Tattoos as a window onto cross-linguistic differences in scalar implicature

Danielle Dionne & Elizabeth Coppock*

Abstract. This paper addresses the question of how to predict which alternatives are active in scalar implicature calculation, and the nature of this activation. It has been observed that *finger* implicates 'not thumb', and a Manner-based explanation for this has been proposed, predicting that if English had the simplex Latin word *pollex* meaning 'thumb or big toe', then *finger* would cease to have the implicature 'not thumb' that it has. It has also been suggested that this hypothetical *pollex* would have to be sufficiently colloquial in order to figure in scalar implicature calculation. This paper makes this thought experiment into a real one by using a language that behaves in exactly this way: Spanish has *pulgar* 'thumb' (< *pollex*), a non-colloquial form. We first use a fill-in-the-blank production task with both English and Spanish speakers to gauge the likelihood with which a speaker will produce a given form as a way of describing a given digit. Production frequency does not perfectly track complexity, so we can then ask whether comprehension follows production frequency or complexity. We do so using a forced choice comprehension task, which reveals cross-linguistic differences in comprehension tracking production probabilities. A comparison between two RSA models – one in which the speaker perfectly replicates our production data and a standard one in which the speaker chooses based on a standard cost/accuracy tradeoff – illustrates the fact that comprehension is much more closely tied to production probability than to the mere existence of sufficiently simple alternatives.

Keywords. scalar implicature; manner implicature; hyponymy, cross-linguistic differences; RSA; computational modelling

1. Introduction. Suppose you heard the following sentence:

(1) She has a tattoo on her finger.

Would you think the tattoo was on the thumb or the ring finger? If you are like most of the participants in our English comprehension study, you will think the ring finger is more likely. The thumb is generally considered a type of finger; people generally agree that we have 10 fingers. So it is arguably not the semantics of *finger* that determines this preference; rather, there is a scalar implicature from *finger* to 'not thumb'. In Gricean terms, the pragmatic reasoning might run as follows, 'Why didn't she choose thumb? It would have been equally short (Manner), more informative (Quantity), and just as relevant (Relevance). Maybe she didn't believe it (Quality).'

Horn (2000) observes that the relationship between *thumb* & *finger* is not parallel to the relationship between *big toe* & *toe*:

(2) a. I hurt my finger. $\rightsquigarrow$ I did not hurt my thumb.
b. I hurt my toe. $\not\rightsquigarrow$ I did not hurt my big toe.

*We are grateful to the audiences at LSA 2020 and ELM 2021 for feedback on this work. Authors: Danielle Dionne, Boston University (ddionne@bu.edu) & Elizabeth Coppock, Boston University (ecoppock@bu.edu).

Horn (2000) concludes that although *thumb* acts as an alternative to *finger* for the purposes of scalar implicature, *big toe* does not act as an alternative for *toe* (p. 308). He explains this in terms of Manner: *big toe* is longer than *toe*, and therefore not a good alternative. Horn (2000, p. 308) writes: "We would predict that if the colloquial language replaced its *thumb* with the polymorphous *pollex* (the Latin and scientific English term for both 'thumb' and 'big toe'), the asymmetry [between *finger* and *toe*] would instantly vanish".

Geurts (2011) zeroes in on Horn's strategic use of the term "colloquial", writing: "It is important to note, however, that the adjective 'colloquial' is doing real work in this statement. It is not enough for an alternative word to be in the language; it has to be sufficiently salient, as well: if the word 'thumb' was rarely used, then presumably the asymmetry between [finger and toe] would vanish too" (p. 122). That is, the prediction is really that if a stronger utterance is present in the language *and* it is sufficiently salient, a scalar implicature will arise when the weaker form is used.

As it turns out, there is a language in which exactly that situation arises: Spanish. Spanish contains a word for 'thumb', namely *pulgar*—the Spanish descendant of Latin *pollex*—but it is less frequently used, and less colloquial. As we will confirm in production studies, there is a great deal of variation in how the thumb is referred to in Spanish. *Pulgar* does not differ from *thumb* in complexity, but it does differ in how prevalent it is in the language, and how salient it is to the speaker as an alternative. Furthermore, *pulgar* 'thumb' is not specific to the hand, just like Latin *pollex*. If Geurts (2011) is right, the asymmetry between *finger* and *toe* that exists in English is predicted to be absent in Spanish, and there should be no implicature from *dedo* 'finger' to 'not thumb' or from *dedo del pie* 'toe' to 'not big toe'.

The four studies reported here test these predictions. We first conducted production studies each in English and Spanish to gauge the salience of available alternatives. We use a fill-in-the-blank production task with both English and Spanish speakers to gauge the likelihood with which a speaker will produce a given form as a way of describing a given digit. We find that production frequency does not perfectly track complexity, so we can then ask whether comprehension follows production frequency or complexity. We do so using a forced choice comprehension task, which reveals cross-linguistic differences in comprehension that tracks production probabilities. Finally, we carry out a comparison between two RSA models, one in which the speaker perfectly replicates our production data and a standard one in which the speaker chooses based on a standard cost/accuracy trade-off. Comparing both of these models to our comprehension data leads us to conclude that the activation of alternatives for the purpose of scalar implicature calculation is much more closely tied to production probability than to the mere existence of sufficiently simple alternatives.

2. Production studies. We conducted two production studies: one in English and one in Spanish. Both tasks contained images of body parts with tattoos (see Figure 1).

2.1. PARTICIPANTS. All participants were recruited on Prolific. All studies involved different groups of participants. In the English production study, all participants were self-reported monolingual native English speakers who were born and currently live in the United States. In the Spanish production study, all participants were self-reported monolingual native Spanish speakers who were born and currently live in Mexico. There were 24 American English speakers and 23 Mexican Spanish speakers in the production studies.

Figure 1: Stimulus Items for Production and Comprehension tasks

2.2. MATERIALS. Participants completed a task in which they were asked to look at a series of pictures. All of the pictures were body parts with a tattoo on them. The tattoos served as an indicator of which digit or body part the speaker was talking about. The target items showed photos of a tattoo on the thumb, ring finger, pinky, big toe, fourth toe and pinky toe. There were six filler items (all different from each other): two photos of tattoos on a leg, two photos of tattoos on an arm, and two photos of tattoos on the back.

2.3. PROCEDURE. Participants were shown a series of images, one by one. With each image, they were asked to fill in the blank of the sentence: *She has a tattoo on* _____, or its translational equivalent, in the case of Spanish. The order of images was randomized. All participants were presented with all six target items and all six filler items.

2.4. NORMALIZING PRODUCTION RESULTS. After the data collection process was completed, all responses for the production study were normalized by hand. This included removing additional words such as "left" or "right" (e.g. "right pinky" became "pinky"). Directional terms ("left"), initial articles, and other non-essential words were stripped away so that all that was remaining was the word or phrase that was used to refer to the digit itself. This removed excess noise from the data and allowed us to group responses together that were essentially identical in form – *dedo de la mano* ('digit of the hand' - *finger*) vs. *dedo* ('digit'), for example. Additionally, responses were coded for specificity — 1 for specific words/phrases that could refer to only one digit (e.g. "thumb" or "pulgar") and 0 for non-specific words/phrases that could refer to more than one digit (e.g. "finger" or "dedo de la mano").

2.5. RESULTS. For the thumb image, 100% of English speakers responded with *thumb* — a specific term. In contrast, Spanish speakers were not unanimous in their responses. Figure 2 presents the production results for digits on the hand. While the single-word translational equivalent to 'thumb', *pulgar*, was preferred, only approximately 42% of participants used it. *Mano* – Spanish for 'hand' – was the second most frequent (17.4%). Spanish speakers preferred using specific terms 63.2% of the time.

For the ring finger images (presented in Figure 2), English speakers preferred the specific term *ring finger* 83% of the time, but some participants (17%) did produce the general term *finger*. Again, Spanish speakers presented more variation in their responses than English participants, with seven unique utterances produced. 35% of Spanish participants produced *dedo anular* ('ring finger'); 26% produced *dedo* ('finger'). Overall, Spanish participants trended like English participants with preference for specific (52.2%) over general terms (47.8%), although the preference was not as strong (see Figure 2).

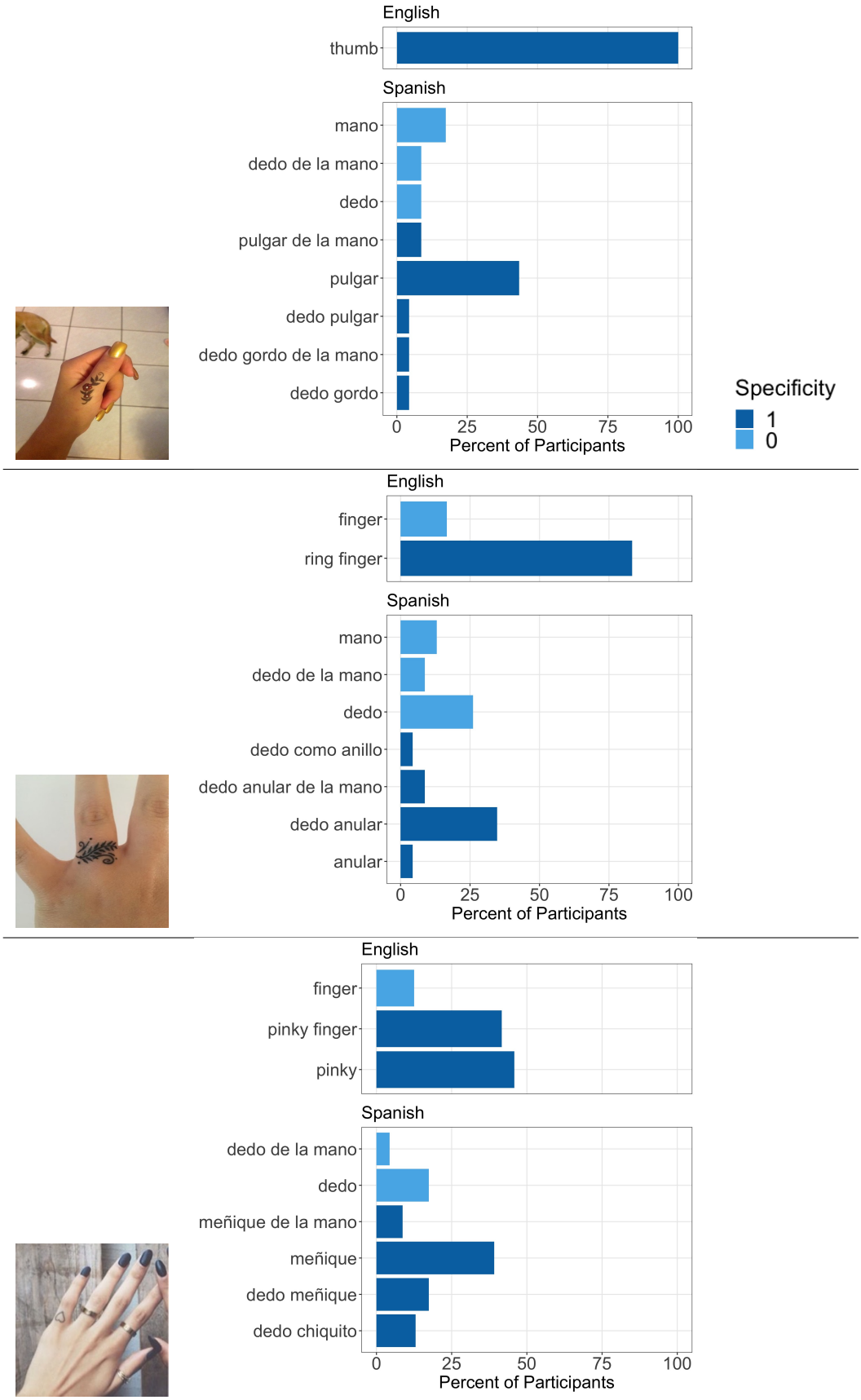

Figure 2: English and Spanish production of finger terms;"specific" terms refer to a single digit

English production for the pinky was similar to English production for the ring finger (see Figure 2): there is variation between specific (e.g. *pinky* – 45.8%, or *pinky finger* – 41.7%) and general term usage (12.5% *finger*). The single word *pinky* is used (and most frequently), but almost as many participants chose to use the two-word alternative *pinky finger*. Spanish speakers also gave varying responses. The most common was the single-word equivalent for 'pinky', *meñique* (39.1%), vs. 21.7% for *dedo* 'finger'.

For the big toe images, English speakers favored the specific term *big toe* (83.3%) over the general term *toe* (16.7%). In contrast, Spanish participants produced six different descriptions (shown in Figure 3). The most common response (69.6%) was the specific term *dedo gordo del pie* (lit. 'fat digit of the foot'), or 'big toe'. The second most common response (34.8%) was the general term *dedo del pie* (lit. 'digit of the foot'), which translates to *toe*.

For the ring toe image, English participants showed increased dispersion in their responses. The majority of participants (58.3%) used the general term "toe". The remaining participants (41.7%) produced various specific terms for the digit ("fourth toe" and "ring toe" to name a few). Spanish speakers had a much higher rate of general term usage, with 82.6% of participants preferring terms like *dedo del pie* 'toe', *dedo* 'digit', or *pie* 'foot'.

Finally, the production results for the pinky toe were similar between English and Spanish. English speakers preferred using the general term *toe* far less than a specific term (20.8% and 79.2%, respectively). There was less dispersion in English production results for the pinky toe than for the ring toe. In Spanish, participants generally preferred a specific term (52.2%) over a general term (47.8%), but the trend was not as strong as in English. In contrast to English, Spanish production data exhibited a much larger amount of dispersion for the pinky toe than the ring toe, as shown in Figure 3.

These production results support Spanish speakers' intuitions that the Spanish single-word alternative *pulgar* 'thumb' is less prevalent than *thumb* is in English. If Geurts (2011) is right, then Spanish speakers and English speakers should differ in scalar implicature calculation due to the differences in prevalence of the alternative forms for 'thumb'.

3. Comprehension Studies. We are now in a position to address our main research question: Upon hearing a general term for a digit (e.g. *finger* or *toe*), what alternatives do English and Spanish speakers use to compute alternatives? Are alternatives activated in accordance with their salience (as measured by production frequency in our production experiments) or in accordance with their complexity (as measured by number of words)? Although these two things are correlated, they are not identical. For example, if Geurts (2011) is right, and the salience of alternatives matters, then English and Spanish will differ with respect to the scalar implicature associated with *finger* due to the difference in production probability for the more specific forms (*thumb* vs. *pulgar*). If complexity is all that matters, then there should be an implicature in both languages, because there is an equally simpler, yet more informative alternative in both languages. Through our comprehension studies, we are able to distinguish among these hypotheses.

3.1. PARTICIPANTS. 45 American English participants and 48 Mexican Spanish participants, recruited via Prolific, completed the comprehension task. English participants were self-reporting American monolinguals that were born and currently reside in the United States. Spanish participants were self-reporting Mexican monolinguals that were born and currently reside in Mexico.

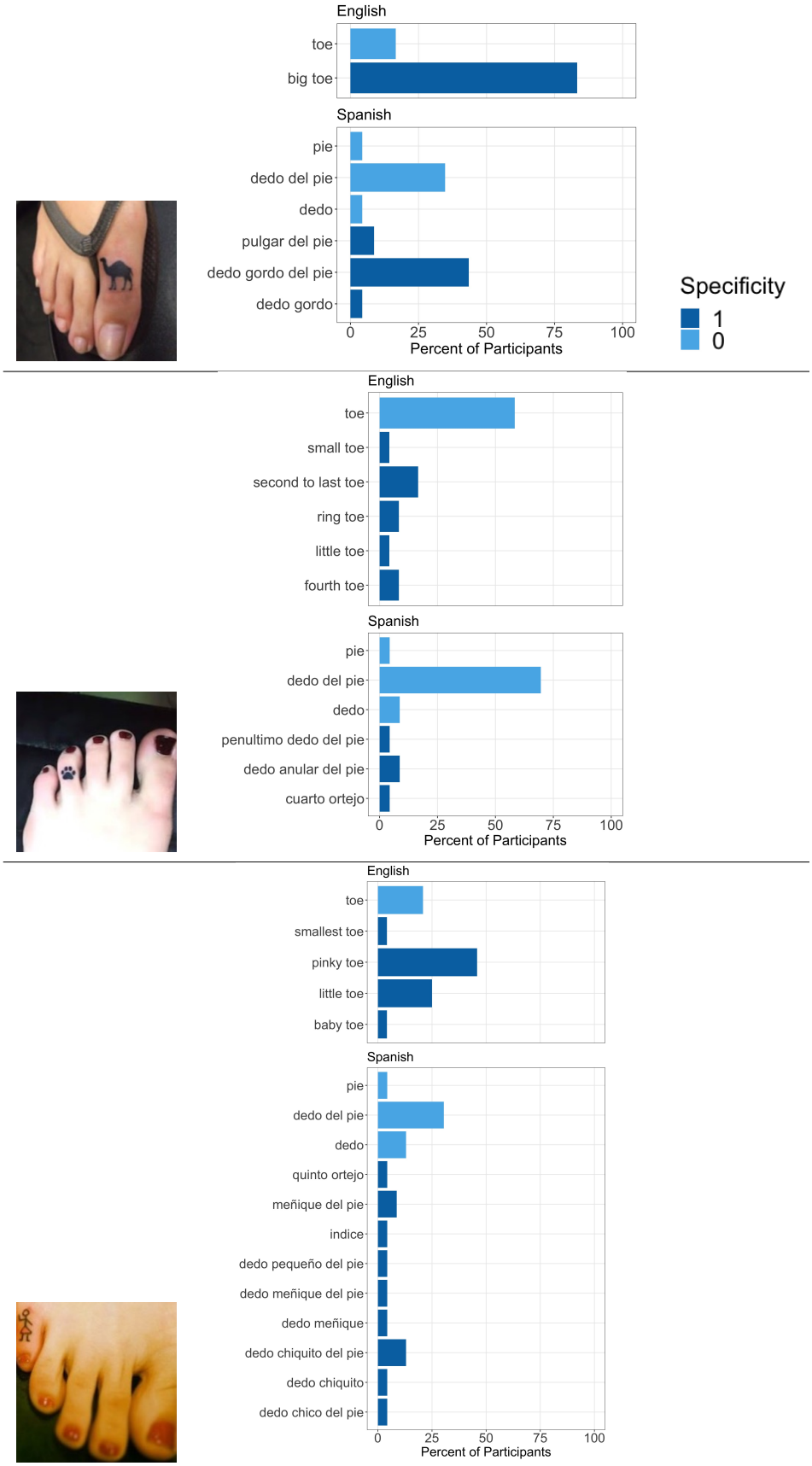

Figure 3: English and Spanish production of toe terms; "specific" terms refer to a single digit

3.2. MATERIALS. The target items for the comprehension studies consisted of 6 image pairs. Three of the pairs were images of hands and three of the pairs were images of feet such that all possible hand combinations and all possible foot combinations were presented. No target pairs consisted of an image of a digit on the hand and an image of a digit on the foot. The images were the same six images from the production study (see Figure 1).

In addition to the 6 target image pairs, participants were also presented with 6 filler image pairs. Three of the filler pairs were "easy", where the utterance clearly matched only one of the images (e.g. *She has a tattoo on her back*, with a pair of images that contained only one back tattoo). The other three filler pairs were considered "hard"; these image pairs contained, for example, two different back tattoos. Filler pairs that were "easy" acted as attention checks, since there was a clear correct response. Participants who failed one or more "easy" fillers were eliminated from the results.

3.3. PROCEDURE. On each trial, a pair of images was presented, both showing a tattoo on a body part. On critical trials, the images showed tattoos on two different fingers, or two different toes: thumb on the left, ring finger on the right, for example. Along with the images, participants read an utterance of the form *She has a tattoo on her X*, where *X* was a general term: *finger* or *toe* or the Spanish translational equivalent (*dedo* or *dedo del pie*). Participants were asked "Which picture are they talking about?" and clicked on an image. Item order and left-right presentation of the images were randomized.

3.4. RESULTS. Responses were simply coded as the image the participant clicked on (e.g. "thumb" for the image with the tattoo on the thumb). The p-values we report are the result of conducting a 1-sample proportion test, where the null hypothesis, or the probability of choosing the correct image is 0.5. The assumption is that the data follow a Bernoulli distribution. We ran a Benjamini-Hochberg adjustment on the p-values.

In the comprehension study, when participants were asked to choose between the thumb image and the ring finger image given the statement "She has a tattoo on her finger", 75% of English participants chose the image of the ring finger, $p = 0.004$ (see Figure 4). In contrast, just over half of the Spanish speakers chose the ring finger image over the thumb image, but the error bar, which depicts a 95% Confidence Interval, distinctly crosses the 50% mark, showing that the Spanish participants' responses are not statistically significantly different from chance ($p = 0.627$).

For the big toe and ring toe image pair, English participants showed a slight preference for the ring toe image, with roughly 63% of participants choosing that image. However, the error bar indicates that this result is not statistically different from chance ($p = 0.145$). Spanish participants actually showed a stronger trend toward the ring toe given the translational equivalent of "She has a tattoo on her toe" ($p = 0.007$).

A full summary of the results is given in Table 1. The estimate is the estimated true proportion in the greater population. We conducted a Benjamini-Hochberg adjustment to obtain the adjusted p-values. What we can take away from these results is that the following implicatures exist in English:

- *finger* $\rightsquigarrow$ not thumb
- *toe* $\rightsquigarrow$ not pinky toe

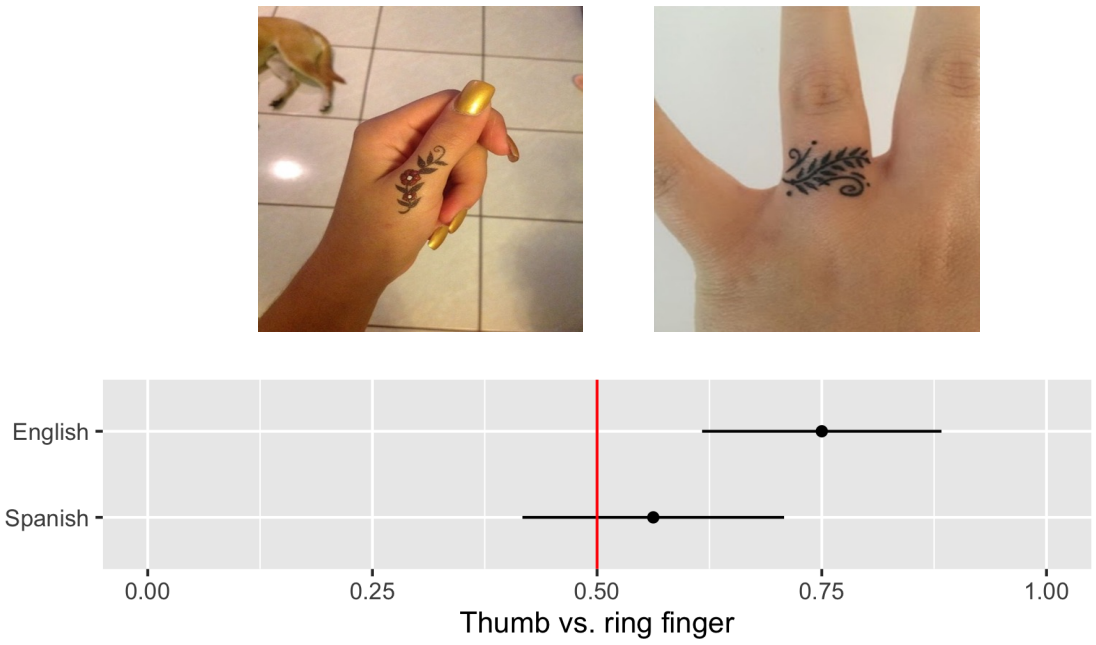

Figure 4: Observed frequency with 95% CI of choosing Thumb or Ring Finger in English and Spanish.

In Spanish, we have:

- *dedo del pie* $\rightsquigarrow$ not pinky toe
- *dedo del pie* $\rightsquigarrow$ not ring toe

Not all of these implicatures are predicted by the assumption that complexity alone is what drives the activation of alternatives. In the next section, we develop precise computational models in order to understand the significance of these results more deeply.

4. Bayesian modeling.

4.1. MODEL DEFINITIONS. To gain a better understanding of complexity and prevalence, and their roles in scalar implicature, we compared two Rational Speech Act models (Frank & Goodman 2012, Goodman & Stuhlmüller 2013; i.a.) that differ in how the speaker is defined. The first model incorporates a traditional speaker model that penalizes longer – more complex – utterances (henceforth referred to as the Complexity model). The second model is a prevalence-based speaker model that has perfect knowledge of speaker production (henceforth referred to as the Production model). For both models, the space of possible states includes six underlying states, corresponding to the six target digits (*thumb, ring finger, pinky, big toe, ring toe,* and *pinky toe*). Literal meanings for each utterance from the production study were hand-specified as a subset of the states.

For the complexity-based speaker model, as presented earlier, the Speaker chooses an utterance based on accuracy and cost. Length is equivalent to length in words, and $L_0(s \mid u)$ is the probability that a literal listener will choose a state s given an utterance u. The model contains

	Condition	Language	Estimate	p-value	adj. p-value
1	Big toe vs. ring toe	Eng	0.64	0.097	0.145
2	Big toe vs. ring toe	Spa	0.72	0.002	0.007*
3	Big toe vs. pinky toe	Eng	0.34	0.05	0.10
4	Big toe vs. pinky toe	Spa	0.50	1.00	1.00
5	Ring toe vs. pinky toe	Eng	0.24	0.0006	0.002*
6	Ring toe vs. pinky toe	Spa	0.21	0.00009	0.0009*
7	Thumb vs. ring finger	Eng	0.75	0.002	0.004*
8	Thumb vs. ring finger	Spa	0.56	0.47	0.627
9	Thumb vs. pinky finger	Eng	0.80	0.0001	0.0009*
10	Thumb vs. pinky finger	Spa	0.63	0.086	0.145
11	Ring finger vs. pinky finger	Eng	0.45	0.651	0.781
12	Ring finger vs. pinky finger	Spa	0.52	0.885	0.965

Table 1: p-values and adjusted p-values for each language/condition pair.

two free parameters. Alpha (α) is the 'rationality parameter', which corresponds to how much the speaker maximizes utility, where utility in this context corresponds to accuracy, that is, probability that the literal listener selects the correct referent. The parameter β is a multiplier on cost, where cost is measured as number of words in the utterance. The cost parameter reflects speakers' degree of preference to be as concise as possible when speaking. Model parameters for the Complexity model were tuned to the thumb/ring finger data point from the experimental results of the English comprehension study: α set at 1 and β set at 2.

$$S(u \mid s) \propto \mathsf{exp}(\alpha \cdot L_0(s \mid u) - \beta \cdot \mathsf{length}(u))$$

A pragmatic listener was then built on top of the Complexity speaker model. As usual in RSA, a pragmatic listener chooses an interpretation using Bayes' Rule, reasoning about the likelihood that a speaker would choose various utterances under various hypotheses about what the speaker intends.

$$L(s \mid u) \propto S(u \mid s) \cdot P(s)$$

In contrast to the Complexity model, the Production model is fed the exact production probabilities for each utterance collected from the production study. The speaker in this model chooses an utterance based on the empirically observed frequencies in my production data. We write $F(u \mid s)$ to denote the frequency with which an utterance *u* was used in the production experiments to describe state *s* (i.e. the finger or toe that had the tattoo).

$$S(u \mid s) \propto F(u \mid s)$$

This ensures that the Production model has full awareness of what utterances are more or less prevalent for speakers – these are utterances speakers actually produced. As in the Complexity model, a pragmatic listener is coded on top of the Production model. In fact, the Pragmatic Listener model is the same for both Speaker models. Because of the difference in the way that the speaker

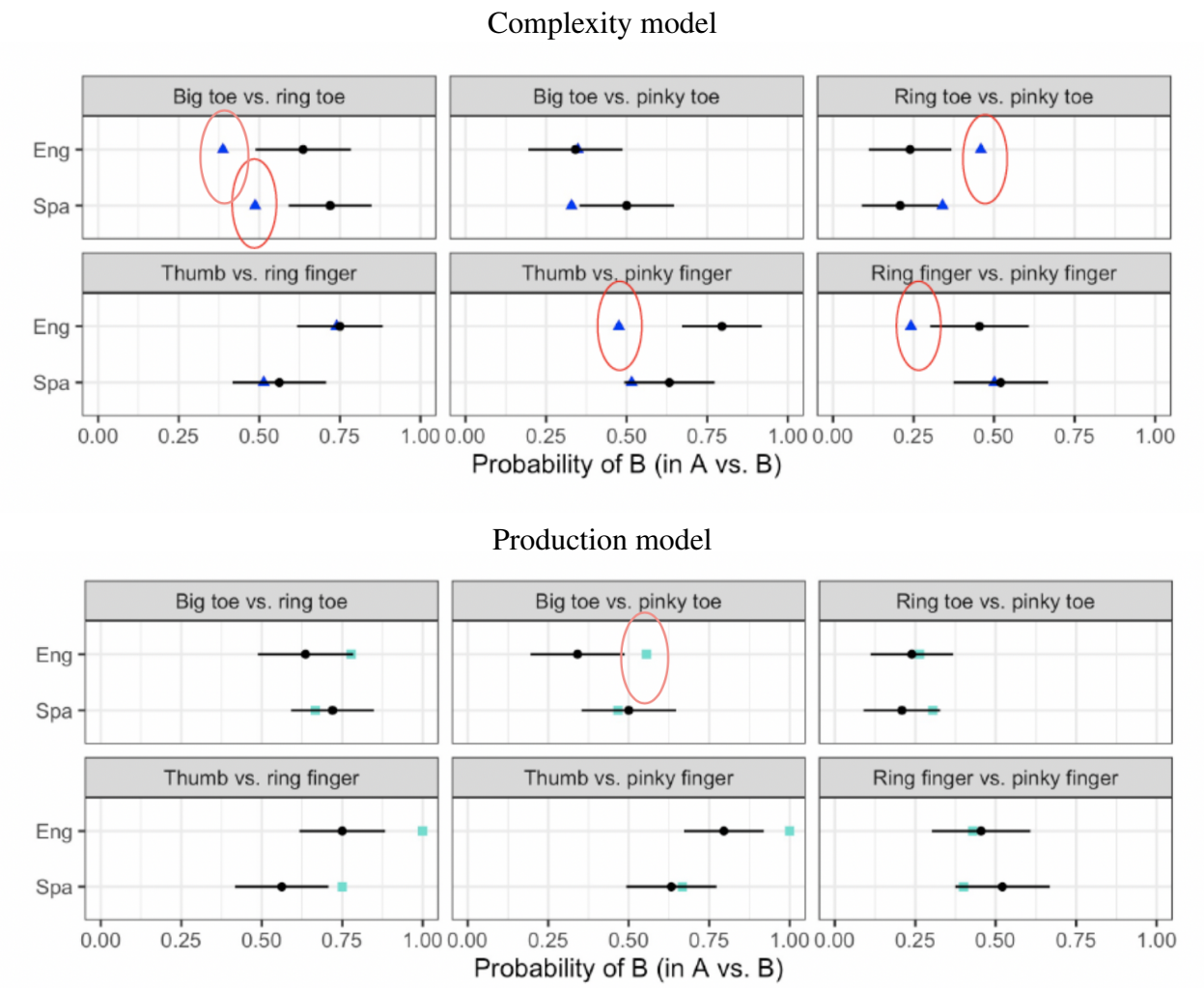

Figure 5: Model predictions plotted against comprehension results; Inaccurate model predictions are circled in red.

is defined, however, the Pragmatic Listener in the Production model has full awareness of actual production probabilities.

4.2. MODEL PERFORMANCE. The model predictions are presented alongside the empirical results in Figure 5. The Complexity model inaccurately predicts no implicature for the thumb/pinky item in English. This is because the model is only considering the fact that these two utterances are equally complex. Additionally, the model incorrectly predicts that there will be an implicature for ring finger/pinky in English since the one-word term *pinky* is an available alternative.

For digits on the feet, the Complexity model incorrectly predicts no implicature for ring toe/pinky toe in English and big toe/ring toe in Spanish – since they are equally complex. However, the production results suggest that they are not equally viable as alternatives. Since *ring toe* and *pinky toe* are equally as complex, if complexity alone determined which alternatives were available to speakers, we would expect no implicature to arise. However, the presence of the implicature *toe*

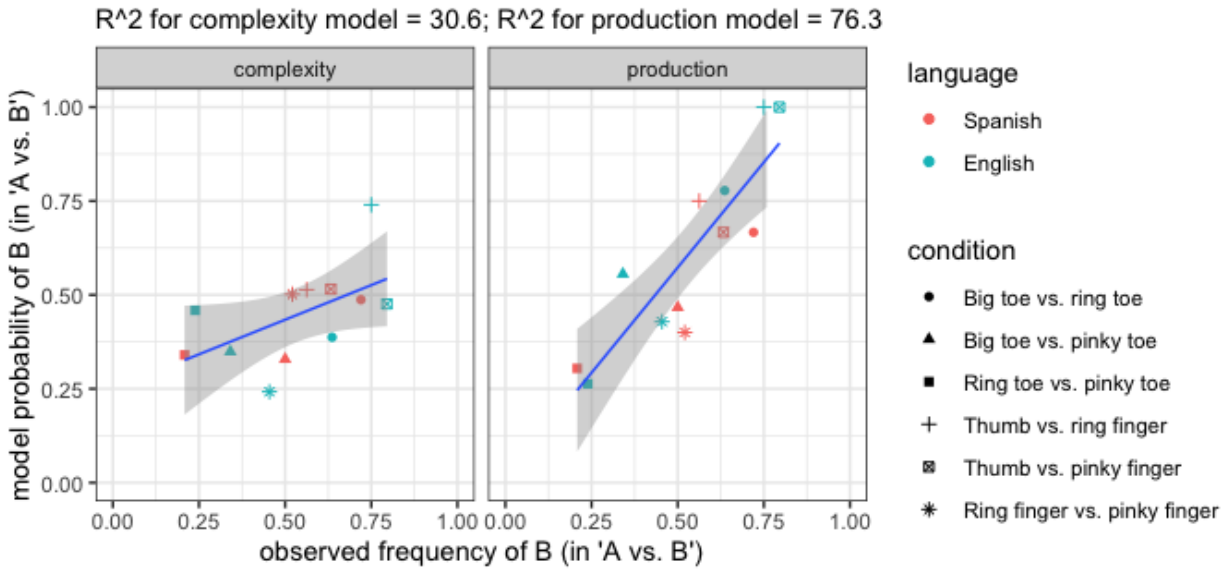

Figure 6: Comparison of Model Results

$\rightsquigarrow$ 'not pinky toe' suggests that *pinky toe* is a more prevalent alternative than *ring toe*. These results suggest that something else is going on in calculating implicatures that complexity alone cannot account for. The Complexity model fails to understand that two alternatives with equal complexity may differ in their prevalence. The Production model, on the other hand, is capable of accounting for this.

Overall, the Production model performs much better than the Complexity model; see Figure 5. The only incorrect prediction it makes is that there would be no implicature for big toe/pinky toe in English. The Production model predicts stronger implicatures for the thumb/ring finger items in English and Spanish, and for the thumb/pinky finger items in English. In other words, where the comprehension results trend rightward, at just over 75%, for the *thumb/pinky item* in English, the Production model predicts 100% of participants selecting the pinky finger over the thumb. Otherwise, the model predictions fall in line with all empirical results for the comprehension experiments in Spanish and English.

Figure 6 plots the rate at which listeners chose the image on the right along the x-axis against the probability assigned to the image on the right by each model on the y-axis. A perfect model would assign probability at the exact same rate as actual production. The R^2 for the Complexity model is only 30.6%. This means that the Complexity model accounts for 30.6% of the variation present in the comprehension studies. The R^2 for the Production model, in contrast, is 76.3%, which is to say that the Production model accounts for 76.3% of the variance in the data. There is a stark contrast in the explanatory power of each model. This shows that listeners have a good mental model of speakers, and that their mental model is not purely complexity-based. In fact, the comparison of these model results suggests that speakers are considering prevalence *over* complexity, since the prevalence-based Speaker model does not include a cost parameter.

5. Conclusions. The results outlined above suggest that Spanish and English speakers do differ with respect to the scalar implicatures associated with *finger* in accordance with the prevalence of the words for 'thumb', 'ring finger' and 'pinky finger'. Our empirical results support the idea that differences across languages in the implicatures associated with general terms are closely tied to differences in production probabilities for more specific terms. Since *pulgar* 'thumb' is not as prevalent in Spanish as *thumb* is in English, it is not available in the set of alternatives to *finger*, which is why speakers do not calculate an implicature. Our modelling results further support the conclusion that alternatives are constrained based on prevalence: the Production model significantly outperforms the Complexity model. While complexity does assist in determining the set of alternatives present for speakers, it is not as explanatory as full awareness of what speakers actually produce. *Prima facie*, these findings go against structural theories, like Katzir (2007) and Horn (2000), that constrain the set of alternatives based on complexity alone. These results also support the idea that activation of alternatives is not an all-or-nothing matter, and this is an idea that is naturally captured in a Bayesian pragmatic model that relies on gradient speaker production probabilities.

References

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998.

Geurts, Bart. 2011. *Quantity implicatures*. Cambridge: Cambridge University Press.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.

Horn, Lawrence R. 2000. From *if* to *iff*: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics* 32(3). 289–326.

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