The below features were selected on the basis of what is known to be informative with respect to transitivity distinctions in sign language. These include features that have been expressly implicated in transitivity coding (e.g., handshape complexity) and those that may have a more indirect, circuitous relationship (e.g., end marking).

The features are divided into different classes, for expository purposes only. To note, a given video may be annotated for two or more features from a given class, or it may not be marked by any feature of that class. Wherever possible, we avoid marking the absence of a feature, as absence of a feature is explicitly modeled by the classifier we selected (see below).[22]

Handshape, finger- and joint-complexity: Depending on level of analysis, *handshape* can be decomposed into component features or into classes, with the former being more phonetic in nature, and the latter more phonological or morphophonological. There has not been an attested link between individual features and transitivity, though handshape classes (handling, transitive; and object, intransitive) are reflected in finger- and joint-complexity measures (e.g., Brentari et al., 2012, 2017).

We coded handshapes using the coding scheme devised by Eccarius and Brentari (2008). From the coding scheme, individual features from Brentari's (1998) model are recoverable (e.g., [closed], [curved], and so on). Further, from these features, joint and finger complexity measures are also recoverable. We therefore wrote a small script to convert these handshape codes into component features, and to assign complexity measures.

As an example, what has often been called the baby-c handshape (MOON, POLICE-OFFICER) is coded as 1Tc;# and decomposed in Tab. 3.7.

Finger complexity was determined by (a) base symbols (i.e., the first digit of the handshape code) or by (b) the number of finger groups (i.e., the number of semicolons, *;*, in the code). With one semicolon in its code, the baby-c handshape has medium handshape complexity. Joint complexity was determined by the presence of the features 'stacked,''crossed' (*complex-joint*); 'flat,' 'curved,' 'bent,' or 'spread' (*medium-joint*) features. If none, the handshape was labeled as *low-joint*. Baby-c, thus, would be considered to have medium joint complexity.

[22] To illustrate, consider two sets of features we annotate for, *contact* and *number of events*. If contact between the hands is present, we code it for whether that contact is initial, medial, or final. If no contact occurs, the video is not annotated for contact. However, with *number of events*, we count how many events there are in a video and simply label the video *monoeventive* is there's only one event, and *multieventive* is there are more than one. We do this instead of only marking, e.g. *multieventive* (from which it can be inferred that the video is not monoeventive), because some sort of event has occurred either way.

Table 3.7.
Example decomposition of handshape codes into component features.'ns' = non-selected fingers.

code snippet	feature
1	one
T	thumb
c	curved open
#	ns-closed ns-flex

Presence of ';' : *medium finger complexity*

Presence of *curved*: *medium joint complexity*

End-marking: The features we include here are *deceleration*, *contact*, *aperture change*, *orientation change*, and *repetition.* Although end-marking marks an event boundary (Wilbur, 2003, 2008), and even non-signers perceive it as such (Strickland et al., 2015), we hypothesize that non-signers may nevertheless associate these cues with transitivity. For instance, if the hand decelerates towards point x, it may be assumed that x is an event participant and/ or boundary. To note, the atelic verbs in Strickland et al.'s (2015) study were largely intransitive (e.g., 'run,' 'float'; but 'discuss') and the telic verbs were largely transitive ('enter,' 'sell'; but 'die').[23] Further note that Hopper and Thompson (1980) suggest that telic events are 'more transitive' than atelic events. As such, telicity measures may be at least indirectly informative for transitivity contrasts.

We would further note that, following Lepic et al. (2016), in some cases of contact the hands represent entities whose relationship is specified by this contact (e.g., *A hits/ touches/ caresses B*).[24] As such, although one hand contacting the other may signify an event reaching its terminus, it may also (or at the same time) signify contact between entities.

We coded cases in which the two hands come into contact with each other as *contact* (HIT). In some cases, we counted contact with the space immediately surrounding the hands, if it was obvious to us that the objects denoted by the two hands were intended to extend past the hands. Contact with a plane was coded as *deceleration* (INSULT). *Orientation change* and *aperture change* were marked when the orientation of the palms changed (e.g., DIE) or when the hand(s) went from a closed handshape to an open handshape or the other way around (DROP). With *aperture change*, our assumption going in was that this would mark grasping if articulated just

[23]Specifically, their design involved participants watching a video of a sign, and then labeling that video with one of two words. One of the words matched in telicity, the other did not. In their design, however, the choice between, e.g., 'run' and 'enter' was not only a choice between *atelic* and *telic*, but also *intransitive* and *transitive*.

[24]In other cases, *contact* might be between figure and ground. As such, on its own, *contact* is an ambiguous cue. However, it is possible that this cue may interact with the movement or stasis of the second hand: an immobile second hand may be taken as a ground, whereas a mobile second hand may be taken as another figure.

once, or perhaps internal movement if repeated. As such, we expect this feature to interact with *repetition*, which we describe next, or local movement (including *wiggle*), which we describe under *path* features. *Repetition* was marked when the production of the pantomime or classifier construction involved a full or partial reduplication of the movement (HAMMER). As described in Wilbur (2009), one form of reduplication of verb root encodes aspectual information, which interacts with telicity at the VP level and thus may be a relevant cue.

Relative Orientation: Two-handed classifier constructions and pantomimes were annotated for the orientation of the palms with respect to each other. Options were *same*, if the palms were pointed in the same direction (PUSH); *opposing*, if the palms faced each other (MANAGE); *opposite*, if the palms faced away from each other (SEPARATE); and *other* for all other cases.

Path: Following Brentari's (1998) discussion of movement types (ibid., Chapter 4), path specifications were *path*, which describes arm movement from the elbow up (DRIVE-TO), *local movement (gross)*, which describes movement from the wrist (PLAY), and *local movement (fine)* which describes movement contained to the fingers (WAIT). From a previous, unpublished analysis on a set of similar classifier constructions and pantomimes, we found that finger movement tended to depict internal movement (legs kicking, flames flickering, etc.) and, thus, intransitive events. Grosser movement in that analysis did not play a vital role in transitivity distinctions, but we attempted it again here anyway.

Tension: We marked videos as tense depending on whether the signer or pantomimer appeared to convey force in his or her movement.

Contact: Here we note where in the representation contact occurs: *initially* (BREAK-UP), *medially* (LAY-OFF), or *finally* (ARRIVE). Contact is defined with respect to the second hand, or a point or plane. We did not explicitly annotate videos for having no contact. We predicted that initial and medial contacts might be interpreted more consistently as grounds, rather than objects of goal-directed objects. As such, we predict that *final* would be more indicative of transitive events, and *medial* and *initial* more indicative of intransitive events.

Relative Movement: We defined relative movement as being between two hands. Single-handed signs were thus not annotated for movement features. (Again, movement features for one-handed productions can be simply described by the path specifications in the *path* category). Movement describes cases in which one hand moves with respect to the other, or cases where both hands move simultaneously. Our categories here were: *towards*, when the hands moved closer towards one another (HIT); *away-from*, when the hands moved away from each other (BREAK); *together*, when the hands moved in the same direction (FEED); and *around/ in* when one hand moves towards and becomes enclosed in the other (JOIN). To note, productions were annotated for these features even if the second hand did not move.

We hypothesized that *static* and *independent* specifications might correlate with transitive events: if the 2nd hand is static, say, and the dominant hand acts on it, there is a transitive parse (again, HIT). If the 2nd hand moves independently, it could be interpreted as an independent agent (such signs out on phonological grounds in ASL). Combined with an agent interpretation of the dominant hand, a felicitous parse would be a transitive one containing two agents. Finally, as Lepic et al. (2016) point out, the two hands can represent a volume. With the hands moving mirror-symmetrically, for instance, a change in volume (intransitive, unaccusative) is a possible parse (DOWNSIZE). With the hands moving together, a volume moving or being is a possible parse (RAIN).

Number of events: Representations of some events were complex, consisting of two or more actions. For instance, in the depiction of a jar of bottle caps spilling over, there is the jar spilling subevent and the bottle caps falling out subevent. Some participants chose to represent both subevents, while others not. We mark representations containing two or more actions as *multieventive*, while simplex representations we mark *monoeventive*. We predict here that transitive (macro)events are more likely to be split into causing and consequent events.

Movement of 2^{nd} hand: For two handed representations, we noted whether the second hand moved or not. If so, we noted how it moved with respect to the dominant hand (hand dominance was assessed on a video-by-video basis). We noted when the 2nd hand was stationary (*static*; like, e.g., FLATTER), moved in some sort of mirror symmetry with respect to the dominant hand (including alternating, rotational, etc. movements; *mirror*; DOWNSIZE), moved in sync with or copied the movement of the dominant hand (*together*; CONTINUE), or moved completely independent of the dominant hand (*independent*; not a property of lexical verbs in ASL, but possible in classifier constructions).

Eye gaze: Eye-gaze was identified by Bahan (1997) and Neidle et al. (2000) as a non-manual object agreement marking strategy, though Thompson et al.'s (2006) results challenge their analysis. In short, object marking through eye gaze is not categorically present (though there are restrictions on where it may occur; see Wilbur, 2013). However, we predict that transitive classifier constructions and pantomimes will exhibit more intentional eye gaze than intransitive events. Here, we make four eye-gaze distinctions: *camera* for when the participant looks at the camera; *hands* for when the participant is looking at his or her hands; *trajectory* for when participants seem to be looking at an imagined trajectory of the object (e.g., looking up and away to simulate tracking a ball before catching it); and *other* for when no other category is appropriate.

Iconicity Score: Although *iconicity scores* are not a phonetic, bottom-up feature, we included it in the model anyway, to see whether (some) top-down information is relevant. Iconicity scores were obtained from Experiment 2b and then binned into three different strata. Items with a score of 5.01 or greater were classified as 'high iconic.' Those with a score less than or equal to 3.0 were classified as 'low iconic.' All others (3.01 - 5) were classified as 'medium iconic.' The prediction here is that 'high iconic' should predict transitive items, as these items were rated as significantly more iconic than intransitive items in Experiment 2b.

Table 3.8.
Table of feature classes and their subordinate features for machine learning analysis. Note that only three features are listed per class. Features, and the literature they were inspired by, are discussed in-text.

Handshape	Fing. Complexity	Joint Complexity	End Marking
curved	complex	complex	deceleration
open	medium	medium	contact
crossed	low	low	acceleration
...			...
Path	**Tension**	**Contact**	**Rel. Mvmt.**
path	tense	initial	towards
localmvmtfine	lax	medial	away-from
localmvmtgross		final	together
			...
Rel. Orientation	**2H Mvmt.**	**Event**	**Eyegaze**
same	static	mono	hand/ trajectory
opposing	mirror	multi	camera
opposite	identical		other
...	...		

Analysis

Preprocessing & Classifier selection: The features described above were annotated in a spreadsheet. We note that some features are highly correlated with each other, while others not. For instance, handshape complexity can be derived from individual handshape features, end-marking features always co-occur with a feature describing where the end marking feature occurs (e.g., *deceleration – final*, *acceleration – initial*), and so on. Other features, like *eye-gaze*, however, are by (the null) hypothesis not correlated with any other feature. While meaningful correlations between features (e.g., *relative movement* and *relative orientation*) are of interest, because of known meaningless correlations, we elected for a model/ solution that does not take correlations between features into consideration at all.

As such, we treat the classification of the selected features as a bag-of-words problem,[25] which assumes independence between features. As we mention above, one known issue about the bag-of-words solution in our case is modeling the presence/ absence of terms. For instance, if a video is marked *monoeventive*, it cannot also be marked *multieventive*. That is, in vocabulary construction (below), both *monoeventive* and *multieventive* are added to the dictionary. A sample would have a 1 for either, and a 0 for the other. In other cases, we do not mark opposite/ other features. For instance, we mark videos for *tense*, but not, e.g., *lax*. So, dictionaries would have an entry for the former, but not the latter. We anticipate that this, however, would have an overall negligible effect on our results.

As just alluded to, the bag-of-words solution to classification does not act directly on raw data (i.e., words), but instead calculates the frequencies of each word in each document for inference.[26] An example schematic is given in Tab. 3.9.

[25] This approach to classification is also used in sentiment analysis and spam filtering, *inter alia*. In these applications, it is not necessarily the ordering of words (or constituents) that predicts good/ bad sentiments or spam vs. non-spam, it is the presence and frequency of certain highly predictive words (e.g., *disgusting* or *Viagra*) that results in best classification.

[26] Under the assumption that many of the features are redundant or otherwise uninformative, several heuristics are performed to whinny down the list of total features to only those that may best result in classifier accuracy. Two such strategies are the removal of stop words (e.g., articles, *an, the*; punctuation, *!, ?*; etc.). One other preprocessing step is to weight words with respect to how often

After vectorization, data vectors were split into k distinct sets for a k-fold leave-one-out cross-validation paradigm, in which classifiers are trained on k - 1 sets of data and tested on the remaining set. This is done in a round-robin fashion, such that each partition of the dataset serves as the test set once. For each iteration, a new classifier was trained, such that knowledge from the last iteration leak into the current iteration and artificially improve performance (i.e., training on the test set). Set sizes varied depending on analysis, but the following ratios of training to test sets were always followed (in order of preference): 7:1, 6:1, and 5:1.

Generally, certain features are not informative with respect to a classifier's decision. For example, if the feature *base* occurs in every sample, irrespective of whether that sample is transitive or intransitive, it does not inform the classification decision–it is equally likely to appear with transitive and intransitive samples. As such, we used a feature extraction method on the training set of each fold to identify those features that were maximally informative. Among the many feature extraction methods available, we had the most success with the *select-k-best* method, where *best* was determined by F values returned by an ANOVA. These were computed automatically using *sci-kit learn*'s native *selectKbest* function (retain k features based on F-values) with the *f_classif* callable (determines F-values). To find the optimal k for each analysis, we iterated through k values from k = 1 to k = 67, the total number of features in the dataset. We simply report the analysis with the k-value associated with best classifier performance. To note, feature extraction was performed on the training set of each fold in each analysis and never on data from the test sets. This ensures that the classifier does not learn patterns based on the entire dataset (i.e., including samples in the test set). However, this means that for each fold in each analysis, a (potentially) different set of k features were most informative. Finally, depending on

they occur across the whole corpus, assigning more importance to relatively infrequently occurring features. This process is called *term frequency, inverse document frequency* weighting. However, this process may assign high importance features that may only be incidentally informative in the desired contrast (intransitive vs. transitive). For instance, if the feature *crossed* only occurs once in the dataset (in an intransitive item), it will be weighted heavily (i.e., 100% of items with this feature are intransitive) even though the feature does not generalize to the rest of the dataset.

the analysis (specifically, how many samples were in each), the optimal k varied from as little as 2 to as great as 52.[27]

Because we are after the most informative transitivity coding features, generally, and not just those that aid in a particular analysis, we counted each time a particular feature was found to be most informative in a particular fold. We iterated each analysis 10 times, compiling all of the most informative features from individual runs. We considered features that were most informative in at least 75% of all folds after 10 iterations to be generalizable. The number 75 does not come from anywhere in particular, but we found that it is both liberal and constraining enough to provide a good, short list of informative features.

As an example, consider the following hypothetical analysis. In each iteration, the data are partitioned into eight sets for an 8-fold cross-validation paradigm. For each fold, feature extraction is performed on the training set (where k can be anywhere from 1 to 67). We add the features extracted from the training set of each fold to a list. If a feature appears as informative in each fold, it will appear eight times on the list. We repeat this process for nine more iterations, each time adding features to the list. At the end, we compute the maximum number of times a feature can appear on the list. In this case, it is 8 folds $\times$ 10 iterations = 80. As such, features appearing on the list ($80 \times 0.75 =$) 60 times or more we take as generally informative in the transitive-intransitive distinction.

We chose a Multinomial Naïve Bayes (MNB) classifier. In general, NB classifiers work well with smaller datasets, which themselves contain only few words each. These classifiers may even outperform more sophisticated classifiers, like Support Vector Machines, up until the dataset reaches a certain size. We chose the MNB classifier over other NB variants (Bernoulli and Gaussian) because the composition of features

[27]The range is dramatic for a simple reason. For smaller and smaller subsets of items, a smaller and smaller subset of features best describe them.

Table 3.9.
Simulated feature vectors: samples are listed at the left, with a subset of the features that describe them to the right. (The classifier only sees the information to the right and never sees information related to an item's class). The numbers represent raw counts of how many times a particular feature appears with a particular sample. From this example, we see that intransitive and transitive items have a completely complementary distribution of features, which a classifier may learn. Note that although these three items may not have the features *bent* or $closed_{nsf}$, these features occur elsewhere in the dataset.

	label	base	bent	camera	closed	closed$_{nsf}$	narrow
walk-backwards	IN	2	0	0	0	0	1
break-stick	TR	0	0	1	2	0	0
cut-bread	TR	0	0	1	2	0	0
...	...	...	...	...	...	...	...

follows a multinomial distribution, given that we transform our bag of words to a vector of counts (e.g., 0, 1, 2, ..., many).[28]

In the MNB solution, for each class, transitive and intransitive, a vector of probabilities is created, giving the likelihood of a particular feature, i, appearing within a sample from a particular class, y. For instance, the probability vector for the intransitive class, I, is given as $\theta_I = (\theta_{I1}, ..., \theta_{In})$, where n is the total number of features that occur in the entire dataset. Each individual probability in θ_I is estimated as in Eq. 3.1:

$$\hat{\theta}_{Ii} = \frac{N_{Ii} + \alpha}{N_I + \alpha n} \tag{3.1}$$

where N_{Ii} is the number of occurrences of feature, i, in the intransitive class, and N_I is the total count of all features in the intransitive class. To note, the equation also factors in training and test sets, and accounts for scenarios in which a feature appears in the test set, but not in the training set. This is achieved through smoothing, or the addition of a small number α, to avoid strictly 0 probabilities. We chose an α of 1 (Laplace smoothing).

Labels used in analyses: To answer our research questions, we ran our classifier analysis using labels derived from different sources or created under different statistical criteria. To answer the question, *Are transitivity distinctions encoded in classifier constructions and pantomimes?*, the labels of the action videos the signer and non-signers watched and then represented were used (PNs: n = 432; CCs: n = 73). To answer the question, *Do non-signers perceive transitivity distinctions in classifier constructions and pantomimes?*, labels were assigned using data from Experiment 2a. Namely, we took the labels of the items that were consistently classed (PNs: n = 43; CCs: n = 44). Finally, for the question, *In cases where encoders and decoders agree on a transitivity parse, are the same features relevant?*, we kept only items where

[28]It is possible to transform the representation of the features from a multinomial to a binomial (Bernoulli) distribution by simply modeling features in terms of their presence (1) or absence (0), but we achieved higher accuracy with the MNB variant.

perceivers were consistent and accurate (PNs: n = 39, n_{IN} = 12, n_{TR} = 27; CCs: n = 37, n_{IN} = 16, n_{TR} = 21). That is, we kept consistent, non-signer-derived labels that matched their corresponding production labels.

For unbiased classification, an equal number of labels is needed.[29] For the labels derived from Experiment 2a, it was the case that an unequal number of intransitive and transitive labels met the criteria we set.[30] As such, we capped the class having more samples by the number of samples contained in the other, smaller class. In some cases, this left us with dataset sizes indivisible by 8, 7 or 6 (i.e., for cross-validation splits). The removal of one sample from each class was thus necessitated. The final number of samples per class were as follows: For consistently labeled pantomimes, n = 32 (16 intransitive, 16 transitive); for consistently labeled classifier constructions, n = 42; for consistently labeled, accurate pantomimes, n = 24; and for consistently labeled, accurate classifier constructions, n = 32.

Metrics used: We used two metrics to assess classifier performance: accuracy and Matthew's Correlation Coefficient. They are defined as follows, assuming a binary classification problem (e.g., *transitive* vs. *intransitive*): *Accuracy* is the number of the number of correctly identified samples of Class A/B out of the total number of samples from Class A/B (*per class*),[31] or, is the number of correctly identified samples of Class A and Class B divided by the total number of samples (*total accuracy*). Read the following as $\text{Predicted}_{Actual}$. So, TR_{IN} should be read as *Predicted transitive when the stimulus was actually intransitive.*

$$Accuracy_{total} = \frac{TR_{TR} + IN_{IN}}{TR_{TR} + IN_{IN} + TR_{IN} + IN_{TR}} \tag{3.2}$$

$$Accuracy_{perClass} = \frac{TR_{TR}}{TR_{TR} + TR_{IN}} \tag{3.3}$$

[29] As an extreme example, if there are 99 intransitive labels and one transitive label, and the classifier predicts 100 intransitive labels, it is still 99% accurate without really learning anything.

[30] To note, we removed one classifier construction, *adjust picture (alternate)*, from the classifier construction production dataset to arrive at an even 72 samples.

[31] In our case, this *per class* accuracy is identical to *precision.*

Accuracy in this way can be thought of the same way as a percentage correct on a multiple choice exam, and suffers from some of the same shortcomings. For instance, many in our generation were simply told to guess "C" if unsure of the answer (if "A", "B", and "D" are also possible answers). If done consistently, you can ensure at least 25% accuracy for those items you do not know. But, of course, this does not demonstrate learning, which is what we (and our teachers) are ultimately trying to tap into. As such, we also used a more balanced measure of classifier performance, Matthew's Correlation Coefficient (MCC).

While accuracy only takes correct identifications (TR_{TR} and IN_{IN}) into consideration, the MCC is more balanced in that it also takes incorrect identifications (TR_{IN} and IN_{TR}), as shown in 3.4. The formula outputs a value in the range of -1 to 1, with 1 being perfect agreement between predicted and observed values, -1 being perfect disagreement, and 0 being random prediction. Explicitly, it is not possible to compare the MCC against chance, at least in the same way as accuracy. For instance, an MCC of 0.5 in a binary problem does not indicate chance performance.

While the MCC is particularly useful in cases where there is an imbalance in the number of classes (say 50 transitive samples, but only 20 intransitive samples), we always include the same number of samples in our analyses (e.g., 20 each of transitive and intransitive samples). Nevertheless, the MCC gives us a metric to use when assessing biases in classifier performance.

$$MCC = \frac{TR_{TR} \times IN_{IN} - TR_{IN} \times IN_{TR}}{\sqrt{(TR_{TR} + TR_{IN})(TR_{TR} + IN_{TR})(IN_{IN} + TR_{IN})(IN_{IN} + IN_{TR})}} \tag{3.4}$$

Results

We had three main questions: (1) *Are transitivity distinctions manifest in the signal?* or *Are transitivity distinctions encoded visually?*; (2) *Irrespective of encoding, are transitivity distinctions perceived?*; and (3) *Do producers and perceivers use the same phonetic features to code transitivity distinctions?* In brief, we can answer

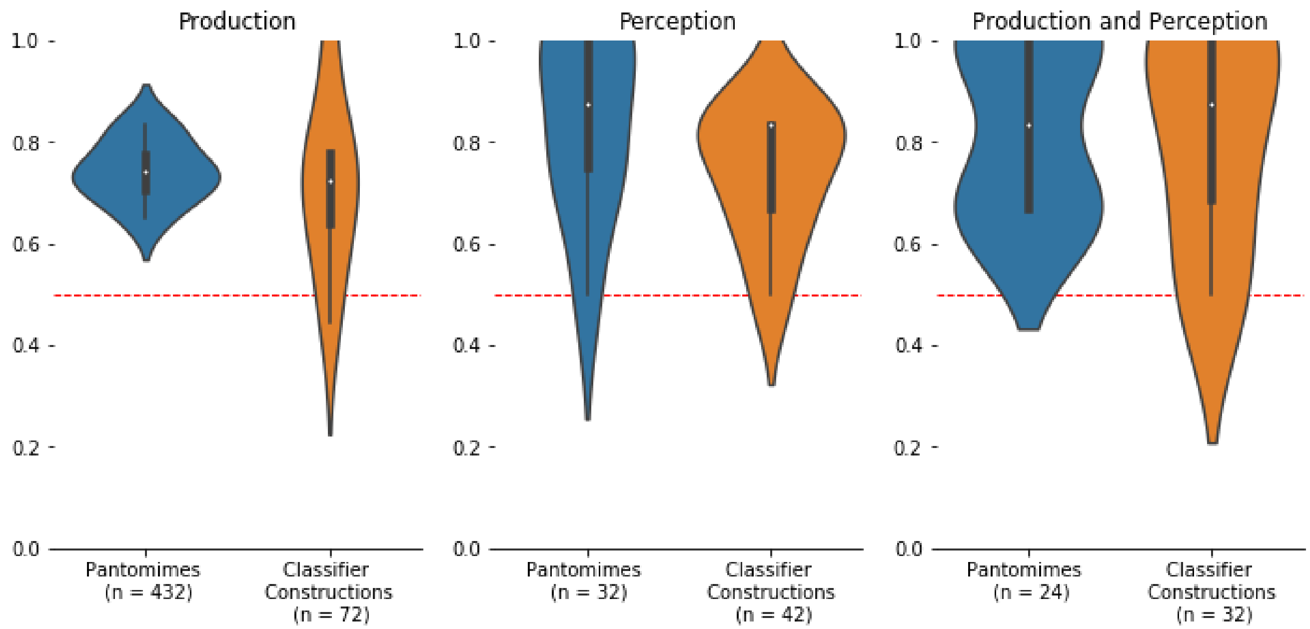


Figure 3.12. **Bottom-up results:** Violin plots showing the distribution of classifier accuracies across a number of different analyses. In each, pantomimes are in orange, classifier constructions are in blue; the red dashed line represents chance, here 50% (transitive or intransitive). Plots are divided by production (transitivity encoding), perception (transitivity decoding), and their intersection (what counts for both encoding and decoding). The n in parentheses indicates how many samples were used in each analysis.

Table 3.10.
Summary of results obtained from machine learning analysis. 'PNs' = Pantomimes. 'CCs' = Classifier Constructions. 'STD' = Standard Deviation. 'MCC' = Matthew's Correlation Coefficient.

	Production		Perception		Production/ Perception	
	PNs	CCs	PNs	CCs	PNs	CCs
Mean	0.7477	0.7083	0.8438	0.7381	0.8333	0.8125
STD	0.0531	0.1565	0.1740	0.1214	0.1667	0.2073
p	<0.0001	0.0005	0.0001	0.0029	0.0015	0.0005
MCC	0.5098	0.4207	0.6888	0.5051	0.6667	0.6299

affirmatively to each. (See Appendix C for more in-depth analyses and Appendix D for a few ancillary analyses).

On the production end, both pantomimers and our signer encoded transitive and intransitive events using the phonetic features we coded for, as evidenced by the respective classifiers achieving their targets significantly above chance levels. Perhaps unsurprisingly, classifier accuracy for pantomime production was slightly higher than for classifier constructions (74.77% vs. 70.83%, respectively) owing to the fact, we argue, that a linguistic system does not impose restrictions on the iconicity of pantomimes. However, this may be an artifact of the size of the datasets, as classifiers generally achieve higher accuracy with more samples and the pantomime dataset was six times larger than the classifier construction dataset. As shown in more detail in Appendix C.1.2, analysis of just the best 72 pantomimes (as determined in Study 1b; §3.2.2) reveals slightly lower performance on pantomime data (67.18%).

The features that were most informative for transitivity distinctions on the production side are listed in Tab. 3.11 for pantomimes and 3.12 for classifier constructions. For pantomimes, features from each of the categories we coded for are represented, indicating that many aspects of the visual signal are needed to describe transitivity distinctions, at least with respect to production. This could also represent the heterogeneity of the pantomime signal, as the pantomimers coded events differently.

Table 3.11.
Common most informative features across production analysis of pantomimes.

Features (Production, Pantomimes)
mediumjoint, localmvmtfine, bent, static, mirror, curved, acceleration, index, thumb, trajectory, closed, final, flex, awayfrom, towards, complexfinger, wiggle, opposing, tense, flex$_{nsf}$, hands, multi, stacked, mono, pivot, base, narrow, nonbase

For classifier constructions, many fewer features were returned as informative, representing only some of the categories we coded for. Handshape, end marking, tension, relative movement, relative orientation, (number of) event(s), and eye-gaze are all represented, while finger- and joint-complexity, path, contact, and 2H movement were not. The relatively constrained set may be due to the size of the classifier construction dataset, with only 72 items (as opposed to 432 items for pantomimes). Or, it could be that the features coding transitivity were more consistent for classifier constructions, owing to linguistic constraints on their form.

Table 3.12.
Common most informative features across production analysis of classifier constructions

Features (Production, CCs)
multi, trajectory, same, tense, deceleration, mono, curved, narrow, nonbase

Finally, Tab. 3.13 lists the features that were commonly identified as informative. Here, we simply take the intersection of the features identified in Tabs. 3.11 & 3.12. Both (number of) event features were identified, as well as eye gaze, tension, and handshape features.

Table 3.13.
Common most informative features across production analyses (pantomimes and classifier constructions).

Features (Production, Pantomimes & CCs)
multi, trajectory, tense, mono, curved, narrow, nonbase

As for our second question, classifiers for pantomimes and classifier constructions again were each significantly accurate at 84.38% and 73.81%, respectively, indicating that transitivity distinctions are made by non-signer perceivers along phonetic grounds. Here again we see that classifiers trained on pantomime data were descriptively more accurate than those trained on classifier constructions, and we offer the same explanation as before. We additionally note that classifier accuracy on non-signer labels is descriptively higher than accuracy on ground truth (i.e., *production*) labels. We offer two explanations here: (1) The dataset is smaller (pantomimes: 432 vs. 32 samples; classifier constructions: 72 vs. 42 samples), and so there are fewer (extraneous) features for the model to fit. Further, the pantomimes in this analysis had already gone through two selection process: to wit, they are the subset of the *best* pantomimes (Study 1b, §3.2.2) that were significantly classed as being transitive or intransitive (Experiment 2a, §3.3). (2) Participants in the pantomime elicitation task were not explicitly asked to represent transitivity in their productions, while Experiment 2a participants were explicitly asked to assess the transitivity of those productions. To note, while our signer naturally encoded transitivity distinctions in her production of classifier constructions, we believe the second point–that Experiment 2a participants were directly asked to assess transitivity–also (partially) explains why classifier accuracy was higher in the perception analysis *vis-a-vis* the production analysis.

Finally, we note that MCC scores across all analyses were fairly robust (see Tab. 3.10), indicating that classifiers were indeed learning the desired pattern based on the features we chose. Generally speaking, there was a general intransitive bias, such that our classifiers were more likely to guess 'intransitive' than 'transitive,' though this was not always the case. See Appendix C for more details. In all cases, the MCC for classifiers trained on pantomimes was higher than those trained on classifier constructions, lending weight to the interpretation presented thus far, namely that pantomimes are not burdened by a linguistic system and are thus able to represent argument structure in a more faithful way, and that this is detectable via classifier analysis.[32]

[32]Conceivably, *accuracy* and the *MCC* could have pointed in different directions, with–say–classifiers trained on pantomimes achieving higher accuracies but lower MCC scores than classifiers trained on classifier constructions. In this case, we might come to the opposite conclusion, and argue that the argument structure of classifier constructions is more transparent, as the accuracy on pantomimes was only artificially higher. This latter possibility might entail that the classifiers guessed in a more targeted way for classifier constructions but in a more 'heuristic' way for pantomimes (e.g., like choosing option 'C' on a multiple choice test.')

Relatively few features were informative in the analysis of the perception of pantomime. In fact, only three were identified (Tab. 3.14), all of which overlap with the production analysis. Of there, two were related to the number of events encoded in a production, and one was related to visible tension.

Table 3.14.
Common most informative features across perception analysis of pantomimes. Items in bold also appeared in the corresponding production analysis.

Features (Perception, Pantomimes)
mono, **multi**, **tense**

As in production, there were nine features identified by our method that contributed most in transitivity distinctions in the perception of classifier constructions, of which six overlapped. These are listed in Tab. 3.15. Two end-marking, one eyegaze, and one path movement feature were identified, as well as the two features corresponding to the number of events encoded in a production.

Table 3.15.
Common most informative features across perception analysis of classifier constructions. Items in bold also appeared in the production analysis.

Features (Perception, CCs)
multi, aperturechange, **trajectory**, localmvmtfine, **tense**, deceleration, **mono**, **curved**, **narrow**

Finally, when comparing the features identified for classifier constructions and pantomimes in the perception analysis together, three features were shared (Tab. 3.16). As before, these features are related to the number of events encoded in a production, and the presence of visible tension in the signal.

Table 3.16.
Common most informative features across perception analysis of pantomimes and classifier constructions.

Features (Perception, Combo)
mono, multi, tense

Finally, to our third question, our results show that classifier accuracy is significantly above chance on items where non-signers accurately chose labels (i.e., the perceiver label matched the production label): 83.33% accurate for pantomimes, 81.25% accurate for classifier constructions. In this analysis, contrary to the others, classifier performance on samples from the pantomime and samples from classifier constructions was nearly identical. We also note that classifier performance on classifier constructions in the *production-perception* analysis is descriptively higher than performance on the same in the *perception analysis.* On the other hand, performance on pantomimes in these analyses is roughly equivalent.

Oddly enough, a nearly disjoint set of features were most informative in pantomimes that were classed correctly vs. those that were just classed consistently. As seen in Tab. 3.17, only the *tense* feature is shared, while two other features, *flex* and *away from*, are not. However, the features *multi* and *mono* do approach our cut-off frequency, which we taken to be significant.[33]

Table 3.17.
Common most informative features across perception analysis of pantomimes. Features in bold are shared with perception analysis.

Features (Accurate, Pantomimes)
flex, awayfrom, **tense**, (multi 70%, mono 63%)

[33]Note that all other features had frequencies of less than 50%, so there may be reason to group *mono* and *multi* in with the frequently occuring informative features.

By contrast, every feature in the analysis of accurately classed classifier constructions was present in the analyses of production and perception.

Table 3.18.
Common most informative features across perception analysis of classifier constructions.

Features (Accurate, CCs)
trajectory, multi, mono

Finally, there was no overlap, strictly speaking, between features identified as informative for accurately classed pantomimes and classifier constructions, though we do note again that *mono* and *multi* almost met our criteria and thus would have overlapped. Although accuracy on pantomimes was not significantly greater than on classifier constructions (§3.3.2), it appears that different features in the signal were responsible.

Table 3.19.
Common most informative features across perception analysis of pantomimes and classifier constructions.

Features (Perception, Combo)
(mono, multi)

Discussion

Classification Accuracy Here we have shown that transitivity coding is manifest in certain phonetic characteristics of pantomimes and classifier constructions. These phonetic characteristics were not arbitrarily picked, but rather were selected among features known to be related to transitivity (or event) encoding in sign languages.

Across the board, our analyses showed that both the encoding of transitivity in pantomimes and classifier constructions and the perception of transitivity distinctions can be predicted from certain phonetic features, though the specific phonetic features may be different. This suggests that transitivity distinctions are iconically coded in both, and iconically decoded in both. Further, given that the features selected for annotation are again derived from the literature on transitivity coding in sign languages, this gives weight to the idea that these linguistic features have their roots in vision/ iconicity.

We also note that classifier performance was not significantly greater for pantomimes over classifier constructions, counter prediction, though they were always descriptively greater. Without the confines of a linguistic system, we argue, pantomimes are generally free to make use of visual imagery in ways that classifier constructions cannot (for phonological/ other constraints on classifier constructions, see e.g. Benedicto & Brentari, 2004; Eccarius & Brentari, 2007, *inter alia*). However, this increased freedom on the form of pantomimes may be related to general event encoding rather than transitivity coding *per se*. For instance, while there was considerable variation (handshape choice, in particular) in our pantomime dataset for the event *adjust picture*, a consistent transitive strategy (here, the use of some handshapes, but not others) allowed for a transitive parse. As such, the pattern to encode transitivity in pantomimes might be more complex. To that end, we ran an ancillary analysis (Appendix D.3) where we trained classifiers on non-signer data and tested it on data from the signer. While accuracy was significantly above chance at 72.22% ($p = 0.0002$), all 67 features were needed for optimal accuracy. (Most other analyses needed far fewer to achieve comparable accuracy). We take this to mean that the non-signer transitivity-coding pattern was buried within all 67 features.

We anticipate the following question: *how can it be the case that classifier accuracy was significantly above chance for both production-derived labels and perception-derived labels, should those labels not be meaningfully overlapping?* First, recall from §3.3 that, by our first measure of accuracy, there was considerable overlap between production- and perception-derived labels. Second, classifier accuracy for the entire dataset was significantly accurate, suggesting that classifiers might be accurate for portions of it. We next anticipate the question, *If classifier accuracy is above chance using the entire dataset, and we might expect it to be high for proper subsets of the data, then isn't the fact that accuracy was high for the consistently classed subset and the accurately classed subset uninteresting?* Not necessarily. We might expect that the transitive-intransitive contrast is made manifest more robustly in one part of the dataset than in another. Consider the following items from our dataset:

		handshapes	endmarking	mvmt.	rel. mvmt.	rel. orient.	# events
move-box	TR	B^T- B^T-	decel.	path	together	opposing	mono
box-move	IN	B^T- B^T-		path	together	opposing	mono

The events *move-box* and *box-move* have the same truth conditions, only that the former additionally includes an agent. In this case, this may be distinguished by the feature *decel(eration)*, but this is just one of many features the classifier has to consider, and may only work for this particular pair of items. These pockets of low variance are difficult for the classifier to separate into categories, which is likely true for human parsers, too. The consistent and accurate subsets of the dataset may have been so chosen because the items within these categories show high variance with respect to argument structure marking.

Features: For each analysis, we report the top k features that aid in the intransitive-transitive distinction. We note with caution though that there could be other features (i.e., those not coded for) or combinations of features (i.e., informative interactions; again, MNB assumes independence of features) that could also be relevant. We note this in particular since the datasets are small generally, and extremely small specifically in the perception analyses (in some cases only 12 samples strong per class). It is also relevant to point out that, as a result, some features appear in some analyses and not in others, making solving the intransitive-transitive solution strictly local. That is, we may not have located a global set of transitivity coding features, but those that encode specific concepts only. Do note, though, that every analysis aside from the analysis using ground truth pantomime labels used complementary sets of concepts in the training and testing sets. The only issue here, then, is that not all concepts filmed were accounted for.

We attempted to address this last issue in an ancillary analysis (§D.2) in which we trained classifiers on a certain number of concepts and tested using a complement set of concepts. We did this solely on the pantomime production side (i.e., using ground truth labels) as this provided the largest dataset. As illustrated in Fig. D.2, classifiers achieved their targets at rates significantly above chance, thus suggesting that there is some meat to the features we coded for.

Just the same, we suggest that in future studies, the proportion of the revealed phonetic features could be experimentally manipulated in order to directly test their individual (or combined) effects. For example, we could manipulate eye-gaze (*look at dominant hand*, *look at non-dominant hand*, *look at camera*, etc.), such that it co-occurs with transitive and intransitive classifier constructions to see whether these cues inform transitivity classing.[34] A similar paradigm to Strickland et al.'s (2015) could be easily employed. Speaking to the two most prevalent features, *mono-* and *multi-eventive*, we could easily accomplish this by holding all other variables constant (handshape, end-marking, eye-gaze, etc.) and add gestural boundaries between some stimuli but not others. Further, we could manipulate what those boundaries look like (e.g., whether they involve deceleration, participation of the second hand, etc.)

Such experiments, however, already exist and provide proof of concept. As we noted in Chapter 2, Hassemer and Winter (2016, 2018) manipulated the flexion of selected and non-selected fingers to see whether they could elicit shape or size parses from their non-signer participants. While the authors use these terms, *shape* and *size*, it seems clear from their description that these might additionally have *intransitive* and *transitive* parses (shape = shape of reference; size = how referent might be manipulated).

[34]In the proposed experiment, for instance, we might expect that eye-gaze at the second hand with a transitive classifier construction might be consistently classed as transitive, but the same eye-gaze paired with an intransitive classifier construction might result in more variability in classing.

Basis for these visual features: A useful metaphor is thinking about what must a computer be taught when viewing human manual actions in order to understand them. Siskind (Siskind, 2003) attempts to do this with a limited set of transitive actions. Here, he models support (which assumes gravity), type of support, friction, and other physical properties inferable from either a single frame of a video or a few consecutive frames. For instance, the "camera" needs to know that hands and tables both offer support to a glass, and that, for instance, a hand may transfer support of the glass to the table. Further, if the table is level with the ground, the glass shouldn't slide. Siskind does this, by his admission, without recourse to handshapes and object affordances, which we concern ourselves with here. Work elsewhere (e.g., Siskind, 2001) models the timing relationships between subevents within *pick up*, *put down*, *assemble* events, among others. Explicitly, the force dynamics available in the signal, which change over time, feed into a semantic representation from which an appropriate verb can be produced.

Siskind works with a (linguistically grounded) understanding of actions and identifies properties of the scene which aid in that understanding. Our work is in the same spirit: the physics of the scenes we're dealing with are not only physical, but phonetic. Visual properties of the referent, combined with experience for how to manipulate them, inform what type of grasp is performed (handshape, but also what surface of the fingers are needed) and what orientation of the palm or wrist is required for successful handling, with–for example–larger objects and slippery objects requiring more fingers than small or rough objects (MacKenzie & Iberall, 1994). The visual system, as well as the intended function of the object in view, *restricts* the number and type of handshapes that can be used in a given situation, much like object NPs provide the restriction for handling classifiers in ASL and presumably other sign languages (Gökgöz, 2013).

3.6 Conclusion & Future Directions

In this chapter, we reported the results of two experimental pre-studies (§3.2.1, §3.2.2) used to create stimuli for our main experiment, presented in §3.3. The results of Experiment 2a suggest that non-signers can consistently class pantomimes and classifier constructions, presumably using vision-mediated strategies. The nature of these strategies, whether they proceed top-down or bottom-up, was taken up in §3.5.1 and §3.5.2, respectively.

The top-down analysis, in which we correlated consistency and accuracy measures with iconicity scores (derived from §3.2.2), did not yield very convincing results. The bottom-up analysis demonstrated that individual visual-phonetic features can encode and decode argument structure in both linguistic and paralinguistic stimuli. While there were nuances in classifier performance between stimulus types, the overall impression is that pantomimes and classifier constructions are on par in utilizing our select set of phonetic features. However, while features identified as being most informative were mostly shared between stimulus types and analysis type (production, perception, and their union), some distinct features emerged. We contend that some of these features are truly related to transitivity coding or decoding, but that some may be related to the coding of specific events. At present, it is difficult to suss out which is which, so we leave this to future studies.

We take our findings as a strong rebuke to top-down analyses of pantomimes, as suggested by McNeill (2005) and others (see §2.5), and tentatively suggest that the components of pantomimes encoding phenomena such as aspect and, here, argument structure, come from concepts evolved in the visual and visual-praxic domains, and likely others. We note that not all of the features we code for can be easily assigned to these domains, such as *eye gaze* (which may be related to joint attention). However, we are optimistic that contributions towards the encoding and decoding of messages from other domains can be found.

We also emphasize that the features we coded for in our corpus are derived from features known to be (even tangentially) implicated in the actual coding of argument structure in ASL and other sign languages, generally (see §2.2.2). The fact that these features hold linguistic status and are used by non-signers to encode pantomimes and make judgments about their 'formal' features suggests their grounding in a cognitive universal, as Strickland et al. (2015); Malaia et al. (2013, i.a.) suggest for deceleration and other kinetic phenomena in telicity marking.

Some future directions: Additional analyses could also be performed on the current dataset. Of particular interest might be the interaction between relative movement and relative orientation in two-handed pantomimes and classifier constructions, which was unanalyzed in our experiment due to the existence of known, spurious correlations between other features. For instance, in our classifier construction dataset, the signer uses two C-handshapes to describe the shape of a rolled-up poster. The hands are touching at the beginning of the construction, but then move outwards. In the next construction, the hands–retaining the same handshapes–move in unison to depict placing the poster on its side. The difference between the two events (the former intransitive, the latter transitive) is the relative movement between the hands: the hands move away from each other when describing the poster, but in unison when describing the (agent's) movement of the poster. However, to show the breaking of a stick, a transitive event, the hands (now in S-handshapes in our dataset, but could conceivably be C-handshapes if breaking something larger) move in mirror symmetry (change in relative movement) and change their orientations relative to each other (facing the same way, then opposing). To capitalize on such interactions between these features (and others), such an analysis as we are proposing would only include relative movement/ orientation features, and might fruitfully employ models that take correlations between features into consideration (SVM or logistic regression, among others).

At the same time, we note that the experiment presented here considers more event types than those used to establish transitivity coding in home sign systems, gesture, and new- and established sign languages (Brentari and colleagues). Specifically, many of the experiments cited here look at one verb of placing and one verb of existing, each with only a few object types (e.g., *place* {*airplane, lollipop*} *here* or {*airplane, lollipop*} *exist here*). Further, these studies limit the number of phonetic (or, likely, morphophonological) features to just handshape complexity measures (finger- and joint-complexity). The results of our experiment show that other phonetic features are relevant to transitivity coding and can capture more than two event types. Further, an ancillary analysis (see Appendix D.4) demonstrates that classifiers trained just on complexity measures achieves low accuracy on production (Pantomimes: 57.41%, $p = 0.0024$[35]; CCs: 56.94%, $p = 0.2888$) and perception (Pantomimes: 40.63%, $p = 0.3771$; CCs: 52.38%, $p = 0.8776$).

In the next chapter, we perform the same experiment as in §3.3 and analyses as in §3.5.1 and §3.5.2, but using ASL lexical verbs.

[35] Although this result is significant, 57.41% is descriptively low, especially compared with the results obtained using other features.

4. TRANSPARENCY OF TRANSITIVITY: ASL LEXICAL VERBS

4.1 Statement of problem, hypotheses, some (more) relevant background

American Sign Language (ASL), as a visual language, affords much more iconicity across multiple linguistic domains than spoken languages. The iconicity can be lexical (e.g., CAT traces the shape of a cat's whiskers), propositional (e.g., event participants are localized according to semi-geographical information; Bradley, 2013), semantic (e.g., sets and their restrictions can be represented visually; Schlenker, 2012), or grammatical (e.g., verbal reduplication can encode, e.g., event iteration; Wilbur, 2009), among other form-meaning or form-structure mappings.

Over time, signs have been documented to lose their iconic properties, at least with respect to their lexical semantics, due to articulatory and phonological pressure (Frishberg, 1975; Napoli et al., 2014). However, other iconic aspects to signs, including iconic grammatical features, are free to remain in the signal. As one example, Lepic et al. (2016) demonstrate that certain plural concepts are mapped onto two-handed signs in an iconic fashion. It is unknown, though, whether the encoding of plural concepts is iconic by analysis (the identity of the sign/ concept cues us into its iconic elements) or whether the form-meaning mapping is guided by some universal principle of perception, representation, or other cognitive domain. Strickland et al. (2015) demonstrate that general properties of vision (including kinematics, e.g., Malaia et al., 2013), showing that non-signers are sensitive to a verb's telicity based on how the sign looks, independent of its meaning. In this domain, there is evidence for a universal mapping bias between visual (phonetic) cues and linguistic structuring.

The lexical feature we examine here is *transitivity*. We aim to tease apart what is iconic by analysis, and what may be universally available mapping biases. Here, we explored whether transitivity distinctions are manifest in the phonetics of ASL lexical signs and, as such, have their basis in perception. If so, transitivity distinctions should theoretically be available to hearing, non-signers. We decompose the problem into several components. We first want to know whether non-signers consistently class verbs as being transitive, ditransitive, intransitive unergative or intransitive unaccusative. If non-signers are consistent in how they class verbs, there's evidence to support that they build a model of transitivity of ASL lexical verbs, despite not having access to their lexical properties.

We then want to know what guides this transitivity classing. We entertain two possibilities here: Classing is done from a top-down perspective, such that the identity of the meaning of the sign informs how non-signers class it. That is, for example, if a non-signer is able to guess that the sign BREAK means *break*, the non-signer may use their knowledge of their own language (or conceptual knowledge–we do not differentiate possibilities here) to guide classification. We assess this possibility by correlating a sign's consistency in classing with its iconicity score, provided in the ASL-LEX corpus (Caselli et al., 2016).

If instead transitivity classing is guided by perceptual features, and thus due to bottom-up processing, we would expect the phonetic characteristics of signs to be predictive of non-signer classing behavior. We address this possibility in a machine learning analysis wherein we use the phonetic characteristics of lexical verbs as features in an algorithm that predicts how non-signers assign transitivity labels to signs. The phonetic features for this analysis also come from ASL-LEX.

Finally, we assess the degree to which the model of transitivity non-signers construct converges with the actual transitivity ASL signs. That is, are non-signers accurate in their transitivity classing? If so, then we have evidence the transitivity of ASL lexical signs may have a basis in iconicity. We rerun the machine learning analysis, using ground truth labels instead of non-signer labels and using only non-signer labels that were accurate. The analysis using ground truth labels tells us whether transitivity distinctions are actually available in the phonetics of ASL lexical verbs.

The analysis using accurate non-signer labels tells us that perhaps only a subset of these phonetic features are salient to non-signers.

We make the following predictions: As lexical signs are generally thought to be less iconic than classifier constructions (with respect to lexical iconicity, Caselli et al., 2016),[1] we predict that their argument structure should be less iconic or even opaque. As such, we expect either (a) participants guess at chance as to what an item's argument structure is (e.g., 25% of participants guess that EAT is transitive, 25% that EAT is ditransitive, and so on for all four classes) and/ or (b) that only 25% of the total stimuli should be consistently classed.

As such, there would be no basis for the bottom-up approach (i.e., no non-signer consensus on a label means that there are no labels to train the classifiers), so we might hope for a top-down explanation. We again predict that consistency, wherever found, should nevertheless correlate with iconicity scores. Finally, we additionally hypothesize that non-signers are largely inaccurate in guessing the argument structure of lexical items.

In a snapshot, this chapter is laid out like so: We first describe our methods (§4.2.1) and report our results for transitivity classing and accuracy (§4.2.2). Our top-down (§4.3.1 and bottom-up analyses follow (§4.3.2). Finally, we discuss our findings in §4.4.

4.2 Experiment 3: Transitivity classing

4.2.1 Methods

We constructed a survey using Amazon Mechanical Turk (AMT) that simply asked participants to choose whether a given sign looks like:

[1] We do not base our hypothesis here on the stimuli's status as (unanalyzable) lexical verbs. As we mentioned in the experiment on classifier constructions and pantomimes, we cannot know ahead of time how non-signers parse these forms. We assume, then, that non-signers consider lexical verbs (and classifier constructions and pantomimes) as 'words' with or without internal structure.

(1) a. Someone/ something is acting on someone/ something else
(ex. grabbing, picking up, hitting, squeezing)

b. An object changes possession or is placed somewhere
(ex.: giving, taking, passing, borrowing, stealing, putting, placing, setting)

c. Something changes shape or location
(ex.: moving, oozing, sliding, deflating, exploding, wilting)

d. Someone is performing an action without an object
(ex.: walking, running, singing, dancing, whistling, sneezing)

Here, (a) was coded as a transitive event, (b) as a ditransitive event, (c) as an unaccusative intransitive event, and (d) as an unergative intransitive event. One example was given. The example included the video BREAK, the classification of BREAK as a transitive verb, and a short justification for why we selected the answer we did. The explanation was meant to calibrate the participants towards how we wanted them to think about the experimental items; participants did not have to provide justifications for their selections. The experiment immediately followed. Finally, participants were asked basic demographic questions (age, vision, English fluency, knowledge of a sign language beyond the manual alphabet and a few signs).

All 197 lexical verbs available from ASL-LEX were used in the survey, each video constituting a single survey item. As the number of survey items was unmanageable for any one person to do, we split the survey into six smaller surveys, five with 33 lexical verbs and one with 32. For each survey, we recruited 24 participants, for a total of 144 participants.

To be sure that participants grasped what was asked of them, we included three comprehension videos. These videos all depicted some real life action. One was intended to be intransitive, and depicted a block tower collapsing. A (c) answer was anticipated here, but (d) was also acceptable. One video was intended to be transitive, and depicted a person hammering a nail into a wooden box. Here the expected response was (a). The last was intended to be ditransitive and showed two people exchanging business cards. Here, we expected a (b) response.[2]

Three lexical verbs per survey (four in the case of the 6^{th} survey) were repeated in another survey. That is, for example, Survey 2 contained three verbs from Survey 1; Survey 3 three verbs from Survey 2; and so on. This was to ensure consistency in rating across surveys and justify treating all the survey takers as a single population, rather than six disjoint populations. Finally, one foil item was included. This item was a video that displayed text on it instructing participants to choose response (b). To be sure that participants could not determine which items contained comprehension or foil videos at first blush, we hid each video with a poster (a light pink `jpeg` image), which disappeared as soon as participants hit play. In total, then, there were 40 items per survey. (Again, 32-33 test items, three comprehension items, three to four repeated videos and one foil video). Participants were given the opportunity to read a consent statement. We considered the submission of the survey to be participants' consent to participating in the survey.

4.2.2 Results

Data processing:

Data were downloaded from AMT and fed through a series of python routines to (a) extract scores for each item, (b) tally them, and (c) compare the frequency of the most frequently selected option for each item against chance. Specifically, the data were first scrubbed of non-target responses and responses from participants who admitted knowledge of a sign language. Participants who did not respond as desired to more than one of the comprehension videos were considered to have not understood the instructions and their responses were excluded. Similarly, if participants chose an answer other than '(b)' for the foil trial, they were considered to be inattentive

[2] These expectations stem from a pre-study we ran using the labels in 1 and live action stimuli. We selected the stimuli from that study that were most consistently classed as (a), (b), or (c) as the comprehension items in this study.

and their answers were discarded.The responses to the comprehension trials and foil trial were similarly discarded, as these responses do not weigh in on the questions at hand.

We use a one-sample t-test of proportion against a chance distribution (=25% of each transitive, ditransitive, etc. responses) as a measure of consistency. Specifically, we first identify the top scoring response (using a simple tally), assign that response a value of 1, and zero-out all remaining responses. We compare this vector of 1's and 0's against the chance vector, which is matched in length but has precisely 25% 1's and 75% 0's. This generates a t-value for each item. This addresses the question *Do non-signers have intuitions on the transitivity of lexical signs?* Each video was assigned the label of the winning response.

Transitivity classing

Data from a total of 113 participants were retained for analysis. Sixteen participants were excluded for admitting knowledge of a sign language or not providing a response to this question. Nineteen participants were excluded for failing the comprehension questions. However, no additional participants failed the foil trail,[3] leaving data from 109 participants for further analysis. Four additional participants were recruited from AMT to make up for the four who failed the foil trial.[4]

We also looked at each participant's response distribution to check for biases. For each retained participant, we identified and counted their most frequent response. We then divided their most frequent response by their total number of responses. We took percentages over 50% to be indicative of bias towards a particular response. Only 10.42% (n = 15) of respondents had a clear bias by this metric. However, they were not uniformly biased towards a singular response: 7/15 (46.67%) chose '1' most frequently, 5/15 (33.33%) chose '4' most frequently, and 3/15 (20%) chose either a '2'

[3]That is, the participants who failed the foil trials (n = 4) were already excluded from analysis for having failed the comprehension trials.

[4]To note: with exclusions, 14 participants were retained in Survey 1, 20 in Survey 2, 21 in Survey 3, 20 in Survey 4, 16 in Survey 5, and 18 in Survey 6.

or '3' response most frequently. As none of these participants met our failure criteria, they were all retained. Further, as none of these participants seemed to form a class with respect to their response distributions, we did not perform any post-hoc analyses on their data separately.

From here, we compared the response distributions of the videos that were shared between surveys. We do this by comparing the labels output by our classing procedure. We counted two labels as dissimilar if (a) they were different (e.g., *trans* vs. *ditrans*) or (b) if they were the same, but one had statistical support and the other was assigned by simple tally (e.g., a *trans* labels was selected significantly more than chance for one item, but only by a simple majority in the other). In some cases, one of the repetitions could not be assigned a label by either method (i.e., it received equally many votes for two or more classes). These were also counted as misses. In total, only eight of the 18 repetition pairs produced identical labels according to our criteria. In only a few cases, this was due to unequal numbers of participants between surveys. In others, this argues for heterogeneity of the survey populations. Full results can be found in Appendix F. Repeated items are marked with '-rep.'

Of the 197 verbs, 127 were classifiable according to our criteria, or 64.47%. Including repeated items, 141 out of 197 verbs were consistently classed, or 65.28%. The distribution of responses is presented in Tab. 4.1. A sample of the results are presented in Tab. 4.2. The remainder of the results, including items that did not pass threshold, are presented in Appendix F. If we assume that non-signers guess randomly, we would expect that only 25% of responses would be consistent. Further, if we assume that the dataset was composed equally of transitive, ditransitive, intransitive unaccusative and intransitive unergative verbs, and we additionally assume that no transitivity-related information is available to non-signers perceptually, then it follows that we might expect non-signers–in aggregate–to assign a roughly equal amount of labels across the dataset (i.e., approximately 49 of each transitive, ditransitive, etc. labels). As such, that 64.47% of verbs were classed is beyond expectation (one sample t-test with hypothetical mean = 0.25; $t(196) = 11.5446$, $p \leq 0.0001$). This suggests

Table 4.1.
Tallies of consistently classed lexical verbs, where consistency is defined as maximum votes that were chosen significantly above chance (at $\alpha = 0.05$). Tallies on the left include repeated items. Tallies in parentheses do not. Nevertheless, a similar rate of consistent classing exists with and without repeated items. Lexical verbs had well over chance (=25%) rates of consistent responses, indicating that participants had some model of transitivity.

	Lexical Verbs	
Transitive	53	(47)
Ditransitive	9	(9)
Intransitive (E)	68	(60)
Intransitive (A)	11	(11)
Total	**141**/216	(**127**/197)
% dataset	65.28%	(64.47%)

that non-signers do have some model of transitivity distinctions based on something available in the visual signal. However, we do not yet know where this transparency of transitivity stems from: it could come from the identity of the sign (and therefore access to linguistic or conceptual knowledge already available to the non-signer) or it could come from individual visual features. The first option represents a top-down approach to transitivity resolution and is explored in the next sections. The second option represents a bottom-up resolution and is explored in §4.3.2.

Accuracy

Next, we ask whether the labels that non-signers assigned to consistently labeled verbs were accurate, that is, whether they matched the actual transitivity of the signs. By assessing non-signers' accuracy, we can zero-in on whether non-signers are figuring out the identity of the signs to guide their transitivity classification.

As in §B.2, we calculate accuracy in two ways. In the first, for each item, we assign a '1' to all correct participant responses and a '0' elsewhere. We then divide the number of 1's by the total number of responses on an item-by-item basis to obtain a percent-correct score for all 216 verbs (197 unique items + 19 repeated verbs). In the second, we compare non-signer labels, as determined statistically, against truth. That is, we compare the 141 the labels that were consistently assigned by non-signers (Study 2a; §4.2.2) to their ground truth labels.

For both methods, we compare the total percent correctly classed against chance. Chance in this case is 25% (*transitive*, *ditransitive*, *unaccusative*, and *unergative*). However, in order to make our results more comparable to those obtained for the analysis of classifier constructions and pantomimes, we additionally ran an analysis in which we group transitive and ditransitive (i.e., one or more objects) items, and unergative and unaccusative (i.e., no objects) together. In this way, there were 145 transitive items and 52 intransitive items (total 197) for the individual accuracy

Table 4.2.

Examples of non-signer classing of ASL verbs into transitive, ditransitive, intransitive unaccusative and intransitive unergative sets. The 'Tallies' columns represent raw counts. The label with the highest tally was compared against chance (=25%). A total of 141 verbs, all thresholded at p-values ≤ 0.05, were percolated into the machine learning analysis (§4.3.2). T-values from all verbs were used in the correlation analysis (§4.3.1).

†Degrees of freedom varied per analysis due to the exclusion of different numbers of participants in each survey. *Df* can be calculated from the tallies. * = $p \leq 0.0001$ (1-tailed)

		Tallies					
item	label	trans.	ditrans.	intrans. (A)	intrans. (E)	t†	Icon. Score
breakdown	intrans (A)	1	0	13	3	4.8536*	3.741
pretend	intrans (E)	4	1	1	15	4.5962*	1.667
read	trans	15	0	3	3	4.5962*	4.571
start	trans	15	1	1	4	4.5962*	3.871
stir	trans	14	1	1	3	4.6906*	6.739
write	trans	14	1	1	3	4.6906*	6.571
imagine	intrans (E)	3	1	2	14	4.2804*	3.8
push	trans	14	5	0	1	4.2804*	3.556
drop	ditrans	2	14	2	3	3.9528*	5.5

analysis and 109 transitive and 32 intransitive items for the consensus analysis (141 total).

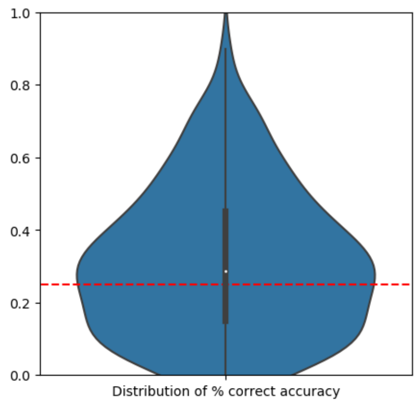
By our counts, the ASL-LEX verb database consists of 53 transitive verbs, 9 ditransitive verbs, 68 intransitive unergative verbs, and 11 intransitive unaccusative verbs, or 62 'transitive' items and 79 'intransitive' items. Ground truth labels were derived in consultation with a native signer who holds an advanced degree in linguistics. See Appendix E for details.

Individual level: For the analysis using all four labels, overall accuracy was low, yet significantly above chance, at 30.37% (STD = 0.1961; 1 sample t test against hypothetical mean, 25%, t(192) = 3.8116, 1-tailed p = 0.0001). The results are visualized in Fig. 4.1, where we can see that the existence of a few highly accurately classed items pulls the group average up. The probability density centers around chance, however.

The analysis using binned labels revealed that accuracy was again significantly different from chance at 52.76% (STD = 0.1994; t(215) = 2.0263, 1-tailed p = 0.022). These results are visualized in Fig. 4.2, where–again–the probability density is centered roughly around chance. Despite both results being significant, they do not appear to be meaningfully different from chance, indicating that the model that non-signers constructed around the visibility of transitivity of ASL lexical verbs is different from the one actually employed by the language, should there be one.

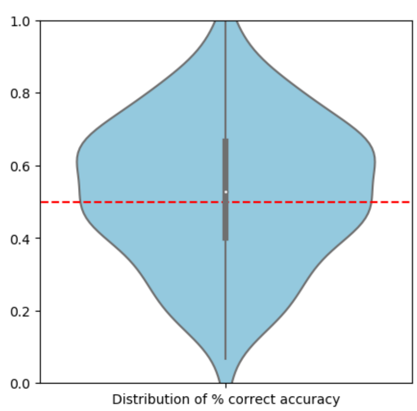
Consensus level: At the consensus level, mean accuracy wasn't much higher. Across all verbs, accuracy using all four labels was 37.59% (t(140) = 3.0753, 1-tailed p = 0.0013). Accuracy using the binned labels only returned marginally significant accuracy at 0.56.74% (t(140) = 1.6091, 1-tailed p = 0.0549). As with the results of the individual level analysis as with here, accuracy was not meaningfully above chance.

The confusion matrix in Fig. 4.2.2 shows that participants generally had a binary classing strategy, providing just about as many transitive labels as unergative labels (with very few ditransitive and unaccusative labels). By contrast, the dataset itself



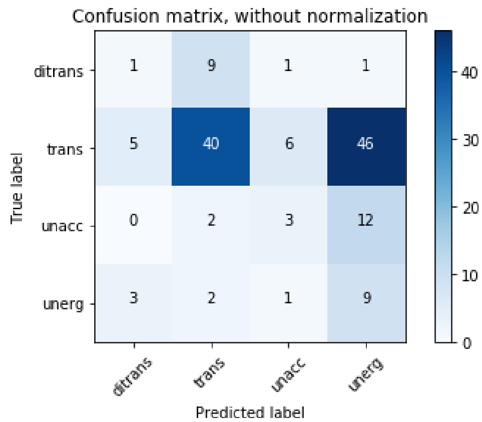
max	0.9 (1)
min	0.0 (8)
mean	0.3104
std	0.1997
median	0.2857

Figure 4.1. **Non-signer classing accuracy; 4 classes**: Violin plot with table of descriptive statistics showing overall *individual level* accuracy using all four transitivity class labels. Numbers in parentheses indicate how many items had the group maximum and minimum accuracy.

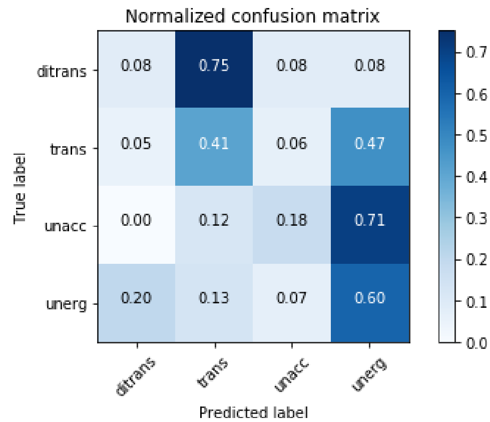


max	1.0 (1)
min	0.0 (1)
mean	0.5276
std	0.1994
median	0.5279

Figure 4.2. **Non-signer classing accuracy; 2 classes**: Violin plot with table of descriptive statistics showing overall *individual level* accuracy using binned ('transitive' - 'intransitive') transitivity class labels. Numbers in parentheses indicate how many items had the maximum and minimum accuracies.



(a)

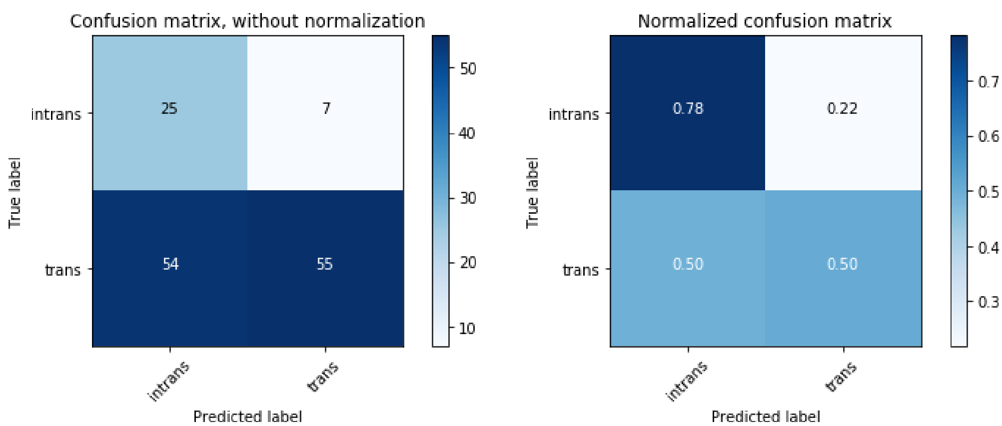


(b)

Figure 4.3. Confusion matrices for accuracy of non-signer labels (='predicted labels') against ground truth labels (='True label'). The figure in (a) shows raw counts; (b) shows counts divided by total number of ground truth labels in dataset (i.e., it is normalized).

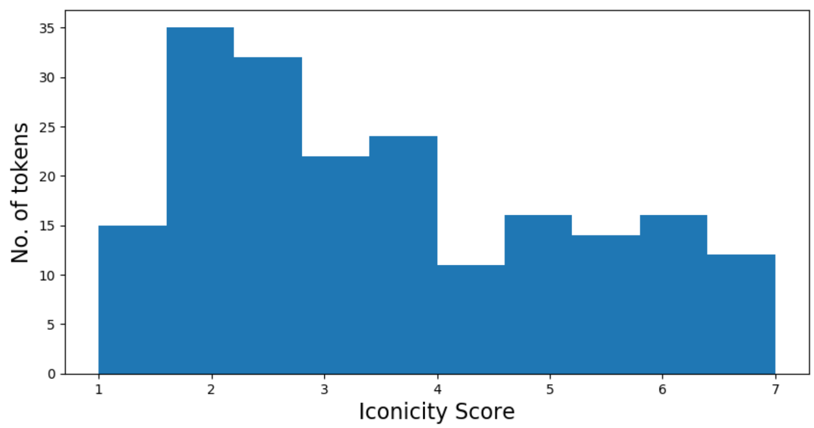
was heavily biased, with vastly more transitive items than any other category. So, although accuracy was above chance, this could be due to these two biases interacting with each other. We calculated the Matthew's Correlation Coefficient (MCC), which is a balanced metric for situations where the number of ground truth labels per class is unequal. The MMC was very low at 0.0928, indicating that there is only a very weak relationship between non-signer judgments and the actual transitivity of the stimuli.

The situation improves when we collapse object-taking items together and non-object-taking items together, likely due to the fact that participants chose primarily two labels, *transitive* and *unergative*. From Fig. 4.4, participants rarely gave transitive items an intransitive label. however, participants were very likely to give transitive items an intransitive label, and only sometimes gave intransitive stimuli the correct label. Still, the MCC improves: MCC = 0.2412. We tentatively suggest that gross object-no-object distinctions are more visually salient than specific transitivity distinctions.



(a) (b)

Figure 4.4. Confusion matrices for accuracy of non-signer labels (='predicted labels') against ground truth labels (='True label'). The figure in (a) shows raw counts; (b) shows counts divided by total number of ground truth labels in dataset (i.e., it is normalized). Plots demonstrate that non-signers generally chose transitive labels irrespective of what they saw. This artificially inflates the total accuracy figure (37.59%).



max	7.0 (1)
min	1.0 (1)
mean	3.5394
std	1.6335
median	3.16

Figure 4.5. Histogram illustrating the tendency for ASL-LEX verbs to be more arbitrary with respect to their lexical iconicity. The adjacent table shows some descriptive statistics. Numbers in parentheses indicate how many items have the max/ min iconicity score.

Iconicity Scores

We did not collect iconicity scores for the ASL-LEX verbs ourselves, as these values are already available in the corpus. Instead, we report here just some summary statistics and visualizations for use in explaining or illustrating the relationship these scores have with consistency and accuracy. A visible skew can be seen in the histogram in Fig. 4.5. A D'Agostino-Pearson normality test, which measures skewness and kurtosis, returned $s^2_{Lex} + k^2_{Lex} = 37.4683$, p < 0.0001, indicating that the distribution of iconicity scores is not normal.

4.3 Main analyses

4.3.1 Top-down: Classing behavior explained by previous linguistic/ conceptual experience

We asked what effect iconicity ratings had on the consistency of transitivity classing. We hypothesized that if participants were generally able to figure out what a sign means, they would be able to access linguistic of conceptual knowledge to resolve the transitivity of the ASL sign. That is, if a participant were to assume that EAT means *eat*, they may assign that verb a transitive label based on how *eat* functions in English. As such, we predicted that the higher the iconicity rating, the more non-signer agreement we'd find.

We are additionally curious about the relationship between accuracy and iconicity scores. That is, it is possible that non-signers converge on a transitivity parse that matches the sign's underlying argument structure. This would indicate that a verb's argument structure is accessible to non-signers through lexical iconicity, and provides a possible explanatory avenue for how argument structure developed in ASL signs through reanalysis or some other top-down strategy. We might expect accuracy and iconicity score to covary, even though accuracy across the dataset was very low. This is because, generally speaking, iconicity scores were generally low (skewed heavily towards arbitrariness).

We also correlate consistency with accuracy, again fully expecting that–as we have defined these measures–they will correlate strongly. That is, as accuracy approaches 0%, consistency approaches 100%; and, as accuracy approaches 100%, consistency again approaches 100%. Consistency should theoretically hit 0% as accuracy is around 50%.

To test these hypotheses, we separately correlated each item's iconicity score with its t-value and its accuracy score using Pearson's correlation coefficient, r. Our calculation of r also returns a p-value, signifying how likely a correlation of strength r could be generated from random data. As we noted in §3.5.1, this p-value is approximate, and improves as datasets get larger. As such, for that analysis we also provide results from randomly generated data. In the present analysis, as no correlation is particularly strong, we forgo additional analyses on randomized data.

Preprocessing

All 197 verbs were included in the analysis, not only those that were assigned a label as a part of the analysis in §4.2.2. The values we compare are characterized as such: we measure consistency by the magnitude of the t-value of an item. T-values were those derived from a 1-sample t-test against a hypothetical mean (25% *transitive*, *ditransitive*, *unergative*, *unaccusative*) as a part of Study 2a (§3.3). In cases where the t-value associated with an item was infinite (which occurred in cases, e.g., where non-signers uniformly picked one label), we simply rounded the maximum t-value occurring in the dataset to the nearest whole number and assigned the new value to that item. For repeated items, t-values from the repeated copy were discarded.

Accuracy was obtained via consultation with a native ASL signer who holds an MA in linguistics. She was given instructions and a questionnaire to fill out. The questionnaire, presented in Appendix E, asked the signer to identify if the verb could take any objects, and–if so–how many. This decides between *transitive* and *ditransitive* verbs. The signer was additionally asked to indicate whether the verb could take noun phrase (NP) or complementizer phrase (CP) objects.[5] Finally, the signer was asked whether the verb could be used in a sentence with the adverbial, WILLING, as this is a diagnostic used by Benedicto and Brentari (2004) to determine agentive subjects. This decides between *intransitive unergative* (agentive) and *intransitive unaccusative* (non-agentive) verbs. More details can again be found in Appendix E.

Accuracy was computed by-item. Items and their 'votes' were put into vectors, as schematized in the table below (Tab. 4.3). A key with the correct number (*1* corresponds to *transitive*, *2* to *ditransitive*, and so on) iterated over the vector and produced a new vector of 1's (hits) and 0's (misses), from which we then derived a percent-correct accuracy score. For repeated items, accuracies were averaged together.

[5]Though this information is not strictly necessary for the present analysis, this question was added for the following reasons: (1) it may be useful in a future analysis and (2) it was intended to prime the signer to think about CP objects as well as NP objects.

Table 4.3.
Accuracy pre-processing schema

accident	
response vector	[1, 3, 2, 3, 3, 2, 3, 4, 4, 3, 1, 1, 3, 3, 1, 1, 1, 4, 1]
key	[1]
result	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1]

Table 4.4.
Table of correlations: Correspondences between top-down measures, accuracy (Acc.), consistency (measure by magnitude of t-values, Tval.), and iconicity scores (Icon.). Linear coefficients (r) are provided with their associated p-values. R^2 values are also provided: note that in the Acc. x Tval case, data were fit with a second order polynomial. Otherwise, data were fit with a linear model.

Comparison		Pearson r	p-value	R^2
Acc.	Tval.	0.0281	0.6972	0.464
Acc.	Icon.	0.0279	0.6995	0.001
Tval.	Icon.	0.3317	<0.0001	0.11

Finally, iconicity scores–of course–were contributed from the ASL-LEX corpus. We did not process these numbers in any way.

To note, though, there was a small mismatch between our dataset and the ASL-LEX corpus. Somewhere in the pipeline, data from the following verbs were lost: *break_1*, *downsize_1*, and *grow*. As such, results from 194/ 197 verbs are reported. We do not anticipate that the missing data would have had a meaningful effect on the results.

Results & Discussion

Since we assume that the relationship between consistency and accuracy is non-linear (for reasons outlined above and more elaborately in §3.5.1), we computed R^2

using a second order polynomial instead of a deriving a Pearson correlation coefficient, which measures linear relationships only. Here, $R^2 = 0.4642$, supporting our hypothesized relationship between accuracy and consistency with medium effect. Incidentally, the Pearson correlation coefficient for this relationship was predictably low at $r = 0.0281$, p = 0.6972. This result is visualized in Fig. 4.6, where t-values are plotted against accuracy scores. Next, we found a small yet significant positive correlation between iconicity scores and consistency: r = 0.3317, p < 0.0001. The relationship is illustrated in Fig. 4.7. Finally, we found no correlation between accuracy and consistency (r = 0.0279, p = 0.6995), as visualized in Fig. 4.8.

The top-down approach would make the prediction that as lexical iconicity increases, so does non-signer consistency in labeling. Further, if the model of transitivity that non-signers construct is the same one that underlies transitivity distinctions in ASL lexical verbs, if there is any, we would expect to see accuracy increase (with consistency) as iconicity scores increase. As surveyed in §4.2.2 and §4.2.2 above, both iconicity scores and accuracy skewed low, perhaps suggesting a connection between accuracy and iconicity (e.g., non-signers were inaccurate because iconicity scores are low). However, we found no relationship between accuracy and iconicity scores. The strength of the correlation between consistency and iconicity scores was stronger, though by no means does it provide a satisfactory explanation of the data.

There were a few salient counterexamples to the general trend. For instance, DOUBT has an iconicity score of 2.185/7 but was consistently classed as *intransitive unergative*: t = 6.3723). The same is true of the sign PRETEND (icon. score = 1.667; t = 4.5962). Conversely, the sign THINK was 100% consistently rated as unergative (t > 10) and ergo 0% accurate (it is transitive), but had a high iconicity score (6.583/7).

Here, we offer that if a sign had no obvious marking such that it could be considered transitive (or ditransitive), it was assigned an intransitive label. That is, *transitive* and *ditransitive* are marked forms, while *intransitive* is an unmarked form. We then argue that out of the two labels, unergative or unaccusative, the unergative label had fewer stipulations in its definition (i.e., it was more neutrally defined): *cf. Something changes shape or location* (unaccusative) vs. *Someone is performing an action without an object*.[6]

Another possible source of the dissociation is in the concepts these three verbs denote: they're all cognitive verbs, or, verbs that do not have a physical or affected object. Although sign languages and gesture are both rich in metaphor, with abstract ideas being 'given' for example (McNeill, 1992), we assume these uses need additional elaboration (e.g., context) to be understood as such. We would expect, then, that verbs denoting physical activities might not exhibit such behavior, but we leave this to future studies.

Non-signers have nevertheless built a consistent model of transitivity, classing 141 out of 197 items, centered around something perceptual, which may be partially explained by the lexical iconicity of the sign as opposed to the transparency of the sign's argument structure. We address whether this *something perceptual* could also be partly grounded in the phonetics of the sign in the next section (§4.3.2).

4.3.2 Bottom-up: Classing behavior explained by perceptual features of the stimuli

That a significant majority of signs are consistently classed as transitive, ditransitive, etc. is only partially explained by iconicity scores (and not at all explained by accuracy). As such, some of the data may be explained by individual visual features that are available in the signal. Here, we intend to suss out what those features may be, if they exist, by feeding phonetic features of the signs to a machine learning algorithm. The winning classes as determined in §4.2.2 were used as labels, the features all come from the ASL-LEX corpus. We additionally used 'ground truth' labels,

[6]To note, unaccusatives have an underlying patient role (with its only argument base generated as a complement to the verb, but then promoted to subject). Perhaps the bias is related to the presence/absence of the patient role.

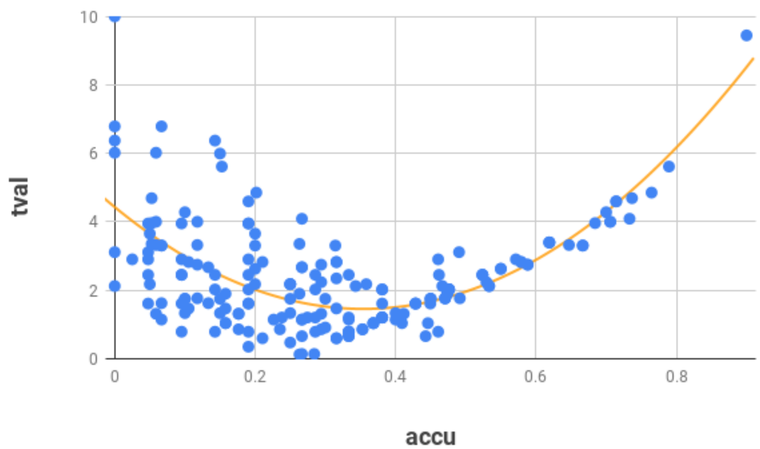


Figure 4.6. Scatter plot showing how accuracy scores and t-values (as a proxy of consistency of transitivity classing) are related by a second order polynomial.

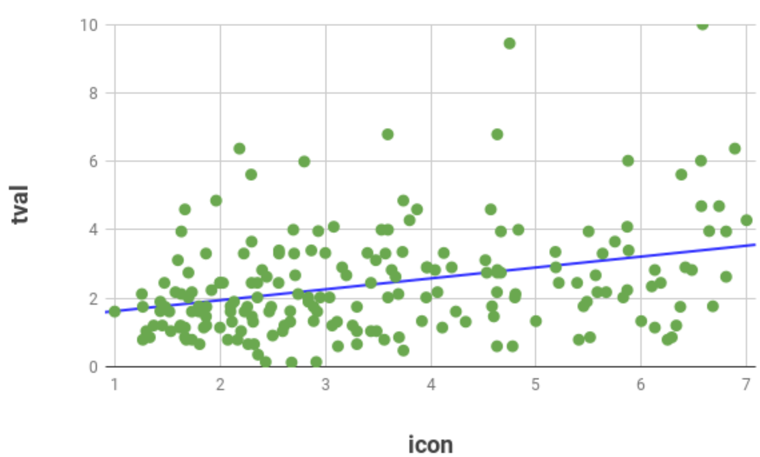


Figure 4.7. Scatter plot showing how t-values, as a proxy of consistency of transitivity classing, vary (weakly) as a function of iconicity score.

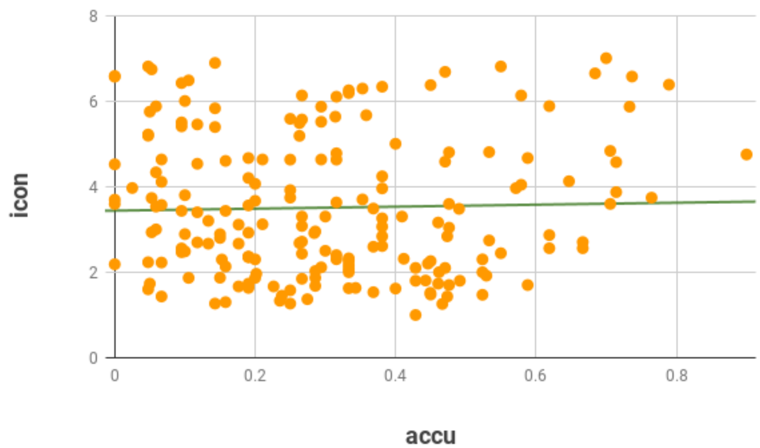


Figure 4.8. Scatter plot showing how accuracy and iconicity scores are not linearly related.

i.e., the actual argument structure of the lexical items to assess whether transitivity information is actually available in the signal.

Phonetic features

Although verbs were coded for 48 different classes of features in the corpus, we chose a list of just eight, including *sign length*, *sign type*, *major location*, *minor location*, *selected fingers*, *flexion*, *movement*, *mean iconicity.* A summary of phonological categories and their features is presented in Tab. 4.9. Within these eight feature categories, there were a total of 59 individual features.

This particular set was chosen in that they represent purely phonetic features, with exception for iconicity ratings. Other features in the corpus are lexical (e.g., sign frequency, neighborhood density, handshape and flexion frequencies, etc.)[7] and are thus not available to non-signers. Mean iconicity ratings were included as iconicity is available to non-signers by definition/ by hypothesis.

Sign type: Sign type is based on Battison (1978)'s observations that one- and two-handed signs are constrained in ASL. The four types of signs he identified are one-handed, two-handed symmetrical/ alternating, two-handed unsymmetrical with same handshape, and two-handed unsymmetrical with different handshapes. Caselli et al. (2016) also code a fifth category, 'other,' for those signs that do not fit the other four categories. This category is included not only because of its visual availability to non-signers, but also due to the iconic properties of two-handed signs in particular. Specifically, Lepic et al. (2016) demonstrate that two handed signs are often plural (which includes reciprocals; e.g., GATHER, MEET) or involve one event participant acting on another (e.g., HIT, FLATTER, HELP).

Location: Location is broken down into two subcategories: Major- and minor-location. Major location describes the general area of the body where the sign is articular (e.g.,

[7]Further, some annotations in the corpus are related to the annotation process (e.g., % gloss agreement) and are thus inappropriate for inclusion in the analysis.

head, trunk). Minor location, by contrast, gives a more specific location (e.g., eye, cheek, chin for verbs with *head* for a major location). There is no specific hypothesis tied to these features.

Selected fingers & Flexion: Select fingers refer to which fingers comprise a particular sign (Brentari, 1998). For instance, the 5-handshape involves all four digits and the thumb, while the 1-handshape only involves the index finger. Flexion described the degree to which the hand is open or closed. Combined, selected fingers and flexion represent features of the internal argument for transitive verbs (e.g., DRINK), and the sole argument for intransitive unaccusative verbs (e.g., GROW). For some ditransitive verbs (BRING, COPY, GIVE), selected fingers may also encode features of the direct object, though likely less iconically. As such, even though these cues are potentially ambiguous between transitive and intransitive unaccusative verbs (and potentially ditransitive verbs), their coding may help the decision between these verbs classes and those remaining. The features within the *selected fingers* category are *index*, *ring*, *mid*, *pinky*, and *thumb*, i.e., the individual fingers of the hand. The features within the *flexion* category were *full extend*, *mid extend*, *low extend*, *half extend/close*, *low close*, *mid close*, *high close*, and *stacked*.[8,9]

Movement: Movement is adapted from the [path] feature described in Brentari (1998), with an additional category ('zigzag') contributed by Hanke (2004). There are no specific hypotheses attached to these features. To note, given that the verbs are in 'dictionary' form[10], elements that may iconically encode transitivity distinctions, such

[8]In the ASL-LEX corpus, signs are annotated, e.g., *im* for 'index mid.' As annotated, each combination of fingers would constitute a separate feature, thus building in correlations between fingers. We instead chose to treat each finger individually, under the assumption that fingers could individually contribute transitivity information. We leave it to future research to see whether finger combinations are more or less informative.

[9]As annotated in the ASL-LEX corpus, *flexion* is denoted by the numbers 1 - 7, representing an open-to-close scale, and the feature, *stacked*. We changed the numerical labels to strings for readability.

[10]NB: It remains unclear whether sign languages have dictionary forms and, if so, what elements of the sign are included. We sampled a small number of verbs of different verb types in the ASL-LEX corpus. Non-body-anchored plain verbs are mostly articulated in neutral space, in front of the signer (e.g. DROP, WRITE, TRAVEL; but: RAKE). Agreement verbs are all articulated moving from proximal to distal loci in neutral space (e.g., ASK, FILM; but: MEET). Spatial verbs and backwards

as directionality (e.g., Börstell, 2017), may be absent from or represented differently in the corpus.

Sign length: We did not have a specific hypothesis surrounding sign length and transitivity classing. However, it is a feature of the signal that is available to non-signers and, as such, is a potential cue. If we may extrapolate, it has been noted that non-signers produce longer gestures depending on the aktionsart of the verb they're trying to express (Duncan, 2002). Aktionsart and transitivity overlap–though not completely–as many events involving objects are either accomplishments (e.g., *I painted the fence*) or achievements (e.g., *I found my lost keys*).

Iconicity: The iconicity scores used in the analysis above were also entered into the algorithm as features, in case one class of verb is more iconic than other classes (perhaps based on features inherent in other categories). We assume that, on average, iconicity score is available to all non-signers, despite not being a purely phonetic characteristic of signs.

Procedure

To make sure that there were an identical number of tokens of each class, we found the label that was the least represented in the dataset and capped the tokens of the other classes. In this case, there were only nine *ditransitive* items, so we randomly drew nine items from the *transitive* (45 items), *intransitive unaccusative* (10 items), and *intransitive unergative* (60 items) datasets. In total then, our dataset was very small at 36 total items. In order to compensate for how small the dataset was and to rule out the effects of specific verbs on classifier accuracy, we iterated through the procedure described below 10 times. Each time, we randomly drew a new set of nine verbs from each category (with replacement). Of course, the same nine ditransitive verbs were used in each iteration.

begin in neutral space, but then move ipsilaterally (e.g., GO, ZOOM-OFF, LEAVE; INTRODUCE, COPY). The reverse is true of spatial verbs with reverse semantics (e.g., COME).

Feature Classes			
sign length	sign type (Battison, 1978)	major location (Brentari, 1998)	minor location (ibid.)
Long ($>$ 1800 ms.)	OneHanded	head	HeadTop
MedLong (1200 - 1800 ms.)	Symm. or Alt.	arm	Forehead
MedShort (600 - 1200 ms.)	Asymm. Same Handshape	torso	Eye
Short ($<$ 600 ms.)	Asymm. Dif. Handshape	2nd hand	CheekNose
	Other	neutral space	UpperLip
			...

Feature Classes CONT'D				
Sel. fingers (Brentari, 1998)	Flexion (ibid.)	Movement (ibid.)	Iconicity (M)	Telicity
index	Fully open	Straight	HiIcon (5 - 7)	telic
middle	Bent (closed)	Curved	MidIcon (3 - 4.99)	atelic
ring	Flat-open	BackAndForth	LoIcon (1 - 2.99)	
pinky	Flat-closed	Circular		
thumb	Curved open	None		
	Curved closed	Other		
	...			

Figure 4.9. Feature classes (headings) and their composite features used in machine learning analysis. All features, except mean iconicity are phonetic.

Classification schema: The schema we follow here is nearly identical to the one described in the last chapter (§3.5.2). Each iteration had 36 verbs in the dataset, nine of each class. The features associated with these verbs were culled from ASL-LEX and then processed through a countvectorizer, which simply creates a dictionary of counts. That is, features from all samples are pooled to created a dictionary. Feature vectors for individual samples are then created. For each feature in the dictionary, a sample's feature vector receives a 0, 1, 2 and so on, depending on how many times a particular feature appears in the sample.

We then use a Multinomial Naïve Bayes (MNB) classifier. We split the dataset into six sets of six for a 6-fold, leave-one-out paradigm. Here, the classifier trains on five sets (or features from 30 verbs) and is tested on the 6^{th} (features from 6 verbs). This results in a singular accuracy score (*X out of 6 correct*). The process is repeated such that every set serves as the test set once. This generates a mean accuracy score (*X out of 36 correct*).

Feature Extraction: Not all features within all feature classes are generally informative for classification. As such, we used the following method to determine which feature class(es) were most informative. We again chose the *select-k-best* solution, using F-scores as our measure of *best*. Feature extraction was performed on the training set of each fold of each iteration. Information about feature informativeness was purged at the end of each fold, such that it could not inform the feature extraction process or classification in subsequent folds (a variation of training on the test set).

Random Analyses: We performed two types of random analyses to make sure that the patterns we observe in our main analysis are truly related to transitivity and not some spurious correlation. In the first, we randomly shuffle labels in the training sets for each fold of each iteration. If the classifier learns a pattern, we have evidence to suggest that there may be some surreptitious factor providing 'transitivity' information.

In the second analysis, we simply use features from the ASL-LEX corpus that non-signers, as non-signers, would not have access to. To wit, we included the feature categories, *Minimal Neighborhood Density*, *Maximal Neighborhood Density*, *Parameter-Based Neighborhood Density*, *Sign Type Frequency*, *Major Location Frequency*, *Minor Location Frequency*, *Selected Fingers Frequency*, *Flexion Frequency*, *Movement Frequency*, *Handshape Frequency*. The descriptions of these categories can be found in Caselli et al. (2016). However, suffice it to say, these categories represent lexical information about ASL verb signs, and thus should be unavailable to our sign naïve participants. That is, participants of Study 2a theoretically could not have used these features to guide their transitivity identifications. If we get comparable results using these features, then, it undermines any positive result we may obtain using non-signer-observable phonetic features.

On a technical note, as these categories contain continuous variables and the bag-of-word solution assumes discrete variables, we binned values within each category. For the former seven categories, we created four bins. For the latter three, we created seven. In total, then, there were ($7 \times 4 + 3 \times 7 =$) 49 features, just shy of the number of relevant, phonetic features we used in the analysis.

Results & Discussion:

Results are presented in Figs. 4.10 and 4.11, and in Tab. 4.5. We do not provide more detailed results as all accuracies were at chance. We also ran the analyses presented in this section using binned transitive-intransitive labels, collapsing transitives and ditransitives together and unergatives and unaccusatives together. While we don't present the details of these analyses, we report that none of them came back significant, with accuracies in the low-to-mid 50's (where chance is, of course, 50%).

On the production end–that is, using ground truth labels–classifiers identified their targets on average 35.31% of the time, significantly above chance (here, 25%; $p < 0.0001$). On the perception end, using labels derived from transitivity classing study (§4.2.2), we obtained a similar result: classifiers were on average 33.33% accurate, and again significantly so ($p = 0.0004$).

The analyses using random labels, as predicted, returned chance results: classifiers were on average 25.78% accurate ($p = 0.6481$) predicting ground truth labels, and 27.5% accurate ($p = 0.2736$) predicting non-signer derived labels.

Finally, in the analysis using lexical features, as opposed to phonetic features, classifiers predicted ground-truth labels achieved 35.31% accuracy ($p < 0.0001$), and those predicting non-signer labels achieved

At first blush, it appears that at least some phonetic correlates of transitivity can be found in both the production (phonetic form) and perception of ASL lexical verbs. Both analyses, the one using ground-truth labels and the one using non-signer derived labels, return 33.33% and 35.31% accuracy, respectively. Both results are significantly above chance. Further, the comparable analyses using randomly shuffled labels return chance performance, indicating that transitivity information is manifest in the signal on both ends.

However, we found that analyses trained on lexical features obtained equally good performance in predicting both ground truth and non-signer labels. With respect to ground-truth labels, it could be the case that transitivity distinctions co-occur with different sublexical properties, be it incidentally or by design (what exactly this design might be, we don't know). The same could be true of non-signer derived labels, too, in that these lexical features could incidentally form a pattern that aligns with non-signer transitivity judgments. Further muddying the results is the fact that random labels assigned to lexical features also provided comparable results, suggesting instead that there is surruptious learning in the analysis on the lexical features.

Ultimately, though, we take this to mean that transitivity distinctions are at best weakly associated with phonetic features underlyingly, and that non-signers only very weakly rely on phonetic features to guide their classification. However, given the analysis using lexical data, information that is by definition unavailable to non-signers, we argue instead that transitivity information is *not* available in the signal, and non-signers do not assume that any phonetic features correspond to particular transitivity classes.

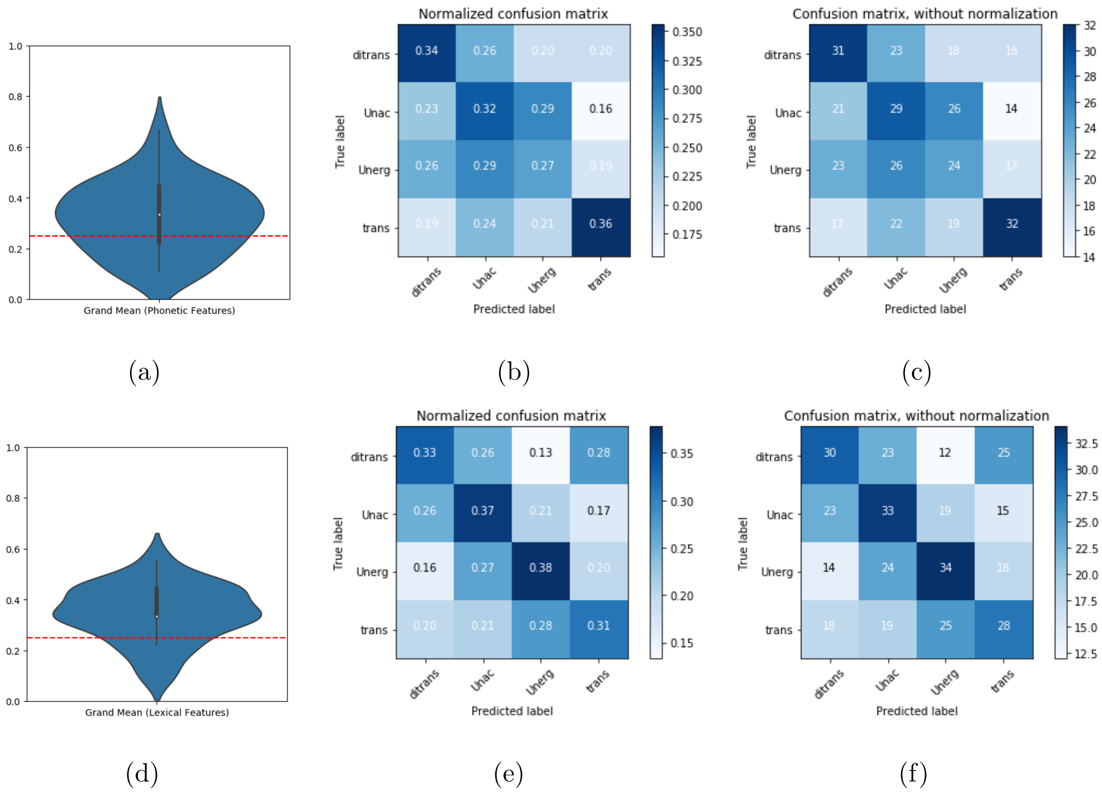
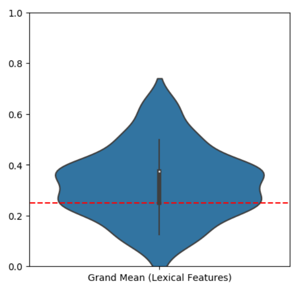


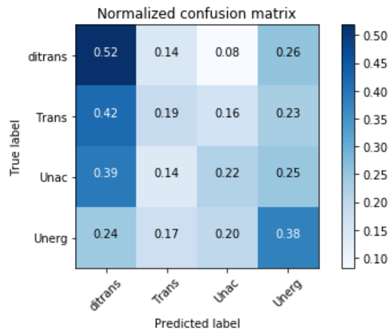
Figure 4.10. **Four classes, non-signer labels.** Violin plot of grand mean classifier accuracy. Plots represents the average of 10 4-fold classifier analyses, each time using samples randomly drawn (with replacement) from the corpus. (Top) Analysis using phonetic features from the ASL-LEX corpus (see Tab. 4.9). (Bottom) Analysis using *lexical* features from the ASL-LEX corpus.

Table 4.5.
Table of results, ASL-LEX machine learning analysis

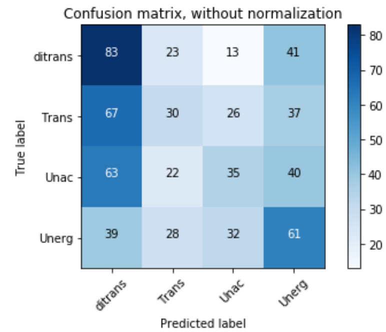
	Production		Perception	
	Phon. Feats.	Lex. Feats.	Phon. Feats.	Lex. Feats.
Mean	0.3265	0.3531	0.3222	0.3472
STD	0.0226	0.0358	0.1356	0.1125
p	<0.0001	<0.0001	0.0023	<0.0001
MCC	0.1047	0.1379	0.0964	0.1297



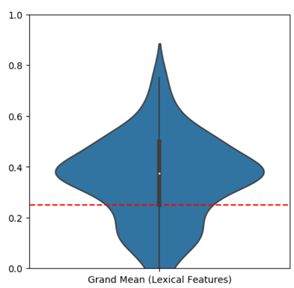
(a)



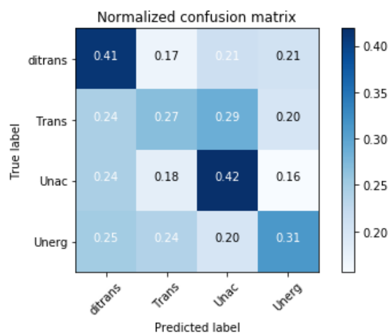
(b)



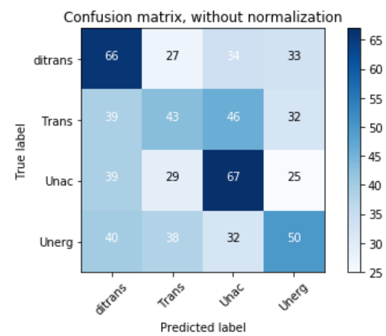
(c)



(d)



(e)



(f)

Figure 4.11. **Four classes, ground truth labels.** Violin plots (a,d) of grand mean classifier accuracy. Plots represent the average of 10 4-fold classifier analyses, each time using samples randomly drawn (with replacement) from the corpus. (Top) Analysis using phonetic features from the ASL-LEX corpus (see Tab. 4.9). (Bottom) Analysis using *lexical* features from the ASL-LEX corpus.

4.4 Discussion

The analysis of non-signer label consistency indicates that non-signers are guided by some property of ASL verbs, such that they more likely than not converge on a transitivity class for each verb in the dataset (141/197 or 71.57% consistently classed, including repeated items). This is unlikely to be due to chance. Our initial hypothesis that there would be no consistent classing, thus, was not supported. This suggests that these lexical verbs are transparent with respect to their *perceived* transitivity. We say *perceived* here in that overall participants were not very accurate when comparing non-signer labels with the actual ground truth transitivity of the verbs (only 39.72% accurate by our less conservative measure). This indicates, then, that the model of transitivity non-signers construct is mismatched with the actual transitivity coding strategy in ASL, should there be one.

Given the reported heterogeneity of the ASL lexicon with respect to iconicity (Lepic & Padden, 2017), we expected that individual phonetic features would not generally predict transitivity classes, although we might expect *local* coverage (i.e., certain phonetic features are informative to a [small] group of lexically related concepts, but not generally). On the other hand, we might instead have expected that lexical iconicity would be a better predictor. Through whatever mechanism or mechanisms, non-signers have indicated that the meanings some verbs are easily guessable.

Non-signers were more consistent in their transitivity classing of verbs with high lexical iconicity scores than those with low scores. Taken alone, this suggests that if non-signers are able to figure out the identity of the sign, they may rely on linguistic or conceptual knowledge to aid in their classing. However, taken together with the non-signers' low accuracy scores, we are left without a tidy explanation. We tentatively suggest that this mismatch is due to the many ways of lexicalizing an event: in a scene where a person breaks a stick, both *The person broke the stick* and *The stick broke* are true in that both are instances of *stick breaking*. We submit that in the absence of evidence (here, in the way of a consistent, identifiable/ iconic agent-marking or transitive-marking feature/ morpheme), non-signers are free to posit any event frame that is consistent with what they can glean from a sign.

To continue our example, while the event of *breaking* or *long-thin-object breaking* may be iconic, the causative or inchoative nature of the event might not be.[11] As a further speculation, we might guess in some cases, too, that events that involve similar kinematics or involve similar objects may be confused for each other. However, evidence for this does not directly fall out from the data that we collected.

Finally, but perhaps most critically, we emphasize that iconicity scores are not a true proxy of *transparency*, as we have been tacitly assuming. That is, iconicity scores are derived from non-signers' assessment of how much a sign looks like its provided meaning. We cannot assume that non-signers will find the same signs iconic in the absence of the their meaning. For instance, the sign WINK has a mean iconicity score of 5.75, was consistently classed (t = 3.6554), but was given a transitive label. In English, there is no alternate, transitive frame for *winking* (cf. **He winked his eye*, *He winked *(at) Joaquin*), with the result that the event was likely misidentified, if identified at all, by participants in our study. This weakens the conclusion that non-signers were more likely to converge on a label if they were able to figure out what a sign means and weakens the interpretation of the result that non-signers were more likely to converge on a label if they were able to figure out what a sign means.

The machine learning analysis explored to what extent non-signers classed a given verb as transitive, ditransitive, etc. based on phonetic features, or put another way, whether non-signers used bottom-up cues to guide transitivity classing. The individual iterations largely performed at chance, irrespective of what features from ASL-

[11] Incidentally, we note that in the production of pantomimes, both agentive and non-agentive strategies were used to convey intransitive events. In one or two cases, agentive and non-agentive strategies were used to convey transitive and intransitive events in our classifier construction dataset. In the sentence-production task, too, there was general variability in the transitivity of the labels assigned to both transitive and intransitive events, with only a percentage of events being unanimously labeled. In the absence of a direct comparison (e.g., *The person broke the stick* and *The stick broke*) or some other determining factor, and assuming that production and perceptions models are similar in this regard, we might expect such variable event framing.

LEX were included, indicating that perceived transitivity information is not inherent in these features.

This indicates that for most sets of verbs, the best classifying features were not themselves significantly good at predicting correct labels. This could mean that transitivity information is (a) not available in the signal, (b) is not carried by these phonetic channels, (c) is carried by these phonetic channels but only weakly (and/ or there are strong, unannotated features that do, too), and so on. The results could also be reflective of a lexicon that contains multiple different ways of encoding transitivity, based around–say–the semantics of a verb (e.g., it is causative, involves transfer, etc.). We expand on this below:

Before then, several caveats are in order. First, and most importantly, the analysis was run on an extremely limited dataset (n = 36 verbs, nine of each transitivity class). There simply may not have been enough tokens for the algorithm to learn an underlying pattern.[12]

At the same time, although the dataset was only 36 items large, we see this as reflective of the larger ASL lexicon in a way. Lexical items are free to be iconic to varying degrees, and many semantically-related verbs may cluster around one strategy or another (Lepic & Padden, 2017). For instance, what makes HIT and FLATTER iconic with respect to their transitivity is the fact that the non-dominant hand in these verbs represents a person on which the dominant hand acts. However, what makes verbs like BREAK and SWEEP iconic is that the hands look like they're holding an object. It is the body-anchoring in EAT, THINK, and TELL (together with handshape and/ or movement) that may give away their argument structure. Yet some verbs make use of path movement to show who is doing what to whom (e.g., GIVE, TELL). That is to say, there may not be any underlying phonetic property that

[12] We attempted another analysis wherein ditransitive and transitive verbs were grouped, and both classes of intransitive verbs were grouped. This regrouping resulted in significant classifier accuracy (56%, where chance = 50%). However, this is not a meaningful improvement over chance. We ran a second analysis, again using the transitive-intransitive grouping, but with randomly generated labels. Here, too, we got significant results (at just 53.25%), indicating that the significance fell out from the number of iterations we performed.

unites these verbs in their iconicity. There are some apparent counterexamples to this, too, wherein the handshape, movement, or other aspect of a sign betrays its apparent lexical iconicity. For instance, we repeat that WINK–an intransitive verb signed in front of the eyes–was given a transitive label, presumably due to the forefingers and thumbs coming into rapid contact with each other. As such, the distribution of iconicity of transitivity distinctions (in reality, or as judged by non-signers) may be misleading, splotchy and variegated.

Some support for this line of reasoning comes from the observation that a new set of features was most informative for classification in the bottom-up analysis every time a new group of verbs was selected.[13] That is, for the verbs selected in Iteration 1, sign length and selected fingers were most informative for transitivity classing. In Iteration 2, however, the transitivity of the selected verbs was best classified by iconicity score and movement.

In sum, then, a future analysis, might break verbs down into semantically motivated classes and look for form-meaning correspondences within. For instance, again, verbs denoting transfer are often directional verbs. A corpus study might break verbs down by the classes identified in Levin (1993) and rerun the analysis on particular groups of verbs (e.g., verbs of production, verbs of bodily emission, and so on). In a way, this probably makes more sense than an analysis of structure-meaning correspondences, as we have been assuming here.

For now, our conclusion runs counter to the iconicity mapping theory championed in Emmorey (2014), which argues for a compositional analysis of iconic features within signs. The data are also inconsistent with Strickland et al. (2015), who show that telicity marking is transparent ASL (and other sign languages), in that non-signers could accurately guess the class of a sign (telic/ atelic) at rates significantly greater than chance, all without knowing the identity of the sign. It is, of course, possible

[13]We do not report or analyze the most informative features for the bottom-up analysis in this chapter, since these features do not result in significant classification.

that some grammatical properties of signs remain transparent while others not (or that some are transparent and others are not and never were).

However, much of the data we leave unexplained. Thus, we are ultimately left with a puzzle. Non-signers were remarkably consistent in parceling ASL lexical items into transitivity classes, yet neither our top-down or bottom-up analyses provide convincing explanations for this behavior: consistency is only weakly correlated with lexical iconicity values with respect to the former, and our machine learning analysis revealed that phonetic features do not predict non-signer classing behavior. In a way, this is a happy result, in that our starting assumption was that iconicity is not as freely available in ASL lexical verbs (as compared to classifier constructions and pantomimes), given phonological and other linguistic pressures on the signal.

5. GENERAL DISCUSSION

In the two subsections below, we first review the results from both the classifier construction/ pantomime experiment and the experiment using lexical verbs (§5.1). We then mirror our discussion in Chapter 2 in light of the results we obtained in this dissertation (§5.2).

5.1 Synthesis of all results

5.1.1 Transitivity Classing

As can be seen in Tab. 5.1, all three stimuli types–pantomimes, classifier constructions, and lexical verbs–had similar rates of consistent classing (61%, 61%, and 64%,

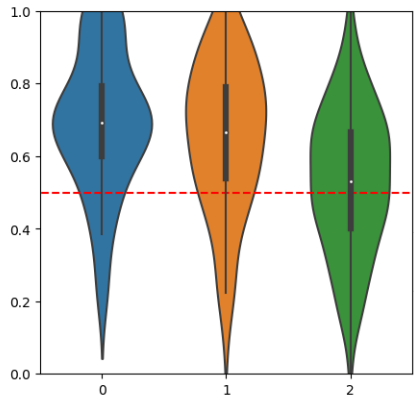
Table 5.1.
Tallies of consistently classed classifier constructions, pantomimes and lexical verbs. All three classes of stimuli had well over chance (=25%) rates of consistent responses, indicating that participants had some model of transitivity. For lexical verbs, numbers in parentheses indicate tallies *excluding* repeated items.

	Classifier Constructions	Pantomimes	Lexical Verbs	
Transitive	25	27	53	(47)
Ditransitive	1	4	9	(9)
Intransitive (E)	5	2	68	(60)
Intransitive (A)	13	10	11	(11)
Total	**44**/72	**43**/71	**141**/216	(**127**/197)
% dataset	61.11%	60.56%	65.28%	(64.47%)

respectively). Given our expectation that we should see roughly 25% consistency per category if participants just assigned answers randomly, all three stimulus types had above chance rates of consistency. That all rates should be in the low to mid 60's, we don't have a specific hypothesis about. However, we may have expected lexical verbs to have been classed less than the other two types given that they are relatively less iconic, generally and with respect to transitivity specifically.

Looking at the distribution of responses, participants generally chose a binary strategy, choosing *transitive* labels and *unaccusative* labels most frequently for pantomimes and classifier constructions, and *transitive* and *unergative* labels most frequently for lexical verbs. Explicitly, while classifier constructions and pantomimes had a large proportion of intransitive unaccusative labels assigned compared to intransitive unergatives, the reverse is true for lexical verbs. We have a two-part explanation. First, considering the intransitive stimuli for the classifier construction/ pantomime experiment were mostly agent-less, participants may have been sensitive to this. This is further confirmed by the result that participants were mostly accurate in their classing.

Second, we take the high proportion of intransitive unergative responses in the ASL-LEX task to be indicative of the 'neutral'-ness of the description of that option: "Someone is doing something without an object." While not revealed in that experiment, behavior of the Errant Group in the classifier construction/ pantomime experiment demonstrated that this option was the most neutral. That is, this group of participants, lacking evidence that a stimulus belongs to any of the other three categories, chose this option nearly exclusively. Again, given that they performed perfectly on the comprehension trials, which tested their understanding of the labels, this seems to be a genuine effect. Finally, the only meaningful relationship we obtained from the top-down analysis of lexical verbs (more on this below) was a correlation between iconicity ratings and consistency, but only for transitive items. Intransitive unergatives showed a small correlation, and the other two categories had noisy results, suggesting an *Aha!-It's-transitive* and *other* response pattern.



Cat.	Mean	1-tailed p
PN	0.6941	< 0.0001
CC	0.6514	< 0.0001
Lex	0.5276	0.0226

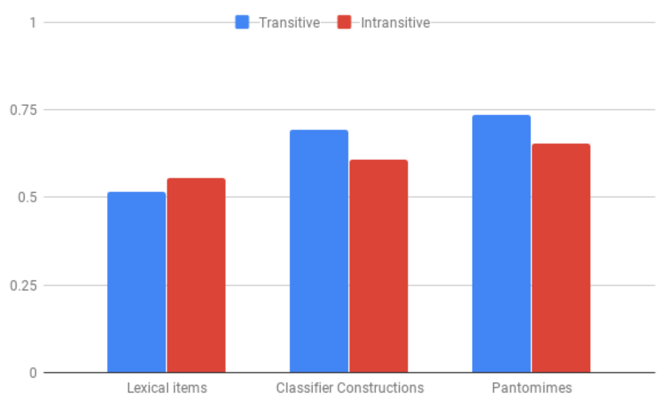
Comparison		Difference	1-tailed p (corr.)
PN	Lex	0.3848	0.0
CC	Lex	0.342	0.0
CC	PN	-0.0427	0.2754

Figure 5.1. Violin plots illustrating the probability density of individual-level accuracy on pantomimes (blue), classifier constructions (orange), and lexical verbs (green).

5.1.2 Accuracy

Collapsing transitives and intransitives together, the pattern that emerges is that the transitivity of pantomimes and classifier constructions was accurately guessed at a rate significantly greater than chance. Lexical items were also guessed accurately at above chance rates, but not meaningfully so at just 52.76% (chance = 50%). The results are presented in Tab. 5.1.footnoteIn this section, we only discuss individual-level accuracy for reasons of brevity.

When assessed separately (see Fig. 5.2), transitives were more accurately classed in pantomimes and classifier constructions than intransitives. For lexical items, transitive items were the only class (out of transitives, ditransitives, unaccusatives and unergatives) to be classed significantly more than chance, though we do not plot these results here. In aggregate, combining the accuracies of transitives and ditransitives on the one hand, and unergatives and unaccusatives on the other, we see that both 'transitive' and 'intransitive' super-categories were classed just about as accurately (with intransitives descriptively being classed more accurately). However, under this



	Trans.	Intrans.	1-tailed p
Lex	0.5171	0.5564	0.1054
CCs	0.694	0.6087	0.0388
PNs	0.7347	0.6523	0.0379

Figure 5.2. Accuracy across all three stimulus types, sorted by transitivity. Table shows numerical means for each transitivity class per stimulus type. The p-value represents difference in means across transitivity type, specifically whether larger mean is significantly greater. Here, since there was an unequal number of transitive and intransitive samples across the board, a Welch's t test was used to determine significance. 'Lex' = lexical verbs; 'CCs' = classifier constructions; and 'PN' = pantomimes.

'binned' analysis, neither transitive or intransitive lexical items were guessed accurately above chance.

At the same time, across pantomimes and classifier constructions, participants were generally more likely to class an item as transitive than they were to class it as intransitive, and the opposite is true for lexical verbs. This bias artificially inflates the accuracy rate to some degree.

In aggregate, as might be expected, the results demonstrate a cline in accuracy dependent on reported iconicity (and/ or the presence of a linguistic system), with the most imagistic stimuli (pantomimes) being classed most accurately, the somewhat less imagistic, partly linguistic stimuli (classifier constructions) being classed less accurately, and the mostly arbitrary stimuli (lexical verbs) being classed the least accurately. Further weight is added to the interpretation of this cline as we compare the iconicity scores of each stimulus type below.

5.1.3 Iconicity Scores

As expected, pantomimes and classifier constructions were on the whole rated more iconic than lexical verbs with respect to their lexical meanings. The comparison between all three stimulus types is visualized in Fig. 5.4. Though not shown here, transitive classifier constructions and transitive pantomimes were both significantly more iconic than intransitives, respectively. However, intransitive lexical verbs were only slightly more iconic than transitive lexical verbs.

The histogram in Fig. 5.3 shows the density of iconicity scores for each stimulus type. To note, the histogram is not normalized and there were more lexical verbs (197) than classifier constructions and pantomimes (73 and 72 items, respectively). As can be seen, iconicity scores of lexical verbs tend to skew low and in fact do not follow a Normal distribution (D'Agostino-Pearson test of normality: $s^2_{Lex} + k^2_{Lex} = 37.4683$, $p < 0.0001$). Both classifier constructions and pantomimes tended to skew towards being iconic ($M_{CC} = 4.4556$, $SD_{CC} = 1.1377$; $M_{Panto} = 4.9293$, $SD_{Panto} = 0.9879$).

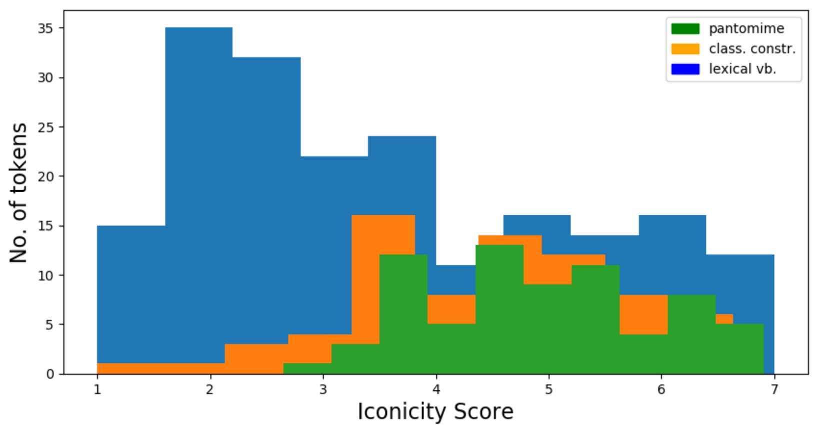
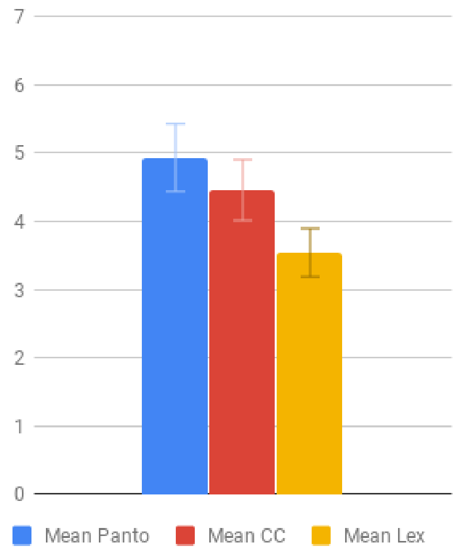


Figure 5.3. Histogram showing the distribution of iconicity scores across pantomimes (green), classifier constructions (orange), and lexical verbs (blue). Lexical verb data from ASL-LEX (Caselli et al., 2016). Note that the lexical verb dataset contains 197 verbs, while the pantomime and classifier construction datasets only contain 72 and 73, respectively, and the histogram is not normalized.



Type	Mean	SD
Pantomimes	4.9293	0.9879
CCs	4.4556	1.1377
Lexical verbs	3.5394	1.6335

Comparison	t	p (uncorr.)
Panto x CC	$t(142) = 2.7675$	$p = 0.0064$
CC x Lex	$t(268) = 5.1545$	$p < 0.0001$
Panto x Lex	$t(266) = 8.3728$	$p < 0.0001$

Figure 5.4. Mean iconicity of pantomimes and classifier constructions (Experiment 2b), and lexical verbs (scores from ASL-LEX). Pantomimes are in blue, classifier constructions in red, and lexical verbs in yellow. Some summary statistics and comparisons between means are presented in the attendant tables.

However, neither distributions of iconicity scores differ from a normal distribution: $s^2_{CC} + k^2_{CC} = 0.8652$, $\mathrm{p}_{CC} = 0.6488$; $s^2_{Panto} + k^2_{Panto} = 4.178$, $\mathrm{p}_{Panto} = 0.1238$.

To recap before moving on to the top-down and bottom-up analyses, we have seen that a comparable proportion of items from all three stimulus types have been consistently classed. However, non-signers were progressively less accurate at guessing the transitivity as we move from pantomimes, to classifier constructions, to lexical verbs. This same cline is found in iconicity ratings. As such, it would appear that iconicity ratings modulate or predict accuracy, which we explore next.

5.1.4 Top-down

In the top-down and bottom-up analyses below, we task to explain why participants were more consistent than expected by chance in deciding an item's transitivity. In the top-down analysis here, we test claims put forth by Klima and Bellugi (1979) and Lepic and Padden (2017) (among others) who argue that the identity of the sign is necessary for the analysis of its parts. We also test claims put forth by McNeill (2003) and Meir (2012) *inter alia* that pantomimes are unanalyzable. Here, we make the additional assumption that, lacking evidence, non-signers view classifier constructions similarly to pantomimes and lexical verbs– as either potentially decomposable or holistic.

Specific to the top-down hypothesis, we make the assumption that transitivity is a part of the phonological make up of a sign, classifier construction, or pantomime. By this analysis, we also assume that this trait (or traits) can be detected once participants guess the meaning of the production. Strong correlations between lexical iconicity (a proxy for the identity of the sign) and consistency (or accuracy) would imply that this process of reanalysis is occurring. We note that a positive result here does not entail a negative result for the bottom-up analysis and a negative result here likewise does not entail a positive one there.

The results of the correlation analyses are presented in Tab. 5.2. As before, the relationship between accuracy and consistency is meaningless in the current analysis, as these two measures are conceptually related to each other (e.g., if participants are 100% accurate, they are also 100% consistent). Of the two relationships of interest, all three stimulus types show a stronger connection between iconicity scores and consistency than iconicity scores and accuracy. We have generally took this to mean that the identity of the sign guides non-signer decisions, but due to the myriad ways to conceptualize an event (both in encoding and decoding), participants did not always settle on the correct answer. To borrow an example from another grammatical distinction, sign languages may choose whether to represent tool nouns using an entity or a handling strategy (Padden et al., 2013). For instance, the sign TOOTH-BRUSH in one language may be signed with the forefinger extended, representing the shape (long, thin) of a toothbrush, and move back and forth across the mouth. In another, the same concept is signed using the handshape required to *hold* the toothbrush. Both are or were verbal predicates that are used to refer to the noun used in tooth-brushing events. We think a similar explanation may be true here.

We note with curiosity that the strongest relationship between iconicity and consistency is exhibited in classifier constructions. We do not at present have a way to explain why this relationship is not actually stronger in pantomimes. For lexical verbs, though, we tentatively suggest that the multitude ways of encoding an event, family resemblances between related signs (Fernald & Napoli, 2000), and other features of a highly heterogeneous lexicon (at least with respect to iconicity; Lepic & Padden, 2017) contribute noise, but that sorting verbs by their phonological neighborhoods and running analyses on each neighborhood might improve results.

Also of note is that, with the exception of pantomimes, all correlations are stronger when just considering transitive stimuli, perhaps suggesting that information that specifically codes or represents *transitive* events is manifest in the signal. If intransitive events/ intransitive items are not explicitly coded, or their potentially iconic parts aren't particularly informative or iconic, then we might expect more variability in non-signer judgment. More succinctly, there might be something in transitive items (assessed by analysis) that identifies an item as transitive, but no such something for (many) intransitive items. On the other hand, all correlations are actually *weaker* among transitive items than intransitive ones when considering pantomimes. This flies against what we're arguing here, and what we'll continue to argue later (in the bottom-up analysis), so at present we offer no explanation for this.

In all, then, we are not very impressed by the results obtained here, and would wait to see the outcome of the bottom-up analyses before ascribing too much importance to them.

5.1.5 Bottom-up

Instead, if transitivity information is available in the signal on its own, and not necessarily via the identity of the sign, we should be able to predict non-signers'

Table 5.2.
Summary of top-down results across all three stimulus types. To note, R^2 is fit to a second order polynomial for the comparison *accuracy x consistency (tvalues)*, but is the square of the Pearson product-moment correlation coefficient elsewhere. r is the Pearson product-moment correlation coefficient everywhere.

		Pantomimes			Classifier Constructions			Lexical Verbs		
Comparison		r	p	R^2	r	p	R^2	r	p	R^2
Acc.	Tval.	0.4539	< 0.0001	0.267	0.5094	< 0.0001	0.456	0.0281	0.6972	0.464
Acc.	Icon.	0.3199	0.0065	0.1023	0.4011	0.0005	0.1609	0.0279	0.6995	0.001
Icon.	Tval.	0.4073	0.0004	0.1659	0.4904	< 0.0001	0.24049	0.3317	< 0.0001	0.11

Table 5.3.
Summary of top-down results across all three stimulus-types, separated out by transitivity type. To note, R^2 is fit to a second order polynomial for the comparison *accuracy x consistency (tvalues)*, but is the square of the Pearson product-moment correlation coefficient elsewhere. r is the Pearson product-moment correlation coefficient everywhere. '*' indicates significance at $p \leq 0.05$.

		Pantomimes				Classifier Constructions				Lexical Verbs			
		trans		intrans		trans		intrans		trans		intrans	
		r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2
Acc.	Tval.	*0.3807	0.145	*0.5283	0.454	*0.6806	0.624	0.2413	0.238	-0.0242	.363	0.2223	.678
Acc.	Icon.	0.2449	0.06	0.3071	0.0943	*0.3829	0.1466	*0.3316	0.1099	0.026	0.0007	0.1793	0.0322
Icon.	Tval.	*0.4153	0.1725	*0.4877	0.2378	*0.4898	0.2399	*0.3886	0.1510	*0.3863	0.1492	0.2298	0.0528

transitivity choices via the phonetic/ phonological makeup of the signs, classifier constructions, and pantomimes. This bottom-up ability was already demonstrated for telicity, as Strickland et al. (2015) demonstrated that non-signers made accurate telicity judgments about signs whose meanings they could not know beforehand.[1] The results would also support Emmorey's (2014) structure-mapping analysis of lexical verbs. The work by Emmorey and Strickland et al. would additionally be extended to classifier constructions and pantomimes. Again, a positive result here does not imply that reanalysis does not also occur, and *vice versa*.

We started with a more basic question, however: are transitivity-related features available in the signal at all? To answer that, we used data from all pantomimes, classifier constructions, and lexical verbs (separately). Mean classifier accuracy was significantly above chance for the former two, but at chance for lexical verbs (Fig. 5.5. T-tests confirmed that performance on classifier constructions and pantomimes was significantly greater than performance on lexical verbs, though performance on these two types did not differ significantly (see Tab. 5.4). As expected, visual features did reliably code transitivity distinctions in the more iconic stimulus types, but were not found to for the comparably more opaque lexical items.

This trend continued when we restricted the items to just those that were consistently classed, and just those that were accurately classed per Experiment 2a. We make the inference going from all items to just those consistently classed (and so on) that the increase in classifier accuracy observed is the result of 'human classifiers' selecting verbs that exhibit the most informative features. More explicitly, by the analysis of all pantomimes and classifier constructions, we demonstrated that transitivity information is available in the signal (i.e., in production). There are also cues relevant to event encoding, but perhaps irrelevant to transitivity classing. Non-signers have consistent (and accurate) judgments about transitivity for just a subset of these items. The distribution of features in these selected items, then, may be different

[1]While the meanings of the lexical signs used in the experiment could technically be guessed, and top-down reanalysis could occur, theoretically no such reanalysis could happen on the nonce stimuli the authors created.

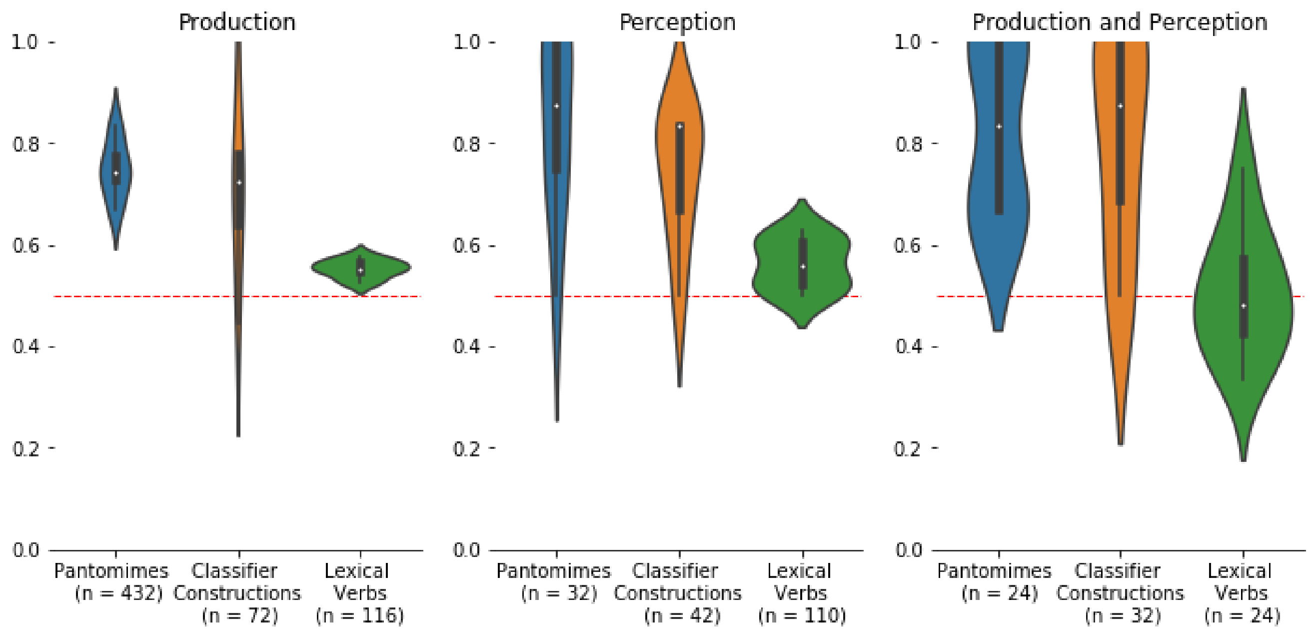


Figure 5.5. **Bottom-up results (PNs, CCs, and Lex):** Violin plots showing the distribution of classifier accuracies across a number of different analyses. In each, pantomimes are in orange, classifier constructions are in blue, and lexical verbs are in green; the red dashed line represents chance, here 50% (transitive or intransitive). Plots are divided by production (transitivity encoding), perception (transitivity decoding), and their intersection (what counts for both encoding and decoding). The n in parentheses indicates how many samples were used in each analysis.

from the distribution of features in the entire dataset. For instance, there may be proportionately more cues *relevant* to transitivity in the smaller set than in the entire set.

Considering the comparable performance of classifiers trained using pantomime and classifier construction data, we were curious to know whether classifiers trained on pantomime would succeed at classing classifier construction data. This might indicate that the information relevant to transitivity distinctions is available to both and perhaps organized the same way. This further suggests that handshape and handshape complexity were not necessarily the only places to look, despite Brentari and colleagues' (Brentari et al., 2012, 2017, 2015) finding that (a) these measures differ between transitive and intransitive classifier constructions and pantomimes, and (b) these measures may be recruited differently by signing and hearing populations. At the same time, we also ran an analysis using *just* handshape features or *just* handshape complexity measures. If handshape and/ or handshape complexity measures are enough to categorize transitive and intransitive events, we should expect significant classifier accuracy. Results of all analyses are presented in Appendix D. But, we will mention here that (a) yes, classifier accuracy was high on the train-on-non-signers-test-on-signer analysis, but (b) no, handshape/ complexity features were not sufficient for successful classification. From (a), our identification of just a few phonetic cues (mono/multi-eventivity, eye-gaze, and tension) as relevant to transitivity distinctions in both stimulus classes is supported. From (b), while we do not question the genuine effects Brentari et al. have demonstrated, we simply wish to note that there are other channels through which transitivity information is coded. What is more the presence-absence of this distinction may not be appropriately assessed just looking at handshape/ complexity measures.

On the other hand, we explain the difference in performance between the iconic stimulus types and lexical verbs as follows, noting that the features used in the analyses of these types differed considerably: as we argued, too, with the poor results obtained for lexical items in the top-down analysis, the heterogeneity of iconic de-

Table 5.4.
Aggregate bottom-up results: Comparisons between classifier performance on each stimulus type, organized by production, perception, and production ∪ perception.

	Comparison		Mean 1	Mean 2	Dif. in means	1-tailed p (corr.)
	CC	Lex	0.7083	0.5526	0.1557	0.0499
Prod.	PN	Lex	0.7477	0.5526	0.1951	0.0
	CC	PN	0.7083	0.7477	-0.0394	0.6151
	CC	Lex	0.7381	0.5618	0.1763	0.0165
Perc.	PN	Lex	0.8438	0.5618	0.2819	0.0049
	CC	PN	0.7381	0.8438	-0.1057	0.2983
	CC	Lex	0.8125	0.5042	0.3083	0.0078
Prod. ∪ perc.	PN	Lex	0.8333	0.5042	0.3292	0.0012
	CC	PN	0.8125	0.8333	-0.0208	0.8044

vices in the lexicon may have made transitivity patterns local to certain families of signs. Further, signs more so than classifier constructions and pantomimes may be under pressure of other sources of iconicity, unrelated to transitivity. EAT must look like *eating* before it looks like *eat something* (compare the sign EAT with the handling classifier CL:EAT-APPLE). Further, lexical signs encode more concepts than *something moves there* and *manipulate something to some effect*, making references to transitivity more variable: *eat something* is differently transitive than *hate someone* and *think something* with respect to iconic encoding. Only the first event might be coded by contact; the second by directionality; the third by no visual means that we can think of. As we advocated before, perhaps splitting the lexicon across semantic or phonological families would uncover cluster-specific transitivity.

5.2 Putting it all together

5.2.1 Argument structure in pantomime, ASL:

In this dissertation, we complement the discussion on object-handling strategies in pantomime (Brentari et al., 2012, 2017, *inter alia.*) by adding perception data. While Brentari and colleagues found variable, yet somewhat stable object-handling distinctions in production looking specifically at handshapes and handshape complexity, we found the same in perception. A significant portion of our pantomime stimuli were able to be consistently classed.

What is more, although Brentari and colleagues seem to avoid using the terms *transitive* and *intransitive*, we found a basis for the use of these terms: the labels we used in all of our experiments were couched in the semantic-driven definition of transitivity outlined by Hopper and Thompson (1980). We do hesitate to invoke more syntactically-defined notions of transitivity, however.

Turning to classifier constructions, we might have expected fewer items to be consistently classed, owing to potential erosion of iconicity by the linguistic system . For instance, one explanation that Brentari and colleagues give to the higher finger complexity exhibited by pantomimers in their handling productions is that their gesturers are making more fine-grained *How-would-I-hold-this* considerations. This, they argue, erodes over time and with sign language exposure. However, a roughly equal amount of classifier constructions as pantomimes were consistently classed. We contend that this is partly due to other features available in the signal, which we turn to in §5.2.2.

Finally, we surprisingly found considerable consistency in the classing of ASL lexical items. In fact, proportionately more lexical items were classified than both pantomimes and classifier constructions. This may be due to the number of participants kept for each study, as 24 / 96 participants were removed from the study using pantomime and classifier construction stimuli. Nevertheless, given the myriad ways lexical items may be iconic and their propensity towards being less iconic than classifier constructions and pantomimes, we may have still expected a lower number of classed lexical verbs. And, at present, we do not have a convincing way to explain why there were so many consistent judgments.

We have a few more notes to mention here: We want to emphasize the importance of considering both production and comprehension, as studies of elicited pantomime (Hall et al., 2013, 2014) and of homesign (Carrigan & Coppola, 2017) have demonstrated that producers and perceivers often have different communicative strategies. For instance, Hall and colleagues show that pantomime produces alternate between word orders, depending on the semantic reversibility of the event. Comprehenders, however, adopted a simple agent-first strategy, and so do not benefit from the word order alternations. On a similar task, Carrigan and Coppola show that the productions of four Nicaraguan homesigners is better understood by native ASL signers (access to a visually-based grammar) than by their own mothers (no access to a visually-based grammar), underscoring the role of the receiver in successful communication.

With successful communication–or accuracy–in mind, we report that non-signers were generally accurate at guessing the transitivity of both pantomimes and classifier constructions. Taking all items into consideration, participants were over 65% accurate in their judgments. However, if we only consider those items that were consistently classed (i.e., items that non-signers as a group had a strong opinion on), accuracy soars higher. This indicates to us that the transitivity of most events was transparent to non-signers, regardless of stimulus type. This is in line with what Hall and colleagues found, but out of step with what Carrigan and Coppola report. The high accuracy suggests to us the communicative use of the hands in the emergence of Language.

5.2.2 Iconicity in formal domains, argument structure:

Transitivity in this case is a superordinate term, covering all ways in which an event may be transitive: what effectors may be involved, what entities are involved (and their physical characteristics), and so on. Many of these different ways of being transitive are overtly or iconically marked in ASL (and probably other sign languages, too). By contrast, in handling classifier constructions, including transitive instrumental classifiers, transitivity may only be marked in a singular way: the shape of the hand has to be consistent with holding an object or a tool. This is likely why transitive events like *to offend someone* do not appear as classifier constructions; there is no intuitive handshape to be used here.

Some other, perhaps arbitrary way, then, is necessary to represent events like *offend*. ASL's solution is to use directionality, though it may surface as a plain verb in other languages as it does not metaphorically related transfer in some neat and tidy way. Yet another solution is to use two hands, as in HIT (also directional) and WRITE (not directional), or the body, as with SHAVE (not two-handed, not directional), and so on. As such, we contend that for lexical signs, knowing the identity of the event helps resolve that event's transitivity.

We see how iconicity can meaningfully interfere with contrasts in other linguistic domains, and do so differently in different pockets of the lexicon: Eccarius and Brentari (2010) show that iconicity meaningfully interferes with the constitution of the ASL lexicon, explaining–for instance–the paucity of distinctive minimal pairs in the language. Iconicity may block metaphorical extensions (e.g., the sign EAT cannot be used in an event like *The acid ate at the metal*; (Meir et al., 2007)). Further, iconiciy can block the emergence of grammatical devices (which may themselves be iconic), as in the case of the 'body-as-subject' metaphor blocking directionality in, e.g., ABSL ((Meir, 2012)). Iconicity can also explain why some forms are susceptible to allophonic variation, while others not. The authors give as an example the observation that [+stacked] and [-stacked] variants exist as allophones for the initialized noun, VERB, but not for the body-part classifier verb, FALL. The possibility of the former, they claim, is counter-iconic (the V-handshape looks less like a V), while the impossibility of the latter is meaning-preserving (the position of the fingers result in two different interpretations: having legs straight or akimbo). The differences are not randomly distributed, however, at least with respect to major lexical categories in sign language lexicons: they propose that iconicity has certain effects among lexical signs, among foreign vocabulary (e.g., initialized signs), and spatial lexicons.

The above was experimentally shown in this dissertation. The only explanatory variable we found in the transitivity classing of lexical verbs was the relationship between iconicity scores and consistency, albeit it was not as strong as we may have hoped for ($r = 0.3317$, $p < 0.0001$, $R^2 = 0.11$; correlation for just transitive verbs was higher: $r = 0.575$, $R^2 = 0.3295$). No phonetic feature or features predicted transitivity class (mean classifier accuracy for all analyses was between 50.52% and 55.26%, where chance was 50%).

This is consistent with Lepic & Padden's (2017) argument for the holistic nature of lexical signs. To reiterate, they claim that the identification of a sign's parts as being iconic or meaningful only derive from the meaning of the whole sign, echoing the work of Klima and Bellugi (1979), who first tested non-signers on the transparency of ASL signs. They, too, found that non-signers were generally insensitive to iconic parts of a sign unless they could obtain the global meaning of the sign. Furthermore, iconicity may only have these effects if it is accessible. Given that the iconicity scores of lexical items as a whole significantly skews low, non-signer access to these top-down strategies, we expected and found, is still rather limited.

On the other hand, the participants seemed to use both top-down and bottom-up strategies in their identification of transitive and intransitive classifier constructions and pantomimes. Which strategy had what effect we are unable to disentangle at this point; we simply note that both types of information are available. We do argue, though, that the bottom-up analyses returned results that are more convincing: classifier accuracy using phonetic features was consistently high across analyses, while we only obtained modest R^2 values for the top-down analysis (the strongest again being between iconicity scores and consistency, discounting the expected accuracy-consistency relationship).

With respect to the most informative, discriminant features, a particular two surfaced in nearly every analysis: *mono* and *multi*, both under the *number of events* category. For illustrative purposes, we reran a select number of analyses with just these two features and found significant classification. For instance, classifiers trained on data from the signer's data, using ground truth labels achieved 65.28% accuracy ($p = 0.0127$). 'Multi' appeared with transitive events 73.91% of the time, while 'mono' described intransitive events 61.22% of the time in this dataset.[2] While we mentioned above that 'multi' and 'mono' could be descriptive of the number of events *or* the number of syllables in a production, we argue that the former is the driving factor. Many of the transitive productions by the signer consisted of a causing event and a subsequent resultant event. For instance, in a video corresponding to *the ball knocked the bottle over*, the signer produced *the ball hit the bottle and the bottle fell.* The pantomimed equivalent used the same strategy. Incidentally, this type of strategy has been reported in established sign languages (HKSL, Tang & Yang, 2007; Danish Sign Language, Engberg-Pedersen, 2010), and also in homesign (one American participant, Rissman & Goldin-Meadow, 2017).

Finally, we have been discussing transitivity coding with respect to transitive items: what strategies do signers and pantomimers use to encode transitive events? However, we haven't said too much about a potential strategy for indicating intransitive events. In fact, for lexical items, the top nine most consistent items were rated as intransitive, of which only four actually were intransitive: TALK (trans), CRY (intrans.), SAW (i.e., 'to saw something'; trans.), DOUBT (trans.), TEAR (i.e. 'to shed a tear'; intrans.), UNDERSTAND (trans.), WORRY (intrans.), LAUGH (intrans.), and THINK (trans.). Some, like TALK and DOUBT, have low iconicity scores ($< 3/7$), while others were quite high (CRY and TEAR both had iconicity scores over $6/7$). We offer, then, that there is no marking for intransitivity. Rather, if the intransitivity of an item could be ascertained via iconicity (the case with CRY and TEAR), they were rated accordingly. However, if the transitivity, generally, of an item couldn't be deduced from iconicity, then a default intransitive label was given. That is, we argue

[2]We hasten to say that we do not imply that 65.28% of the data are explained by or 65.28% of the total accuracy is attributable to just these two features.

that this is why TALK and DOUBT, both low iconicity transitives, were classed as intransitive. Further, as suggested above, we infer from the Errant Group from Experiment 2a that the intransitive unergative label was seen as a default label. From Experiment 3, on lexical verbs, we see that this is the predominant intransitive label (unaccusative guesses were much more rare). This default strategy blurs what may have been actually driving unergative guesses, if there should be such a strategy.

By contrast, no obvious pattern of the likes of the above emerged from looking at the classifier construction and pantomime data. A more careful consideration of the phonetic corpus we created is, thus, necessary to look at coding differences between transitive and intransitive stimuli, which we leave to future research.

5.2.3 Emergence of grammatical features in visual communication systems: Holistic of compositional?

In Chapter 2 we outlined two grammatical processes in sign languages that have emerged via two different sources. Telicity marking (§2.3.2) on the one hand seems to borrow heavily from the visual system, recruiting kinematic features that underlie general event segmentation. On the other hand, the emergence of agreement marking in ISL, ABSL, and NSL (§2.4.1) appears to have occurred from the reanalysis of (a) holistic verb forms (ISL, ABSL) into component parts or (b) spatial markers into agreement markers (NSL). As such, we noted that the emergence of grammatical devices in sign languages isn't attributable to just one device; both processes are potential routes. On theoretical grounds, we also do not rule out the possibility for some grammatical features to be derived from both processes.

In the present work, we reported evidence that both processes may be active in the resolution of argument structure in classifier constructions and pantomimes, but not in lexical verbs, where top-down processes were more explanatory. Our findings stand in contrast to theories arguing that pantomimes, homesigns, and other developmentally early or paralinguistic forms are holistic, indivisible units across the board; a more nuanced look at these forms, with specific hypotheses about individual (grammatical) features is warranted.

Perhaps the most convincing way we have achieved this is by showing that our analyses on both pantomimes and classifier constructions returned similar results. Classifier constructions have been treated as multimorphemic for decades (e.g., Frishberg, 1975; Supalla, 1983, 1986), and even researchers arguing for a gestural account of these constructions (e.g., Liddell, 2003; Schembri et al., 2005, and–generally–researchers who use the term 'depicting construction') admit that they are composed of at least a handshape and a movement. The similarity in results in both the top-down and bottom-up analyses invites us to conclude that pantomimes can also be decomposed into meaningful components. This reinforces speculation by Wilbur and Malaia (2008) that (at least handling-type) co-speech gesture can be given a linguistic treatment. However, we want to say that we remain neutral with respect to gesture (in any instantiation) being a part of or deeply connected/ intertwined with the linguistic system, as our results do not necessarily bear on this question. We note though that there are substantial differences between gesture and sign (even in young sign languages and homesign systems; see §2), which suggest something *extra* needs to occur before a gestural system can be considered linguistic.

A critical limitation here is that we only looked at a subset of events: events of manipulation and of movement (although many sub-types of events were explored). We did not explore pantomimic representations of events that may not be as easily mapped to the body, such as verbs of *criticizing, wondering*, and so on. We also did not explore verbs from other domains that *can* be easily mapped to the body (e.g., verbs of emotion). Finally, our instructions to the pantomimers discouraged against full-body pantomimes, though a few surfaced anyway. As such, we cannot directly comment on the use of full-body pantomime observed in young sign languages (e.g., Aronoff, Meir, Padden, & Sandler, 2010) and homesigners (e.g., Goldin-Meadow et al., 1995). While we would still contend that even these pantomimes are decompositional, we cannot offer direct evidence of this.

On the other hand, the results for lexical verbs illustrate that perhaps the iconicity of transitivity, should it exist at all, is only accessible from the top down. As non-signers did agree on the transitivity of a great number of items, we have an indication that argument structure is iconic. But, as we only obtained meagre results from the top-down analysis and chance results from the bottom-up analysis, this consistency in judgment awaits a more convincing explanation. We ultimately take this to argue for Klima & Bellugi's (1979) and Lepic & Padden's (2017) point that 'sub components' of a sign's meaning–which we extend here to its argument structure–are only accessible through the meaning of the entire sign.

The results of the bottom-up analysis for pantomimes/ classifier constructions and those obtained for ASL lexical verbs cannot be directly compared, as different sets of phonetic/ phonological features were used in each analysis. For example, both analyses included handshape information: selected fingers and handshape were common to both analyses, but individual component features (e.g., [spread], [flat], and [base]) were only available to the analysis of classifier constructions and pantomimes. End-marking features, again, only partially overlapped, with more detailed information in the classifier construction and pantomime dataset (the ASL-LEX dataset was simply annotated for *telic* and *atelic*[3]). Some features, however, were simply unavailable to the ASL-LEX dataset. Take, for instance, the features *mono(eventive)* and *multi(eventive)*, which, again, may be rephrased as number of syllables (*1* or *more than 1*). ASL lexical verbs are predominantly monosyllabic (Brentari, 1998) and verbs (generally) typically express only a single event. As these two features in particular were informative in the analysis of classifier constructions and pantomimes, we may tentatively infer that the monosyllabicity of ASL lexical signs was partially responsible for low classifier performance. In this case, ASL's tendency to conspire towards monosyllabic signs erases this transitivity cue.[4]

[3]These are our annotations and are not available on ASL-LEX.

[4]Of course, we do not currently have any way of knowing whether transitivity was actually coded via number of syllables/ events. We can, however, check non-signer inferences by presenting them with one- and two-syllable signs and asking them which ones are transitive. This would at least answer whether there is some bias (among non-signers) in interpreting multisyllabic forms as transitive.

5.3 Future directions

Here we recapitulate some of the analyses, experiments/ experimental designs, and so on that we might wish to do or do over. We also provide some speculative support for our findings in the neurological/ behavioral separation of grasps and grasping on the one hand, and reaching on the other, both of which we assume are relevant (a) to the coding of transitive pantomimes and (b) to the coding of transitive pantomimes *bottom-up.*

5.3.1 Potential improvements to the current studies

1. Sort items by semantic type, and run top-down/ bottom-up analyses on each subtype separately. For instance, we might sort items into two groups, one of which denote transfer and the other not. Much of the discussion in this dissertation has revolved around notions of syntactic transitivity, although it may have been more fruitful to frame the discussion around semantic notions of transitivity (e.g., Hopper & Thompson, 1980), given that we have been concerned primarily with form-meaning correspondences (perhaps on the way to discovering form-meaning-structure correspondences).

2. By the same token, we might sort lexical verbs into phonological families supposing that those families have transitive and intransitive members. (An example family is WAR, OPPOSE, ARGUE, etc. which are all articulated with the fingers of the hands opposed to each other; example due to Lepic & Padden, 2017). Or, we could compare transitive families versus intransitive families, being careful to control the confound *family.*

3. Use lexical stimuli that are not in 'citation' form (e.g., are instead inflected for agreement, etc.). For instance, *directionality* as a potential cue was totally unavailable, though it marks recipients and themes with some regularity in sign languages (Gökgöz, 2013; Börstell, 2017). Further, it has been experimentally

demonstrated that hearing non-signers keep track of referents established in space (including who is doing what to whom) in both perception (Cassell et al., 1999; Schlenker & Chemla, 2018) and production (e.g., So, Coppola, Licciardello, & Goldin-Meadow, 2005; Perniss & Özyürek, 2015).

4. Use more objective measures than what we coded classifier constructions and pantomimes for. Ultimately, having a more accurate measure of relative position and orientation of the hands, displacement and velocity of the hands, aperture of the fingers, and so on would provide a more solid perceptual basis on which to make our claims (on par with, say, Malaia et al., 2013 and Hassemer & Winter, 2018). Likewise, coding the lexical stimuli for the same features would make direct comparison between all three stimulus types possible.

5. Code for features that may be relevant for intransitive stimuli. For instance, across all items in the classifier construction dataset, transitive items had in total 438 features, while intransitive items had 403. In the pantomime dataset, however, there were 2,433 features among transitive items while only 2,085 for intransitive items. This may have had an impact on identifying intransitive items in the classifier analysis, and our concept of just what is 'intransitive coding.' Potential relevant features may be total displacement (especially for intransitive verbs of motion) and gesture duration (related to intransitivity via aspect/ telicity).

5.3.2 Reaching, grasping

Here we offer some (very) speculative behavioral and neurological support for our hypothesis that pantomimes can be seen as internally complex, but that these subcomponents do not necessarily have to stem from the linguistic system *per se*. Take this as a direction where we want to go. We offer that these subcomponents may have their origins in the execution, perception, and comprehension of grasping behavior. We fully note that the same facts presented below have led some researchers (here, we're thinking of Arbib, 2010 specifically) to argue for a holistic analysis of proto-pantomime, which we take at the moment to be inseparable from modern notions of pantomime.

A single grasp is internally complex and generally proceeds along the following timescale: visual fixation on the target, a preparatory handshape phase, and a trajectory phase, which are all monitored in case updates or refinements to the path and handshape need to be made on the fly (Jeannerod, 1984). That is, while a grasp may look holistic, it can be broken down into several epochs, or schemata (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995). Further, these epochs can be separately impaired (reaching OK, grasp impaired: Jeannerod, Decety, & Michel, 1994), demonstrating that they do not necessarily form a holistic action sequence in the brain either.

On this last point, we should mention that there are two major types of grasps (MacKenzie & Iberall, 1994), each with constituent grasps.One class is the class of precision grasps, or those grasps which maximize stability, maximize surface area of the sensitive finger pads in contact with the object, etc. Among precision grabs are, e.g., those used for grasping the stem of a flower, holding a pencil (as if to write), or holding a cellphone to your ear. The other class is the class of power grabs, which sacrifice fine touching for a more stable grip. Examples of power grips are, e.g., the one used to hold on to the handrail of a moving bus, hoisting a wheelbarrow, or holding a beer.

Note that the choice of grasp may depend on the intended function of the object grasped, such that the same object (of course) may be grasped in different ways (e.g., a pencil used for writing, versus puncturing). Size of the referent object is also relevant, as a precision grip would be ineffective for, e.g., hoisting a heavy barrel (even if it did have a handle). To note, non-signers are aware of these factors as they gesture, as has been experimentally demonstrated in, e.g., Ortega and Özyürek (2016) and Masson-Carro, Goudbeek, and Krahmer (2016). In essence, we surmise, the handshape tells you about the object (its size and shape) and its function (holding for writing, holding for puncturing), and so both the object and the verb are potentially identified.

However, there are some notable limits to what we currently see as an explanatory root in the composition of a grasp. We offer two: the characterization of intransitive pantomimes and the extension of a simple grasp to a more complex system of movements that encode, e.g., *breaking*, tool use, and other manual actions.

First, as for intransitive pantomimes, at the moment, we can only offer the following explanation: if a pantomime cannot receive a transitive parse, it is intransitive by default. This is slightly infelicitous as we are forced to admit the following: (a) intransitives are default with respect to overt marking, yet (b) we assume that transitive actions/ transitive gestures were ontologically basic and evolved first (Arbib, 2005). However, we have the impression that this infelicity works its way into Language anyway, as transitives generally involve more morphology than intransitives (by way of Case marking, causative morphology, etc.; (Bybee, 1985)) but seem to us to be considered 'prototypical' sentences.

Second, while there is currently some understanding of the neurological and behavior correlates of the execution and observation of grasping and tool use, we are unaware of studies demonstrating the same for the expression of different or more complex actions.

Finally, to this we will append a small critique of Arbib's work (Arbib, 2005, 2010, *inter alia*), though in a few ways we depend on his ideas for our extrapolation here. Arbib defends holophrasis as a necessary step towards compositionality in protosign, the stage of human communication when the brain was 'ready' for Language. In his view, protosigns were used to refer to sometimes even complex events (e.g., a single gesture for 'Throw the spear when the prey animal is within range.'), a point Tallerman disputes (Tallerman, 2010), citing cognitive limits on tracking event participants in the visual system. And although he develops a sophisticated model for piecemeal action recognition, which we will elaborate a bit below, he still offers a McNeillian argument for holophrasis: "If I pantomime 'he is opening the door' there will [...] be no natural separation of noun and verb" (Arbib, 2010, p. 156). In this, he likely intends that there is a symbolization of the form-meaning correspondence between the proto-sign and the meaning *he is opening the door*, and that such symbolization forms a holophrase. We do not have a specific hypothesis that refutes this, and we may not even disagree, but we do want to focus on what makes the pantomime 'he is opening the door' appropriate, meaningful and communicative. At any rate, this way of thinking preempts the need to 'carve out' pieces of individual meaning (here, Arbib means words, but we also intend morphemes smaller than the word) from holophrases, since those pieces may have been coded into the form in the first place.

Arbib's preoccupation, and he's not alone here, is with word-hood and the stringing of words together into a proto-utterance. In other cases, like Aronoff et al. (2010) (same volume), the problem boils down to the lack of duality of patterning in pantomime (and, by extension, proto-sign). In this dissertation, we have shown that pantomime (and, by that same extension, proto-sign) can be thought of as compositional, only that the pieces come from extra-linguistic domains. We note that the bottom-up analyses of pantomime would have failed if this were not so (discounting for the moment other explanations), the top-down analysis would have been stronger, but we would have expected comparable or better non-signer classing results.

Turning now to these 'pieces,' in Arbib's terms, they may be pre-assembled in the praxic domain and recognized via the Mirror System. An identical or similar action schema can be selected and, in certain cases, meaningfully modified (where 'meaning' means to suit some specific purpose), as Arbib writes below.

> The ability to recognize another's performance as a set of familiar movements and then repeat them, but also to recognize that such a performance combines novel actions that can be approximated by (i.e., more or less crudely be imitated by) variants of actions already in the repertoire. (Arbib, 2005, 108).

We concede that the way forward we propose does not easily end up at an open-ended semantics. Our discuss is fairly limited to verbs of manipulation, such that events and their participants are identified by handshape, movement, and so on of a single 'word.' We briefly mentioned extensions, such as the use of the event *hold a baby* to refer to *baby* (as in ASL BABY; Napoli, 2017) or the use of entity and handling handshapes in naming tools (Padden et al., 2013), but we cannot offer much other supporting arguments at the moment.

5.4 Conclusion

In this dissertation, we compared lexical signs and classifier constructions with pantomimes produced by hearing non-signers with respect to how each type encodes transitivity information. We further asked whether that code can be inferred from the form of signs and pantomime, or, whether transitivity information is transparent. To our knowledge there is currently no formal study of this aspect of iconicity in sign or gesture. Intuitively, though, classifier constructions and pantomimes should be transparent in this respect due to their noted iconicity in other domains, whereas lexical signs may tend to be more opaque. Our exploration adds to our knowledge of grammatical features in (sign) Language that are simultaneously linguistic and iconic.

To start, we collected transitivity judgments from non-signers, asking whether a given pantomime, classifier construction or lexical verb was transitive, ditransitive, unergative, or unaccusative. We found that non-signers were largely consistent in their judgments across all three stimulus classes. This indicates that there is transitivity-related information in the signal.

We then wanted to know what type of reasoning underlies this consistency in classing. We argued that from a non-signer's perspective, we cannot assume any linguistic access to any of these stimulus types. As such, all three start on equal footing as potentially decomposable or holistic 'words.' Thus, the identification of a stimulus' transitivity could proceed top-down, from the identity of the stimulus' meaning. Or, non-signers could resolve transitivity information from the presence (or absence) of a certain (set) of characteristic phonetic feature(s) of the stimulus. We tested the former hypothesis by correlating consistency of classing with iconicity ratings, arguing that a meaningful correlation between these two measures indicates top-down access to transitivity. We tested the latter by annotating our stimuli for phonetic features and using a text classification algorithm (or classifier) to try to predict transitivity from these features. We reasoned that if the classifier performs well, participants may be using phonetic features to make transitivity decisions.

In so doing, we took to task claims that pantomimes and lexical verbs are holistic, nondecompositional wholes. Regarding pantomimes, we demonstrated the contrary, that pantomimes *can* be decomposed into pieces, and–further–that this process proceeds in both top-down and bottom-up processes, though the bottom-up process has more explanatory power. Further, the results we obtained for pantomimes were consistent with those observed for classifier constructions, drawing a parallel between how pantomimes and classifier constructions are encoded and decoded for transitivity. We note that classifier constructions have quite independently been known to be decomposable (e.g., Supalla, 1986; Zwitserlood, 2003; Benedicto & Brentari, 2004). In doing so, we found a set of consistent phonetic features that reliably code transitivity distinctions in both, noting that these features were drawn from the literature on transitivity, telicity, and agreement encoding in sign languages. Thus, we extrapolate and say that these features may have arisen from other cognitive domains (specifically, vision and praxis) and been encoded into transitivity. This weighs in on the discussion of the formalization of cognitive predispositions into (extant) languages and, we argue, the same into Language as it emerged in the species.

On the other hand, regarding lexical verbs, we found that bottom-up processing of transitivity in lexical signs was not possible: no feature of combination of features seemed to explain (a) the actual encoding of transitivity (if there is such a consistent strategy) *or* (b) non-signer transitivity judgments. However, we provided evidence that the identity of a sign could cue non-signers in to its (perceived) transitivity, in accordance with Klima and Bellugi (1979) and Lepic and Padden (2017).

In sum, we hoped to have shown that grammatical phenomena, like transitivity, can be linguistically and iconically encoded in ASL. We also hoped to have shown that pantomime is amenable to linguistic analysis on par with sign language. We do not wish to have argued for a conflation of linguistic sign and paralinguistic pantomime, however.

REFERENCES

REFERENCES

Arbib, M. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and brain sciences*, *28*(02), 105–124.

Arbib, M. (2010). Holophrasis and the protolanguage spectrum. In *The emergence of protolanguage: Holophrasis vs compositionality* (pp. 153–167). John Benjamins Publishing Company.

Arbib, M. (2012). *How the brain got language: The mirror system hypothesis.* New York, New York: Oxford University Press.

Argyle, M. (1975). *Bodily communication.* New York: International Universities Press.

Aronoff, M., Meir, I., Padden, C., & Sandler, W. (2003). Classifier constructions and morphology in two sign languages. *Perspectives on classifier constructions in sign languages*, 53–84.

Aronoff, M., Meir, I., Padden, C., & Sandler, W. (2005). Morphological universals and the sign language type. In *Yearbook of morphology 2004* (pp. 19–39).

Aronoff, M., Meir, I., Padden, C. A., & Sandler, W. (2010). The roots of linguistic organization in a new language. In *The emergence of protolanguage: Holophrasis vs compositionality* (pp. 133–153). John Benjamins Publishing Company.

Auracher, J. (2017). Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance. *PloS one*, *12*(11).

Bahan, B. J. (1997). *Non-manual realization of agreement in American Sign Language.* (Unpublished doctoral dissertation). Boston University.

Baker, S. A., Idsardi, W. J., Golinkoff, R. M., & Petitto, L.-A. (2005). The perception of handshapes in American Sign Language. *Memory & cognition*, *33*(5), 887–904.

Battison, R. (1978). *Lexical borrowing in American Sign Language.* Silver Spring, MD: Linstok Press.

Benedicto, E., & Brentari, D. (2004). Where did all the arguments go?: Argument-changing properties of classifiers in ASL. *Natural Language & Linguistic Theory*, *22*(4), 743–810.

Benedicto, E., Cvejanov, S., & Quer, J. (2007). Valency in classifier predicates: A syntactic analysis. *Lingua*, *117*(7), 1202–1215.

Börstell, C. (2017). *Object marking in the signed modality: Verbal and nominal strategies in Swedish Sign Language and other sign languages* (Unpublished doctoral dissertation). Stockholm University.

Bradley, C. (2013). *Motion events and event segmentation in ASL* (Unpublished master's thesis). Purdue University. West Lafayette.

Brentari, D. (1998). *A prosodic model of sign language phonology.* MIT Press.

Brentari, D., & Coppola, M. (2013). What sign language creation teaches us about language. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 201–211.

Brentari, D., Coppola, M., Cho, P. W., & Senghas, A. (2017). Handshape complexity as a precursor to phonology: variation, emergence, and acquisition. *Language Acquisition*, *24*(4), 283–306.

Brentari, D., Coppola, M., Mazzoni, L., & Goldin-Meadow, S. (2012). When does a system become phonological? Handshape production in gesturers, signers, and homesigners. *Natural Language & Linguistic Theory*, *30*(1), 1–31.

Brentari, D., & Eccarius, P. (2010). Handshape contrast in sign languages. In D. Brentari (Ed.), *Sign languages* (p. 284—311). Cambridge: Cambridge University Press.

Brentari, D., & Padden, C. (2001). Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. *Foreign vocabulary in sign languages*, 87–120.

Brentari, D., Renzo, A. D., Keane, J., & Volterra, V. (2015). Cognitive, cultural, and linguistic sources of a handshape distinction expressing agentivity. *Topics in cognitive science*, *7*(1), 95–123.

Bybee, J. (1985). Diagrammatic iconicity in stem-inflection relations. *Iconicity in syntax*, 11–48.

Carrigan, E. M., & Coppola, M. (2017). Successful communication does not drive language development: Evidence from adult homesign. *Cognition*, *158*, 10–27.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2016). ASL-LEX: A lexical database of American Sign Language. *Behavior research methods*, 1–18.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, *7*(1), 1–34.

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80.

Cogill-Koez, D. (2000). Signed language classifier predicates: Linguistic structures or schematic visual representation? *Sign Language & Linguistics*, *3*(2), 153–207.

Cormier, K., Quinto-Pozos, D., Sevcikova, Z., & Schembri, A. (2012). Lexicalisation and de-lexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language & communication*, *32*(4), 329–348.

Costello, B. (2016). *Language and modality: Effects of the use of space in the agreement system of Lengua de Signos Española (Spanish Sign Language)* (Unpublished doctoral dissertation). Universidad del País Vasco-Euskal Herriko Unibertsitatea.

Croft, W. (1998). Event structure in argument linking. *The projection of arguments: Lexical and compositional factors*, *2163*.

de Quadros, R. M., & Quer, J. (2008). Back to back (wards) and moving on: On agreement, auxiliaries and verb classes in sign languages. *Sign languages: Spinning and unraveling the past, present and future*, 530–551.

Dingemanse, M. (2011). Ezra pound among the Mawu. *Semblance and signification*, *10*(2011), 39.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, *19*(10), 603–615.

Dryer, M. (2007). Word order. *Language typology and syntactic description*, *1*, 61–131.

Dryer, M. (2011a). The evidence for word order correlations. *Linguistic Typology*, *15*(2), 335–380.

Dryer, M. (2011b). Order of subject, object and verb. In M. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/chapter/81`

Dukić, L., Milković, M., & Wilbur, R. B. (2010). Evidence of telicity marking by nonmanuals in Croatian Sign Language (HZJ). In *Poster presented at the 10th international conference on theoretical issues in sign language research (TISLR 10).* Purdue University: West Lafayette, IN, USA.

Duncan, S. D. (2002). Gesture, verb aspect, and the nature of iconic imagery in natural discourse. *Gesture*, *2*(2), 183–206.

Eccarius, P. (2008). *A constraint-based account of handshape contrast in sign languages* (Unpublished doctoral dissertation). Purdue University.

Eccarius, P., & Brentari, D. (2007). Symmetry and dominance: A cross-linguistic study of signs and classifier constructions. *Lingua*, *117*(7), 1169–1201.

Eccarius, P., & Brentari, D. (2008). Handshape coding made easier: A theoretically based notation for phonological transcription. *Sign Language & Linguistics*, *11*(1), 69–101.

Eccarius, P., & Brentari, D. (2010). A formal analysis of phonological contrast and iconicity in sign language handshapes. *Sign Language & Linguistics*, *13*(2), 156–181.

Emmorey, K. (2014). Iconicity as structure mapping. *Philosophical Transactions of the Royal Society B*, *369*(1651), 20130301.

Emmorey, K., Grabowski, T., McCullough, S., Damasio, H., Ponto, L., Hichwa, R., & Bellugi, U. (2004). Motor-iconicity of sign language does not alter the neural systems underlying tool and action naming. *Brain and language*, *89*(1), 27–37.

Emmorey, K., & Herzig, M. (2003). Categorical versus gradient properties of classifier constructions in asl. *Perspectives on classifier constructions in signed languages*, 222–246.

Emmorey, K., McCullough, S., & Brentari, D. (2003). Categorical perception in American Sign Language. *Language and Cognitive Processes*, *18*(1), 21–45.

Engberg-Pedersen, E. (2010). Expressions of causation in Danish Sign Language. *Sign Language & Linguistics*, *13*(1), 40–67.

Ergin, R., & Brentari, D. (2017). Handshape preferences for objects and predicates in Central Taurus Sign Language. In M. LaMdendola & J. Scott (Eds.), *Proceedings of the 41st annual Boston University conference on language development* (pp. 222–235). Somerville, MA: Cascadilla Press.

Fernald, T. B., & Napoli, D. J. (2000). Exploitation of morphological possibilities in signed languages: Comparison of american sign language with english. *Sign Language & Linguistics*, *3*(1), 3–58.

Fischer, S., & Gough, B. (1978). Verbs in American Sign Language. *Sign Language Studies*(18), 17–48.

Frishberg, N. (1975). Arbitrariness and iconicity: Historical change in American Sign Language. *Language*, 696–719.

Fujimori, A. (2012). The association of sound with meaning: The case of telicity. In A. Di Sciullo (Ed.), (pp. 141–166).

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological science*, *24*(7), 1079–1088.

Givón, T. (1980). The binding hierarchy and the typology of complements. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, *4*(3), 333–377.

Gökgöz, K. (2013). *The nature of object marking in American Sign Language* (Unpublished doctoral dissertation). Purdue University.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldin-Meadow, S. (2015). The impact of time on predicate forms in the manual modality: Signers, homesigners, and silent gesturers. *Topics in cognitive science*, *7*(1), 169–184.

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, *40*.

Goldin-Meadow, S., Brentari, D., Coppola, M., Horton, L., & Senghas, A. (2015). Watching language grow in the manual modality: Nominals, predicates, and handshapes. *Cognition*, *136*, 381–395.

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279.

Goldin-Meadow, S., Mylander, C., & Butcher, C. (1995). The resilience of combinatorial structure at the word level: Morphology in self-styled gesture systems. *Cognition*, *56*(3), 195–262.

Goldin-Meadow, S., Özyürek, A., Sancar, B., & Mylander, C. (2008). Making language around the globe: A crosslinguistic study of homesign in the United States, China, and Turkey. In *Crosslinguistic approaches to the psychology oflanguage: Research in the tradition of dan isaac slobin* (p. 27—39). New York: Psychology Press.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168.

Grose, D., Wilbur, R. B., & Schalber, K. (2007). Events and telicity in classifier predicates: A reanalysis of body part classifier predicates in ASL. *Lingua*, *117*(7), 1258–1284.

Hall, M. L., Ferreira, V. S., & Mayberry, R. I. (2014). Investigating constituent order change with elicited pantomime: A functional account of SVO emergence. *Cognitive Science*, *38*(5), 943–972.

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition*, *129*(1), 1–17.

Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *Proceedings of the language resources and evaluation conference 4* (Vol. 4, pp. 1–6).

Haspelmath, M. (2015). Transitivity prominence. *Valency classes in the world's languages*, *1*, 131–148.

Hassemer, J., & Winter, B. (2016). Producing and perceiving gestures conveying height or shape. *Gesture*, *15*(3), 404–424.

Hassemer, J., & Winter, B. (2018). Decoding gestural iconicity. *Cognitive science*, *42*(8), 3034–3049.

He, J., & Tang, G. (2018). Causativity and transitivity in classifier predicates in Tianjin Sign Language: A case study. In *Formal and experimental advances in sign language theory (FEAST 7).* University of Venice: Venice, Italy.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97.

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in grammar and discourse. *language*, 251–299.

Horton, L., Rissman, L., Brentari, D., & Goldin-Meadow, S. (2017). The emergence of agent-marking strategies in child homesign systems. In *Formal and experimental advances in sign language theory (FEAST 6).* University of Iceland: Reykjavik, Iceland.

H-Y. Tai, J. (1985). Temporal sequence and chinese word order. In *Iconicity in syntax: Proceedings of a symposium on iconicity in syntax, Stanford, June 24-6, 1983* (Vol. 6, p. 49).

Inkelas, S. (2014). Non-concatenative derivation: Reduplication. In R. Lieber & P. Stekauer (Eds.), *The oxford handbook of derivational morphology.* Oxford: Oxford Handbooks in Linguistics.

Jeannerod, M. (1984). The timing of natural prehension movements. *Journal of motor behavior*, *16*(3), 235–254.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in neurosciences*, *18*(7), 314–320.

Jeannerod, M., Decety, J., & Michel, F. (1994). Impairment of grasping movements following a bilateral posterior parietal lesion. *Neuropsychologia*, *32*(4), 369–380.

Kegl, J. (1990). Predicate argument structure and verb-class organization in the ASL lexicon. *Sign language research: Theoretical issues*, 149–175.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press.

Kimmelman, V. (2012). Word order in Russian Sign Language. *Sign Language Studies*, *12*(3), 414–445.

Kimmelman, V. (2016). Transitivity in RSL: A corpus-based account. In *Proceedings of the 7th workshop on the representation and processing of sign languages: Corpus mining, language resources and evaluation conference (lrec)* (pp. 117–120). Portorož, Slovenia.

Kimmelman, V. (2018). Basic argument structure in Russian Sign Language. *Glossa: a journal of general linguistics*, *3*(1).

Kimmelman, V., Pfau, R., & Aboh, E. (2016). Argument structure of Russian Sign Language classifiers: Insights from corpus data. In *12th international conference on theoretical issues in sign language research (TISLR 12).* La Trobe University: Melbourne, Australia.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, *48*(1), 16–32.

Klima, E. S., & Bellugi, U. (1979). *The signs of language.* Harvard University Press.

Kuhn, J. (2015). ASL loci: Variables or features? *Journal of Semantics*, *33*(3), 449–491.

Lam, C. (2013). Reduplication across categories in Cantonese. In *Proceedings of PACLIC 27.* Taipei, Taiwan.

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive psychology*, *60*(4), 291–318.

Lepic, R., Börstell, C., Belsitzman, G., & Sandler, W. (2016). Taking meaning in hand: Iconic motivations in two-handed signs. *Sign Language & Linguistics*, *19*(1), 37–81.

Lepic, R., & Padden, C. (2017). A-morphous iconicity. In C. Bowern, L. Horn, & R. Zanuttini (Eds.), *On looking into words (and beyond)* (p. 489-515). Berlin: Language Science Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago press.

Levin, B., & Hovav, M. R. (2005). *Argument realization.* Cambridge University Press, Cambridge, MA.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Liddell, S. K. (2003). *Grammar, gesture, and meaning in American Sign Language.* Cambridge University Press.

Lillo-Martin, D., & Klima, E. S. (1990). Pointing out differences: ASL pronouns in syntactic theory. *Theoretical issues in sign language research*, *1*, 191–210.

Lourenço, G. (2018). *Verb agreement in Brazilian Sign Language: Morphophonology, syntax & semantics* (Unpublished doctoral dissertation). Universidade Federal de Minas Gerais.

MacKenzie, C. L., & Iberall, T. (1994). *Advances in psychology, vol. 104. the grasping hand* (Vol. 104). Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers.

Malaia, E. (2014). It still isn't over: Event boundaries in language and perception. *Language and Linguistics Compass*, *8*(3), 89–98.

Malaia, E., Ranaweera, R., Wilbur, R. B., & Talavage, T. M. (2012). Event segmentation in a visual language: Neural bases of processing American Sign Language predicates. *Neuroimage*, *59*(4), 4094–4101.

Malaia, E., & Wilbur, R. B. (2012a). Kinematic signatures of telic and atelic events in ASL predicates. *Language and Speech*, *55*(3), 407–421.

Malaia, E., & Wilbur, R. B. (2012b). Telicity expression in the visual modality. *Telicity, change, and state: A cross-categorial view of event structure*, 122–136.

Malaia, E., Wilbur, R. B., & Milković, M. (2013). Kinematic parameters of signed verbs. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1677–1688.

Martell, C. (2005). *Form: An experimental in the annotation of the kinetics of gesture* (Unpublished doctoral dissertation). University of Pennsylvania.

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2016). Can you handle this? the impact of object affordances on how co-speech gestures are produced. *Language, cognition and neuroscience*, *31*(3), 430–440.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago press.

McNeill, D. (2000). *Language and gesture* (Vol. 2). Cambridge University Press.

McNeill, D. (2003). Aspects of aspect. *Gesture*, *3*(1), 1–17.

McNeill, D. (2005). *Gesture and thought.* University of Chicago Press.

Meir, I. (1998). Syntactic-semantic interaction in Israeli Sign Language verbs: The case of backwards verbs. *Sign Language & Linguistics*, *1*(1), 3–37.

Meir, I. (2010). Iconicity and metaphor: Constraints on metaphorical extension of iconic forms. *Language*, *86*(4), 865–896.

Meir, I. (2012). The evolution of verb classes and verb agreement in sign languages. *Theoretical linguistics*, *38*(1-2), 145.

Meir, I. (2016). Grammaticalization is not the full story: a non-grammaticalization account of the emergence of sign language agreement morphemes. In *Mediterranean morphology meetings* (Vol. 10, pp. 112–124).

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., ... Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, *158*, 189–207.

Meir, I., Padden, C., Aronoff, M., & Sandler, W. (2013). Competing iconicities in the structure of languages. *Cognitive Linguistics*, *24*(2), 309—343.

Meir, I., Padden, C. A., Aronoff, M., & Sandler, W. (2007). Body as subject. *Journal of Linguistics*, *43*(3), 531–563.

Milković, M., & Malaia., E. (2010). Event visibility in Croatian Sign Language: Separating aspect and aktionsart. In *Poster presented at the 10th international conference on theoretical issues in sign language research (TISLR 10).* Purdue University: West Lafayette, IN, USA.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, *55*(4), 259–305.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869.

Napoli, D. J. (2017). Iconicity chains in sign languages. *On looking into words (and beyond)*, 517.

Napoli, D. J., Sanders, N., & Wright, R. (2014). On the linguistic effects of articulatory ease, with a focus on sign languages. *Language*, *90*(2), 424–456.

Napoli, D. J., Spence, R. S., & de Quadros, R. M. (2017). Influence of predicate sense on word order in sign languages: Intensional and extensional verbs. *Language*, *93*(3), 641–670.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in psychology*, *5*, 376.

Neidle, C., Shepard-Kegl, J., MacLaughlin, D., Robert, G. L., & Bahan, B. (2000). *The syntax of American Sign Language: Functional categories and hierarchical structure.* MIT Press, Cambridge, MA.

Newmeyer, F. J. (2000). On the reconstruction of 'proto-world'word order. *The evolutionary emergence of language: Social function and the origins of linguistic form*, 372–388.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, *41*(1), 1–16.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In *Sound symbolism* (p. 325-–47). Cambridge England; New York, NY: Cambridge University Press.

Ortega, G., & Özyürek, A. (2016). Generalisable patterns of gesture distinguish semantic categories in communication without language.

Özyürek, A., Kita, S., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2008). Development of cross-linguistic variation in speech and gesture: Motion events in English and Turkish. *Developmental psychology*, *44*(4), 1040.

Padden, C. (1988). *Interaction of morphology and syntax in ASL* (Unpublished doctoral dissertation). University of California, San Diego.

Padden, C., Meir, I., Aronoff, M., & Sandler, W. (2010). The grammar of space in two new sign languages. In D. Brentari (Ed.), *Sign languages: A cambridge survey.* New York: Cambridge University Press.

Padden, C., Meir, I., Hwang, S.-O., Lepic, R., Seegers, S., & Sampson, T. (2013). Patterned iconicity in sign language lexicons. *Gesture*, *13*(3), 287–308.

Parrill, F. (2010). Viewpoint in speech–gesture integration: Linguistic structure, discourse structure, and event structure. *Language and Cognitive Processes*, *25*(5), 650–668.

Perniss, P. (2007). Achieving spatial coherence in German Sign Language narratives: The use of classifiers and perspective. *Lingua*, *117*(7), 1315–1338.

Perniss, P., & Özyürek, A. (2015). Visible cohesion: A comparison of reference tracking in sign, speech, and co-speech gesture. *Topics in cognitive science*, *7*(1), 36–60.

Perniss, P., Özyürek, A., & Morgan, G. (2015). The influence of the visual modality on language structure and conventionalization: Insights from sign language and gesture. *Topics in cognitive science*, *7*(1), 2–11.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, *41*(1-3), 47–81.

Quer, J. (2011). When agreeing to disagree is not enough: Further arguments for the linguistic status of sign language agreement. *Theoretical Linguistics*, *37*(3-4), 189–196.

Quer, J. (2017). The right place for locative agreement in sign languages. In *The 53rd Annual Meeting of the Chicago Linguistic Society.*

Quinto-Pozos, D., & Parrill, F. (2015). Signers and co-speech gesturers adopt similar strategies for portraying viewpoint in narratives. *Topics in cognitive science*, *7*(1), 12–35.

Rathmann, C., & Mathur, G. (2011). A featural approach to verb agreement in signed languages. *Theoretical Linguistics*, *37*(3-4), 197–208.

Rissman, L., & Goldin-Meadow, S. (2017). The development of causal structure without a language model. *Language Learning and Development*, *13*(3), 286–299.

Rissman, L., Horton, L., Flaherty, M., Brentari, D., Goldin-Meadow, S., Senghas, A., & Coppola, M. (2016). Strategies in gesture and sign for demoting an agent: Effects of language community and input. In *Evolang: Proceedings of the 11th international conference.*

Sandler, W., & Lillo-Martin, D. (2006). *Sign language and linguistic universals.* Cambridge University Press.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*(7), 2661–2665.

Schalber, K. (2006). Event visibility in Austrian Sign Language (ÖGS). *Sign Language & Linguistics*, *9*(1), 207–231.

Schembri, A. (2003). Rethinking 'classifiers' in signed languages. In K. Emmorey (Ed.), *Perspectives on classifier constructions in sign languages* (pp. 13–44). Psychology Press.

Schembri, A., Jones, C., & Burnham, D. (2005). Comparing action gestures and classifier verbs of motion: Evidence from Australian Sign Language, Taiwan Sign Language, and nonsigners' gestures without speech. *Journal of Deaf Studies and Deaf Education*, *10*(3), 272–290.

Schlenker, P. (2012). Complement set anaphora and structural iconicity in ASL. *Snippets*, *25*, 15–17.

Schlenker, P., & Chemla, E. (2018). Gestural agreement. *Natural Language & Linguistic Theory*, 1–39.

Sendra, V. C., Kaland, C., Swerts, M., & Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, *47*(1), 1–13.

Senghas, A. (1995). The development of Nicaraguan Sign Language via the language acquisition process. In *Proceedings of the 19th annual boston university conference on language development* (pp. 543–552).

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development*, *18*(4), 511–531.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological science*, *12*(4), 323–328.

Senghas, A., Newport, E. L., & Supalla, T. (1997). Argument structure in Nicaraguan Sign Language: The emergence of grammatical devices. In E. Hughes & A. Greenhill (Eds.), *Proceedings of the 21st annual Boston University Conference on Language Development.*

Senghas, A., Ozyurek, A., & Goldin-Meadow, S. (2013). Homesign as a way-station between co-speech gesture and sign language: The evolution of segmentation and sequencing. *Oxford studies in the evolution of language: The evolutionary emergence of human language*, 62–76.

Simper-Allen, P. (2016). *"Cut and break"-beskrivningar i svenskt teckenspråk: Barns och vuxnas avbildande verbkonstruktioner* (Unpublished doctoral dissertation). Institutionen för lingvistik, Stockholms universitet.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of artificial intelligence research*, *15*, 31–90.

Siskind, J. M. (2003). Reconstructing force-dynamic models from video sequences. *Artificial Intelligence*, *151*(1-2), 91–154.

So, W. C., Coppola, M., Licciardello, V., & Goldin-Meadow, S. (2005). The seeds of spatial grammar in the manual modality. *Cognitive science*, *29*(6), 1029–1043.

Stokoe, W. C. (1960). Sign language structure. *Studies in linguistics: Occasional papers*, *8*.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *PNAS*, *112*(19), 5968–5973.

Supalla, T. (1983). *Structure and acquisition of verbs of motion and location in American Sign Language* (Unpublished doctoral dissertation). University of California, San Diego.

Supalla, T. (1986). The classifier system in American sign language. *Noun classes and categorization*, 181–214.

Svantesson, J.-O. (2017). Sound symbolism: The role of word sound in meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(5).

Svenonius, P. (2005). Slavic prefixes inside and outside VP. *Nordlyd*, *32*(2).

Tallerman, M. (2010). Holophrastic protolanguage: Planning, processing, storage, and retrieval. In *The emergence of protolanguage: Holophrasis vs compositionality* (pp. 83–99). John Benjamins Publishing Company.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Tang, G., & Yang, G. (2007). Events of motion and causation in Hong Kong Sign Language. *Lingua*, *117*(7), 1216–1257.

Taub, S. F. (2001). *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge: Cambridge University Press.

Thompson, R. (2006). *Eye gaze in American Sign Language: Linguistic functions for verbs and pronoun* (Unpublished doctoral dissertation). UC San Diego.

Thompson, R. (2011). Iconicity in language processing and acquisition: What signed languages reveal. *Language and Linguistics Compass*, *5*(9), 603–616.

Thompson, R., Emmorey, K., & Kluender, R. (2006). The relationship between eye gaze and verb agreement in American Sign Language: An eye-tracking study. *Natural Language & Linguistic Theory*, *24*(2), 571–604.

Thompson, R., Vinson, D. P., & Vigliocco, G. (2009). The link between form and meaning in American Sign Language: Lexical processing effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 550.

Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., & Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods*, *40*(4), 1079–1087.

Wilbur, R. B. (2003). Representations of telicity in ASL. In *Proceedings from the annual meeting of the Chicago Linguistic Society* (Vol. 39, pp. 354–368).

Wilbur, R. B. (2008). Complex predicates involving events, time and aspect: Is this why sign languages look so similar. *Theoretical issues in sign language research*, 217–250.

Wilbur, R. B. (2009). Productive reduplication in a fundamentally monosyllabic language. *Language Sciences*, *31*(2-3), 325–342.

Wilbur, R. B. (2013). The point of agreement: Changing how we think about sign language, gesture, and agreement. *Sign Language & Linguistics*, *16*(2), 221–258.

Wilbur, R. B., & Malaia, E. (2008). Contributions of sign language research to gesture understanding: What can multimodal computational systems learn from sign language research. *International journal of semantic computing*, *2*(01), 5–19.

Wilbur, R. B., & Patschke, C. (1999). Syntactic correlates of brow raise in ASL. *Sign Language & Linguistics*, *2*(1), 3–41.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216.

Zheng, C. (2012). *Path verbs of motion in SwaTawWe serial verb constructions* (Unpublished master's thesis). Purdue University.

Zwitserlood, I. (2003). *Classifying hand configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands)* (Unpublished doctoral dissertation). Universiteit Utrecht.

Zwitserlood, I. (2008). Morphology below the level of the sign-frozen forms and classifier predicates. In J. Quer (Ed.), *Signs of the time: Selected papers from tislr 8 (2004)* (pp. 251–272). Hamburg: Signum Press.

Zwitserlood, I., & Gijn, I. V. (2006). Agreement phenomena in Sign Language of the Netherlands. *Arguments and Agreement*, 195–229.

APPENDIX

A. DETAILS, STUDY 1

A.1 Study 1a, action list

Item	Transitivity	Item	Transitivity
ball bounce	IN (A)	adjust picture	TR
ball drop	IN (A)	approach coat hanger	TR
ball fly into box	IN (A)	ball hit water bottle	TR
ball roll	IN (A)	bounce ball	TR
balloon deflate	IN (A)	break stick	TR
book fall over	IN (A)	close microwave door	TR
bottle caps spill	IN (A)	coat rack approach	TR
bowl break	IN (A)	crush can	TR
box move	IN (A)	cut bread	TR
bread spin	IN (A)	cut paper	TR
car pass tower	IN (A)	dip finger in jar	TR
cards scatter	IN (A)	draw on whiteboard	TR
door microwave close	IN (A)	drop ball no bounce	TR
fan turn	IN (A)	hammer nail	TR
fizzle down	IN (A)	hit bottle with ball	TR
hanger swing	IN (A)	knock over tower	TR
lid blow off	IN (A)	lift teapot pinky	TR
light turn on	IN (A)	light candle	TR
march	IN (E)	measure book	TR
measuring tape retract	IN (A)	move box	TR
newspaper tear	IN (A)	plug in	TR
overflow	IN (A)	pop balloon	TR

paper airplane land	IN (A)	pour bottle caps	TR
paper drop	IN (A)	pull out measuring tape	TR
pen move	IN (A)	push button microwave	TR
person bend over	IN (E)	push toy in box	TR
picture shift	IN (A)	put book on side	TR
poster roll up	IN (A)	put cup on coaster	TR
race car into box	IN (A)	remove cork	TR
shaving cream spray	IN (A)	reorient pole	TR
stick break	IN (A)	roll out	TR
string relax	IN (A)	shake shaker	TR
string taughten	IN (A)	spin bread	TR
tape measurer bend	IN (A)	spray shaving cream	TR
tower fall	IN (A)	swat balloon	TR
toy appear in box	IN (A)	take keys out of box	TR
toy crawl	IN (A)	take lid off	TR
toy skitter	IN (A)	taughten string	TR
walk backwards	IN (E)	tear paper	TR
whirly gig drop	IN (A)	turn fan	TR

A.2 Study 1b, full results

Best pantomimer	Item	t(df)	p
IP	The man adjusted a picture.	t(16) = -1.725	0.0525
HO	The ball hit the bottle.	t(29) = -2.9599	0.0031
IP	The man bounced a ball.	t(17) = -4.6048	0.0001
NP	The man broke the stick.	t(26) = -4.3239	0.0001
HO	The man closed the microwave door.	t(23) = -4.7718	0
NP	The man crushed the can.	t(28) = -2.3032	0.0146
NP	The man cut the bread in half.	t(21) = -3.8537	0.0005
NP	The man dipped his finger into the jar.	t(51) = -7.9089	0
NP	The man dropped the ball.	t(25) = -3.4647	0.001
RVN	The man hammered the nail.	t(56) = -7.085	0
NP	The man turned the fan.	t(27) = -6.1325	0
NP	The man swatted the balloon.	t(25) = -3.6498	0.0006
HO	The man hit the bottle with a ball.	t(28) = -5.4306	0
IP	The man knocked over the block tower.	t(25) = -3.1455	0.0022
NP	The man lifted a kettle with his finger.	t(26) = -2.6551	0.0068
RVN	The man lit a candle.	t(50) = -3.9822	0.0001
CM	The man measured the book	t(28) = -3.4104	0.001
CS	The man moved the box across the table	t(19) = -3.5477	0.0011
HO	The man plugged in the charger.	t(27) = -6.6938	0
CM	The man poured buttons out of the jar.	t(26) = -1.3943	0.0877
RVN	The man popped the balloon.	t(21) = -3.3469	0.0016
NP	The man uncorked the wine bottle.	t(24) = -6.6842	0
NP	The man pulled out the measuring tape.	t(29) = -6.0937	0
HO	The man pushed a button ...	t(25) = -5.1657	0
IP	The man shoved a toy into a box.	t(44) = -2.7177	0.0047

Best pantomimer	Item	t(df)	p
NP	The man set the book on its side.	t(24) = -2.4224	0.0119
NP	The man put the cup on a coaster.	t(25) = -5.0883	0
NP	The man set a tube on its side.	t(21) = -3.5161	0.0011
NP	The man spun the bread.	t(27) = -4.8364	0
NP	The man rolled the tube back and forth.	t(27) = -5.5794	0
NP	The man shook the shaker.	t(20) = -5.2796	0
NP	The man sprayed shaving cream.	t(25) = -2.5753	0.0083
IP	The man took the lid off of a jar.	t(29) = -9.3467	0
NP	The man took the keys out of the box.	t(26) = -10.1114	0
IP	The man tightened the string.	t(25) = -7.8335	0
CM	The man tore the paper in half.	t(24) = -2.832	0.0047
NP	The coat rack moved toward the man.	t(28) = -5.3245	0
NP	The ball bounced.	t(26) = -6.6966	0
NP	The ball dropped.	t(19) = -4.3519	0.0002
HO	The ball rolled.	t(27) = -3.1215	0.0022
CS	The balloon deflated.	t(25) = -3.4686	0.001
HO	The book fell over.	t(29) = -4.152	0.0001
HO	The jar of bottle caps spilled over.	t(24) = -8.9842	0
IP	The bowl broke.	t(25) = -5.5807	0
HO	The box moved across the table.	t(24) = -11.9345	0
IP	The bread spun.	t(54) = -7.6679	0
IP	The paper whirled down onto the table.	t(29) = -3.2669	0.0014
IP	The race car drove past the tower.	t(24) = -5.717	0
NP	The cards scattered everywhere.	t(50) = -7.2427	0
HO	The fan oscillated.	t(26) = -7.4386	0
NP	The drink bubbled down.	t(25) = -4.0332	0.0002
NP	The hanger swung back and forth ...	t(28) = -10.4003	0

Best pantomimer	Item	t(df)	p
IP	The lid blew off the jar.	t(24) = -1.9458	0.032
HO	The light turned on.	t(25) = -4.3339	0.0001
HO	The box slid across the table.	t(23) = -9.625	0
CM	The measuring tape bent in half.	t(26) = -3.5235	0.0008
IP	The man marched.	t(26) = -2.8416	0.0044
IP	The microwave door closed.	t(23) = -4.6126	0.0001
HO	The paper airplane landed on the table.	t(54) = -14.659	0
HO	The paper floated down onto the table.	t(21) = -3.5813	0.0009
HO	The pen moved along the table.	t(27) = -1.4811	0.0753
IP	The man bowed at the waist.	t(24) = -23.7382	0
NP	The picture shifted.	t(29) = -3.9726	0.0002
HO	The poster rolled up.	t(25) = -7.5056	0
RVN	The toy skittered on the table.	t(26) = -4.5455	0.0001
CM	The toy race car drove into the box.	t(29) = -4.0289	0.0002
HO	Shaving cream sprayed onto the table.	t(26) = -1.9217	0.0331
RVN	The stick broke.	t(26) = -3.8695	0.0003
HO	The block tower fell.	t(23) = -3.3483	0.0015
NP	The toy came up out of a box.	t(26) = -3.9108	0.0003
HO	The toy crawled up the incline.	t(21) = -8.7881	0
CM	The man walked backwards.	t(25) = -7.1641	0

B. DISCUSSION OF FULL PARTICIPANT GROUP (CC AND PANTOMIME EXPERIMENTS)

In the main text, the analyses in Chapter 3 were run only on the data from the identified Target Group, or, those participants who performed on Study 2a (§3.3). Here below we report abbreviated results from the same analyses, but performed on data from *all* participants (Whole Group). That is, this pool also contains data from the Target Group. We do not perform any analyses on just the Errant Group (or the group that did not perform Study 2a as intended), as–again–all 24 of these participants chose just a single response (23 all 4's, one all 1's). The results of any analysis on *just* the Errant Group we might thus be able to presage without going through all the extra trouble. We present the top-down analysis of the Whole Group next (B.3), followed by the bottom-up analysis in B.4.

B.1 (Whole Group) Study 2a results:

However, there were considerable changes in the proportion of transitive, ditransitive, and intransitive labels in the new dataset. Among classifier constructions, the number of transitive labels increased from 18 to 25, and intransitive unergative labels decreased from 19 to 13. Among pantomimes, the number of transitive labels increased from 17 to 27, and intransitive unergative labels decreased from 17 to 10.

Of the 144 pantomimes and classifier constructions, 84 were classifiable as transitive, ditransitive, intransitive unergative or intransitive unaccusative according to our criteria. Of the 85, 42 were classifier constructions and 43 were pantomimes. As such, significantly more items were classified than chance (chance = 0.25; $\mu = 0.59$, SD = 0.5, t(142) = 8.1672, one-tailed p $\leq$ 0.0001). Individually, significantly more classifier constructions ($\mu = 0.5833$, SD = 0.5, t(71) = 5.6971, p $\leq$ 0.0001) and pan-

Table B.1.
(Whole Group) Tallies of consistently classed classifier constructions, pantomimes and lexical verbs, where consistency is defined as maximum votes that were chosen significantly above chance (at $\alpha = 0.05$). Both classes of stimuli had well over chance (=25%) rates of consistent responses, indicating that participants had some model of transitivity.

	Classifier Constructions	Pantomimes
Transitive	18	17
Ditransitive	1	2
Intransitive (E)	19	17
Intransitive (A)	4	7
Total	**42**/72	**43**/71
% dataset	58.33%	60.56%

Table B.2.
Post-hoc cf. Whole Group: Same as Tab. B.1, but only includes data from the target group. Green up-arrows represent an increase in the tally of a particular class, compared to the aggregate data in Tab. B.1; red down-arrows the reverse. Note also that the proportion of consistently classed classifier constructions increased

	Classifier Constructions		Pantomimes	
Transitive	25	↑	27	↑
Ditransitive	1		4	↑
Intransitive (E)	13	↓	10	↓
Intransitive (A)	5	↑	2	↓
Total	**44**/72		**43**/71	
% dataset	61.11% ↑		60.56%	

tomimes ($\mu = 0.5915$, SD = 0.5, $t(70) = 5.8135$, $p \leq 0.0001$) were classed than we would expect by chance. This suggests that classifier constructions and pantomimes are iconic w.r.t. their transitivity. We explore how this may be the case in subsequent sections.

The breakdown of classifier constructions and pantomimes into transitive, ditransitive, etc. classes is presented in Tab. B.1. As most of the input videos (i.e., the action videos) were either transitive or intransitive unergative, with few intransitive unaccusative and zero ditransitive videos, it is unsurprising that the classifier constructions and pantomimes depicting these actions were generally classed as transitive and intransitive unergative.

B.2 (Whole Group) Accuracy:

To assess accuracy, we decided to bin responses into 'transitive' and 'intransitive' categories, where *transitive* meant an action involving one or more objects (transitive and ditransitive stimuli) and *intransitive* meant an action that does not involve an object (intransitive unergative and intransitive unaccusative stimuli). We did this due to the low incidence of unergative stimuli in the action video dataset (3) and the zero incidence of ditransitive stimuli. Further, given that participants consistently chose unergative labels for unaccusative stimuli, perhaps due to an agency bias, accuracy might be artificially low.

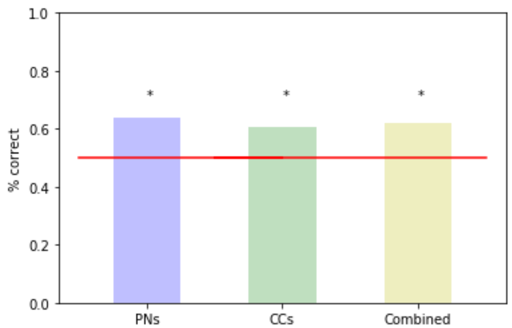
Accuracy was measured in two ways. In the first, individual responses for an item were classed as 'hit' (1) or 'miss' (0), then the responses were averaged to get a percent correct figure. In the second method, we assessed accuracy by consensus: items were categorized as 'transitive' or 'intransitive' (a) by simple tally (i.e., an item gets more 'transitive' than 'intransitive' labels or *vice versa*), and (b) by our measure of consistency (i.e., an item gets significantly more 'transitive' than 'intransitive' labels or *vice versa*).

Individual level: By the first method, mean accuracy across the entire dataset was 62.09%, which is significantly greater than chance (SD = 0.17301; 1-sample t-test of proportion against hypothetical mean, 0.50: t(142) = 8.3561, 1-tailed $p < 0.0001$). For pantomimes, mean accuracy was 63.62% across all items, which was significantly greater than chance (t(70) = 7.0082, $p < 0.0001$). Finally, for classifier constructions, mean accuracy was 60.59%, which was again significantly greater than chance (t(71) = 4.9456, $p < 0.0001$). These results are summarized in Fig. B.1.

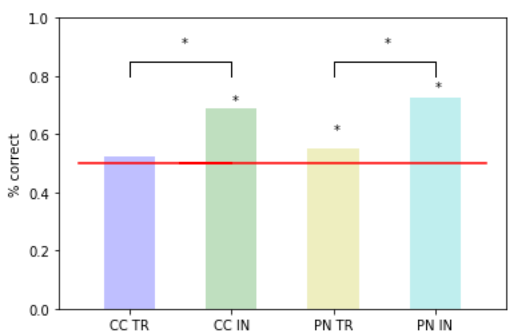
Further, across both pantomimes and classifier constructions, intransitive items were more accurately identified than transitive ones. See Fig. B.2. For pantomimes, mean accuracy on intransitive items was 72.55% (SD = 0.1664; t(34) = 8.0180, $p < 0.0001$). While mean accuracy on transitive items was only 54.92%, this is still significantly above chance (SD = 0.1047; t(35) = 2.8249, p = 0.0039). The difference between accuracies is significant (t(69) = 5.3573, $p < 0.0001$). As for classifier constructions, accuracy on intransitive items was significantly greater than chance at 69.03% (SD = 0.1697, t(35) = 6.7267, $p < 0.0001$) while accuracy for transitive items was not (52.15%, SD = 0.1532, t(35) = 0.8409, p = 0.2031). There was also a significant difference between these two measures (t(70) = 4.4293, $p < 0.0001$).

Consensus level: For the consensus analysis, we again had two inclusion criteria, one inclusive (no statistical thresholding) and one exclusive (thresholded statistically). For the first analysis, we did however need to weed out items that did not have a clear winner, namely, those items that garnered an equal number of transitive and intransitive votes. As such, eight pantomimes and seven classifier constructions were excluded from the analysis, leaving just 63 pantomimes and 66 classifier constructions.

Mean accuracy across this dataset was significantly greater than chance at 72.09% (t(128) =5.5726, $p < 0.0001$). Mean pantomime accuracy across all 63 pantomimes was significantly above chance at 76.19% (t(62) = 4.8419, $p < 0.0001$) For all 66 classifier constructions, mean accuracy was also significantly above chance at 68.18% (t(65) = 3.1472, p = 0.0025). Participants were not more likely to accurately class pantomimes over classifier constructions (t(127) = 1.0098, p = 0.3145).



Comparison	t(df)	p
PN vs. chance	t(70) = 7.0082	< 0.0001
CC vs. chance	t(71) = 4.9456	< 0.0001
Combined vs. chance	t(141) = 8.3561	< 0.0001



Comparison	t(df)	p
CC TR vs. chance	t(35) = 0.8409	0.2031
CC IN vs. chance	t(35) = 6.7267	< 0.0001
PN TR vs. chance	t(35) = 2.8249	0.0039
PN IN vs. chance	t(34) = 8.0180	< 0.0001
CC TR vs. CC IN	t(70) = 4.4293	< 0.0001
PN TR vs. PN IN	t(69) = 5.3573	< 0.0001

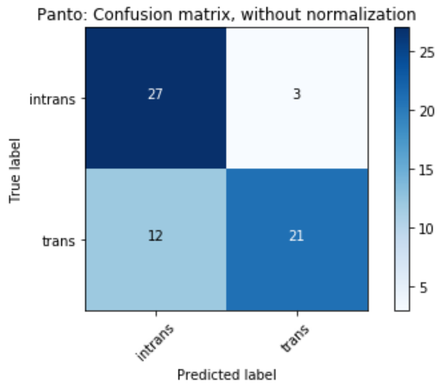
Figure B.1. Accuracy of pantomimes and classifier constructions, Whole group, individual analysis. (a) Average accuracy on classifier constructions juxtaposed with average accuracy on pantomimes. Both are significantly greater than chance, but neither is significantly greater than the other. (b) Same as (a) but with accuracy on intransitive and transitive items displayed separately, revealing that accuracy on intransitive items was significantly greater than on transitive items. 'CC' = *classifier construction*; 'PN' = *pantomime*; 'TR' = *transitive*; 'IN' = *intransitive*.

We next measured the accuracy of just those items that were consistently classed. Again, these numbered 85, with 43 pantomimes and 42 classifier constructions. Including both pantomimes and classifier constructions, non-signers achieved 75.29% accuracy, which is significantly above chance ($t(84) = 5.375$, $p < 0.0001$). For just consistently classed pantomimes, mean accuracy rose slightly to 76.74% from 76.19% and was significantly greater than chance ($t(42) = 4.1027$, one-tailed $p = 0.0002$). Finally, as for consistently classed classifier constructions, mean accuracy rose to 73.81% from 68.18%, and was again significantly greater than chance at ($t(41) = 3.4674$, one-tailed $p = 0.0012$). Participants were not significantly more accurate in classing pantomimes than classifier constructions among consistently classed items ($t(84) = 0.3101$, $p = 0.7573$).

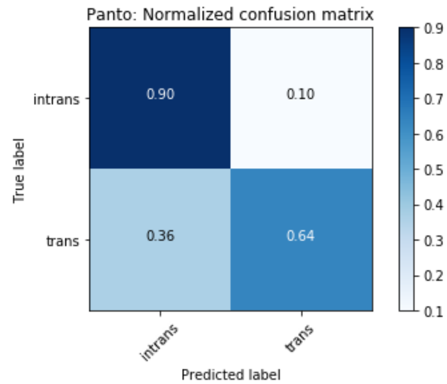
We noticed from the individual-level analysis that participants were significantly more accurate in classing intransitive actions than in classing transitive ones. To gain a better understanding of how participants classed classifier constructions and pantomimes, for this analysis we not only looked at accuracy, but also where and how participants guessed incorrectly. This more detailed information is presented in Figs. B.2 & B.3.[1] Taken together, the plots illustrate a bias towards intransitive guesses: if a participant made a mistake, it was likely in classing a transitive item as intransitive, rather than the other way around.

Although accuracy among consistently classed items appears to be much higher than accuracy within the dataset generally, there are two issues that muddle the interpretation of this measure: (1) The number of transitive and intransitive pantomimes and classifier constructions were unequal, and (2) non-signers were biased towards an 'intransitive' response. Specifically, there were 48 true transitive items (of which 25

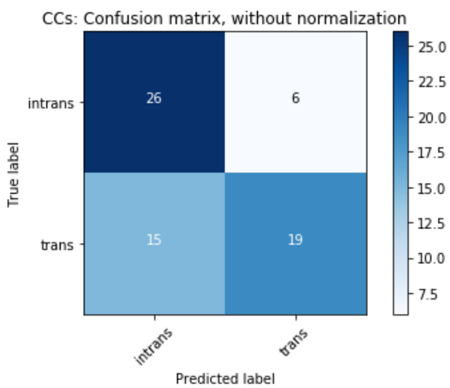
[1]A note on reading confusion matrices: Rows represent ground truth labels. The sum of each row is the total number of ground truth labels for a particular class. Columns represent non-signer derived labels, with column totals equalling the total number of non-signer labels for a particular class. The principal diagonal (top-left to bottom-right) represents correct predictions, or when non-signers accurately classed an item as, e.g., intransitive when it was actually intransitive. The minor diagonal (top-right to bottom left) represents error, or when, say, a participant labeled an intransitive stimulus as transitive. Cells are colored according to the frequency of cases relative to the total number of cases. The higher frequency a case, the darker its cell in the matrix. As such, the desired outcome is a dark colored band on the principal diagonal and a light-colored band on the minor diagonal.



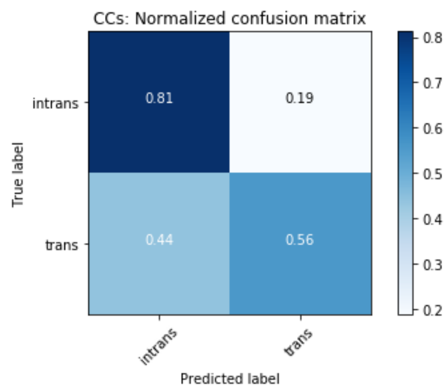
(a)



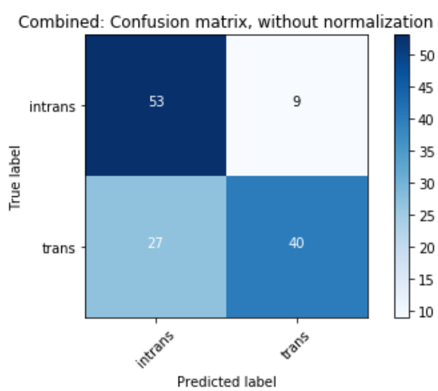
(b)



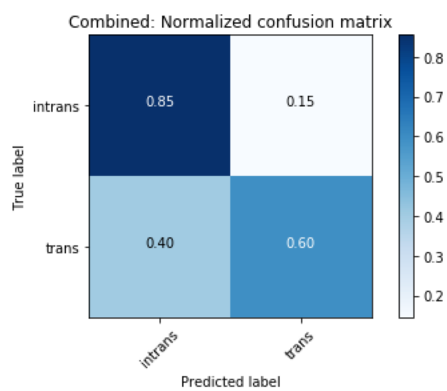
(c)



(d)

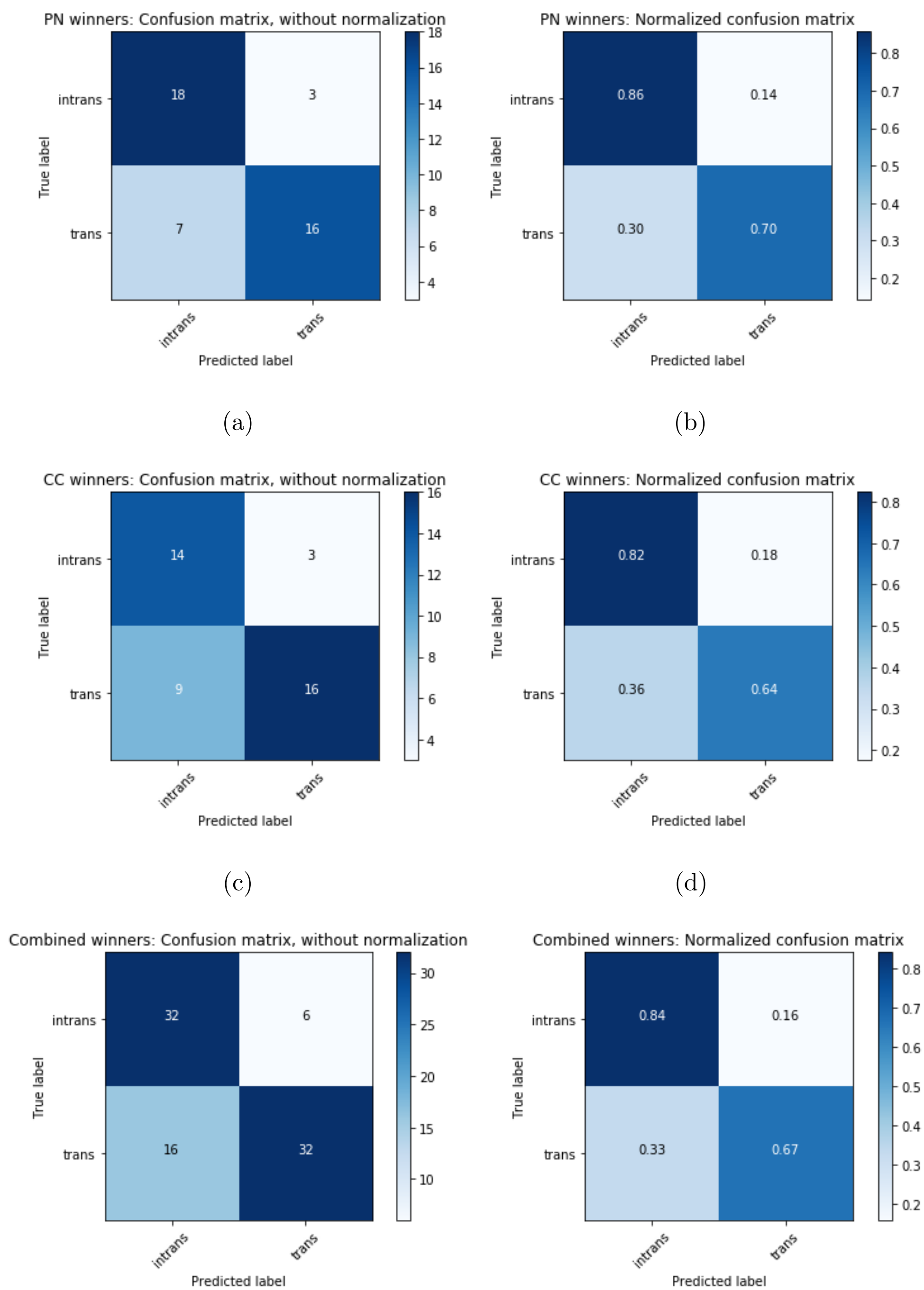


(e)



(f)

Figure B.2. Confusion matrices showing non-signers' pantomime (a,b) and classifier construction (c,d) classing accuracy. Pooled results are in (e,f). The plots on the left show raw counts, whereas the plots on the right are normalized.



(a) (b) (c) (d) (e) (f)

Figure B.3. Same as Fig. B.2, but only including items that were consistently classed.

Table B.3.
Raw label counts. 'GT' = *ground truth*; 'NS' = *non-signer label*; 'CC' = *classifier construction*; 'PN' = *pantomime.*

	Intransitive		Transitive	
	GT	NS	GT	NS
CCs	18	23 ↑	24	19 ↓
PNs	20	24 ↑	23	19 ↓
Total	38	47 ↑	47	38 ↓

were classifier constructions and 23 pantomimes) and 38 true intransitive items (of which 17 were classifier constructions and 21 were pantomimes). Of non-signer labels, there were 38 transitive labels (of which 19 were classifier constructions and 19 were pantomimes) and 48 intransitive labels (of which 23 were classifier constructions and 25 were pantomimes). This information is summarized in Tab. B.3.

From Tab. B.3 we can see that ultimately more underlyingly transitive stimuli were classified than intransitive stimuli (47 transitive vs. 38 intransitive). Despite this, participants provided more intransitive labels than transitive labels (47 intransitive vs. 38 transitive). This likely represents the largest source of misclassification.

As such, we ran another measure to gauge non-signer performance: Matthew's correlation coefficient (MCC). The statistic takes performance on both correct an incorrect identifications and rejections into consideration (accuracy only looks at correct identifications and correct rejections), and returns a value between -1 and 1.

Similarly with other measures of correlation, -1 is a perfect dissociation, 1 is a perfect association, and 0 is chance association.

The MCC for pantomimes and classifier constructions together is 0.5088.The MCC for just classifier constructions is 0.4571, and for pantomimes is 0.5574. We interpret these as strong scores, suggesting that participants were generally accurate despite their intransitive bias.

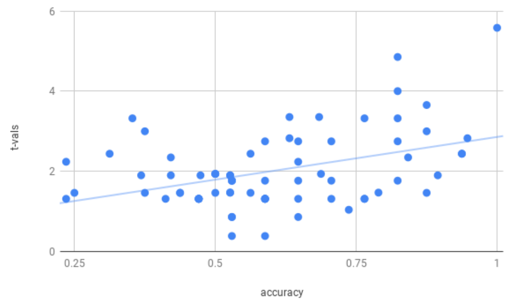
Results here show that at an individual level, participants were fairly successful at accurately classing transitive and intransitive pantomimes and classifier constructions. However, participants were more successful *en masse* when class labels were decided by the majority (simple tally) and even more so when decided by significantly many votes. This gradual increase in accuracy from considering participants as individuals, to weak agreement, to consistency indicates a relationship between consistency and accuracy, which we explore more fully in §B.3.

B.3 (Whole Group) Consistency, explained by iconicity & accuracy:

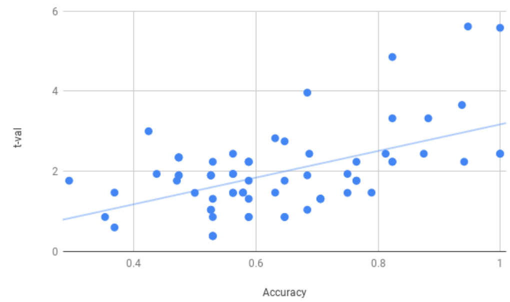
Again, if the top-down approach is on the right track, we might expect to see consistency (i.e., agreement on a label) and/ or accuracy increase with lexical iconicity score. A correlation between consistency and iconicity scores would indicate that agreement on a label is modulated by some aspect of the perceived event. A correlation between accuracy and iconicity scores would indicate further that this aspect of the perceived event is or is closely related to transitivity.

Again, *consistency* corresponds to the magnitude of the t-value associated with an item. This t-value was derived, again, by comparing the frequency of the most selected option against chance. The t-values we used were for all four labels (*transitive*, *ditransitive*, *intransitive unergative*, etc.).

W.r.t. consistency and accuracy, we observed a moderate correlation among all items ($r = 0.4614$, $p < 0.0001$), just classifier constructions ($r = 0.4101$, $p = 0.0003$), and just pantomimes ($r = 0.5278$, $p < 0.0001$). When separating out transitives and intransitives, we see that the correlation between accuracy and consistency holds: Among transitive pantomimes, for instance, $r = 0.58653$ and, among intransitive pantomimes $r = 0.6192$. We leave the rest unreported. Accuracy is plotted against consistency in Fig. B.4 to illustrate the effect. This is wholly expected given that consistency and accuracy are related to one another, as explained in the main text.



(a) Classifier Constructions

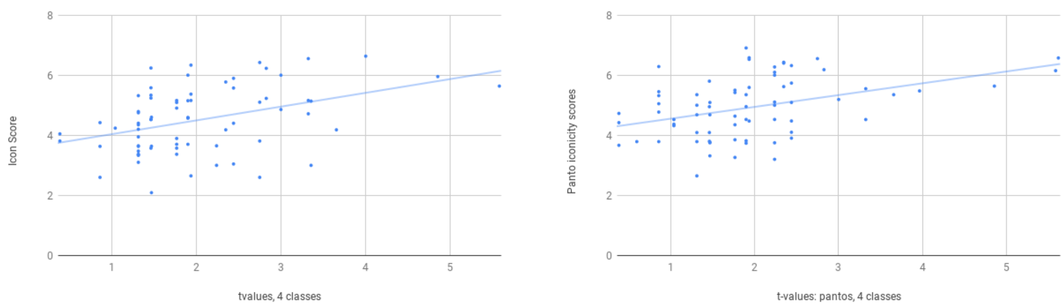


(b) Pantomimes

Figure B.4. **Consistency x Accuracy:** Scatter plots with least squares regression line illustrating the relationship between t-values generated from Study 2a and accuracy scores, where t-values are interpreted as strength of participant agreement for a given label for a given item. T-values are from analysis of all four labels.

As for consistency and iconicity scores, results of a correlation analysis again indicate a modest correlation (Pantomimes: r = 0.4073, p = 0.0004; Classifier constructions: r = 0.4015), p = 0.0005; combined: r = 0.3887, p < 0.0001). Similarly, the results hold by transitivity class (e.g., transitive pantomimes: r = 0.4153; intransitive pantomimes: r = 0.4877; Further correlations are significant, though unreported). Iconicity is plotted against consistency in Fig. B.5 to illustrate the effect. To note, there was no correlation between accuracy and iconicity score (Combined: r = 0.0239, p = 0.7767; CCs: r = 0.0024, p = 0.9838; PNs: r = 0.0110, p = 0.9274). We assume that this is the direct effect of our Errant Group choosing all 4's, irrespective of stimulus.

Two further points are worth noting: On average, participants (Whole Group) were more likely to accurately guess the transitivity of pantomimes over classifier constructions. This is true for both intransitive and transitive items, further suggesting a link between lexical iconicity and transitivity classing. However, accuracy among intransitives was higher than transitives, despite the latter, counter-intuitively, being more iconic than the former (recall the results of Study 2b, §3.4).



(a) Classifier Constructions (b) Pantomimes

Figure B.5. **Consistency x Iconicity:** Scatter plots with least squares regression line illustrating the relationship between t-values generated from Study 2a and mean iconicity scores from Study 2b, where t-values are interpreted as strength of participant agreement for a given label for a given item. T-values are from analysis of all four labels

This lead us to believe that the observed effect has its roots in how participants classed the stimuli. As noted above, participants were biased towards providing intransitive labels. We suggested, thus, that participants had an *intransitive until proven otherwise* classing strategy. This is underscored by the fact that many inherently transitive stimuli were classed as intransitive.

Evidence for this may also come from how we calculated accuracy. There was no penalty for misses, s.t. the following scenario could be true: if participants provided, e.g., intransitive labels to all stimuli, they would be 100% accurate for intransitive stimuli but 0% accurate for transitive stimuli. As we reported in §B.2, the MCC for pantomimes, classifier constructions, and their combination suggests that participants' classing ability is not as robust as the reported accuracy scores suggest.

Taking stock, we are still left with a small puzzle. Non-signers agreed *more* on inherently transitive items over intransitive items. However, they were more likely to agree that an item was intransitive than transitive. Yet, they also rated transitive items as being more iconic than intransitive ones.

The correlation analysis described above lead us to discover the existence of the Errant Group, as again it was unclear how non-signers were (a) generally inaccurate, ascribing intransitive labels to most every item, despite (b) rating transitive items as being more iconic than intransitive items w.r.t. their lexical meanings. After the separation of the Errant Group, the analyses on the Target Group came in line as predicted.

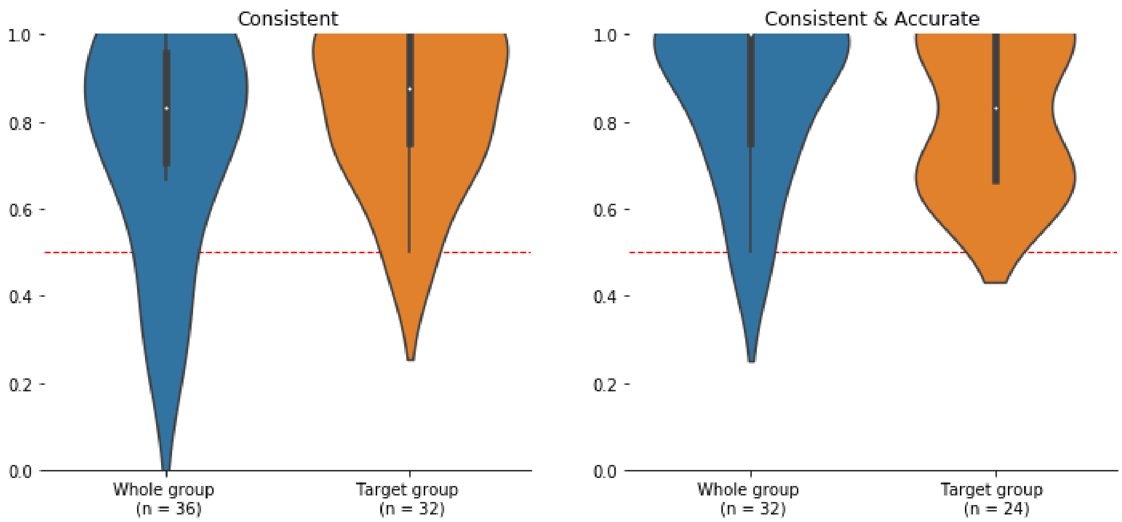
B.4 (Whole Group) Consistency, explained phonetic features

Here we compare classifier performance using two sets of non-signer-derived labels. One set is derived from all participants of Study 2a (§3.3), and the other is derived from only those participants who completed the task as intended (i.e., the group identified in §3.3.2).

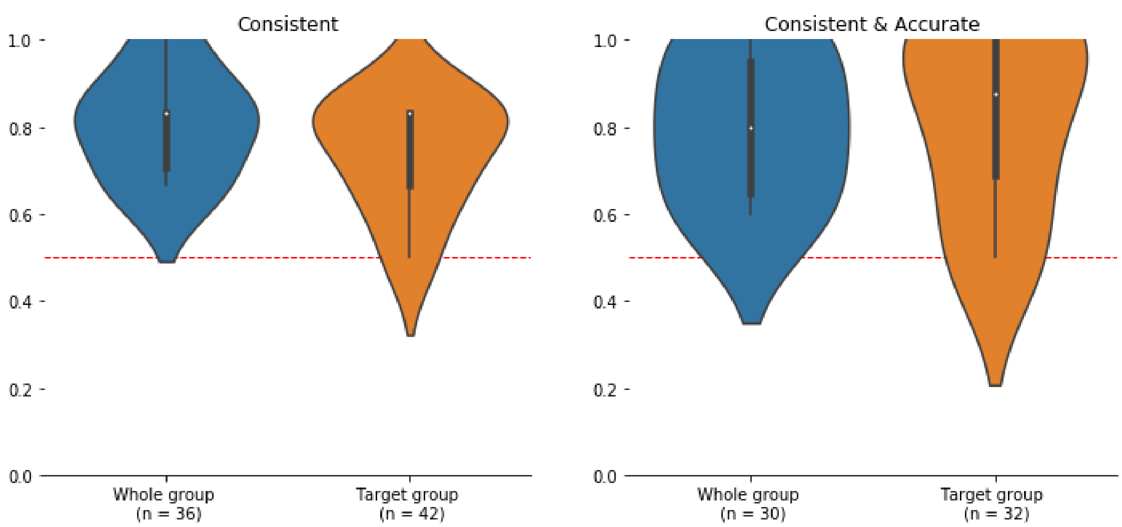
Ultimately, the difference between groups is the number of samples used in each analysis, with the whole group contributing more samples than the target group (e.g., as seen in Fig. B.6(a), the Whole Group contributed six more samples than the Target Group in the analysis using consistently classed samples).

It is also the case that the Whole Group returned more intransitive labels than did the Target Group. As such, we may have expected to see classifier performance dip or become biased, as we saw in the top-down analysis. And while the capping procedure may have nullified the effects of this bias, in that an equal number of transitive and intransitive samples were always used in the analyses, features that should have correlated with the transitive class would still have been in the intransitive class. However, the effect seems surprisingly small, especially when compared to the dramatic effects seen in the top-down analysis of the Whole Group (above; B.3).

We do see differences in classifier performance between Groups when conducting our ancillary analyses. For example, a Whole Group analysis only using handshape complexity measures achieved 68.75% accuracy (marginally significant at $p = 0.0501$), while the comparable Target Group analysis yielded chance performance (47.5%, $p = 0.8746$; difference between means 21.25%, $t(7) = -2.6011$, $p = 0.0209$).



(a) Pantomimes



(b) Classifier Constructions

Figure B.6. Violin plots showing classifier performance on labels derived from the full group (all Study 2a participants) and target group (only those participants who completed the task as intended). All accuracies are significantly above chance. Analyses using labels from the full group were in no case significantly more accurate than analyses using target group labels, though in most cases accuracy was descriptively higher.

C. FULL RESULTS: ML ANALYSIS OF CLASSIFIER CONSTRUCTIONS AND PANTOMIMES

In what follows, we provide extra information concerning our machine learning analyses of pantomimes and classifier constructions. For each analysis, we provide three violin plots, showing performance of classifiers before- and after feature extraction, and performance of classifiers trained on random labels and tested on unshuffled labels. Label shuffling occurred for each fold. No feature selection was performed in the random-label analyses. In some cases, significant results were obtained from random label analyses. An additional test was performed in these cases, and is described in detail below. Finally, we further provide two confusion matrices per analysis, one normalized and one with raw counts. In all cases, confusion matrices show performance of analyses *using* feature extraction. Past the first analysis, we do not provide additional text interpretation where graphs and tables tell a clear enough story. We provide commentary only in special situations.

C.1 Production

C.1.1 Pantomimes: Ground Truth

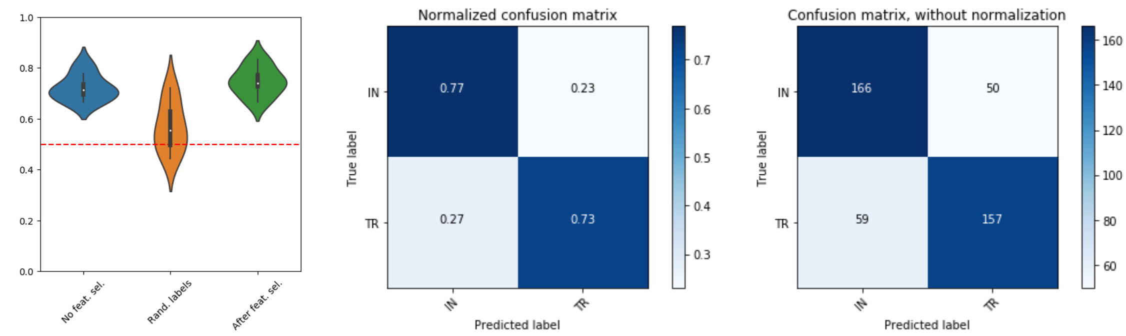
Results are presented in Fig. C.1. Performance was equally good (and equally significant) for analyses with and without feature selection (difference in means: 0.0254; t(7) = 0.9397, 2-tailed p = 0.3632). The number of informative features identified in the analysis using feature extraction was $k = 47$ (out of 68 total). The analysis using random labels was significantly greater than chance at 56.94% (1-tailed p = 0.0458), but not meaningfully so. However, we iterated the analysis 10 times and found that the analysis using random labels approaches chance levels (accuracy =

Table C.1.
Common cross-fold extracted features. Features were common across all eight folds except where indicated: * = common to 6/8 folds. Rarer features are unreported.

Features (Pantos, GT labels)
mediumjoint, closed_{nsf}, bent, independent, local mvmt (fine), pinky, static, together, mirror, curved, throughout, acceleration, index, thumb, trajectory, complexjoint, closed, final, flex, awayfrom, towards, complexfinger, wiggle, base, tense, deceleration, hands, unopposed, multi, stacked, mono, orientationchange, pivot, medial, opposing, narrow, nonbase, flex_{nsf}, open*, same*, aperturechange*, on*, faceecho*, crossed*, $\text{loop}^{\dagger}$,

49.13%; one-sample t against hypothetical mean, 50%, $t(9) = -0.5379$, 2-tailed $p = 0.6053$). On the other hand, the analysis using feature extraction stayed around 75% (accuracy = 0.7451; one-sample t against hypothetical mean, 50%, $t(9) = 187.80$, 2-tailed $p < 0.0001$).

We present the features that were most informative for the analysis in Tab. C.1. To note, in the main text, we report features that appeared 75% of the time after 10 iterations of the analysis. Here, we simply report features that were common to each fold of one iteration.



(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.7222	0.5694	0.7476
Std	0.0481	0.0915	0.0531
p	< 0.0001	0.0045	<0.0001
MCC	-	-	0.4958

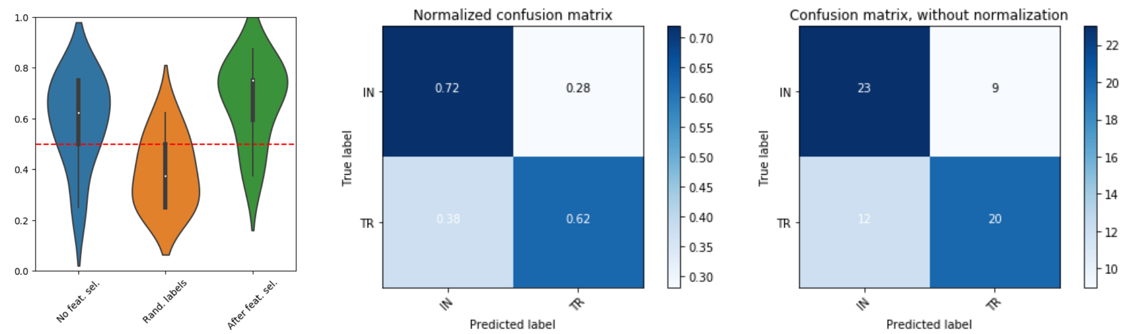
Figure C.1. **Pantomimes, ground-truth labels:** Plots illustrating performance of classifiers on ground truth labels for all 432 pantomime productions. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers were just as accurate in identifying transitive items as they were intransitive items. Summary statistics are presented in the table.

C.1.2 Pantomimes: Ground Truth, best productions

For this analysis, we only included data from the 72 pantomimes that were chosen in Study 1b (§3.2.2) as the best exemplars of the event they represent. Performance is descriptively worse here than in the analysis using all 432 productions. We attribute this to (a) the size of the dataset (there are more samples in the analysis on all 432 samples) and (b) the possibility that event-related or event-specific information is boosting classifier performance in the analysis of all 432 samples. (See Appendix D.2 for detailed argumentation). As such, only the analysis with feature extraction meets significance at 67.18% (p = 0.0081).

Table C.2.
Common cross-fold extracted features. Features were common across all eight folds except where indicated: * = common to 6/8 folds. Rarer features are unreported.

Features (Pantos, GT labels, best productions)
flex, stacked, mono, local mvmt (fine), tense, deceleration, closed, multi, final, bent*



(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.5938	0.3906	0.6718
Std	0.1624	0.1317	0.1523
p	0.1686	0.1034	0.0081
MCC	-	-	0.3452

Figure C.2. **Pantomimes, ground-truth labels from best 72 productions:** Plots illustrating performance of classifiers on ground truth labels for only the 72 best pantomime productions. (That is, for each concept, one of the six total productions was chosen). (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during training (testing used unshuffled labels). Dashed red line represents chance (50%, either transitive or intransitive). (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers showed a bias towards ‘intransitive’ labels.

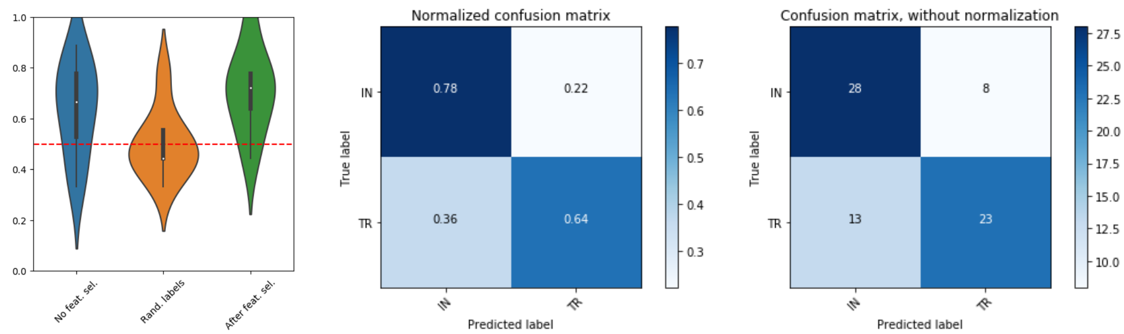
C.1.3 Classifier Constructions: Ground Truth

Results are presented in Fig. C.3. Analyses with and without feature extraction are significantly accurate at 70.83% ($p = 0.0005$) and 63.89% ($p = 0.0245$), respectively. Descriptively, both these figures are better than those obtained using data from just the 72 best pantomime productions, though not statistically so (2 sample $t(7) = 0.4417$, 2-tailed $p=0.6655$). Recall that, in the latter analysis, accuracy with and without feature extraction was 67.18% ($p = 0.0081$) and 59.38% (n.s.), respectively. We take this to indicate that when the number of samples of pantomimes and classifier constructions are equal and as both datasets become larger, transitivity information is actually more transparent in classifier constructions than in pantomimes, contrary to our assumptions that the linguistic system puts pressure on the former to be more arbitrary.

The features that were identified as most informative across folds are presented in Tab. C.3. The two event-related features, *mono(eventive)* and *multi(eventive)*, were common to all eight folds in this analysis, as well as *tense* (presence of tension in the production), *trajectory* (eyes look at path dominant hand takes, or the 2nd hand), *narrow* (related to flexion; *aka curved-open*, a grasping handshape). *Deceleration*, related to end marking, and *curved*, a movement type, were also informative in most folds.

Table C.3.
Common cross-fold extracted features. Features were common across all eight folds except where indicated: * = common to 7/8 folds, † = common to 6/8 folds. Features that were more rare are unreported.

Features (CCs, GT labels)
multi, mono, tense, trajectory, narrow, deceleration†, curved‡



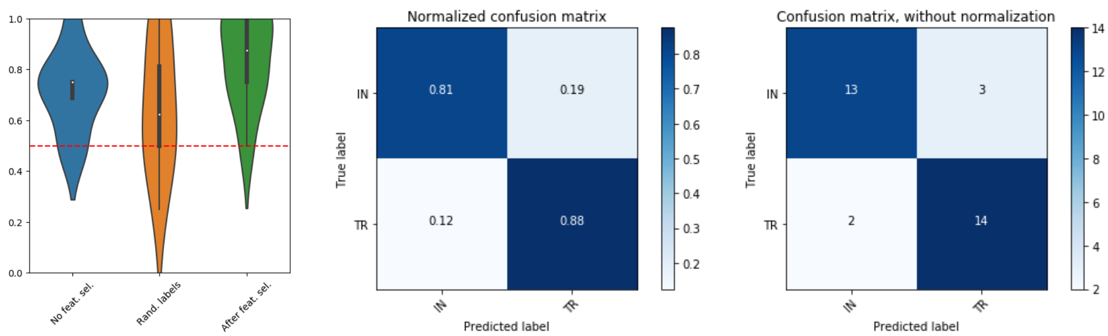
(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.6389	0.5	0.7083
Std	0.1734	0.1242	0.1565
p	0.0245	1.0	0.0005
MCC	-	-	0.4207

Figure C.3. **Classifier constructions, ground-truth labels:** Plots illustrating performance of classifiers on ground truth labels for classifier construction productions. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed line represents chance (50%, either transitive or intransitive). (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers demonstrated an intransitive bias.

C.2 Perception

C.2.1 Pantomimes: Winning non-signer labels



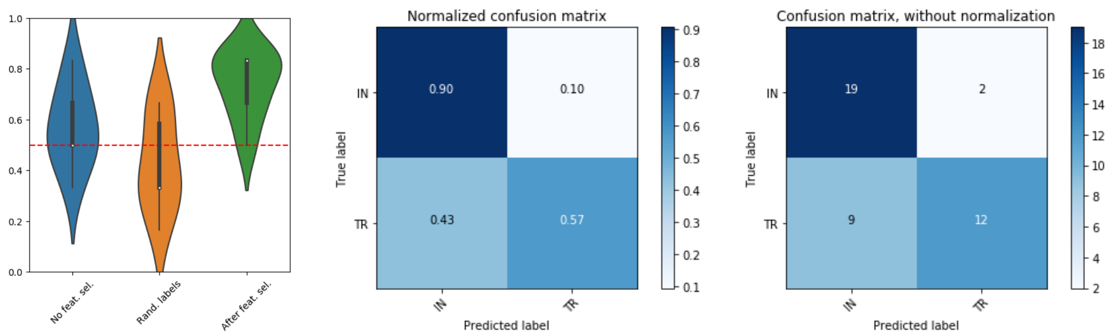
(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.7188	0.6563	0.8438
Std	0.1499	0.2480	0.1740
p	0.0201	0.1102	0.0001
MCC	-	-	0.6888

Figure C.4. **Pantomimes, non-signer labels:** Plots illustrating performance of classifiers on labels selected by participants in Study 2a. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive. (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers were equally accurate at classing intransitive and transitive samples.

C.2.2 Classifier Constructions: Winning non-signer labels

Results are presented in Fig. C.5. Only the analysis using feature selection meets significance at 73.81% (p = 0.0029). Even so, classifiers showed a distinct intransitive bias, suggesting that the high total accuracy we see is driven more by intransitive guesses.



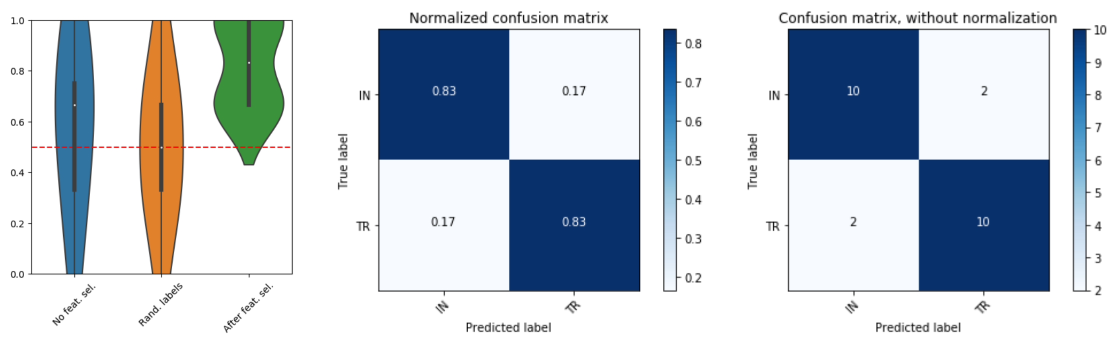
(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.5714	0.4286	0.7381
Std	0.1506	0.1750	0.1214
p	0.4408	0.4408	0.0029
MCC	-	-	0.5051

Figure C.5. **Classifier constructions, non-signer labels:** Plots illustrating performance of classifiers on labels selected by participants in Study 2a. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive. (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers showed a sizable intransitive bias.

C.3 Production & Perception

C.3.1 Pantomimes: Accurate winning non-signer labels



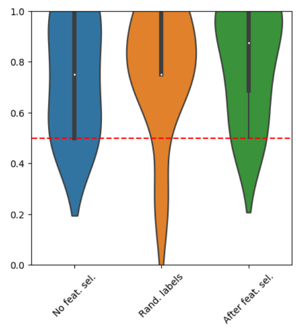
(a) (b) (c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.5833	0.5	0.8333
Std	0.3227	0.2886	0.1667
p	0.5413	1.0	0.0015
MCC	-	-	0.6667

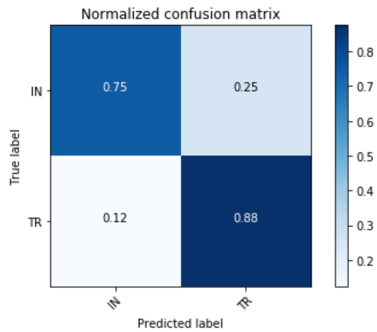
Figure C.6. **Pantomimes, accurate non-signer labels:** Plots illustrating performance of classifiers on only the labels selected accurately by participants in Study 2a. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive. (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers were equally accurate at classing intransitive and transitive samples.

C.3.2 Classifier Constructions: Accurate winning non-signer labels

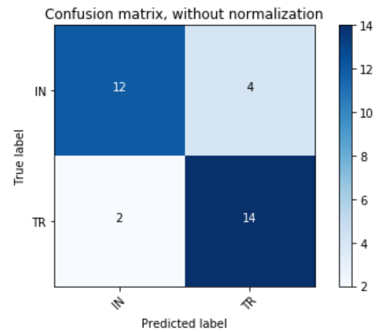
Accuracy with and without feature selection significant at 91.66% ($p < 0.0001$) and 75% ($p = 0.0227$), respectively. To note, the analysis using random labels returns a significant result, with an accuracy on par (or even higher) than some we have reported above. We have reason to believe that this is a fluke, considering such performance on random labels was never achieved in any other analysis, despite all analyses being run from the same script. As such, we iterated the analysis 10 times, each time drawing a different set of transitive samples (the transitive dataset was larger) with replacement. As can be seen in Fig. C.7d, the random analysis converges towards chance (grand mean = 59.02), while the analysis with feature extract attains 85.07% accuracy.



(a)



(b)



(c)

	No Feat. Sel.	Rand. Labels	Feat. Sel.
Mean	0.75	0.7813	0.8125
Std	0.2165	0.2318	0.2073
p	0.007	0.0021	0.0005
MCC	-	-	0.6299

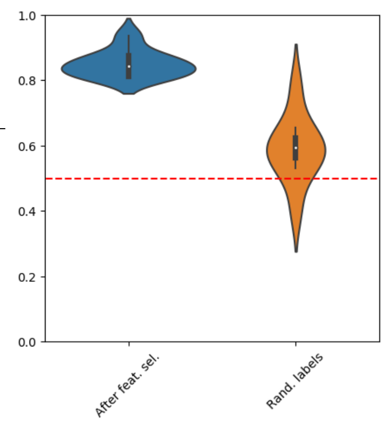


Figure C.7. **Classifier constructions, accurate non-signer labels:** Plots illustrating performance of classifiers on only the labels selected accurately by participants in Study 2a. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive. (b), (c) Normalized and raw confusion matrices (respectively) showing how the classifiers identified their targets. Descriptively, classifiers were more or less equally accurate at classing intransitive and transitive samples. (d) Plot of the analysis iterated 10 times, showing that analysis using random labels converges towards chance.

D. ANCILLARY MACHINE LEARNING ANALYSES (CLASSIFIER CONSTRUCTIONS AND PANTOMIMES):

Because of the flexibility machine learning analyses provide (from the choice of labels, to the partitioning of the data, etc.), we ran some clarifying analyses. In the below, we explore cross-subject differences in transitivity coding (D.1), the effect of the identity of the event on transitivity coding (D.2), whether data from pantomime samples can predict transitivity in classifier constructions (D.3), and whether transitivity distinctions are discernible from just handshape complexity measures (D.4).

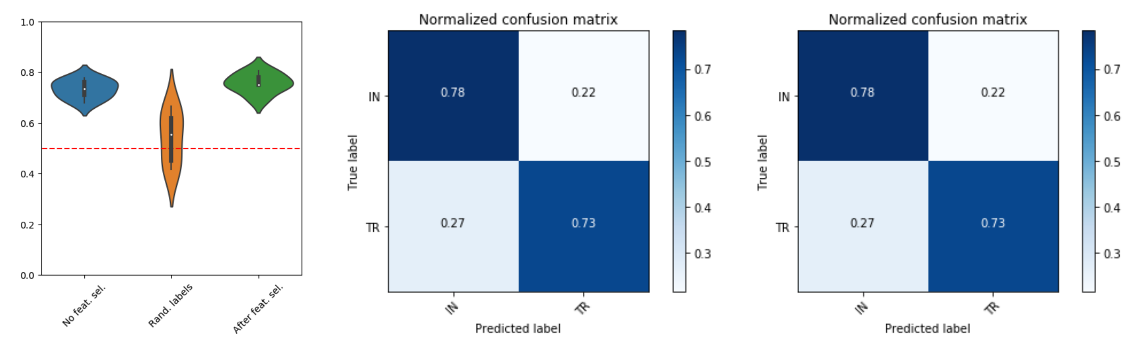
D.1 Train five pantomimers, test sixth:

Here we wanted to explore whether there are commonalities in transitivity coding between seemingly heterogeneous event representations in the pantomimes we collected. That is, for many of the events, pantomimers chose distinct aspects to represent, and/ or mapped those aspects to different articulators. Further, as indicated in the results of Study 1b, certain pantomimers were less successful than others in conveying these events. Nevertheless, through the noise, certain regularities w.r.t. transitivity coding may emerge.

To this end, we reran our pantomime production analysis, sorting the data s.t. samples from five of the six pantomimers made up the training set of each fold, and samples from the sixth pantomimer made up the test set. Aggregate information is presented in Fig. D.1, but as we can see in Tab. D.2, per-fold accuracy is nearly identical across the board, suggesting that transitivity coding was detectable across subjects. What's more, many of the same phonetic features were informative across folds, indicating that pantomimers used similar devices to express transitivity distinctions.

Table D.1.
Common cross-fold extracted features. Features were common across all six folds except where indicated: * = common to 5/6 folds, † = common to 4/6 folds.

Features
localmvmtfine, bent, static, mirror, pivot, acceleration, trajectory, closed, final, flex, awayfrom, towards, complexfinger, curved, wiggle, opposing, tense, multi, mono, base, nonbase, mediumjoint*, thumb*, complexjoint*, hands*, stacked*, narrow†



(a) (b) (c)

Figure D.1. **Train five pantomimers, test sixth** Plots illustrating performance of classifiers on ground truth labels, sorted s.t. classifiers were trained on data from five of the six pantomimers and then tested on data from the remaining one. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive.

Compared to the main production analysis, in which samples were randomly percolated into training/ test sets, mean classifier accuracy was not significantly different (75.69% vs. 74.76%; Welch's $t(12) = -0.3611$, 2-tailed $p = 0.3622$), but descriptively greater.

Table D.2.

Ancillary analysis: Train five pantomimers, test sixth: (Top) Per-fold accuracies, each one representing test accuracy on a distinct subject. (Bottom) Aggregate statistics are also given.

Test on	RVN	CM	CS	HO	IP	NP
Accuracy	0.8056	0.75	0.75	0.7917	0.75	0.6944

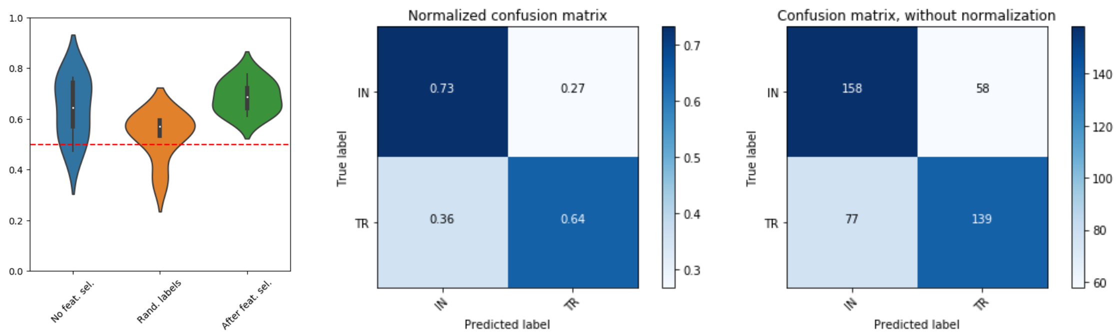
Average	STD	p	F1	MCC
0.7569	0.0356	<0.0001	.7506	0.5146

D.2 Train X concepts, test Y concepts:

To test the degree to which transitivity coding varies as a function of whatever concept or event is being depicted, we ran an additional analysis training classifiers on a certain set of concepts, and testing on a complementary set of concepts. An example scheme is presented in Tab. D.3. To note, for all analyses run on just the signer's data, this test is already factored in, as the signer contributed only one instance of each concept. Further, for analyses on perception (i.e., using non-signer labels) and production-perception agreement (i.e., using only accurate non-signer labels), this test is also already factored in. Thus, this solely applies to pantomime production data.

Accuracy with and without feature selection significant at 68.75% ($p < 0.0001$) and 64.12% ($p < 0.0001$), respectively, though the analysis with feature selection did not yield significantly greater accuracy ($t(5) = 0.8348$, 1-tailed $p = 0.2117$). Accuracy for the analysis using random labels was 53.70% ($p = 0.1357$). These are represented graphically in Fig. D.2a. Normalized and raw confusion matrices (Fig. D.2b, c respectively) show how the classifiers identified their targets. (Matrices represent samples with feature extraction.) Descriptively, classifiers demonstrated a slight bias in predicting intransitive labels ($F1 = 0.6731$; $MCC = 0.3765$).

When compared to the results of the analysis presented in Appendix C.1.1, mean classifier accuracy for this analysis is significantly lower (68.75% vs. 74.76%; Welch's $t(12) = -1.8408$, t-tailed $p = 0.0474$). However, when the results here are compared to the results of classifiers trained on only data from the best 72 pantomimes (one token of each of 72 concepts) or trained on the classifier construction production data (also one token of each of 72 concepts), results are similar: 68.75% vs. 67.18% vs. 70.83%, respectively. This suggests that concept-specific information may have artificially improved classifier accuracy in the analysis of pantomime production data. The general take-home message, however, remains the same, as classifier accuracy in the train X concepts, test Y concepts analysis is still significantly greater than chance.

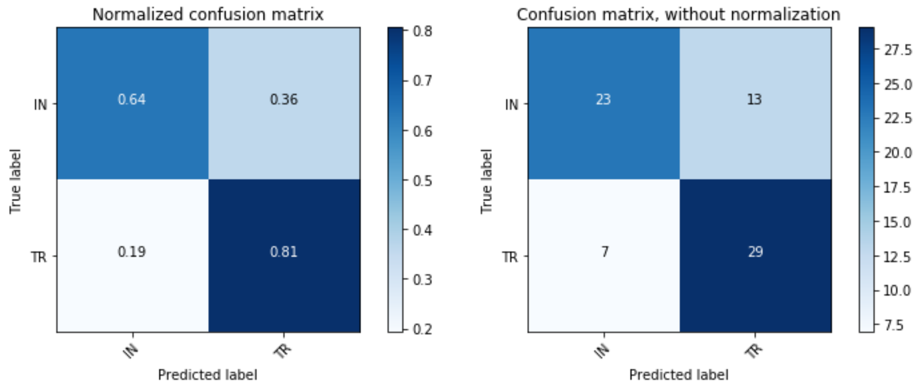


(a) (b) (c)

Figure D.2. **Train X concepts, test y concepts:** Plots illustrating performance of classifiers on ground truth labels, sorted s.t. training and test sets always contained completely complementary sets of concepts/ events. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during both training and testing. Dashed red line represents chance, 50%, either transitive or intransitive.

Table D.3.
Mock-up of train on X-concept, test on Y-concept scheme

Train on:	Test on:
crush-can	*cut-bread*
person-bend	*walk-backwards*
...	...
drop-ball	*hit-bottle-with-ball*



(a) (b)

Figure D.3. **Train pantomimes, test classifier constructions:** Confusion matrices demonstrating a small transitive bias.

D.3 Train pantomimes, test classifier constructions:

We were also curious to know whether the features used in transitivity coding in pantomimes are the same used in the coding of classifier constructions. If so, this adds weight to the hypothesis that there are universal mapping biases, s.t. linguistic forms are related to pre-, non-, or para-linguistic forms.

For this analysis, pantomime data (432 samples) made up the training set, and classifier construction data (72 samples) comprised the test set. It did not make sense to perform cross-validation, considering the training and testing samples come from two separate datasets. We used the same feature extraction method, same classifier, and same hyperparameters as in all other analyses.

Classifier accuracy was 72.22% (p = 0.0002).[1] Performance was always better in the analysis without feature extraction, except when k = 67 (i.e., all features were included). As such, we do not report here most informative features. Confusion matrices are presented in Fig. D.3. The matrices demonstrate a bias towards predicting transitive labels (F1 score ('micro'): 0.7222; MCC: 0.4507).

[1]Incidentally, the analysis using random labels performed at chance: 50% accuracy (p = 1.0).

It is interesting to note, though, that all features were required for maximum accuracy for this analysis, but in no other analysis we performed and it is not immediately obvious why this should be. However, considering that accuracy in this analysis is on par with the accuracies obtained in other analyses, we have evidence to suggest that the features that encode transitivity distinctions in pantomime are informative for this distinction in classifier constructions. We take this to mean that iconicity strongly underlies transitivity coding in both, and that this iconicity has grammaticalized into ASL. Of course, data are from just six pantomimers and one signer. So, more data are clearly needed to make this observation generalizable.

D.4 Informativeness of handshape complexity measures

Here we tested the claim put forth in Brentari et al. (2012, 2017); Goldin-Meadow et al. (2015), *inter alia*, that signers and non-signers use different coding strategies in transitive and intransitive environments. For instance, Brentari et al. (2012) found that non-signers show more finger complexity in the production of handling handshapes (transitive events) than in the production of object handshapes (intransitive event), contrary to what they found for the signing group. Brentari et al. (2017) also report that for both their signer and non-signer groups, higher joint complexity, specifically, was found in handling handshapes than object handshapes, though for most non-signer groups finger complexity did not differentiate transitive from intransitive productions.

This pattern was generally not repeated in our dataset, as can be seen in Figs. D.4 & D.5. In the production data, more transitive pantomimes exhibited high finger- and joint-complexity than intransitive pantomimes. Transitive pantomimes were also more likely to have medium joint complexity than intransitive pantomimes, but a greater number of transitivity pantomimes had low finger complexity as well. Intransitive pantomimes, however, were more likely to have medium finger complexity (by only one case, though) and low joint complexity. Although we do not run statistics on the proportion of in/transitive items having high- medium- or low finger- or joint-complexity, the pattern observed for pantomimes is consistent with previous findings.

However, the pattern is absent among classifier constructions. For instance, no transitive classifier construction had a complex joint specification, while four intransitive classifier constructions did. Further, for three measures–medium finger complexity, low finger complexity, and low joint complexity–there was no meaningful difference between transitive and intransitive stimuli.

A roughly similar pattern can be seen in the perception data. For pantomimes, again, there were more transitive items with complex finger specifications, and to a lesser extent complex joint, while intransitive items were more likely to have low finger and joint specifications. On the other hand, we see that there is no jarring difference in frequencies of complexity measure for classifier constructions, except perhaps for medium joint complexity, which was more common among transitive items.

Stepping back and looking at the perception and production tallies of both pantomimes and classifier constructions, we see generally that (a) low finger and joint complexities are common to both intransitive and transitive items, but (b) transitive items seem to have more complex finger and medium joint specifications. Perhaps this pattern (or some more detailed one) is learnable by our machine learning algorithm.

This is patently not the case. Among pantomimes, Fig. D.6, handshape complexity measures were generally uninformative in both production (57.87%; $p = 0.0012$)[2] or perception analyses (58.33%, $p = 0.5413$). The same is true of classifier constructions, Fig. D.7: Accuracy on production labels was 59.72% ($p = 0.1249$), and accuracy on perception labels was 62.5% ($p = 0.1249$).

There may be several reasons why we do not find transitivity distinctions in handshape complexity, as Brentari and colleagues do. As oft repeated elsewhere, our dataset contains more verb-object pairs than what these authors consider. Another

[2]Although the p-value is well below 0.05, we nevertheless do not believe that 58% represents a meaningful result, especially compared with more impactful results found in our analyses.

could be different uses of coding schema. We bin our measures into three levels (high, medium, and low), as do Brentari et al. (2017), but we additionally apply these to each handshape in the production. On the other hand, Brentari et al. (ibid.) simply add another ‘point’ to such productions. The result is that our scale for a given production is potentially 1 - 6, while Brentari and colleagues’ is 1 - 4. Another marked difference between our studies is the number of productions recorded: here, we recorded 432 pantomimes, and 73 classifier constructions (the number are much smaller for perception analyses), while Brentari and colleagues (ibid.) capture 2,537 data points. Our classifiers may just not have enough data.

One possibility for the differences between our and Brentari and colleagues’ results is our inclusion of multiple handshapes per production, as the latter only analyzed one handshape per production. Forty-two transitive productions use more than one handshape, but only 26 intransitive productions do so, too.[3]

Another source of discrepancy could be the number and type of events chosen, or the objects used in those events. Brentari and colleague’s paradigm includes a limited set of objects (e.g., lollipops and toy airplanes) occurring in only a few event types (e.g., verbs of *putting* or *being*). On the other hand, our study uses multiple different objects participating in 72 events. The choice of events and objects may bias (in their case) or wash out (in ours) differences in complexity measures.

With respect to our perception results, the caveat here is of course that human perceivers have decision functions that are quite distinct from a classifier’s. We reiterate here again that our classifier only considers single factors at a time, or rather, it does not assume interaction between features.[4]

[3]There are some rare instances where 3 or more handshapes were used. These were simply tallied as “using more than one handshape,” so the actual number of handshapes produced in transitive and intransitive contexts is only an estimate.

[4]Though note that there are some classifiers that do.

Pantomimes

	complexfinger	complexjoint	mediumfinger	mediumjoint	lowfinger	lowjoint
IN	21	20	140	132	189	198
TR	63	41	139	178	196	179
Difference	-42	-21	1	-46	-7	19

Classifier Constructions

	complexfinger	complexjoint	mediumfinger	mediumjoint	lowfinger	lowjoint
IN	11	4	34	26	20	35
TR	17	0	33	36	21	35
Difference	-6	4	1	-10	-1	0

Figure D.4. Raw counts of handshape complexity measures in intransitive and transitive (a) pantomimes and (b) classifier constructions in the production dataset. Red-shaded cells indicate where a given feature is more prevalent in transitive items, and green-shaded cells indicate where a feature is more prevalent in intransitive items. Unshaded cells indicate that a feature is equally prevalent in both datasets.

Pantomimes

	complexfinger	complexjoint	mediumfinger	mediumjoint	lowfinger	lowjoint
IN	19	16	88	83	109	117
TR	47	19	86	94	83	103
Difference	-28	-3	2	-11	26	14

Classifier Constructions

	complexfinger	complexjoint	mediumfinger	mediumjoint	lowfinger	lowjoint
IN	7	3	18	15	11	18
TR	9	0	16	22	11	14
Difference	-2	3	2	-7	0	4

Figure D.5. Raw counts of handshape complexity measures in intransitive and transitive (a) pantomimes and (b) classifier constructions in the perception dataset. Red-shaded cells indicate where a given feature is more prevalent in transitive items, and green-shaded cells indicate where a feature is more prevalent in intransitive items. Unshaded cells indicate that a feature is equally prevalent in both datasets.

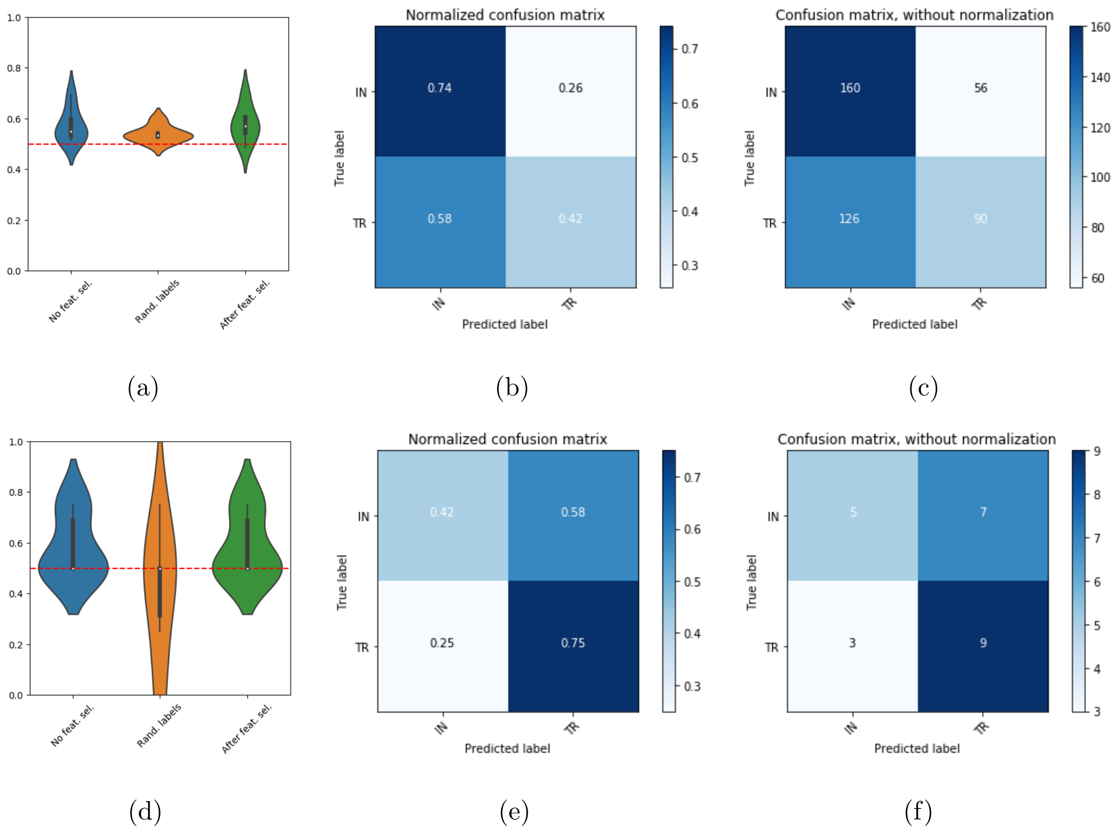


Figure D.6. **Handshape complexity features only (pantomimes):** Plots illustrating performance of classifiers, using only handshape complexity measures (finger and joint complexity). Analyses are of handshape complexity of dominant hand only. Plots in top row use production labels. Plots in bottom row use perception labels. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during training. Dashed red line represents chance, 50%, either transitive or intransitive.

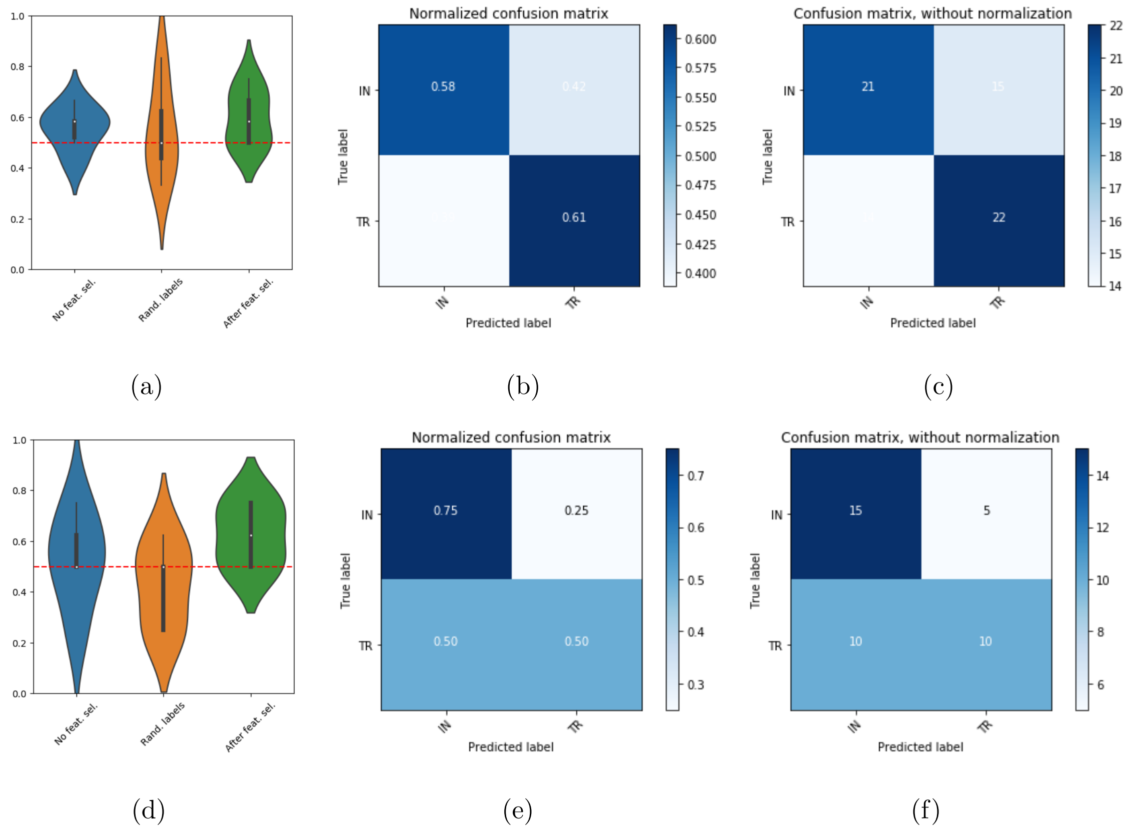


Figure D.7. **Handshape complexity features only (classifier constructions):** Plots illustrating performance of classifiers, using only handshape complexity measures (finger and joint complexity). Analyses are of handshape complexity of dominant hand only. Plots in top row use production labels. Plots in bottom row use perception labels. (a) Violin plots showing classifier accuracies before (blue) and after (green) feature selection. Orange blob represents analysis in which labels were randomly assigned to samples during training. Dashed red line represents chance, 50%, either transitive or intransitive.

E. GROUND TRUTH LABELS FOR ASL-LEX VERBS

Verbs from ASL-LEX and their transitivity labels. The first column is the item in question. The second column asks whether the item can (y) or cannot (no) take an object, and–if so–how many (1,2). The third column asks whether the verb can take an object noun phrase (NP); the fourth, an object complementizer phrase (CP). To note, both columns should contain an 'n' if the second column contains an 'n.' The fifth column asks if the verb can be used with the adverbial, WILLING, a diagnostic used by Benedicto and Brentari (2004) to test for an agentive subject.

Labels (the sixth column) were derived from the questionnaire in the preceding columns. A 'y, 1' indicates a 'transitive' verb, and a 'y, 2' a 'ditransitive' verb. An 'n' in the second column and an 'n' in the fifth indicate an 'unaccusative' verb, and an 'n' in the second column and a 'y' in the fifth indicate an 'unergative' verb.

Finally, the last column indicates a few things. In some cases, the signer marked that a verb does not take any objects (an 'n' in the second column) yet marked that this same verb may take an NP or CP complement– a contradiction. Likewise, the converse sometimes occurred: an item was marked as having one or more objects (a 'y' in the second column), but the signer indicated that this/ these objects were neither an NP or a CP (not a contradiction *per se*, only that there are likely no other category of argument that could co-occur with this set of verbs). The former case is marked with *Intrans w/ obj*, and the latter with *Trans w/o obj*. Finally, in some cases the signer indicated that an NP object performed a locative function by entering 'y, 1' into the second column and 'loc' in the third. We edited these entries to 'n,' as we are unsure of the syntactic function of locative arguments in ASL, but we assume they are *not* direct objects. We justify this assumption immediately below.

One complication in the assessing the argument structure of lexical verbs (in particular) is the lack of overt morphology, particles or otherwise to distinguish between arguments (in/direct objects), and obliques and adjuncts (here, locatives), should locatives be adjuncts in the first place (e.g., Kimmelman, 2018 considers the NP following a verb like LIVE to be a direct object, thus making LIVE a transitive verb). This is relevant not only to the discussion of argument structure (marking) in sign languages, but also to how we determined the ground truth labels for the ASL-LEX verbs (see below).

The following examples, due to Fischer and Gough (1978), illustrate the problem:

(1) a. ME ARRIVE/ GO-TO NEW-YORK
'I arrived in/ went to New York'
b. VASE FALL-OFF TABLE
'The vase fell off the table'
c. WE NOT INVITE-HER OUR PARTY
'We will not invite her to our party'

The examples in 1 illustrate that locative NPs may directly follow verbs without any overt marking (e.g., Case) to distinguish them from non-locative NPs. To our knowledge, there have been only a few solutions to this problem offered. Rathmann and Mathur REF show that locatives and (what we'll call) arguments are licensed by different verbs. 2b demonstrates that the verb BRING can licence four nominals, PAPER, JOHN, HOME, and SCHOOL. While verbs may accommodate up to four arguments, this is cross-linguistically rare. That ASL should have so many four-place predicates (akin to BRING) would also make it quite special. It is more harmonious from this perspective to assume that the nominals, HOME and SCHOOL, are adjuncts and not arguments. This is further corroborate by the fact that BRING and GIVE have two different interpretations, the first spatial and the second relational. In relational contexts, a four-place predicate parse is not available (2a). Finally, the distribution of the question words WHO and WHERE–which target arguments and adjuncts, respectively–demonstrate the selectional requirements of the verbs. Although both GIVE and BRING are transitive, similar tests could be used to suss out arguments and adjuncts in cases that are ambiguously transitive or intransitive (e.g., ENTER/ ARRIVE SCHOOL).

(2) a. *PAPER JOHN$_i$ BILL$_j$ MARY$_k$ $_j$GIVE$_k$
'John gave paper from Bill to Mary.'
b. PAPER JOHN$_i$ HOME$_a$ SCHOOL$_b$ $_a$BRING$_b$
'John brought paper from home to school.'

(3) a. WHO/ *WHERE JOHN$_i$ $_i$GIVE PAPER
'Who/ *where did John give paper to'
b. *WHO/ WHERE JOHN$_i$ BRING$_a$ PAPER
'*Who/ where did John bring paper to'

Finally, using corpus data from five sign languages, Börstell and colleagues (forthcoming) demonstrate that treating locative arguments as direct objects has a profound effect on transitivity prominence both among sign languages and between signed and spoken languages. Transitivity prominence refers to the proportion of languages that realize direct objects (etc.) for a particular verb. Verbs with high transitivity prominence, thus, are those that cross-linguistically select for a direct object. Kimmelman (2016) further defines transitivity prominence as the proportion of times a verb in a particular language realizes a direct object. For example, *break* and *eat* are optionally transitive in English, but one use might be more frequent than another.

Returning to Börstell et al.'s study, the authors calculated the proportion of cases where a select set of 12 verbs were each immediately followed by a direct object NP. They compared these proportions across the surveyed sign languages, but also against transitivity prominence of corresponding verbs in spoken languages (Haspelmath, 2015). When locative arguments were treated as direct objects, there were few strong relationships among between-language transitivity prominence rankings. There was also no correspondence between prominence of individual sign languages and the spoken language rankings. This is further evidence suggesting that NPs with locative interpretations are adjuncts and the verbs they co-occur with are indeed intransitive.

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
accident	y, 1	y	n	n	trans	
accomplish	y, 1	y	n	n	trans	
act	n	n	n	n	unacc	
announce	y,1	y	y	y	trans	
appear	y, 1	loc	n	y	trans	
arrive	y, 1	loc	n	n	trans	
ask	y, 2	y	n	y	ditrans	
bake1	y, 1	n	n	n	trans	trans w/o obj
bake2	y, 2	y	n	y	ditrans	
beat	y, 1	y	n	y	trans	
believe	y, 1	y	y	y	trans	
borrow	y, 2	y	n	y	ditrans	
brag	n	n	n	n	unacc	
break1	y, 1	y	n	n	trans	
break2	y, 1	y	n	y	trans	
breakdown	n	n	n	n	unacc	
breathe	n	n	n	n	unacc	
buy	y, 1	y	n	y	trans	
callattention	y, 1	y	n	n	trans	
calltty	y, 2	y	n	y	ditrans	
call	y, 2	y	n	y	ditrans	
can	n	n	n	n	unacc	
cause	y, 1	y	n	n	trans	
chat	y, 1	y	n	y	trans	
cheat1	y, 1	y	n	n	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
cheat2	y, 1	y	n	n	trans	
check	y, 1	y	n	y	trans	
comb	y,1	loc	n	y	trans	
come	n	loc	n	y	unerg	
continue	y, 1	y	y	y	trans	
cook	y, 1	y	n	y	trans	
copy	y, 2	y	n	y	ditrans	
coverup	y, 1 possibly 2	y	n	y	trans	
crawl	n	n	n	n	unacc	
create	y, 1	y	n	y	trans	
cry	n	n	n	n	unacc	
decide1	y, 1	n	y	n	trans	
decide2	y, 1	n	y	n	trans	
deny	y, 1	y	n	n	trans	
dice	y, 1	y	n	y	trans	
die	n	n	n	n	unacc	
disappear	n	loc	y	y	unerg	intrans w/obj
dive	n	loc	n	y	unerg	*
dontmind	n	n	n	n	unacc	
doubt	n	n	n	n	unacc	
download	y, 1	y	n	y	trans	
downsize1	y, 1	y	n	n	trans	
downsize	y, 1	y	n	y	trans	
draw	y, 1	y	n	y	trans	
drop	y, 1	loc	n	y	trans	
drown	n	loc	n	n	unacc	*
earn	y, 1	y	n	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
eat1	y, 1	y	n	y	trans	
eat2	y, 1	y	n	y	trans	
embarrass	y, 1	y	y	n	trans	
enjoy	y, 1	y	y	n	trans	
evaluate	y, 1	y	n	y	trans	
fall2	n	loc	n	n	unacc	*
feel	y, 1	y	y	n	trans	
fight	y, 2	y	n	y	ditrans	
figure	y, 1	y	y	y	trans	
film	y, 2	y	n	y	ditrans	
find	y, 1	y	n	n	trans	
fingerspell	y, 1	y	n	y	trans	
finish	y, 1	y	n	n	trans	
fly	n	n	n	y	unerg	
forbid	y, 1	y	n	y	trans	
forfeit	y, 1	y	n	y	trans	
frustrate	y, 1	y	y	n	trans	
get	y, 2	y	n	y	ditrans	
go	n	loc	n	y	unerg	*
graduate	y, 1	y	n	y	trans	
guess1	n	n	n	n	unacc	
guess2	y, 1	y	y	y	trans	
happen	n	n	n	n	unacc	
have	y, 1	y	n	n	trans	
health	n	n	n	n	unacc	
hearing	y, 1	y	n	n	trans	
help	y, 1	y	n	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
hope	y, 1	y	y	n	trans	
hunt	y, 1	y	n	y	trans	
ignore	y, 1	y	n	y	trans	
imagine	n	n	y	n	unacc	intrans w/obj
impact	y, 1	y	n	n	trans	
inject	y, 1	y	n	y	trans	
insult	y, 1	y	n	n	trans	
introduce	y, 2	y	n	y	ditrans	
invite	y, 2	y	n	y	ditrans	
juggle	y, 1	y	n	y	trans	
jump	n	loc	n	y	unerg	*
kill	y, 1	y	n	y	trans	
kneel	n	loc	n	y	unerg	*
knitting	y, 1	y	n	y	trans	
know	y, 1	y	y	y	trans	
laugh	n	n	n	y	unerg	
learn	y, 1	y	y	y	trans	
leave	n	loc	n	y	unerg	*
letknow	y, 1	y	y	n	trans	
lie	n	n	n	y	unerg	
live1	n	loc	n	y	unerg	*
live2	n	loc	n	y	unerg	*
lookappearance	y, 1	y	n	n	trans	
lookat	y, 1	y	n	y	trans	
lookfor	y, 1	y	n	y	trans	
losegame	y, 1	y	n	y	trans	
lose	y, 1	y	n	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
make	y, 1	y	n	y	trans	
match	y, 2	y	n	y	ditrans	
mean2	y, 1	n	y	n	trans	
meet	y, 1	y	n	y	trans	
misunderstand	y, 1	y	y	n	trans	
mock	y, 1	y	n	y	trans	
move	y, 1	loc	n	y	trans	
need	y, 1	y	n	n	trans	
offend	y, 1	y	n	y	trans	
owe	y, 1	y	n	y	trans	
own	y, 1	y	n	y	trans	
paint	y, 2	y	n	y	ditrans	
parachute	n	loc	n	y	unerg	*
peg	y, 1	loc	n	y	trans	
play	y, 1	y	n	y	trans	
pop	y, 1	y	n	y	trans	
prefer	y, 1	y	y	n	trans	
pretend	y, 1	y	y	y	trans	
promise	y, 1	y	y	y	trans	
promote	y, 1	y	n	y	trans	
pull	y, 1	y	n	y	trans	
punish	y, 1	y	n	y	trans	
push	y, 1	y	n	y	trans	
rake	y, 1	y	n	y	trans	
read	y, 1	y	n	y	trans	
recording	y, 1	y	n	n	trans	
relax	y, 1	n	y	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
remember	y, 1	y	y	y	trans	
rest	n	n	n	y	unerg	
run	n	n	n	y	unerg	
sail	n	loc	n	y	unerg	*
saw	y, 2	y	n	y	ditrans	
see	y, 1	y	y	y	trans	
select	y, 1	y	n	y	trans	
setup	y, 1	y	y	y	trans	
sew	y, 1	y	n	y	trans	
shave	y	y	n	y	trans	
shop1	y, 1	y	n	y	trans	
show	y, 2	y	y	y	ditrans	
sing	y, 1	y	n	y	trans	
sit	n	loc	n	y	unerg	*
skate	n	loc	n	y	unerg	*
skateboarding	n	loc	n	y	unerg	*
sleep	n	loc	n	y	unerg	*
smoking	y, 1	y	n	y	trans	
sneeze	n	loc	n	n	unacc	*
start	y, 1	y	y	y	trans	
steal	y, 1	y	n	y	trans	
stir	y, 2	y	n	y	ditrans	
stop	y, 1	y	y	y	trans	
subtract	y, 1	y	n	n	trans	
summarize	y, 1	y	n	y	trans	
surf	n	loc	n	y	unerg	*
swallow	y, 1	y	n	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
sweep	n	loc	n	y	unerg	*
swing	n	n	n	n	unacc	
talk	y, 1	y	n	y	trans	
teach	y, 1	y	n	y	trans	
tear	y, 1	y	n	y	trans	
tease	y, 1	y	n	y	trans	
tell	y, 1	y	y	y	trans	
text	y, 2	y	n	y	ditrans	
thief2	y, 1	y	n	y	trans	
thinkover	y, 1	y	n	y	trans	
think	y, 1	n	y	n	trans	
throw	y, 1	y	n	y	trans	
translate	n	n	n	y	unerg	
travel	n	loc	n	y	unerg	*
try	y, 1	y	n	y	trans	
understand	y, 1	y	y	y	trans	
upload	y, 1	y	n	y	trans	
use	y, 1	y	n	y	trans	
vacuum	n	loc	n	y	unerg	
vomit	n	n	n	n	unacc	
wait	n	n	n	y	unerg	
walk	n	n	n	y	unerg	
wander	n	n	n	y	unerg	
want	y, 1	y	n	n	trans	
warn	y, 1	y	y	n	trans	
waste	y, 1	y	n	y	trans	
wear	y, 1	y	n	y	trans	

VERB	Can take object?	Object NP?	Object CP?	Can use with WILLING?	Label	Note
weigh	y, 1	y	n	y	trans	
whip	y	y	n	y	trans	
win	y, 1	y	n	y	trans	
wink	n	n	n	n	unacc	
wonder	y, 1	n	y	n	trans	
work	n	n	n	y	unerg	
worry	n	n	n	n	unacc	
write	y, 1	y	n	y	trans	
zoomin	n	n	n	n	unacc	
zoomoff	n	n	n	n	unacc	

F. FULL RESULTS OF AMT EXPERIMENT, LEXICAL VERBS

For each item, a label, and the number of class assignments it received are presented. Class labels were decided by raw tally. In cases where there were an equal number of votes for two or more categories, the label is meaningless but assigned in the following order: Trans(itive) > Ditrans(itive) > (Intransitive) unerg(ative) > (Intransitive) unacc(usative). The strength of inter-rater agreement was determined via a 1-sample t-test against chance (25%), performed on the class that had the most number of votes. The suffix '-rep' indicates that the item was repeated in a different survey. P-values are two-tailed.

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
accident	trans	7	2	7	3	1.0415	0.3114
accident-rep	trans	9	6	3	2	1.7523	0.0958
accomplish	unerg.	2	2	0	13	4.8536	0.0002
accomplish-rep	unerg.	6	1	2	12	2.9047	0.0088
act	unerg.	4	4	0	12	3.1141	0.0057
act-rep	unerg.	6	0	2	13	3.3986	0.0029
announce	unerg.	4	3	2	12	2.9047	0.0088
appear	unacc.	4	5	6	6	0.3536	0.7274
arrive	trans	5	5	0	5	0.6614	0.5191
ask	unerg.	6	3	1	9	1.9007	0.0735
bake-1	trans	9	0	2	6	2.2392	0.0397
bake-2	ditrans	3	9	7	1	1.7523	0.0958
beat	ditrans	7	8	3	3	1.206	0.2419
believe	unerg.	4	5	1	11	2.4518	0.0235

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
borrow	trans	7	1	1	6	1.625	0.1265
brag	unerg.	4	2	1	12	3.3561	0.0035
break-1	trans	17	0	2	1	7.3244	0.00
break-2	trans	15	1	4	1	4.5962	0.0002
breakdown	unacc.	1	0	13	3	4.8536	0.0002
breathe	unerg.	7	0	3	11	2.4518	0.0235
buy	ditrans	4	11	0	0	4.0896	0.0011
call	trans	11	2	0	6	2.8267	0.0112
call-attention	trans	10	3	0	4	2.749	0.0143
call-tty	unacc.	4	5	9	2	1.7523	0.0958
can	ditrans	5	7	6	3	0.7906	0.4385
cause	ditrans	6	11	1	3	2.4518	0.0235
chat	unerg.	3	0	2	10	3.3072	0.0052
cheat-1	trans	7	0	5	7	1.0415	0.3114
cheat-2	unacc.	4	3	6	4	0.8616	0.4016
check	unerg.	2	5	5	8	1.3346	0.1978
comb	unerg.	7	2	1	11	2.4518	0.0235
come	unerg.	6	2	6	7	0.7906	0.4385
continue	trans	8	2	0	5	2.125	0.0519
cook	ditrans	4	6	6	3	0.6005	0.5557
copy	trans	7	5	2	3	1.3148	0.2071
cover-up	unacc.	3	0	8	8	1.3346	0.1978
crawl	trans	4	4	3	4	0.141	0.8899
crawl-rep	trans	8	3	7	1	1.4699	0.1589
create	trans	7	1	7	4	1.0415	0.3114
create-rep	trans	11	2	6	2	2.4518	0.0235
cry	unerg.	3	0	0	14	6.0178	0.00

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
decide-1	unerg.	4	2	1	13	3.6555	0.0017
decide-2	ditrans	6	8	5	2	1.206	0.2419
deny	unerg.	8	2	1	10	2.0254	0.0564
dice	unacc.	4	4	6	1	1.1456	0.2711
die	unacc.	2	6	9	2	1.9007	0.0735
disappear	trans	11	3	2	1	3.3235	0.0043
dive	ditrans	2	10	4	4	2.1794	0.0421
dont-mind	unerg.	8	0	2	11	2.4518	0.0235
doubt	unerg.	4	0	0	17	6.3723	0.00
download	ditrans	2	7	5	1	1.625	0.1265
downsize	unacc.	5	3	9	2	1.9007	0.0735
downsize-1	unacc.	4	0	12	1	4.0021	0.001
draw	trans	8	2	8	2	1.3346	0.1978
drop	ditrans	2	14	2	3	3.9528	0.0008
drown	unacc.	7	4	8	2	1.206	0.2419
earn	trans	6	6	2	1	1.1456	0.2711
eat-1	unerg.	6	2	1	10	2.3479	0.0305
eat-2	unerg.	5	3	0	9	2.2392	0.0397
embarrass	unerg.	4	0	5	11	2.6285	0.0165
enjoy	unerg.	8	3	0	10	2.0254	0.0564
evaluate	unerg.	6	1	4	10	2.0254	0.0564
fall-2	unerg.	3	1	2	9	2.6732	0.0182
feel	unerg.	6	2	0	11	2.8267	0.0112
fight	trans	7	1	2	7	1.3148	0.2071
figure	unacc.	3	6	9	2	1.7523	0.0958
film	trans	10	4	2	5	2.0254	0.0564
find	trans	9	3	3	6	1.6137	0.1223

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
fingerspell	unerg.	4	2	3	6	1.1456	0.2711
finish	unerg.	1	2	3	13	3.9632	0.0009
fly	unerg.	1	2	6	8	1.7678	0.0962
forbid	trans	12	2	0	7	2.9047	0.0088
forbid-rep	trans	7	2	5	6	0.9139	0.3722
forfeit	unerg.	5	1	1	8	2.125	0.0519
forfeit-rep	unerg.	6	2	0	9	2.2392	0.0397
frustrate	unerg.	2	2	0	15	5.6142	0.00
frustrate-rep	unerg.	3	0	0	12	5.1448	0.0001
get	trans	10	2	2	3	2.749	0.0143
go	ditrans	1	10	4	5	2.1794	0.0421
graduate	unacc.	3	5	7	6	0.7906	0.4385
guess-1	trans	9	7	1	4	1.6137	0.1223
guess-2	ditrans	4	5	3	3	0.6614	0.5191
guess-2-rep	trans	13	5	0	3	3.3986	0.0029
happen	unerg.	5	4	2	8	1.4699	0.1589
have	unerg.	1	4	0	12	4.0021	0.001
health	trans	10	3	1	6	2.1794	0.0421
hearing	trans	9	2	1	9	1.6137	0.1223
help	trans	11	5	2	3	2.4518	0.0235
hope	unerg.	7	0	0	8	2.125	0.0519
hunt	trans	6	6	3	4	0.6005	0.5557
ignore	unerg.	2	1	2	12	4.0021	0.001
imagine	unerg.	3	1	2	14	4.2804	0.0004
impact	trans	10	3	0	8	2.0254	0.0564
inject	trans	8	4	3	6	1.206	0.2419
insult	unerg.	4	0	2	9	2.6732	0.0182

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
introduce	ditrans	3	7	6	3	1.0415	0.3114
invite	unerg.	1	3	6	7	1.3148	0.2071
juggle	trans	9	1	5	5	1.7523	0.0958
jump	trans	7	7	5	2	0.7906	0.4385
kill	unacc.	2	7	11	1	2.4518	0.0235
kneel	ditrans	5	6	3	1	1.1456	0.2711
knitting	trans	11	1	3	4	2.8267	0.0112
know	unerg.	5	1	1	10	2.749	0.0143
laugh	unerg.	1	0	1	18	9.4443	0.00
learn	trans	8	4	4	5	1.206	0.2419
leave	unacc.	3	7	9	2	1.6137	0.1223
let-know	unerg.	1	2	2	10	3.3072	0.0052
lie	unerg.	4	2	4	9	1.9007	0.0735
live-1	trans	7	2	5	7	0.7906	0.4385
live-1-rep	unerg.	1	2	4	10	2.749	0.0143
live-2	unerg.	5	2	3	11	2.4518	0.0235
live-2-rep	unerg.	3	2	4	6	1.1456	0.2711
look-appearance	unerg.	3	0	2	10	3.3072	0.0052
look-appearance-rep	trans	9	2	2	8	1.6137	0.1223
look-at	unerg.	5	1	1	12	3.3561	0.0035
look-for	unerg.	2	1	3	11	3.3235	0.0043
lose	trans	9	4	2	5	1.7523	0.0958
lose-game	trans	7	5	5	4	0.7906	0.4385
make	trans	13	1	5	2	3.3986	0.0029
match	trans	10	1	3	1	3.3072	0.0052
mean-2	unacc.	3	4	7	5	1.0415	0.3114
meet	unacc.	2	3	8	4	1.7678	0.0962

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
misunderstand	unerg.	2	1	8	9	1.7523	0.0958
mock	trans	10	2	3	6	2.0254	0.0564
move	unacc.	1	9	11	0	2.4518	0.0235
need	unerg.	5	2	2	6	1.1456	0.2711
offend	trans	5	5	4	5	0.1268	0.9005
owe	trans	6	5	2	4	0.8616	0.4016
own	unerg.	5	5	2	8	1.3346	0.1978
paint	trans	12	1	4	4	2.9047	0.0088
parachute	ditrans	2	10	6	3	2.0254	0.0564
peg	unacc.	4	3	5	3	0.6614	0.5191
play	unerg.	6	1	1	11	2.8267	0.0112
pop	trans	11	1	1	4	3.3235	0.0043
prefer	unerg.	5	3	2	10	2.1794	0.0421
pretend	unerg.	4	1	1	15	4.5962	0.0002
promise	trans	10	0	2	9	2.0254	0.0564
promote	ditrans	0	8	6	1	2.125	0.0519
pull	trans	13	4	1	1	3.9632	0.0009
punish	ditrans	3	7	2	5	1.3148	0.2071
push	trans	14	5	0	1	4.2804	0.0004
rake	trans	13	0	4	4	3.3986	0.0029
read	trans	15	0	3	3	4.5962	0.0002
recording	trans	5	2	4	3	0.6614	0.5191
relax	trans	12	0	1	7	3.1141	0.0057
relax-rep	unerg.	8	0	0	13	3.3986	0.0029
remember	unerg.	6	2	5	8	1.206	0.2419
remember-rep	unerg.	5	1	4	9	1.9007	0.0735
rest	unerg.	6	2	1	12	2.9047	0.0088

Tallies

item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
run	unerg.	3	2	4	6	1.1456	0.2711
run-rep	ditrans	7	8	3	1	1.4699	0.1589
sail	unacc.	2	3	8	6	1.4699	0.1589
saw	trans	14	1	1	1	6.0178	0.00
see	unerg.	5	1	4	10	2.1794	0.0421
select	ditrans	8	9	2	2	1.6137	0.1223
set-up	trans	8	4	5	4	1.206	0.2419
sew	trans	11	1	1	2	4.0896	0.0011
shave	trans	11	0	1	7	2.8267	0.0112
shop-1	trans	7	5	0	5	1.3148	0.2071
show	unerg.	4	6	3	7	0.9139	0.3722
sing	trans	13	1	2	5	3.3986	0.0029
sit	trans	7	6	4	4	0.7906	0.4385
skate	trans	4	3	4	4	0.141	0.8899
skateboarding	ditrans	3	6	4	6	0.6005	0.5557
sleep	unerg.	3	2	0	12	4.0021	0.001
smoking	trans	11	2	0	7	2.6285	0.0165
sneeze	unerg.	3	3	1	14	3.9528	0.0008
start	trans	15	1	1	4	4.5962	0.0002
steal	trans	10	1	3	1	3.3072	0.0052
stir	trans	14	1	1	3	4.6906	0.0002
stop	trans	12	0	0	5	4.0021	0.001
subtract	ditrans	6	9	4	1	1.7523	0.0958
summarize	trans	10	3	6	2	2.0254	0.0564
surf	trans	9	5	3	4	1.6137	0.1223
swallow	unerg.	4	0	2	9	2.6732	0.0182
sweep	trans	8	3	5	3	1.4699	0.1589

		Tallies					
item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
swing	unacc.	4	2	6	5	0.8616	0.4016
talk	unerg.	3	0	1	16	5.9935	0.00
teach	unerg.	5	4	4	8	1.206	0.2419
tear	unerg.	3	0	1	17	6.3723	0.00
tease	unerg.	7	2	0	8	1.7678	0.0962
tease-rep	trans	12	2	3	4	2.9047	0.0088
tell	unerg.	9	0	1	10	2.1794	0.0421
tell-rep	unerg.	4	0	0	11	4.0896	0.0011
text	trans	12	0	1	8	2.9047	0.0088
text-rep	trans	12	1	0	7	3.1141	0.0057
thief-2	trans	9	1	8	3	1.6137	0.1223
thief-2-rep	trans	8	2	2	5	1.7678	0.0962
think	unerg.	0	0	0	15	10	0
think-over	unerg.	4	2	2	11	2.8267	0.0112
throw	ditrans	5	6	1	5	0.8616	0.4016
translate	unacc.	3	6	9	2	1.7523	0.0958
travel	unerg.	5	5	2	9	1.6137	0.1223
try	unerg.	4	2	1	14	3.9528	0.0008
understand	unerg.	1	0	0	13	6.7876	0.00
upload	ditrans	3	7	5	4	1.0415	0.3114
use	unerg.	3	5	3	6	0.8616	0.4016
vacuum	trans	12	6	2	0	3.1141	0.0057
vomit	unerg.	5	2	2	12	2.9047	0.0088
wait	unerg.	5	1	4	11	2.4518	0.0235
walk	unerg.	2	0	5	8	2.125	0.0519
wander	trans	7	1	4	7	1.0415	0.3114
want	trans	8	3	3	3	1.7678	0.0962

Tallies

item	label	trans.	ditrans.	unacc.	unerg.	t-value	p
warn	trans	11	2	3	4	2.6285	0.0165
waste	trans	11	2	2	6	2.4518	0.0235
wear	trans	7	6	3	5	0.7906	0.4385
weigh	trans	10	2	1	2	3.3072	0.0052
whip	trans	15	1	0	3	5.6142	0.00
win	trans	8	3	5	1	1.7678	0.0962
wink	trans	13	1	1	5	3.6555	0.0017
wonder	unerg.	4	2	1	14	3.9528	0.0008
work	trans	9	6	2	4	1.6137	0.1223
worry	unerg.	2	0	0	13	6.7876	0.00
write	trans	14	1	1	3	4.6906	0.0002
zoom-in	unacc.	4	0	10	3	2.749	0.0143
zoom-off	ditrans	5	6	5	4	0.4756	0.6398