Modelling the Integration of Co-Speech Gestures into Sentence Meaning Composition

Ludmila Reimer (ludmila.reimer@rub.de)
Department of Philosophy II, Ruhr University Bochum, Universitätsstraße 150
44870 Bochum, Germany

Markus Werning (markus.werning@rub.de)
Department of Philosophy II, Ruhr University Bochum, Universitätsstraße 150
44870 Bochum, Germany

Abstract

To investigate how co-speech gestures modulate linguistic understanding, we experimentally test two Bayesian Pragmatic models. We identify the semantic effect of a spoken or gestural utterance with the change it makes in listener's probabilistic predictions of the speaker's communicative intentions. We focus on action-expressing gestures and the respective verbs that correspond to action-affording instruments or, respectively, their denoting nouns. Combining Pustejovsky's' Generative Lexicon approach with Gibson's affordance theory, we ask: (1) Does a co-speech gesture make any difference for semantic comprehension and the corresponding probabilistic prediction? (2) Is the semantic effect of a gesture similar or identical to the one of the corresponding verb? (3) To which extent does the gesture's semantic effect depend on the listener's recognition of the gesture as an expression of the corresponding verb? (4) Does the comprehended affordance predict the instrument better than the co-occurrence statistics (GloVe) regarding the verb and the noun?

Keywords: co-speech gestures; sentence meaning composition; Bayesian Pragmatics; Generative Lexicon; affordances; enacted cognition

Introduction

While traditional theories of semantics assume that the meaning of a sentence is compositionally determined by the meanings of the individual words and the way they are syntactically combined, the dynamic view considers the evolvement of meaning as an online process unfolding in time. While listening to an utterance word by word, at every moment in time, the listener generates a prediction about the communicative intention that the speaker is about to express. With every word, this prediction might change. Whereas, according to traditional compositional semantics the meaning of a word is whatever it contributes to the meaning of a discourse, in the dynamic picture the meaning of a word can be regarded as the change it makes in the listener's prediction of the speaker's communicative intention. The gestures produced by the speaker can provide a second stream of information for the listener, especially in the case of iconic co-speech gestures. They can be more nuanced than spoken utterances, e.g., providing the shape of a spoken about object or representing aspects of a described action, such as speed, orientation, and even indicate what kind of tool was used during the action. We focus on this latter kind, i.e., iconic co-speech gestures that accompany action verbs and provide more detailed information on the actions and the tools used in comparison to the spoken stand-alone utterances.

To shed some light on how the different relationships between co-speech gesture and utterance affect the semantics in the act of communication, we investigate the semantic predictions generated by the speaker in the listener. The dynamic picture allows us to investigate the semantic effect of the co-speech gesture g by looking at the way it influences the listener's probabilistic prediction of the upcoming word w_{n+1}: $P(w_{n+1}|g, w_1 \dots w_n)$.

We will use a Bayesian Pragmatic Framework (Werning et al., 2019), which is closely related to Rational Speech Act Theory, RSA (Frank & Goodman, 2012). The general idea of Bayesian Pragmatics is to account for the rational cooperation between speaker and listener in an act of communication by modelling the listener's probabilistic predictions about the speaker's communicative intentions with the use of Bayesian probability theory. Bayesian pragmatics has been successfully used to explain results of behavioral experiments on simple referential games (Frank & Goodman, 2012), scalar implicatures (Degen et al., 2015; Goodman & Stuhlmüller, 2013), gradable adjectives (Lassiter & Goodman, 2013; Qing & Franke, 2014), modal expressions (Lassiter & Goodman, 2015), and figurative meaning (Kao et al., 2014). Predictive processing is widely acknowledged in cognitive and neuroscience as a general mechanism by which the subject at every point in time generates the most probable prediction of the next event on the basis of ongoing perceptual input and learned statistical regularities (Clark, 2013; Hohwy, 2013). The idea of predictive processing has also been applied to language comprehension (Huettig, 2015; Kuperberg & Jaeger, 2016, for review). Often it is implemented by Bayesian pragmatics, which also offers itself as a model in what has been introduced as the Predictive Completion Task of communication (Cosentino et al., 2017).

Generative Lexicon and Affordances We combine one of the ideas found in Pustejovky's Generative Lexicon approach (Pustejovsky, 1998) with Gibson's notion of affordances (Gibson, 1977, 1979). According to the Generative Lexicon approach, lexical entries of concrete nouns contain "Qualia Structures" that specify certain aspects (or components) of a word's meaning. These components are retrieved in sentence meaning composition. We focus on the telic component, i.e., the function or purpose of an object denoted by a noun. For

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

example, the noun *knife* contains the telic components *cut* or *spread*. Typically, the retrieval is triggered by verbs like *use*. Therefore, sentences like (1) typically are understood as (2):

(1) They used the spoon for the soup.

(2) They used the spoon for eating the soup.

According to Gibson (1979, 1986) many objects come with subject- and situation-dependent affordances. These are dispositional properties (e.g., sit-ability) that relate to actions to be potentially performed on that object (Werning, 2010). They can either be generic (e.g., sit-ability for chairs) or ad hoc, i.e., a particular affordance depending on the situation, agent, and object (e.g., the chair can be a hiding spot for children during hide-and-seek). Generic affordances are often stored as telic components in the lexicon of nouns. Looking at the example above, the interpretation of (1) is ambiguous between (2) and (3):

(3) They used the spoon for stirring the soup.

An iconic co-speech gesture (e.g., an eating or stirring gesture) may be used to disambiguate the interpretation of (1) intended by the speaker as meaning either (2) and (3).

Holle and Gunter (2007) provide evidence that gestures help listeners disambiguate homonyms, e.g., ball in *She controlled the ball* to be interpreted either as an object used in football or as an event involving dancing. Furthermore, the presence of a gesture facilitates the processing of lesser frequent interpretations of homonyms. According to McNeill (McNeill, 1992) and his collected sample narrations, co-speech gestures are the most common gestures used. Thus, it is likely that speakers use them frequently to trigger the intended telic component in a listener, facilitating speech comprehension, and cutting down on words, making the overall communication smoother and more efficient. Since co-speech gestures usually happen automatically and without planning, the intended telic component that is "acted out" or coded in the co-speech gesture could be activated unconsciously by the speaker's lexicon and the linked affordance. The lexicon, as we assume it, does not only store lexical or semantic entries, but also either stores or links to corresponding sensory-motor information. Other researchers also consider gestures to be "demonstrative of the embodied nature of the mind" (Hostetter and Alibali, 2008), as they are found to prime or activate internal action representations (Goldin-Meadow & Beilock, 2010; Krauss et al., 2000; Pouw et al., 2014). Additionally, there is an overlap in the way how speakers use gestures to support cognitive processes and how cognizers might manipulate their environment (Pouw et al., 2014). Pouw et al. (2014) thus assume that gestures provide the cognitive system with an external, physical, and visual presence that provides a means to think with – making gestures not only a mere by-product of speech and imagistic thinking but give them the status of embedded or extended cognition.

For an in-depth review of various methodologies used to investigate gestures, see (Kandana Arachchige et al., 2021). In their paper, the authors offer a wide overview on the current research, and it becomes clear that the paradigm to use either gestures that are either congruent or incongruent with the utterances of a speaker is widely used. However, most studies limit themselves to whether the integration of information provided by gestures is an automatic process and to the question which brain regions might be involved. So far, there seem to be no approaches that try to pry apart the contribution of a gesture in semantic sentence composition in a detailed manner, nor approaches that apply Bayesian Pragmatics to model the integration of gestural information with spoken utterances.

Table 1:
Solid lines indicate congruent continuations (matches), dashed lines incongruent continuations (mismatches).

Context Sentences	Target Sentences
General *Das Kind ist dabei, die Kekse zu backen.* (The child is baking cookies.)	**Noun I** *Es benutzte dazu das Förmchen.* (To do so, they used the cookie cutter.)
Specific I *Das Kind ist dabei, die Kekse auszustechen.* (The child is cutting out cookies.)	**Noun II** *Es benutzte dazu den Pinsel.* (To do so, they used the brush.)
Specific II *Das Kind ist dabei, die Kekse zu bestreichen.* (The child is glazing cookies.)	

Material Preparation

For all experiments conducted, the material was constructed in German with uniform constraints for all experiments. First, we constructed sets of context and target sentences in a systematic way:

a) General context: "The agent x is v_{gen}-ing object y."
b) Specific context: "The agent x is v_{spec}-ing object y."
c) $Noun$ target: "To do so, they (x) are using n."

Note that in German all sentences used for $Noun$ target ended with a noun n denoting the tool used to perform the action denoted by the verb v_{spec}, as can be seen in Table 1. The target sentences were constructed in pairs, featuring two different instruments that were coherent continuations of the General context sentences – but only coherent for those Specific context sentences that featured the corresponding (matching) verb of the noun used in the target sentence. I.e., v_{spec} describes the function of the instrument denoted by the noun n and thus expresses the noun's n telic component. For example, the main telic component of a *Förmchen* ('cookie cutter') is *ausstechen* ('to cut out') and for *Pinsel* ('brush') it is *bestreichen* ('to glaze'). Both the cookie cutter and the

brush are also congruent (matching) continuations of the sentence:

(4) Das Kind ist dabei, die Kekse zu backen.
'The child is baking cookies.'

But obviously not for each other's Specific context sentences, since cookie cutters are usually not used for brushing or glazing, and brushes are not used for cutting out cookies.

All target pairs were controlled for frequency, grammatical gender in the case of nouns (anticipating future use in online processing tasks, the noun n would not be primed a gendered article), and GloVe values (between (1) both context sentences and the noun n used in $Noun$ target, (2) context verb and target noun, both for matches and mismatches). GloVe is a measure of semantic similarity. We used an implementation of GloVe (Global Vectors, Pennington, Socher, & Manning, 2014) based on all articles from German Wikipedia and ca. three million news articles from Leipzig Corpora Collection (Goldhahn, Eckart, & Quasthoff, 2012).

Video Recordings and Post-Processing

Videos for the underspecified sentences, combined with three gesture conditions (match, mismatch, no gesture) were recorded in full HD resolution (1920x1080 pixel) and a frame rate of 50fps. The speaker was recorded from the knees up, allowing for framing with the hands to be visible at all times. The speaker was not informed about the purpose of the experiment prior to filming; the only instructions received were to (1) read out the first sentence of a sheet while trying to convey additional information present in the second sentence and (2) to read out the same sentence once without gesturing. The gestures were not rehearsed, but spontaneously produced by the speaker.

In post-processing, the speaker was centered, and the video format was set to 4:3 to accommodate this change. The videos were cut with ca. 600ms (30 frames) before gesture or speech onset and 600ms after offset. Speech on- and offsets were determined by inspecting the audio track, and gesture on- and offsets were defined by the hands leaving and returning to the resting position (hands hanging loosely down). This resulted in 3 videos with the same sentence uttered; to minimize pronunciation variation effects across these 3 sentences, all audio was rated in terms of clarity of speech and uniform speed by three German native speakers. The best-scoring audio was used for all 3 videos. Next, the speaker's face was blurred to, first, mask discrepancies between the new audio and mouth movements, and second, to eliminate the influence of facial expressions of the speaker on the listener. For the online questionnaires, the videos were resized to 640×480 pixels, thus decreasing loading times and ensuring smooth playback on participants' devices.

Building our Model

To build our model, we use the following values:

$\boldsymbol{Tel(v,n)}$ is a binary value, 1 or 0, indicating whether the instrument denoted by the noun n has the generic affordance denoted by the specific verb v or not, in other words, whether the telic component of the noun n is described by the specific verb v or not.

$\boldsymbol{GloVe(n,v)}$**,** ranging from -1 to 1, measures the semantic similarity between a noun n and a verb v, based on corpus linguistic co-occurrence statistics.

$\boldsymbol{Recog(g,v)}$ (Recognizability Rating), rated on a Likert scale from 1 to 7 in Experiment 1, reflects how well a verb v describes a shown (silent) gesture and thus how well the gesture g is recognized as the action denoted by the verb v.

$\boldsymbol{NRatingText(n,v)}$, rated on a Likert scale from 1 to 7 in Experiment 2, measures the likelihood of the noun n in the written target sentence following the context sentence (either General or Specific) containing the verb v.

$\boldsymbol{NRatingVideo(n,g(v))}$, rated in Experiment 3, measures the likelihood of the written sentences containing the noun n in the targets following a video presenting the General Sentence auditorily and containing the gesture $g(v)$, expressing the action denoted by the verb v.

Bayesian Pragmatics

Previous experiments have employed Bayesian pragmatics successfully to model context effects on word meaning (Werning et al., 2019; Werning & Cosentino, 2017). Here, the Gricean idea of communication resulting from reciprocal intentionality between speaker and listener is modelled as the relation between a subjective probability function P^L, describing the listener's probabilistic predictions, and an objective probability function P^S, describing the speaker's intentions and behaviors. P^L, in the case of rationality, is supposed to track P^S such that we can assume $P^L = P^S = P$. To capture the effect of a co-speech gestures g on the listener's predictive probability for a noun n, we apply Bayes' rule in the following way – with k being a normalizing factor:

$$(5)\ P_v\ (n|g) = k \cdot P_v\ (g|n) \cdot P_v\ (n)$$

By indexing the probability function P with the verb v, we acknowledge that both the gesture g and the noun n depend on the verb v. This is because (i) the probability of choosing the object denoted by n as an instrument depends on its appropriateness to afford the action denoted by v, and (ii) the probability of choosing the gesture g depends on how well it expresses the action denoted by v. We assume that the listener updates their prior $P_v\ (n)$, regarding the noun n, to the posterior $P_v\ (n|g)$, by taking into account the speaker's gesture g. Thereby, the listener considers the likelihood $P_v\ (g|n)$ for the speaker's choice of the gesture g, given the speaker's intention to communicate the use of the instrument denoted by the noun n (Werning et al., 2019). Applying the conditionalization rule to (1) we get:

$$(6)\ P(n|g,v) = k\ \cdot P(g|n,v) \cdot P(n|v)$$

The Model RTN estimates the listener's probabilistic prediction $P(n|g,v)$ with a monotonously increasing function f_1 of $NRatingVideo\big(n,g(v)\big)$:

$$(7a)\ P(n|g,v) = f_1\left(NRatingVideo\big(n,g(v)\big)\right)$$

Further, we assume that in order to correctly consider the speaker's intentions, the listener has to recognize the gesture g as an expression of the specific verb v, as is reflected in $Recog(g,v)$ and additionally the verb v has to be a telic component of the noun n, as is reflected in $Tel(v,n)$. Thus, $P(g|v,n))$ can be estimated with the product of the monotonously increasing functions f_2 and f_3 of $Recog(g(v),v)$ and, respectively, $Tel(v,n)$:

$$(7b)\ P(g|v,n)) = f_2\big(Recog(g(v),v)\big) \cdot f_3(Tel(v,n))$$

The prior $P(n|v)$ can be estimated with a monotonously increasing function f_4 of $NRatingText(n,v)$:

$$(7c)\ P(n|v) = f_4(NRatingText(n,v))$$

After inserting (3a)-(3c) into (2) and subsequent logarithmization and linear approximation of the logarithmized functions, we derive the following linear regression model RTN:

$$(8)\ NRatingVideo\big(n,g(v)\big) = \begin{array}{l} z + a \cdot Recog(g(v),v) \\ +b \cdot Tel(v,n) \\ +c \cdot NRatingText(n,v) \end{array}$$

The Model RTG is derived by the same estimations except for the last; here we estimate $P(n|v)$ with a monotonously increasing function f_5 of $GloVe(n|v)$:

$$(3c')\ P(n|v) = f_5(GloVe(n|v))$$

Analogously to (4), this yields (5):

$$(9)\ NRatingVideo\big(n,g(v)\big) = \begin{array}{l} z + a \cdot Recog(g,v) \\ +b \cdot Tel(v,n) \\ +c \cdot GloVe(n|v) \end{array}$$

Experiments

We chose 47 sets of context and target sentences for this series of experiments. We used General context, Specific context, and $Noun$ target, as well as the verbs v_{spec}.

Experiment 1 – $Recog(g(v),v)$

Participants 40 monolingually raised German native speakers were recruited via Prolific. One participant was excluded because they did not finish their questionnaire. Of the remaining 39 participants 19 identified as female, 19 as male, one as other. The mean age was 33,36 years (SD=13,60). The participants were equally distributed between the two material lists used.

Materials The videos of the source material were stripped of their audio. Only videos showing a gesture were used. These videos were combined with their matching specific target verb their mismatching specific target verb, and three filler items (randomly assigned specific verbs used as targets in other verb pairings). This resulted in 470 single trials that were split into two lists, to cut down on the time needed to finish the questionnaires. Every participant saw 235 trials.

Procedure Ratings were done using online questionnaires, via Qualtrics. The experiments were preceded by three practice items with the option to repeat them (one participant chose to do so). Participants had to rate on a 7-point scale (ranging from "very well" to "not well at all") whether the displayed verb described the gestures well.

Results Matching target verbs scored 5,22 points (SD=0,60), mismatching targets scored 2,38 points (SD=0,61), and fillers scored 1,90 points (SD=0,57). All comparisons were highly significant (All differences between the scored turned out to be highly significant ($p<.0001$). Notably, mismatches were rated as better descriptions than filler items; a possible explanation might be that for a match and mismatch pair the manipulated object stayed the same, putting both types of gestures into a similar gesturing space. Think of the cookies being cut out and brushed on: both gestures were performed in a similar space in front of the body, about the height a table or countertop would have. This makes the gestures spatially more similar than, e.g., a gesture representing hanging up a boxing bag. Overall, all match target verbs were rated as good descriptions of the shown gestures, indicating that participants perceived a meaning or semantic component in the gestures and matched them to the target verbs.

Experiment 2 – $NRatingText(n,v)$

Participants 30 German native speakers were recruited via Prolific, none of them had taken part in the previous studies.15 identified as female, 15 as male. The average age was 33,5 years (SD=10,60).

Materials No videos were used, only General context, Specific context, and $Noun$ target (see Table 3). This resulted in This resulted in 282 single trials per participant.

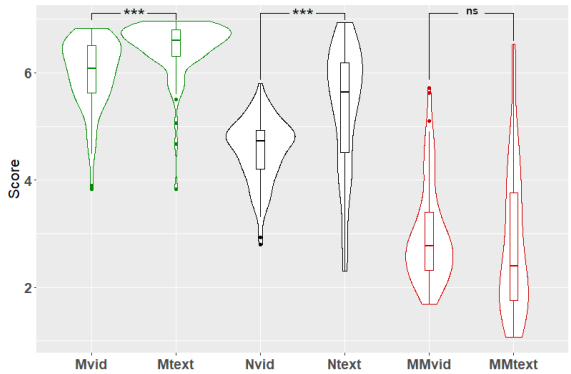

Figure 1: Violin and box plots of the scores; M= match, N= neutral, MM=mismatch; text=Exp2; vid= Exp3.

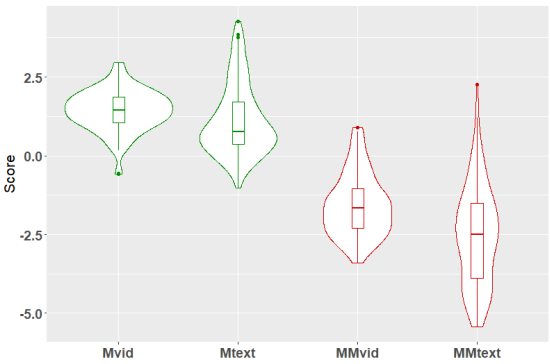

Figure 2: Violin and Box Plots of Normalized Scores; M= match, MM=mismatch; text=Exp2; vid= Exp3.

Table 2: Model comparison between Model RTN and Model RTG

Model	N	df	RSE	R^2	R^2_{adj}	BIC	ΔBIC	AIC	AICc
Model RTN	188	184	0,6776	0,8449	0,8424	409,31	0	393,12	393,45
Model RTG	188	184	0,779	0,795	0,7917	461,74	52,44	445,56	445,89

Model RTN			**Model RTG**		
	p	β		p	β
$\boldsymbol{NRatingText(n,v)}$	$1{,}86e^{-13}$	0,3869272	$\boldsymbol{GloVe(n\|v)}$	0,0855	0,7833378
$\boldsymbol{Tel(v,n)}$	$3{,}78e^{-13}$	1,5952457	$\boldsymbol{Tel(v,n)}$	$< 2e^{-16}$	2,9167161
$\boldsymbol{Recog(g(v),v)}$	0,00175	0,1275582	$\boldsymbol{Recog(g(v),v)}$	0,0132	0,1174883

Procedure. Ratings were done using online questionnaires, via Qualtrics. The experiments were preceded by three practice items with the option to repeat them (one participant chose to do so). Participants had to rate on a 7-point scale (ranging from "very likely" to "not likely at all") whether a sentence (Target *Noun*) was likely to follow after a given first sentence (either General context or Specific context).
Results Matching target verbs scored a mean rating of 6,48 points (SD= 0,50), mismatching ones 2,82 points (SD= 1,35), and neutral targets scored 5,36 points (SD= 1,16). This puts the matches in between the area of "very likely" and "likely", the mismatches in "rather unlikely", and the neutrals just between "rather likely" and "likely". All differences between the scored turned out to be highly significant (p<.0001).

Experiment 3 – $\boldsymbol{NRatingVideo(n,g(v))}$

Participants 35 monolingually raised German native speakers were recruited via Prolific, none of them had taken part in the previous study. Five participants were excluded because they did not finish their questionnaire. Of the remaining participants, 10 participants identified as female, 20 as male. The average age was 32,27 years (SD=13,28).
Materials We used all three video variations of the source material, each was combined with the two corresponding *Noun* targets (see Table 3), resulting in 282 single trials.
Procedure Ratings were done using online questionnaires, via Qualtrics. The experiments were preceded by three practice items with the option to repeat them (one participant chose to do so). Participants were shown a video of our speaker uttering the unspecific context sentence, either performing a matching, mismatching or a neutral gesture. Participants had to rate on a 7-point scale (ranging from "very likely" to "not likely at all") how likely they thought a displayed sentence (Target *Noun*) would follow the video.
Results Matching target verbs scored a mean rating of 5,99 points (SD= 0,65), mismatching ones 2,98 points (SD= 0,92), and neutral targets scored 4,56 points (SD= 0,58). This puts the matches in the area of "likely", the mismatches in "rather unlikely", and the neutrals just between "rather likely" and "neither likely nor unlikely". All differences between the scored turned out to be highly significant (p<.0001).

Discussion: Behavioral Experimental Results

Since we consider iconic co-speech gestures to be useful tools when it comes to disambiguating spoken utterances, we expected the matches' scores to be higher in Exp. 3 ($NRatingVideo$) than in Exp. 2 ($NRatingText(n,v)$), and the scores of mismatches in Exp. 3 to be lower than Exp. 2.

The underlying assumption is that gestures, even though more ambiguous in nature when not combined with speech, increase specificity of the contexts, especially when the stand-alone verbs permit multiple tools; and thus, participants can match the corresponding tools better, leading to a higher acceptance of matching nouns and a higher rejection of the mismatching nouns. However, it seems at first glance, that the effects were reversed for matches and negligible for mismatches (see Figure 1): The differences between the match conditions of Exp. 2 and 3 is only -0.489 points however this difference is highly significant (df = 174.46, p<.0001, Cohen's d= 0.843) as well as the difference between the neutral conditions, which is -0.80 points (df = 137, p<.0001, Cohen's d = 0.87). The difference between the mismatch conditions is 0,16 points and did not turn out to be significant (df = 164.01, p = .346, Cohen's d = 0.138).

Looking at the violin plots in Figure 1, the vastly different distribution of scores is revealed. The highly significant difference between the means of the neutral conditions is especially striking since one would assume that these should not differ: the only difference in the experimental set-up was that the context sentence was presented auditorily in a video in Experiment 3 but displayed to be read in Experiment 2. There are two possible explanations: (1) video material is simply processed differently than purely written input, e.g., videos might be engaging additional cognitive processes that

Table 3: In Exp. 2 & 3, for each condition (M, MM, N), contexts (v^1_{spec} 🟢, v^2_{spec} 🟡) are combined with their targets (🟢,🟡)

	Context Exp. 2 $\boldsymbol{NRatingText(n,v)}$ "The agent $\boldsymbol{x}$ is… object $\boldsymbol{y}$."		**Context Exp 3.** $\boldsymbol{NRatingVideo(n,g(v))}$ "The agent $\boldsymbol{x}$ is… object $\boldsymbol{y}$."		**Target (Exp. 2 & 3)** "To do so, they ($\boldsymbol{x}$) are using…"	
Match (M)	v^1_{spec}-ing	v^2_{spec}-ing	v_{gen}-ing + g(v^1_{spec})	v_{gen}-ing + g(v^2_{spec})	n^1	n^2
Mismatch (MM)	v^1_{spec}-ing	v^2_{spec}-ing	v_{gen}-ing + g(v^1_{spec})	v_{gen}-ing + g(v^2_{spec})	n^2	n^1
Neutral (M)	v_{gen}-ing	v_{gen}-ing	v_{gen}-ing	v_{gen}-ing	n^1	n^2

interfere with the sentence meaning comprehension; we can assume that the processes in our experimental set ups are closely related to gesture comprehension, and this additional costs would not only explain the different values in the neutral conditions but also for the higher ratings of matches in Exp. 2; (2) the video material added a clue for the listener, namely a contrast condition in the form of the presence or absence of the gesture; every time a gesture was absent, the listener could infer that now less specific information was given, and thus they rated the following nouns n closer to the neutral "neither likely nor unlikely".

Even though this would not lead to a perfect normalization between the two experiments, we subtracted the neutral scores from the corresponding matches and mismatches (see Figure 2) to derive a rough approximation of how a normalized comparison could look. Here, we observe the trend that we predicted for the matches, but not for the mismatches. The higher rating of the mismatches is even more pronounced in this case; the participants might have considered the mismatching gesture as additional information and relied on their general knowledge of which instruments can be used to manipulate the object in the context sentence, regardless of the action represented by the gesture g, to judge the likelihood of the instrument noun n.

These results show that using a control condition in video materials that is supposedly behaving "just like spoken/written language" can get us into hot water and are not easily disentangled. For a fair comparison between the scores in Exp. 2 and 3, we would need a transformation that does not only bring the means of the controls to the same level but that additionally changes the form, i.e., the density distribution represented by the form of the violin plots. As of now, we can only conclude for this issue that the absence of a gesture apparently does not mean that a listener only attends to the linguistic input in that case.

On the bright side, the difference between match and mismatch conditions within Experiments 2 and 3 were highly significant. This means that listeners did not only consider the General context in Exp. 3 but did pay attention to the shown gesture g and even were able to recognize it as an expression of an action denoted by verb v_{spec}– or reject it as such in a mismatch condition. However, the processing of the combined General context and the gesture g must be different from the processing of the Specific context. To disentangle whether this was due to the mode of presentation (auditorily in Exp. 3, written in Exp. 2), we would have to conduct another experiment introducing video material with no gesture and the speaker uttering the Specific context.

Discussion: Model Fit

As can be seen in Table 2, Model RTN offers the best fit for our data: it has the lower BIC, AIC, and AICc values, as well as a higher R^2 value than Model RTG. Additionally, in the Model RTN $NRatingText(n,v)$ (β=0.39, p<.0001), $Tel(v,n)$ (β=1.60, p<.0001), and $Recog(g(v),v)$ (β=0.13, p<.005) were all highly significant predictors, whereas in Model RTG only $Tel(v,n)$ (β=2.92, p<.0001) is a highly

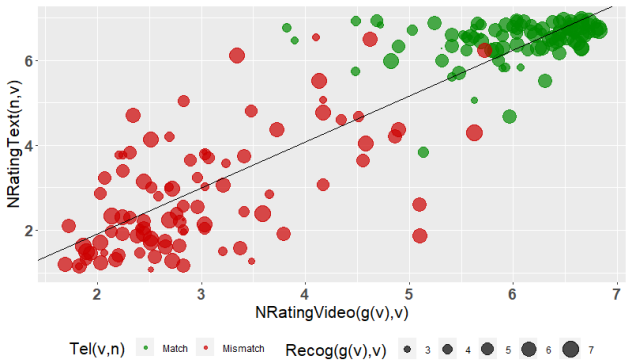

Figure 3: Bubble plot of Model RTN. Scored ratings of $NRatingVideo(n,g(v))$, $NRatingText(n,v)$; colour split for match (green) and mismatch (red) conditions, i.e., weighted for $Recog(g(v),v)$.

significant predictor, while $Recog(g(v),v)$ (β=0.12, p=.0132) and $GloVe(n|v)$(β=0.78, p=.0855) are not statistically significant predictors. To test whether the differences between Model RTN and Model RTG were significant, we calculated the z statistic (Clogg et al., 1995), and found the difference to be highly significant (p<.0001). It is especially noteworthy that $GloVe(n|v)$ does not turn out to be a significant predictor in Model RTG. Even though the material's matches and mismatches were constructed considering their GloVe values to each other (matching verb-noun pairs having higher GloVe values towards each other than mismatching pairs), $GloVe(n|v)$ is not adding any advantage in estimating $NRatingVideo(n,g(v))$. This could be due to the nature of GloVe, which only considers co-occurrences as a measure of semantic similarity, and while nouns denoting tools (e.g., *knife*) often co-occur with verbs denoting their telic components (*cut*), they also often, as objects, co-occur with verbs (*sharpen*) that express the telic component of other instruments (*grindstone*). Since the nouns were clearly marked as instruments, participants in Exp. 3 did not consider them as manipulated objects, which is a distinction not reflected by GloVe. This distinction, however, is reflected in $NRatingText(n,v)$. Thus, it is no surprise that Model RTN outperforms Model RTG.

Conclusion

We found that an iconic co-speech gesture makes a difference for a listener's probabilistic prediction regarding an upcoming instrument noun, and thus has a semantic effect on linguistic comprehension. Even though that semantic effect of the gesture differs significantly from the one of the corresponding action verb, the gesture's semantic effect strongly correlates with the one of the corresponding action verb, especially when one takes into account how well the expressed action is recognized in the gesture. A model (RTN) that includes as a predictor how well the listener understands the action (denoted by the verb) as afforded by the instrument (denoted by the noun) better predicts our data than a model (RTG) that instead includes as a predictor the co-occurrence statistics regarding the verb and the noun (GloVe).

References

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, *100*(5), 1261–1293.

Cosentino, E., Baggio, G., Kontinen, J., & Werning, M. (2017). The time-course of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in Psychology*, *8*(818). https://doi.org/10.3389/fpsyg.2017.00813

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, *2*, 548–553.

Frank, Michael. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998. https://doi.org/10.1126/science.1218633

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, *1*(2), 67–82.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gibson, J. J. (1986). Gibson Theory of Affordances. In *Chapter Eight The Theory of Affordances* (S. 127–136).

Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on psychological science*, *5*(6), 664–674.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2007.19.7.1175

Huettig, F. (2015). Four central questions about prediction in language processing. In *Brain Research* (Bd. 1626, S. 118–135). https://doi.org/10.1016/j.brainres.2015.02.014

Kandana Arachchige, K. G., Simoes Loureiro, I., Blekic, W., Rossignol, M., & Lefebvre, L. (2021). The Role of Iconic Gestures in Speech Comprehension: An Overview of Various Methodologies. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.634074

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. In *Proceedings of the National Academy of Sciences of the United States of America* (Nummer 7, S. 1–6). https://doi.org/10.1073/pnas.1407479111

Krauss, R. M., Chen, Y., & Gotfexnum, R. F. (2000). 13 Lexical gestures and lexical access: A process model. *Language and gesture*, *2*, 261.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positiveform adjectives *. *Proceedings of SALT*, *23*, 587610.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, *136*, 123–134. https://doi.org/10.1016/j.cognition.2014.10.016

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. The University of Chicago Press.

Pouw, W. T. J. L., de Nooijer, J. A., van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology*, *5*(APR), 1–14. https://doi.org/10.3389/fpsyg.2014.00359

Pustejovsky, J. (1998). *The generative lexicon*. MIT press.

Qing, C., & Franke, M. (2014). Gradable adjectives , vagueness , and optimal language use : *Proceedings of SALT*, *24*, 23–41.

Werning, M. (2010). Complex first? On the evolutionary and developmental priority of semantically thick words. *Philosophy of Science*, *77*(5), 1096–1108. https://doi.org/10.1086/656826

Werning, M., & Cosentino, E. (2017). The interaction of Bayesian pragmatics and lexical semantics in linguistic interpretation: Using event-related potentials to investigate hearers' probabilistic predictions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Hrsg.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (S. 3504–3509). Cognitive Science Society.

Werning, M., Unterhuber, M., & Wiedemann, G. (2019). Bayesian Pragmatics Provides the Best Quantitative Model of Context Effects on Word Meaning in EEG and Cloze Data. In A. Goel, C. Seifert, & C. Freska (Hrsg.), *Proceedings of the 41th Annual Conference of the Cognitive Science Society* (S. 3085–3091). Cognitive Science Society.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number GRK-2185/1 (DFG Research Training Group Situated Cognition), and 367110651 (PI: MW), within the Priority Program XPrag.de (SPP1727).