Complexity vs. salience of alternatives in implicature: A cross-linguistic investigation

Word count: 10185

Danielle Dionne
Boston University
ddionne@bu.edu

Elizabeth Coppock
Boston University
eecoppock@bu.edu

Abstract Scalar implicature depends on the activation of alternatives. For instance, in English, *finger* implicates 'not thumb', suggesting that *thumb* is an activated alternative. Is this because it is more specific (Quantity) and equally short (Manner)? Indeed, *toe* doesn't imply 'not big toe', perhaps because *big toe* is longer. As L. Horn points out, this Quantity/Manner explanation predicts that if English had the simplex Latin word *pollex* meaning 'thumb or big toe', then the asymmetry would disappear. But would it suffice for that word to exist in the language, or would the word also have to be sufficiently salient? We explore this question in four languages that are sometimes said to lack a single-word alternative for thumb: Spanish (which does have *pulgar* 'thumb or big toe' (< *pollex*), though it is a non-colloquial form), Russian, Persian, and Arabic. To gauge the salience of various ways of describing digits, we use a fill-in-the-blank production task. We then measure the availability of implicatures using a forced choice comprehension task. We find cross-linguistic differences in implicature, and moreover that implicature calculation tracks production probabilities more closely than structural complexity of the alternatives. A comparison between two Rational Speech Act models – one in which the speaker replicates our production data and a standard one in which the speaker chooses based on a standard cost/accuracy trade-off – shows that comprehension is more closely tied to production probability than to the complexity of 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.

Forced to choose, would you guess that the tattoo was on the thumb or the ring finger? If you are like most of the participants in the English comprehension study we will report on in this paper, you would probably guess the ring finger. The thumb is technically a finger; otherwise how could it be true (as it is) that people generally have 10 fingers? So it is 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* and *finger* is not parallel to the relationship between *big toe* and *toe*:

(2) a. I hurt my finger. ⇝ I did not hurt my thumb.
b. I hurt my toe. ⇝̸ I did not hurt my big toe.

Horn 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. He then makes the following prediction: "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" (fn. 17, p. 308).

The idea that implicature is affected by the structural complexity of the alternatives involved – in other words, that Grice's "Be Brief" submaxim of the Maxim of Manner plays an important role – is not a new one, nor is it specific to Horn. In morphology, the concept is called 'blocking', and it's a well-established idea (e.g. Aronoff 1976; Kiparsky 1982). For example, the more lexicalized *decency* blocks the more productive **decentness*. The phenomenon of 'partial blocking', where a more lexicalized form prevents a more productive form from covering the same meaning space, as in for example *informant* vs. *informer*, suggests that blocking is not a purely structural phenomenon, but is rather about how speakers choose to express a given meaning: more simply, if possible. The more complex form is only used to express the meaning that is not carved out by the simpler form (Kiparsky 1982; Horn 1984). The same phenomenon can be observed on a syntactic level, as in the famous *kill* vs. *cause to die* example from McCawley (1978). Although it is not a new idea, the maxim of manner has received comparatively little explicit attention in the recent pragmatics literature, as Rett (2015) discusses, although there are exceptions including Blutner (1998; 2000), van Rooy (2003), Jäger (2000; 2012), and Mazzarella and Gotzner (2021); see also Rett (2020) for an overview.

Although not explicitly discussed under the 'Manner' heading, structural complexity does play an important role in modern theorizing about scalar implciature. For instance, Katzir (2007) and Fox and Katzir (2011) propose that alternatives for a given sentence are constructed through deletions, contractions and replacements based on a *substitution source*, which is the lexicon of the language. Alternatives are ordered by relative structural complexity, and this ordering determines whether an alternative is activated. The constraint on complexity helps to solve the so-called 'symmetry problem', which in a nutshell is to explain why, for example, *John read three books* implicates the negation of *John read four books*, when there is no *a priori* reason not to assume that both *John read four books* and *John read exactly three books* are salient alternatives. From the Katzir/Fox perspective, the reason is that the latter is too structurally complex. In the Rational Speech Act framework, complexity is incorporated in the form of a cost penalty that lowers the probability of an alternative being chosen by a speaker the more structurally complex it is, where in principle various choices can be made with regard to how structural complexity is measured (Frank and Goodman 2012, Bergen et al. 2016, Bennett and Goodman 2018, Degen et al. 2020).

Returning to fingers and toes: Geurts (2011) zeroes in on Horn's (2000) 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 both equally short and sufficiently salient, a scalar implicature will arise when the weaker form is used.

Whether this conjecture holds is our central research question here.

Many authors, including Fox and Katzir, have espoused the idea that relevance or salience constrains the set of alternatives considered for the purposes of scalar implicature calculation. For instance, Matsumoto (1995) observes that the following sentence carries a scalar implicature that it was *not* "a little bit more than warm" yesterday:

(3) It was warm yesterday, and it is a little bit more than warm today.

To account for this observation, Katzir (2007) assumes that the substitution source may include words and phrases in the surrounding text. There are a number of other frameworks for the analysis of scalar implicature in which some notion of relevance constrains alternatives, as discussed by Zondervan (2010); relevant works include Krifka 1995, van Kuppevelt 1996, van Rooij 2002 and van Rooij and Schulz 2003, Chierchia 2004; 2006, and Chierchia 2013: pp. 109–110. There is also experimental evidence for the role of relevance in scalar implicature; see Zondervan 2010, Cummins and Rohde 2015, Skordos and Papafrago 2016, Franke et al. 2017, Gotzner 2017; 2019, and, for an overview, Gotzner and Romoli 2022. Among the findings in this literature is that prosodic emphasis on a term can activate its focus alternatives, and that the activation of focus alternatives can reduce the activation of other, mentioned alternatives. The effect of this is seen in implicature: Whether or not a weak scalar item is strengthened to exclude a stronger alternative depends on the activation level of that stronger alternative. But being activated in the local discourse context either through mention or through focus is conceptually a bit different from the notion that Geurts invokes, using the term "colloquial". The latter is a more stable, language-wide sort of "salience", which we refer to as BASELINE SALIENCE (the idea being that words may differ in their baseline likelihood of being used by a speaker, independently of the specific discourse context).

Work on 'scalar diversity' (e.g.Doran et al. 2009, Beltrama 2013,Van Tiel et al. 2016, McNally 2017, Gotzner et al. 2018, Sun et al. 2018, Simons and Warren 2018, van Tiel et al. 2019, Westera and Boleda 2020, Ronai and Xiang 2020, Pankratz and van Tiel 2021) shows that alternatives that are low on a scale implicate the negation of a higher alternative with varying robustness, and has led to mixed findings with respect to the question of what drives this variation. Doran et al. (2009) show that scalar implicatures are computed at different rates between cardinals, quantificational items, ranked orderings, and gradable adjectives. Various potential factors have been explored as a way of explaining this diversity, including semantic distance, boundedness of a scale, extremeness of the strong alternative, polarity, and availability of alternatives. Of these, the notion of "availability" is most closely related to Geurts's notion of how "colloquial" an alternative is; presumably, the more colloquial an alternative is, the more available it is. Doran et al. (2009) proposed that the diversity they observed was due to a difference in salience of alternatives, and this view is consistent with finding by de Carvalho et al. (2016) that weak scalar terms differ in their likelihood of priming stronger scalar alternatives, but the empirical studies bearing directly on the role of salience have given mixed results. Van Tiel et al. (2016) found that the distinctness of scalemates played a role, but actually found no clear role for the availability of scalar alternatives. On the basis of a different, corpus-based methodology, Pankratz and van Tiel (2021) conclude that relevance does play a role after all. Their measure of "relevance" was based on scalar constructions that draw on an explicit contrast, as in *It's warm but not hot.* But this method measures a relation between two terms, and therefore does not serve to operationalize Geurts's notion, which was not relational. Geurts's conjecture is that cross-linguistic variation in scalar implicature depends (at least in part) on the baseline salience of the higher scalar alternative.

In this paper, we investigate Geurts's conjecture with the help of four languages that are sometimes said to lack a word for "thumb": Spanish, Russian, Persian, and Arabic. In fact, Spanish even resembles Horn's hypothetical version of English with *pollex* instead of *thumb* insofar as it contains the word *pulgar*—the Spanish descendant of Latin *pollex*, a "polymorphous" word that can in principle refer to either the thumb or the big toes. But several Spanish speakers have expressed informally to us that 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 structural complexity, but it does differ in how likely it is to be used, and, we infer, its baseline salience for the speaker as an alternative.

Cross-linguistic research on scalar implicature has supported the idea that when a more informative scalar alternative is lacking in the language, a less informative scalar alternative is not strengthened as it would be in a language where the stronger alternative is present. Deal (2011) shows that in Nez Perce, which has an existential-only modal system, the meaning of the weak modal is *not* strengthened to the negation of a strong modal meaning. This is just as predicted under the view that English *may* implies 'not must' due to the existence of *must* as a stronger alterantive in the language. There is, in addition, some evidence that absence of a presuppositionally stronger alternative in a language removes the inference to its negation, as predicted by the principle *Maximize Presupposition* (Heim 1991): Collins (2016) argues that the non-uniqueness inference typically associated with indefinites is absent in Tagalog just in contexts where definiteness-marking is not available. On the other hand, Chemla (2007) shows that despite the fact that French lacks a single word for *both*, the weaker alterantive *all* is still interpreted in a way that excludes a 'both' meaning. So the evidence in this domain is not univocal.

Even in cases where the stronger alternative exists in the language, there is some evidence for cross-linguistic variation in the rate of implicature calculation. In a broad cross-linguistic investigation on the meaning and acquisition of quantifiers, Katsos et al. (2016) report a small degree of variation across languages in the rate at which both children and adults give "false" judgments for uses of weak quantifiers like *some* in contexts where they are underinformative (e.g. a situation where *all* is true). Stateva et al. (2019) report cross-linguistic differences in the rate at which implicatures of *some* and its counterparts are computed, and argue that the variation observed is due to the work of different processes of pragmatic enrichment, relying on different pragmatic principles. But as far as we know, it has not been addressed whether cross-linguistic variation in the the degree of strengthening for a weak scalar element is modulated by how colloquial the stronger alternative is.

Inspired by the case of *pulgar*, we argue that baseline salience does play an important role in scalar implicature. Data from English, Spanish, Russian, Persian and Arabic allow us to assess how consistently and how much baseline salience matters, at least when it comes to reference to digits on the hand and the foot. To measure baseline salience, we use a fill-in-the-blank production task, in which participants were shown an image of a finger or a toe with a tattoo on it, and asked to complete the sentence, "She has a tattoo on _____.". Implicature is measured using a forced choice comprehension task in which participants are asked to choose between two possible digits (marked with tattoos in the images they are presented with).

Since production frequency does not mirror structural complexity exactly, we can make a meaningful comparison between production-based models of implicature and complexity-based models. In the Rational Speech Act (RSA) framework (Frank and Goodman 2012; Scontras et al. 2018), a listener assigns a probability to a given interpretation of an

utterance based on the likelihood that a speaker would produce the utterance when intending to communicate the interpretation in question. A role for salience can naturally be integrated into an RSA model, so that more salient alternatives are associated with higher speaker likelihoods. But this framework is also compatible with a view in which salience does not play a role, and beyond the literal meaning, only the complexity of the utterance (implemented via the 'cost' of the utterance to the speaker) determines a speaker's choice of utterance and the listener's concomitant implicature calculations. So the research question suggested by Geurts's conjecture can be formalized as a choice between two sorts of models within the RSA framework. We do exactly this, by computationally implementing two types of RSA models: one in which the speaker perfectly replicates our production data and another in which the speaker chooses based on the usual 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 goes beyond structural complexity, and mirrors production probabilities more closely.

2 Methods

In this section, we describe our methods for both the production and the comprehension studies, so that we can later present the results side-by-side. A juxtaposition of the production and comprehension results will reveal the extent to which implicatures (in this arena) depend on the probability of speakers using a specific alternative, as opposed to the structural complexity of a specific alternative.[1]

2.1 Methods for production studies

Our production studies took the form of a fill-in-the-blank task, in which participants were shown a part of the body with a tattoo and asked to complete the sentence, "She has a tattoo on ____" (or translational equivalent, described in more detail below). The resulting data provided information about the range of strategies that native speakers actually use to describe the fingers and toes in question, as well as the likelihood of using each of these strategies.

2.1.1 Materials and procedure

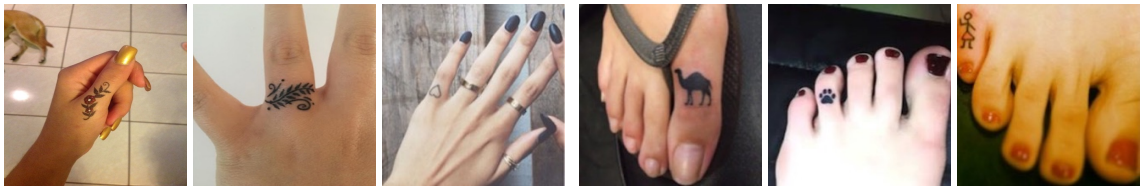
The images for the production studies (as well as the comprehension studies discussed below) are shown in Figure 1. The tattoos served as an indicator of which digit or body part the speaker was talking about. There were six filler items (all different from each other): two photos of tattoos on a leg, two photos of tattoos on an arm, and two photos of tattoos on the back. The target items showed photos of a tattoo on a finger or a toe. We included three digits of the hand (the thumb, ring finger, pinky), and three corresponding digits of the foot (big toe, ring toe and pinky toe).

Participants were shown a series of images, one by one, through a Google Form. With each image, they were asked to fill in the blank in the sentence: *She has a tattoo on* _____, or its translational equivalent in five different languages:

- English: "She has a tattoo on..."
- Spanish: "Tiene en tatuaje en...."

[1] All studies were carried out under IRB Protocol #5254X (exempt), approved by the Boston University Institutional Review Board on August 13, 2019.

Figure 1: Stimulus items for production and comprehension tasks



- Russian (romanized orthography): U neë tatuirovka na...
- Persian (romanized orthography): Ladayha washim ’alai ___ hā
- Arabic (romanized orthography): ū roy ___ khālkūbī karde āst

See Figure 2 for examples and how it appeared in the official script. The order of images was randomized. All participants were presented with all six target items and all six filler items.

2.1.2 Participants

All participants were recruited on Prolific. Participants were pre-screened for language and country of birth. For language, we screened based on both first language and fluent languages, and we included only participants who were raised monolingual. All of the demographic information is self-reported. The numbers and countries of birth are shown in Table 1. All studies involved different groups of participants.

2.1.3 Normalization of responses

After the data collection process was completed, 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*). Initial articles were also stripped off, and repetitions of the prompt were stripped away so that all that was remaining was the word or phrase that was used to refer to the digit itself. Thus *La mano izquierda* ‘the left hand’ was normalized

Figure 2: Sample production items for Russian, Persian, and Arabic

Table 1: Demographic breakdown for participants in the production study

n	First language	Country of Birth
24	English	United States
23	Spanish	Mexico
24	Russian	Russia
23	Persian	Iran
25	Arabic	Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mali, Saudi Arabia, Syria, UAE, Yemen

to *mano* 'hand', and *ella tiene un tatuaje en el cuarto ortejo* 'she has a tattoo on the fourth finger' was normalized to *cuarto ortejo* 'fourth finger'. Case and gender distinctions were also neutralized. Thus in Russian, for example, *bezymyannom palets* ('ring finger', with prepositional case) was normalized to *bezymyannyĭ palets* ('ring finger', nominative case).

For Russian, Persian, and Arabic, the normalized responses were also romanized for readability by an English-speaking audience. For Persian and Arabic, the responses were romanized using the International Journal of Middle East Studies (IJMES) transliteration system.[2] For Russian, the romanization was based on the Library of Congress romanization table.[3] The full datasets (anonymized), including the original responses, are available at https://github.com/eecoppock/tattoos.

Additionally, responses were coded for specificity — 1 for specific words/phrases that could refer to only one digit (e.g. *thumb* or *pulgar* 'thumb') and 0 for non-specific words/phrases that could refer to more than one digit (e.g. *finger* or *dedo de la mano* 'finger of the hand'). A more detailed semantic annotation was furthermore carried out for computational modelling purposes: Each normalized response was associated with a set of digits (among the three fingers and three toes under consideration) that it could truthfully apply to (formalized as a six-element vector of TRUEs and FALSEs). The literal semantics is described in more detail below as we describe the RSA models, and is given in the supplemental file ("Appendices").

2.2 Methods for comprehension studies

2.2.1 Materials and procedure

In our comprehension studies, participants were given a general description (*finger* or *toe*) and asked to choose between two images. The description and the images appeared through a web interface, with one description / image pair per screen. 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";

[2] https://www.cambridge.org/core/services/aop-file-manager/file/57d83390f6ea5a022234b400/TransChart.pdf

[3] https://www.loc.gov/catdir/cpso/romanization/russian.pdf

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.

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 like *She has a tattoo on her X*, where *X* was a general term: *finger* or *toe* or the translational equivalent.[4] 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. Responses were recorded as the image the participant clicked on (e.g. "thumb" for the image with the tattoo on the thumb). In some cases, participants went back to a previous page and entered an answer for the same stimulus twice. In such cases, only the final response was used in the analysis.

2.2.2 Participants

Like in the production study, all participants were recruited on Prolific, and participants were pre-screened for language and country of birth. For language, we screened based on both first language and fluent languages, and we included only participants who were raised monolingual. All of the demographic information is self-reported, and all studies involved different groups of participants. The numbers and countries of birth for the comprehension studies are shown in Table 2. Participants who did not complete the study or failed one or more "easy" fillers were eliminated from the results. In the table, n is the number of participants whose data figured into the analysis, k is the total number of individuals who participated at all, i is the number eliminated due to incomplete participation, and e is the number eliminated due to failing an attention check.[5]

Table 2: Demographic breakdown for participants in the comprehension study

$n = k - i - e$	First language	Country of Birth
$94 = 100 - 0 - 6$	English	United States
$100 = 102 - 0 - 2$	Spanish	Mexico
$49 = 52 - 0 - 3$	Russian	Russia
$38 = 46 - 4 - 4$	Persian	Iran
$32 = 39 - 4 - 3$	Arabic	Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mali, Saudi Arabia, Syria, UAE, Yemen

[4] Since the stimuli were presented in a written modality, participants were left to use their own imagination about the prosodic intonation contour, which is known to affect the activation of alternatives (Franke et al. 2017; Gotzner 2019). The role of prosody is an issue that could be studied in future work.

[5] In previous work on this topic (Dionne and Coppock 2021), we found an 18-point difference between English and Spanish in the thumb vs. ring finger condition. Despite the magnitude of the difference, statistical tests did not allow us to reject the hypothesis that the variation was due to noise; it was only marginally significant. That study was under-powered; in order to detect a difference of that size with 80% power, we need 100 participants in each group, whereas we only had around 50. We therefore re-ran the study with around 100 new participants in both English and Spanish, and we report those results here. Reassuringly, the estimates are extremely similar across these datasets, e.g. close to 50% for thumb vs. ring finger in Spanish, and close to 75% for the same condition in English. Smaller numbers of participants were involved in the comprehension studies for the other languages, so our estimates there are less certain, but none of our conclusions rely crucially on specific contrasts among these languages.

3 Results

We are now in a position to start addressing our main research question: Are alternatives activated in accordance with their baseline salience (as measured by production frequency in our production experiments) or in accordance with their structural complexity (as measured by number of words)? Although these two things are correlated, they are not identical.

A detailed breakdown of the variants we found in production, their relative frequency, and how we coded them for specificity is given in the supplemental file ("Appendices"). In this section, we report only the rate at which a specific expression was chosen to describe a given digit. Except for the thumb in English, for every digit, in every language, there was variation as to whether the digit was described specifically (i.e. in a way that distinguished the digit from other ones), or generally (in a way that did not distinguish between the digits). Overall, across languages, the ring finger and the ring toe were least likely to be described in a specific manner; perhaps these are inherently the least distinctive of the digits we tested. It may also be worth keeping in mind that we find a fair amount of variation across languages with respect to the number of different unique expressions used to describe a given digit, ranging from one (in the case of English *thumb*) to 13; Arabic speakers produced 13 different ways of referring to the big toe. Overall, the greatest amount of dispersion among responses was found in Arabic, followed by Spanish, then Russian, then Persian, with English in last place.

To analyze the comprehension data, for each image pair, we carried out a statistical test in order to determine the presence or absence of an implicature in relation to that pair. The data associated with each image pair is a set of selections among the two images. For example, if the choice is between the thumb and the ring finger, then it will be a series of observations consisting of some number of "thumb" selections and some number of "ring finger" selections. We are interested in whether the choice between images favored one over the other, or if the selection rate was at chance, with no statistically significant preference for one image over the other. We therefore conducted a 1-sample proportion test for each digit pair (using `prop.test` in R, which delivers p-values based on a chi-square statistic), where the null hypothesis, or the probability of choosing either digit is 0.5, assuming that the data follow a Bernoulli distribution. This procedure resulted in six significance tests per language (three for each finger pair, and three for each toe pair). We therefore adjusted our significance threshold to reduce the risk of false positives, using a Benjamini-Hochberg adjustment, which controls the false discovery rate (the ratio of false positives to positives).[6]

A full summary of the results is given in Table 3. The pair in the first column is of the form 'A vs. B'. 'A Spec.' is the rate at which specific descriptions were produced for A; likewise for B. 'Pref. for B' is the degree of preference for B. A preference of zero means that 50% of the selections were for B. A negative preference is a preference for A, so less than 50% of the selections were for B. The CI is a 95% confidence interval around the estimate of the preference for B. The p value estimates the probability that there is no preference for one over the other (adjusted via a Benjamini-Hochberg correction).

[6] Another option would have been the Bonferroni correction, which simply divides the significance level alpha by the number of significance tests. This option is more conservative, in that it more strongly reduces the risk of false positives, but it also carries a greater risk of false negatives. The Bonferroni option is appropriate in cases where the results are expected to be uncorrelated with each other. Here, the tests are not independent, because they are in sets of threes of the form A vs. B, B vs. C, and A vs. C. Bonferroni therefore seems too stringent for our purposes. However, it would result in the same set of significant contrasts for this particular dataset.

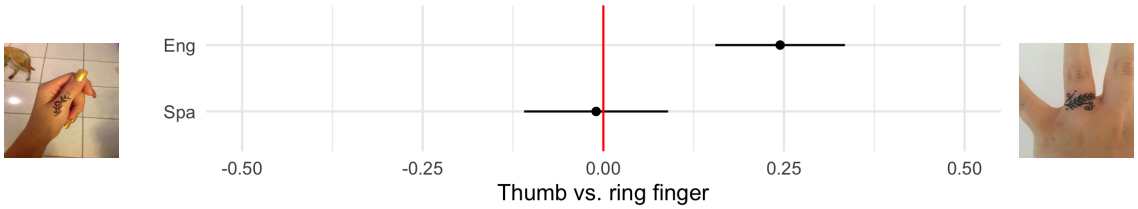
Table 3: Summary of production and comprehension results for all five languages

Pair (A vs. B)	Lg	—Production— A Spec.	B Spec.	—Comprehension— Pref. for B	(95% CI)	p (adj.)
Thumb vs. ring finger	**Eng**	**1.00**	**0.83**	**0.24**	**(0.14,0.33)**	**<0.01**
	Spa	0.65	0.52	–0.01	(–0.11,0.09)	0.95
	Rus	0.88	0.84	0.01	(–0.13,0.15)	1.00
	Per	0.87	0.22	–0.05	–0.21,0.12)	0.67
	Ar	0.62	0.46	0.06	(–0.12,0.23)	0.67
Thumb vs. pinky finger	**Eng**	**1.00**	**0.88**	**0.28**	**(0.18,0.35)**	**<0.01**
	Spa	**0.65**	**0.78**	**0.15**	**(0.05,0.24)**	**0.01**
	Rus	0.88	0.94	–0.11	(–0.24,0.04)	0.26
	Per	0.87	0.78	0.11	(–0.07,0.26)	0.40
	Ar	0.62	0.62	0.06	(–0.12,0.23)	0.67
Ring finger vs. pinky finger	Eng	0.83	0.88	0.04	(–0.06,0.14)	0.61
	Spa	**0.52**	**0.78**	**0.17**	**(0.07,0.26)**	**<0.01**
	Rus	0.84	0.94	0.05	(–0.1,0.19)	0.67
	Per	0.22	0.78	–0.13	(–0.28,0.04)	0.25
	Ar	0.46	0.62	–0.09	(–0.26,0.09)	0.53
Big toe vs. ring toe	Eng	0.83	0.42	0.11	(0,0.2)	0.11
	Spa	**0.57**	**0.17**	**0.17**	**(0.07,0.26)**	**<0.01**
	Rus	0.94	0.44	0.07	(–0.08,0.21)	0.53
	Per	**0.87**	**0.13**	**0.29**	**(0.12,0.4)**	**<0.01**
	Ar	0.42	0.19	–0.09	(–0.26,0.09)	0.53
Big toe vs. pinky toe	Eng	0.83	0.79	–0.10	(–0.19,0.01)	0.16
	Spa	**0.57**	**0.52**	**–0.14**	**(–0.23,–0.04)**	**0.02**
	Rus	0.94	0.97	–0.15	(–0.28,0)	0.11
	Per	0.87	0.83	–0.05	(–0.21,0.12)	0.67
	Ar	0.42	0.42	–0.16	(–0.31,0.03)	0.21
Ring toe vs. pinky toe	**Eng**	**0.42**	**0.79**	**–0.32**	**(–0.39,-0.22)**	**<0.01**
	Spa	**0.17**	**0.52**	**–0.33**	**(–0.4,–0.24)**	**<0.01**
	Rus	**0.44**	**0.97**	**–0.42**	**(–0.47,–0.3)**	**<0.01**
	Per	**0.13**	**0.83**	**–0.24**	**(–0.36,–0.07)**	**0.02**
	Ar	**0.19**	**0.42**	**–0.38**	**(–0.46,–0.2)**	**<0.01**

Focusing in on our motivating example, we see that Geurts's prediction is borne out for that case. 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", 74.5% of English participants chose the ring finger ($\chi^2 = 21.5$, $p < 0.0001$) (see Figure 3). In contrast, 49% of the Spanish speakers chose the ring finger image over the thumb image. In that graph, the red line at 0 represents no preference (50-50 distribution); more positive signifies greater preference for ring finger. The error bar, which depicts a 95% confidence interval, distinctly crosses this 50% mark, showing that the Spanish participants' responses are not statistically significantly different from chance; in fact, the null hypothesis is quite likely ($\chi^2 = 0.01, p = 0.95$).

The contrast between English and Spanish here is statistically significant according to a logistic regression model with image-pair, language, and their interaction as fixed effects and standard errors clustered by participant (using the lm_robust package in R). The logistic regression model is summarized in Table 4. In that table, each row is a term in the

Figure 3: Comprehension results for English (*finger*) and Spanish (*dedo* 'finger') in thumb vs. ring finger condition, with 95% CI



logistic regression model formula.[7] The reference level for language was Spanish, and the reference level for image-pair was 'Thumb vs. ring finger'. This was a condition with an estimate very close to 50%, so positive departures from this estimate indicate preferences for the 'B' member in an 'A vs. B' pair, and negative coefficients indicate preferences for the 'A' member. As the table shows, the coefficient for English is 0.25, which means that there is an estimated 25-point difference between English and Spanish for the 'Thumb vs. ring finger' condition, such that English participants were more likely to select the ring finger than Spanish participants by 25 points. This difference is significant at the 0.01 level. There were a number of other differences detected within and between languages, which we will not discuss individually.

4 Computational modeling

4.1 Complexity vs. Production: Defining the models

To understand the significance of the results, we turn now to formal modeling. Our main comparison is between two types of Rational Speech Act (RSA) models (Frank and Goodman 2012; Goodman and Stuhlmüller 2013: i.a.) that differ in how the speaker is defined. The first type incorporates a traditional speaker model that penalizes longer – more structurally complex – utterances (henceforth referred to as the Complexity models). The second type of model incorporates a speaker with perfect knowledge of speaker production (henceforth referred to as Production models).

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. Given an utterance u, the pragmatic listener L_1 assigns a probability to a state s, written $L_1(s \mid u)$, in proportion to the probability that the speaker S would use u to characterize state s, written $S(u \mid s)$, multiplied by the prior probability of state s, written $P(s)$.

(4) **Pragmatic listener**: L_1
$L_1(s \mid u) \propto S(u \mid s) \cdot P(s)$

This part is common across all of the comprehension models we will consider. Our comprehension models differ on how the speaker model they embed—the $S(u \mid s)$ part—is determined.

[7] Est.: estimated coefficient of the term corresponding to the row; t: t-statistic; SE: standard error; CI: 95% confidence interval around the estimate; df = degrees of freedom.

Table 4: Logistic regression model for comprehension results

term	est.	SE	t	p	CI.low	CI.high	df
(Intercept)	0.49	0.05	9.75	<**0.01**	0.39	0.59	99.00
Thumb vs. pinky finger	0.16	0.06	2.75	**0.01**	0.04	0.28	99.00
Ring finger vs. pinky finger	0.18	0.07	2.51	**0.01**	0.04	0.32	99.00
Big toe vs. ring toe	0.18	0.07	2.62	**0.01**	0.04	0.32	99.00
Big toe vs. pinky toe	–0.13	0.07	–1.92	0.06	–0.26	<0.01	99.00
Ring toe vs. pinky toe	–0.32	0.06	–5.48	<**0.01**	–0.44	–0.20	99.00
Eng	0.25	0.07	3.77	<**0.01**	0.12	0.39	191.26
Rus	0.02	0.09	0.23	0.82	–0.15	0.19	95.45
Per	–0.04	0.10	–0.44	0.66	–0.23	0.15	66.85
Ar	0.07	0.10	0.71	0.48	–0.13	0.28	52.34
Eng:Thumb vs. pinky finger	–0.13	0.08	–1.65	0.10	–0.28	0.02	191.26
Eng:Ring finger vs. pinky finger	–0.38	0.10	–3.65	<**0.01**	–0.59	–0.18	191.26
Eng:Big toe vs. ring toe	–0.32	0.09	–3.40	<**0.01**	–0.50	–0.13	191.26
Eng:Big toe vs. pinky toe	–0.21	0.10	–2.18	**0.03**	–0.40	–0.02	191.26
Eng:Ring toe vs. pinky toe	–0.24	0.09	–2.83	**0.01**	–0.41	–0.07	191.26
Rus:Thumb vs. pinky finger	–0.28	0.09	–2.97	<**0.01**	–0.47	–0.09	95.45
Rus:Ring finger vs. pinky finger	–0.14	0.13	–1.07	0.29	–0.40	0.12	95.45
Rus:Big toe vs. ring toe	–0.12	0.11	–1.09	0.28	–0.34	0.10	95.45
Rus:Big toe vs. pinky toe	–0.03	0.10	–0.32	0.75	–0.24	0.17	95.45
Rus:Ring toe vs. pinky toe	–0.11	0.10	–1.07	0.29	–0.31	0.09	95.45
Per:Thumb vs. pinky finger	–0.00	0.11	–0.02	0.99	–0.23	0.22	66.85
Per:Ring finger vs. pinky finger	–0.26	0.15	–1.77	0.08	–0.55	0.03	66.85
Per:Big toe vs. ring toe	0.16	0.11	1.47	0.15	–0.06	0.38	66.85
Per:Big toe vs. pinky toe	0.13	0.14	0.91	0.36	–0.15	0.41	66.85
Per:Ring toe vs. pinky toe	0.14	0.13	1.07	0.29	–0.12	0.39	66.85
Ar:Thumb vs. pinky finger	–0.16	0.10	–1.65	0.11	–0.35	0.03	52.34
Ar:Ring finger vs. pinky finger	–0.34	0.13	-2.55	**0.01**	–0.60	–0.07	52.34
Ar:Big toe vs. ring toe	–0.34	0.14	–2.44	**0.02**	–0.61	–0.06	52.34
Ar:Big toe vs. pinky toe	–0.09	0.15	–0.60	0.55	–0.39	0.21	52.34
Ar:Ring toe vs. pinky toe	–0.12	0.13	–0.89	0.38	–0.38	0.15	52.34

One feature that is constant across speaker models is that they take accuracy into consideration. The notion of accuracy is encoded in the literal listener L_0, who chooses among the states consistent with the literal meaning of a given utterance, in proportion to the prior probability of the state:

(5) **Literal listener**: L_0
$L_0(s \mid u) \propto [\![u]\!](s) \cdot P(s)$

The space of possible states corresponds 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 (see second appendix in the supplemental file, "Appendices").

As in all RSA models, the speaker in a Complexity model chooses an utterance in proportion to its utility, modulo the 'rationality parameter' α, which modulates how much the speaker maximizes utility.

(6) **Complexity speaker**: S_{cplx}
$S_{\text{cplx}}(u \mid s) \propto \exp(\alpha \cdot U_{\text{cplx}}(u,s))$

In a Complexity model, utility for an utterance reflects a trade-off between accuracy and cost. Accuracy is measured based on the probability that a literal listener L_0 will choose a state s given an utterance u. Cost is measured by length in words.[8]

A cost parameter β reflects speakers' degree of preference to be as concise as possible when speaking. Utility for the Complexity speaker—$U_{\text{cplx}}(u,s)$, the utility of utterance u given state s—is thus defined as follows:

(7) **Utility for Complexity speaker**: U_{cplx}
$U_{\text{cplx}}(u,s) = \log(L_0(s \mid u) - \beta \cdot \text{cost}(u))$

In words: Utility for a Complexity speaker model is accuracy minus cost (modulo a tranformation into log space). This class of models is quite ordinary, as RSA models go.

In contrast, a Production speaker chooses an utterance based on the empirically observed frequencies in the production data. As in other RSA models, the speaker in a Production model chooses an utterance in proportion to its utility, modulated by the rationality parameter α.

(8) **Production speaker**: S_{prod}
$S_{\text{prod}}(u \mid s) \propto \exp(\alpha \cdot U_{\text{prod}}(u,s))$

But for a Production speaker, the main determinant of utility for a given utterance as a way of communicating a given state—$U_{\text{prod}}(u,s)$—is the empirically observed frequency with which that utterance was used to communate that state in our production studies. The exact counts are smoothed via a parameter ϵ, which is a small artificial "count" assigned to unobserved utterances. In natural language processing, smoothing is known to improve language models, making them less bound to the training data and more able to generalize to new data (see Jurafsky and Martin 2019; Heafield et al. 2013; Chen and Goodman 1998; Kneser and Ney 1995; Gale and Church 1994 for detailed overviews of different smoothing methods). In this setting, too, it is reasonable to suspect that utterances not observed in the production study could emerge if the production study were conducted a second time with entirely new participants, and including a non-zero epsilon parameter holds open that possibility. We write $F_\epsilon(u \mid s)$ to denote the smoothed 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), where ϵ is the artificial count assigned to unobserved utterances. Given this, utility for a Production speaker is defined as follows:

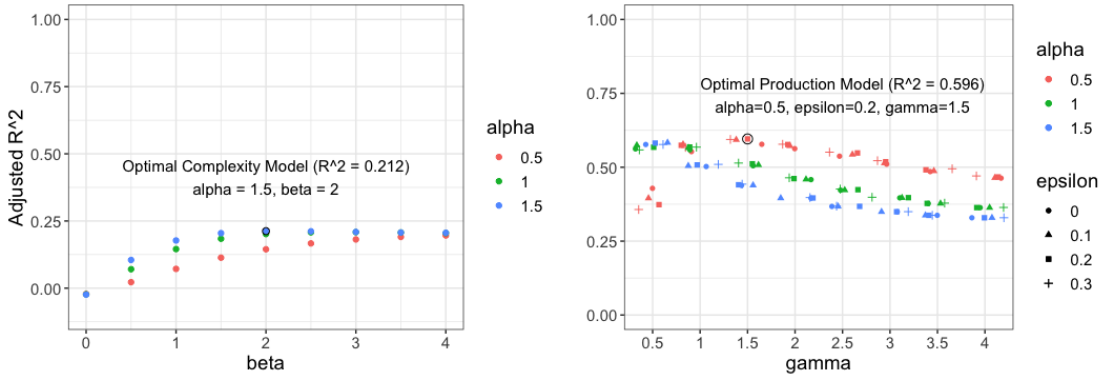
(9) **Utility for Production speaker**: U_{prod}
$U_{\text{prod}}(u,s) = \log(L_0(s \mid u) + \gamma \cdot F_\epsilon(u \mid s))$

where $F_\epsilon(u \mid s)$ is defined as follows:

(10) **Smoothed frequency**: F_ϵ
$$F_\epsilon(u \mid s) = \begin{cases} F(u \mid s) & \text{if } F(u \mid s) > 0 \\ \epsilon & \text{otherwise.} \end{cases}$$

[8] We also tried implementing a model in which cost was measured by the number of characters in the string, in light of the findings by Rohde et al. (2012), Degen et al. (2013), and Degen et al. (2020) that length in characters affects the likelihood that a speaker will choose a given utterance. Interestingly, this measure of cost was much worse than the measure of cost in terms of words, peaking with an R^2 of around 16%, vs. 22%.

Figure 4: Parameter optimization for Complexity (left) and Production (right) models



The γ parameter in (9) is the analogue of the cost parameter β in the Complexity models: It modulates the importance of smoothed frequency.

In addition to smoothed frequency, a Production speaker also takes accuracy into account, scaling the utility by the probability that the literal listener selects the correct referent. That probability is given by literal listener $L_0(s \mid u)$. Since all referents are assumed to be equally probable *a priori*, what these definitions boil down to is that the speaker in a Production model does not consider utterances that do not fit with the literal semantics, and otherwise chooses an utterance based on smoothed frequency.

4.2 Complexity vs. Production: Model performance

We turn now to the question of how these models fare in relation to the comprehension data and in relation to each other. The first step is to find the optimal settings of the free parameters for both classes of models. This will allow us compare the two models in their best respective lights. Our evaluation metric will be the amount of variance in the comprehension data that the model explains (the R^2 value), taking the mean ratings in the comprehension study as the target data. With five languages, and six image-pairs, there are 30 data points to predict.

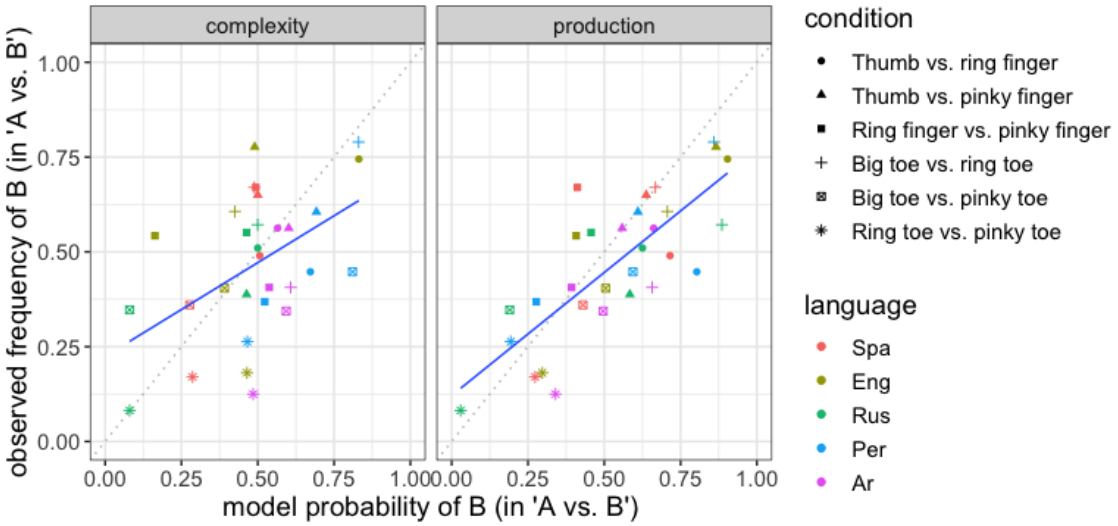
As mentioned above, both models have multiple free parameters, one being the rationality parameter α. The Complexity model also includes a cost parameter β, and the Production model includes a frequency parameter γ (modulating the importance of frequency) and a smoothing parameter ϵ (the artificial non-zero frequency for unobserved but semantically valid utterances). Figure 4 visualizes the effect of varying the relevant parameters for both models. The points circled in black represent the optimal models.

For Complexity models, we find that the optimal settings are $\alpha = 1.5$ and $\beta = 2$. Although these parameter settings are optimal for this class of models, the R^2 value is not particularly high: only 0.212. The optimal Production model has an R^2 of 0.596, with $\alpha = 0.5$, $\gamma = 1.5$, and $\epsilon = 0.2$.[9] Notice also that regardless of how the parameters are set, Production models tend to yield higher R^2 values than Complexity models.

As Figure 4 already suggests, the optimal performance of the Production model is clearly superior to that of the Complexity model. Figure 5 visualizes the relative performance of the optimal Complexity model vs. the optimal Production model.

[9] Observe that there is a trade-off between α and γ among lower values of γ: The higher γ gets, the lower α should be; higher values of α reduce the quality of the model. This is not surprising in light of the fact that they are doing a very similar job, both modulating the importance of frequency.

Figure 5: Comparative evaluation of Complexity (left) and Production (right) models



On the x-axis is the rate at which listeners chose the image on the right. The y-axis represents the probability assigned to the image on the right (*B* in '*A* vs. *B*') by the model. Error bars represent 95% confidence intervals around the empirical estimates from the comprehension studies. These estimates are plotted on the x-axis, and model predictions are plotted on the y-axis. A perfect model would assign a probability to the picture on the right that matches the rate at which it was chosen, so all of the points would be aligned on the dotted grey line, which has an intercept of 0 and a slope of 1. The blue lines shown on the plots represent a linear model with the best fit to the data, and indeed, the blue line on the panel illustrating performance for the Production model is close to a perfect diagonal. As this figure illustrates, the correlation between model predictions and human comprehension is much better for the Production model.

Let us consider some particular cases in which the Complexity model makes incorrect predictions, in order to understand where it is failing. There are two types of errors: failure to predict a preference, and predicting a preference that isn't there. For English, the Complexity model inaccurately fails to predict any preference for the pinky with the thumb/pinky item. This is because the model is only considering the fact that these two utterances are equally complex, failing to take into consideration the fact that the thumb is much more likely to be described as such than the pinky is. Similarly, since *ring toe* and *pinky toe* are equally complex, the Complexity model does not predict a preference for the ring toe, given a choice between the ring toe and the pinky toe. But since speakers are morely likely to use *pinky toe* to describe the eponymous digit than to use *ring toe* for the digit it names, the Production model does correctly predict a preference for the ring toe given a choice between those two, hearing *toe*. The Production model makes the correct prediction in this case. These cases illustrate the general fact that speakers do not always reach for the alternative that is least costly in terms of word length, and comprehenders know this.

There are certainly cases in which the Production model makes incorrect predictions as well; for example, it predicts a very strong preference for the ring toe in the 'big toe vs. ring toe' condition in Russian, whereas the actual proportion was around 50%, as predicted by the Complexity model. It also inaccurately predicts a strong preference

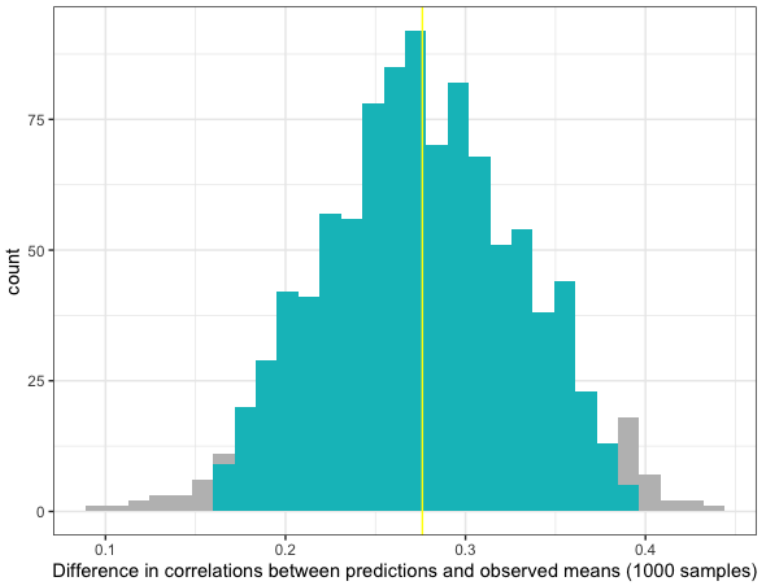


Figure 6: Histogram of correlation differences for Production vs. Complexity models

for the ring finger in the 'thumb vs. ring finger' condition in Persian. The Complexity model accurately predicts no preference here. Nevertheless, overall, the Production model performs quite decently, and substantially better than the Complexity model.

To give statistical support to the claim that the Production model outperforms the Complexity model, we carried out a statistical comparison between their respective correlations with the data. To estimate the uncertainty around the correlations for each model, we used a bootstrapping method, sampling with replacement at the participant level. We constructed 1000 alternative datasets by sampling n participants from among those in our study, where n was the number of participants in the comprehension study. For each dataset, the correlation between it and both models was computed, along with the difference between these two correlations. The resulting distribution over correlation differences was normally distributed, as shown in the histogram in Figure 6. That graph also plots the mean (the yellow vertical bar) and the 95% confidence interval of the mean, which is the region in blue. Datapoints ploted in grey lie more than 2 standard errors from the mean. Crucially, the 95% confidence interval excludes zero, allowing us to rule out the null hypothesis, according to which the two models do not differ. In other words, the Production model has a reliably higher correlation with the data than the Complexity model.

4.3 Mixed speaker model: Definition and evaluation

It is natural to wonder whether it helps to combine the Production and Complexity models together, taking into consideration both speaker production probability and cost. (Indeed, a reviewer wondered this.) We therefore constructed a Mixed speaker model, combining the Complexity speaker and the Production speaker such that the Mixed speaker chooses an utterance based on accuracy, cost, and empirically observed production data. The Mixed speaker model contains four free parameters: α, β, γ, and ϵ. As in the previous models, α is the 'rationality parameter' that maximizes speaker utility. The β parameter modulates the effect of cost. The Mixed speaker model also considers the frequency of the utterance, or $F_\epsilon(u \mid s)$, which takes a smoothing parameter ϵ, and is multiplied by a

coefficient γ that modulates the effect of frequency. The speaker for the Mixed model is therefore defined as:

(11) **Mixed speaker**: S_{mixed}
$S_{\text{mixed}}(u \mid s) \propto \exp(\alpha \cdot U_{\text{mixed}}(u,s))$

where $U_{\text{mixed}}(u,s)$ is defined as follows:

(12) **Utility for Mixed Speaker**: U_{mixed}
$U_{\text{mixed}}(u,s) = \log(L_0(s \mid u) - \beta \cdot \text{cost}(u) + \gamma \cdot F_\epsilon(u \mid s))$

where $F_\epsilon(u \mid s)$ is defined as in (10).

We carried out a model optimization procedure and found that the optimal Mixed model was one where $\beta = 0$, and α, γ and ϵ had their optimal settings for the Production model. Setting β to 0 renders a Mixed model equivalent to a Production model. We therefore conclude that it is not beneficial to add a cost parameter to a production model.

5 Conclusion and outlook

Our results suggest that the activation of alternatives depends on how colloquial they are, as conjectured by Geurts. Broadly, our results support the idea that differences both within and across languages in the implicatures associated with general terms are closely tied to differences in production probabilities for more specific terms (a measure of *baseline salience*). This conclusion was supported statistically by comparing two types of RSA models to the data: Complexity models, on which alternatives are selected by a speaker on the basis of a cost/accuracy trade-off, where cost is measured in number of words, and Production models, which use production probabilities as the primary basis for selecting an utterance. The fact that Production models provided a significantly better fit to the comprehension data fits well with a Bayesian perspective on interpretation, according to which listeners choose interpretations by reasoning about the likelihoods of various alternative actions a speaker could have taken, and it suggests that listeners have a very keen understanding of how speakers behave.

This work opens up a number of questions. One is the extent to which it is possible to improve upon the predictions of Complexity models, or in other words, to derive speaker production behavior rather than taking it as a given. One possible locus of improvement for Complexity models is in the prior probability distribution over referents. The fact that expressions for toes tend to be longer and more marked than expressions for fingers (e.g. *dedo* 'digit' vs. *dedo del mano/pie* 'digit of the hand/foot' – the former, shorter expression is much more often used to describe fingers, and the latter, longer type of description is much more often used to describe toes) suggests that toes are more marked as referents (meanings) than fingers. We conjecture that a better fit to the data could be obtained by a Complexity model with an uneven prior over referents such that toes are less likely as referents *a priori* than fingers. We leave it to future work to explore this possibility and to study the variation among models along this dimension. Even the Production model could stand to improve; there is a healthy amount of variation in the comprhension results that remains mysterious. Another issue is how to understand variation across languages in the degree of dispersion among possible production alternatives they exhibit. In some languages, there are very few options for how to describe a given digit, and in others, there are many. The role of dispersion among many lexical alternatives remains an open question for future investigation, but there is some evidence that it plays a role: In a study on the acquisition of quantifiers across 31 languages, Katsos et al. (2016) suggests that

the number of competing expressions may contribute to the cross-linguistic differences they found. We also wonder how bilingualism might affect implicature calculation: Do bilingual speakers compute implicatures based on alternatives in languages other than the one being spoken? Finally, in what other domains might we find cross-linguistic pragmatic differences that arise due to the salience of alternatives? We hope that this work serves as an impetus to explore this kind of variation further.

Supplementary file

Supplementary file: Appendices for "Complexity vs. salience of alternatives in implicature: A cross-linguistic investigation". Provides details of production results and literal meaning tables used by the RSA models.

Acknowledgements

The authors are grateful to the audiences at LSA 2020 and Experiments in Linguistic Meaning 2021 for feedback on this work. We would especially like to thank Laurence Horn and Florian Schwarz for their encouragement and guidance.

Competing interests

The authors have no competing interests to declare.

Author Contributions

DD: conceptualization, methodology, investigation, data curation, formal analysis, visualization, writing — original draft, and writing — review & editing
EC: conceptualization, methodology, investigation, data curation, formal analysis, visualization, writing — original draft, writing — review & editing, funding acquisition, supervision, and project administration.

References

Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press, Cambridge, MA.

Beltrama, A. (2013). Is good better than excellent? an experimental investigation on scalar implicatures and gradable adjectives. In *Sinn und Bedeutung 17*, pages 81–98.

Bennett, E. and Goodman, N. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*, *178*:147–161.

Bergen, L., Levy, R., and Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics & Pragmatics*, *9*(20):1–83.

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, *15*(2):115–162.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, *17*(3):189–216.

Chemla, E. (2007). French *both*: A gap in the theory of antipresupposition. *Snippets*, *15*:4–5.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Tech. Rep. TR-10-98, Computer Science Group, Harvard University.

Chierchia, G. (2004). A semantics for unaccusatives and its syntactic consequences. In Alexiadou, A., Anagnostopoulou, E., and Everaert, M., editors, *The Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface*, pages 22–59. Oxford University Press.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry*, *37*(4):535–590.

Chierchia, G. (2013). Free choice nominals and free choice disjunction: the identity thesis. In Fălăuş, A., editor, *Alternatives in Semantics*. Palgrave Macmillan.

Collins, J. (2016). Definiteness and implicatures in Tagalog. Ms., Stanford University.

Cummins, C. and Rohde, H. (2015). Evoking context with contrastive stress: effects on pragmatic enrichment. *Frontiers in Psychology*, *6*(1779):1–11.

de Carvalho, A., Reboul, A. C., der Henst, J.-B. V., Cheylus, A., and Nazir, T. (2016). Scalar implicatures: The psychological reality of scales. *Frontiers in Psychology*, *7*(1500):1–9.

Deal, A. R. (2011). Modals without scales. *Language*, pages 559–585.

Degen, J., Franke, M., and Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Degen, J., Hawkins, R. X. D., Graf, C., Kreiss, E., and Goodman, N. D. (2020). When redundancy is rational: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4):591–621.

Dionne, D. and Coppock, E. (2021). Tattoos as a window onto cross-linguistic differences in scalar implicature. In Beltrama, A., Schwarz, F., and Papafragou, A., editors, *Proceedings of ELM 1*, volume 1, pages 147–158.

Doran, R., Baker, R., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, *1*:1–38.

Fox, D. and Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*(1):87–107.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084):998.

Franke, M., Schlotterbeck, F., and Augurzky, P. (2017). Embedded scalars, preferred readings and prosody: An experimental revisit. *Journal of Semantics*, *34*(1):153–199.

Gale, W. A. and Church, K. W. (1994). What is wrong with adding one? In Oostdijk, N. and de Haan, P., editors, *Corpus-Based Research into Language*. Rodopi.

Geurts, B. (2011). *Quantity Implicatures*. Cambridge University Press, Cambridge.

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

Gotzner, N. (2017). *Alternative sets in language processing: How focus alternatives are represented in the mind*. Springer.

Gotzner, N. (2019). The role of focus intonation in implicature computation: a comparison with only and also. *Natural Language Semantics*, *27*:189–226.

Gotzner, N. and Romoli, J. (2022). Meaning and alternatives. *Annual Reviews in Linguistics*, *8*(1).

Gotzner, N., Solt, S., and Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*(1659):1–13.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified kneser-ney language model estimation. *ACL*, pages 690–696.

Heim, I. (1991). Artikel und Definitheit. In von Stechow, A. and Wunderlich, D., editors, *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 487–535.

Mouton de Gruyter, Berlin.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Schiffrin, D., editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42. Georgetown University Press, Washington, DC.

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

Jäger, G. (2000). Some notes on the formal properties of bidirectional optimality theory. *ZAS Papers in Linguistics*.

Jäger, G. (2012). Game theory in semantics and pragmatics. *in Maeinborn et al. (2012)*, pages 2487–2425.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing*, volume 3rd draft ed. Pearson.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kuvač Kraljević, J., Hrzica, G., Grohmann, K. K., Skordi, A., Jensen de López, K., Sundahl, L., van Hout, A., Hollebrandse, B., Overweg, J., Faber, M., van Koert, M., Smith, N., Vija, M., Zupping, S., Kunnari, S., Morisseau, T., Rusieshvili, M., Yatsushiro, K., Fengler, A., Varlokosta, S., Konstantzou, K., Farby, S., Guasti, M. T., Vernice, M., Okabe, R., Isobe, M., Crosthwaite, P., Hong, Y., Balčiuniene, I., Ahmad Nizar, Y. M., Grech, H., Gatt, D., Cheong, W. N., Asbjørnsen, A., Torkildsen, J. v. K., Haman, E., Miękisz, A., Gagarina, N., Puzanova, J., Andjelković, D., Savić, M., Jošić, S., Slančová, D., Kapalková, S., Barberán, T., Özge, D., Hassan, S., Chan, C. Y. H., Okubo, T., van der Lely, H., Sauerland, U., and Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*(33):9244–9249.

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

Kiparsky, P. (1982). Word formation and the lexicon. In Ingemann, F., editor, *Proceedings of the 1982 Mid-America Linguistic Conference*, pages 3–32, University of Kansas, Lawrence, KS.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. *ICASSP-95*, *1*:181–184.

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguistic Analysis*, *25*:209–257.

Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguistics and philosophy*, *18*(1):21–60.

Mazzarella, D. and Gotzner, N. (2021). The polarity asymmetry of negative strengthening: dissociating adjectival polarity from facethreatening potential. *Glossa: a journal of general linguistics*, *6*(1):47.

McCawley, J. (1978). Conversational implicature and the lexicon. In Cole, P., editor, *Syntax and Semantics, volume 9: Pragmatics*, pages 245–259. Academic Press.

McNally, L. (2017). Scalar alternatives and scalar inference involving adjectives: a comment on van tiel, et al. 2016. In *Asking the right questions: essays in honor of Sandra Chung*, pages 17–27. Linguistics Research Center, Santa Cruz, CA.

Pankratz, E. and van Tiel, B. (2021). The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, *FirstView*:1–33.

Rett, J. (2015). *The Semantics of Evaluativity*. Oxford University Press, Oxford.

Rett, J. (2020). Manner implicatures and how to spot them. *International Review of Pragmatics*, *12*(1):44 – 79.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., and Kaufmann, S. (2012). Communicating with cost-based implicature: a game-theoretic approach to ambiguity. In *Proceedings*

of the 16th Workshop on the Semantics and Pragmatics of Dialogue, pages 107–116.

Ronai, E. and Xiang, M. (2020). Pragmatic inferences are qud-sensitive: an experimental study. *Journal of Linguistics*, *FirstView*:1–30.

Scontras, G., Tessler, M. H., and Franke, M. (2018). *Probabilistic language understanding: An introduction to the Rational Speech Act framework*. Retrieved 2021-1-17 from https://www.problang.org.

Simons, M. and Warren, T. (2018). A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology*, *71*(1):272–279.

Skordos, D. and Papafrago, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, *153*:6–18.

Stateva, P., Stepanov, A., Déprez, V., Dupuy, L. E., and Reboul, A. C. (2019). Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, *10*:957.

Sun, C., Tian, Y., and Breheny, R. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, *9*:2092.

van Kuppevelt, J. (1996). Inferring from topics. *Scalar implicatures as topic-dependent inferences*, *19*:393–443.

van Rooij, R. (2002). Relevance only. In Bos, J., Foster, M. E., and Matheson, C., editors, *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002)*, pages 155–60.

van Rooij, R. and Schulz, K. (2003). Exhaustification. In Bunt, H., van der Sluis, I., and Morante, R., editors, *Proceedings of the Fifth International Workshop on Computational Semantics*, pages 354–98. University of Tilburg, Tilburg.

van Rooy, R. (2003). Relevance and bidirectional OT, in reinhard blutner & henk zeevat (eds.). *Pragmatics in optimality theory*, pages 173–210.

van Tiel, B., Pankratz, E., and Sun, C. (2019). Scales and scalarity: processing scalar inferences. *Journal of Memory and Language*, *105*:93–107.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1):137–175.

Westera, M. and Boleda, G. (2020). A closer look at scalar diversity using contextualized semantic similarity. In *Proceedings of Sinn und Bedeutung 24*, pages 439–454.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. PhD thesis, Utrecht University.