An Alternatives Account of
‘Most’ and ‘More Than Half’

Fausto Carcassi Jakub Szymanik
Institute for Logic, Language and Computation
Department of Linguistics, University of Amsterdam
Spuistraat 134, 1012 VB Amsterdam
fausto.carcassi@gmail.com jakub.szymanik@gmail.com

Abstract
While ‘most’ and ‘more than half’ are generally assumed to be truth-conditionally equivalent, the former is usually interpreted as conveying greater proportions than the latter. Previous work has attempted to explain this difference in terms of pragmatic strengthening or variation in meanings. In this paper, we propose a novel explanation that keeps the truth-conditions equivalence. We argue that the difference in typical sets between the two expressions emerges as a result of two previously independently motivated mechanisms. First, the two expressions have different sets of pragmatic alternatives. Second, listeners tend to minimize the expected distance between their representation of the world and the speaker’s observation. We support this explanation with a computational model of usage in the Rational Speech Act framework. Moreover, we report the results of a quantifier production experiment. We find that the difference in typical proportions associated with the two expressions can be explained by our account.

1 Introduction

According to the standard analysis of ‘most’ and ‘more than half’, the sentences ‘most cats sleep’ and ‘more than half of the cats sleep’ are truth conditionally equivalent. More generally, ‘most As are B’ and ‘more than half of the As are B’ are verified by the same As and Bs. ‘Most As are B’ is analysed as conveying that the size of $A \cap B$ is greater than the size of $A - B$, whereas ‘More than half of As are B’ is analysed as conveying that
the size of \( A \cap B \) is greater than half the size of \( A \) \cite{Hackl2009}:

\[
\llbracket \text{most} \rrbracket (A)(B) \iff |A \cap B| > |A - B| \tag{1}
\]

\[
\llbracket \text{MTH} \rrbracket (A)(B) \iff |A \cap B| > \frac{1}{2} |A| \tag{2}
\]

In contrast to this assumption, the behaviours of ‘most’ and ‘more than half’ differ. The main difference, which will be the focus of this paper, is that ‘most’ tends to be used to convey proportions higher than ‘more than half’. More specifically, while ‘more than half’ is usually used for proportions right above 50%, ‘most’ is used for proportions that are significantly higher than 50%.

The difference between ‘most’ and ‘more than half’ call for an explanation. Early work has focused on the different behaviour of the two expressions with respect to their upper bounds \cite{Ariel2003} or their cognitive encoding, which has been argued to lead to different verification procedures \cite{Hackl2009}. More recent work has focused on the more general fact that ‘most’ typically conveys proportions higher than ‘more than half’. Following \cite{DenicSzymanik2020}, we can categorize the explanations for this difference in two classes.

First, the \textit{pragmatic strengthening hypotheses} claim that while the two expressions are truth-conditionally identical, ‘most’ is pragmatically strengthened, resulting in a threshold higher than ‘more than half’. This strengthening can happen, e.g., through a scalar implicature or through an R-implicature. \cite{Solt2016} argues for the latter option and claims that the reason why ‘most’ receives a different interpretation to begin with is a (non truth-conditional) difference in the types of scales underlying the two expressions. On the other hand, \textit{lexical meaning hypotheses} attempt to explain these differences in terms of a truth-conditional difference in their logical forms, which can come from, e.g., conventionalization of implicatures or from ‘most’ being a vague quantifier. More recent work has produced experimental evidence supporting the hypothesis of a semantic difference between the two expressions. \cite{Ramotowskaetal2019} observe a difference in decision times and behaviour of subjects verifying sentences with the two quantifiers that are consistent with the model in which the threshold for ‘more than half’ is 50% and the threshold for ‘most’ is higher. \cite{DenicSzymanik2020} report that the thresholds of ‘most’ does not change under downward monotone environments, and argue that this finding suggests that the difference between the two quantifiers is due to semantics (see original paper for a fuller explanation).

In this paper, we propose a novel explanation of the difference between ‘most’ and ‘more than half’. We argue that two independently needed

---

\footnote{In the following, we assume that ‘A’ and ‘B’ refer to two finite sets \( A \) and \( B \). For a slightly more general definition in terms of measure functions see \cite{Solt2016}.}
mechanisms in the interpretation of quantifiers suffice to predict the difference between the two expressions, without assuming a difference in scale structures or truth conditions. The first mechanism is the tendency of the listener to guess points central to a category, in order to minimize the expected distance between their own guess and the speaker’s observation. The second mechanism is the structural theory of conceptual alternatives, which lets the alternative set of an utterance depend on the structure of the concept conveyed by the utterance. We show that these mechanisms make the correct predictions with a computational model of pragmatics, the Rational Speech Act model. We support our proposal with experimental data, showing that a hierarchical Bayesian model implementing our account can fit the production quantifier data.

The paper is structured as follows. First, in section 2 we give an overview of the most detailed account of the difference in the literature, Solt [2016]. Then, in section 3 we discuss two independently needed mechanisms in the pragmatics of quantifiers. In section 4, we present the computational framework of Rational Speech Act (RSA) modelling and show how it can model the two discussed mechanisms. In sections 5, 5.1, and 5.2 we apply the developed model to the case of ‘most’ and ‘more than half’ and show that it makes qualitatively correct predictions when implemented in the RSA framework. We then present a replication of some of the results of a quantifier production experiment Pezzelle et al. [2018], and show in section 7 that our model can fit the experimental data better than a model without the structural account of alternatives. Finally, in section 8 we compare our modelling result with the previous account by Solt.

2 Solt’s account

Solt [2016] offers a comprehensive review of the the difference between ‘most’ and ‘more than half’. Solt [2016] considers all appearances of the two expressions as quantifiers in the nominal domain in the Corpus of Contemporary American English (COCA) Davies [2017]. While various differences emerge when comparing the two expressions, in order to compare the typical proportions for which the two expressions are used those appearances were selected which included a specific percentage ($n = 54$ for ‘more than half’ and $n = 141$ for ‘most’). The corpus data shows that (1) ‘more than half’ is mostly used for percentages in the 50%-65% range, (2) ‘most’ has a much flatter distribution which covers the whole 50%-100% range, but rarely below 60%.

While the precise proportions reported in Solt [2016] are noisy estimates, the crucial observation is that the two expressions differ in the way they are used with respect to both their lower bounds and their upper bounds. The lower bound of ‘more than half’ is close to 50%, while the lower bound of
‘most’ is close to the upper bound of ‘more than half’. The upper bound of ‘more than half’ is much lower than 100%, while the upper bound of ‘most’ is close to 100.

Solt [2016] also proposes an account of the difference between the two expressions. In this account, the difference between the upper bounds and that between the lower bounds of the typical sets of ‘most’ and ‘more than half’ receive separate explanations. In order to explain the difference in lower bounds, Solt proposes that the scales that underlie the two expressions are different. Namely, according to Solt ‘more than half’ uses a ratio scale, while ‘most’ uses a semi-ordered scale [Stevens, 1946]. Two points on a ratio scale, such as the scale of weights, can be compared precisely to each other, and each of them can be compared to a proportion of the other. Therefore, if a language user’s measure of $A \cap B$ and $B$ are on a ratio scale, $A \cap B$ can be compared precisely to half of the measure of $A$. Where such precise comparisons are possible, a speaker can utter expressions such as ‘more than half of the As are B’. This in turn allows the speaker to use ‘more than half’ for cases where the proportion of As that are B is close to 0.5.

In contrast to ratio scales, a point on the type of semi-ordered scale Solt discusses can be represented as a whole distribution, which encodes uncertainty about the precise value in some underlying precise scale. For instance, the measure of an object’s weight that is obtained just by observing the object can be represented as a distribution on the physical scale of weights. Two points on such a scale can be distinguished from each other only when the distributions do not overlap excessively. An expression such as ‘most As are B’ only requires us to determine whether the size of $A \cap B$ is greater than the size of $A - B$. However, since perceptible differences in semi-ordered scales require substantial differences between the points, ‘most As are B’ will only be uttered when the measure of $A \cap B$ is substantially greater than the measure of $A - B$. In sum, since ratio scales allow arbitrarily precise comparisons between points while semi-ordered scales require substantial differences, Solt’s account predicts the lower bound of ‘more than half’ to be closer to 0.5 than the lower bound of ‘most’, as observed in the corpus data.

Solt accounts for the difference in the upper bounds of the two expressions with scalar implicatures. Solt points out that ‘more than half’ has a rich set of alternative utterances, including ‘more than two thirds’ and ‘more than three quarters’. On the other hand, the alternative utterances to ‘most’ are more sparse, including ‘all’. Since the set of alternative utterances is more fine-grained for ‘more than half’ than for ‘most’, scalar implicatures constrain the upper bound of the former to be lower than the latter. Solt proposes that the reason why the two expressions have different sets of alternatives is that each expression only alternates with expressions that use the same scale type.

---

2This type of semi-ordered structure has been proposed as a cognitive model of perception of quantities [Gescheider, 2013], preferences [Luce, 1956], and numbers [Dehaene, 1999].
In the next section, we introduce an alternative account that explains the difference between 'most' and 'more than half' without assuming a difference in the scales underlying the two expressions.

3 Two mechanisms in the interpretation of quantifiers

In this section, we present our account in non formal terms. Our account explains the difference in typical set between 'most' and 'more than half', and is based on two phenomena relating to the interpretation of quantifiers. The first is the idea that the listener attempts to minimize the difference between their own guess and the speaker’s observation, the second is the fact that different conceptual structures cause different sets of alternatives. We next consider these two mechanisms in turn.

3.1 Distance-minimizing listeners

The members of some semantic domains, such as the domain of numbers, colors, or proportions, enter in relations of similarity to each other. For instance, two shades of blue are closer to each other than either of them is to a shade of red. On the other hand, some semantic domains, such as nationality, football teams, or personal identity, are not usually structured by relations of similarity. For instance, it is nonsensical to claim that Billy the Kid is closer, in terms of his identity, to Jesse James than Doc Holliday.

In many cases, when communication happens in domains structured by distance, and the listener’s task is to construct a representation of the world state given a description produced by the speaker, communicative success is not simply a function of whether the listener’s representation is identical or not to the true world state. Rather, success is an inverse function of the similarity between the true state of the world and the listener’s guess. In other words, the closer the listener’s guess to the true world state, the more successful the communication.

In this perspective, it is a sensible strategy for a listener to not simply sample from the set of possible world states given their probability after receiving the message, but rather to attempt to minimize the expected distance between their guess and the true world state. For instance, if the speaker utters ‘blue’, the listener might select a shade of blue that is located around the center of the blue category, because a point near the center of the category will have a lower expected distance to the true world state than a

---

3There are feature with respect to which two individuals might be more or less close to each other, but this does not concern their identity as such.

4This communicative setup is called descriptive by [Franke, 2014], who opposes it to referential communicative games. In the following, we limit ourselves to discussions of descriptive communication.
point that is around the margin of the category. Previous literature supports this idea that listeners should tend to guess the center of a category when communicative success depends on the similarity between true state and listener's guess, e.g. Jäger et al. [2011] showed that the optimal strategy for such so-called simmax signaling games involves a listener that guesses the central point in the category.

Consistently with previous literature (See chapter 2 of Carcassi [2020] for an overview), we will assume in the rest of this paper that communication with quantifiers happens on a semantic domain structured by a distance, namely the scales of proportions and numbers. Moreover, we claim that in communication with quantifiers, communicative success is of the graded type presented above. For instance, if 1/2 of the As are B, the communication is more successful if the listener guesses $|A \cap B|/|A| = 0.6$ than if the speaker guesses 0.9. This implies that a rational listener does not guess a proportion simply by sampling from the posterior over proportions conditional on the received signal, but rather they attempt to minimize the expected distance between their guess and the true state of the world.

The tendency for the listener to guess a state that minimizes the expected distance to the speaker’s observation, when in a scalar semantic domain, is not only a result about rational agents, but also aligns with the way we use quantifiers in practice. For instance, imagine receiving the signal ‘between 50 and 100’, and creating a representation of the world state. Even within the part of the scale of integers covered by the expression—e.g. numbers between 50 and 100—the guess does not happen uniformly. Rather, we intuitively tend to guess an integer that is around the center of the category, i.e. around 75. In other words, we are less likely to select a number close to the category boundaries, such as 99. As we discuss in more detail below, the situation is subtler when multiple possible utterances are involved.

### 3.2 The structural account of alternatives

In this paper, we propose an alternative explanation for the difference in the sets of alternative utterances which avoids introducing mechanisms that are specific to ‘most’ and ‘more than half’. In particular, we rely on the structural account of conceptual alternatives [Chemla 2007, Buccola et al., 2018] to explain why ‘most’ and ‘more than half’ have different sets of alternative utterances.

The conceptual account of alternatives built on the structural account of alternatives [Katzir 2007, Fox and Katzir, 2011, Trinh and Haida, 2015], which was initially proposed by Katzir [2007] to solve the symmetry problem of Gricean pragmatics (see e.g. Breheny et al. [2018] for an overview of the symmetry problem). One instance of the problem goes as follows. According

---

5 In the following, we will focus on the scale of proportions, but what we say can be easily generalized to the scale of integers.
to classic Gricean pragmatics reasoning, ‘some’ implicates ‘not all’ because if ‘all’ had been true, the speaker would have chosen to utter ‘all’ instead of ‘some’. This reasoning relies on the assumption that the live alternatives are ‘some’ and ‘all’. However, a symmetric line of reasoning arrives at the conclusion that ‘some’ implicates ‘all’. If ‘some but not all’ had been true, the speaker would have uttered ‘some but not all’ rather than ‘some’. Therefore, an utterance of bare ‘some’ implicates that ‘some but not all’ is false, i.e. ‘not some or all’ is true. Therefore, under the assumption that the speaker is truthful and can decide to utter ‘some but not all’, ‘some’ ought to implicate ‘all’. This contradicts the fact that ‘some’ implicates ‘not all’ rather than ‘all’.

A way to solve this mechanism is to break the symmetry between ‘all’ and ‘some but not all’, by excluding the latter from the set of alternatives. The way that the structural account of alternatives achieves this is to restrict the set of alternatives to ‘some’ to only those utterances that have a structure at most as complex as ‘some’, thus including ‘all’ while excluding ‘some but not all’.

Formally, the structural theory of alternatives starts with the idea of a structural alternative. $\psi$ is a structural alternative to $\phi$ ($\psi \preceq \phi$) iff $\psi$ is structurally at most as complex as $\phi$, i.e. $\psi$ can be obtained from $\phi$ through a “finite series of deletions, contractions, and replacements of constituents of $\phi$” with constituents of the same category taken from the lexicon [Katzir, 2007]. The core idea is to define the set $A_{\text{str}}(\phi)$ of utterances alternative to $\phi$ as follows:

$$A_{\text{str}}(\phi) = \{ \psi | \psi \preceq \phi \} \quad (3)$$

In words, the set of utterances that enter in the calculation of implicatures for $\phi$ is the set of utterances that are structurally at most as complex as $\phi$.

While the original criterion for alternatives in [Katzir, 2007] is a syntactic one, there is emerging theoretical and experimental evidence that it is best characterized as acting not on the syntactic structure, but rather on the conceptual structure of utterances [Chemla, 2007, Buccola et al., 2018]. In the following, we limit ourselves to applying the basic idea in equation 3 to conceptual structure rather than syntactic structure, as it is all we need for present purpose.

We make three crucial assumptions about the way alternatives are generated for the expressions under consideration. First, not every expression of the form ‘$a \, b$’ is considered, where $a$ is a cardinal number (e.g. ‘three’) and $b$ an ordinal number (e.g. ‘fourth’) such that $a \leq b$. If every $a$ and $b$ were considered, the set of alternatives to ‘one half’ would be the set of rational numbers in the unit interval. Various factors plausibly restrict the set of considered numbers. First, the listener can generally assume that the speaker has noisy measurement of the true proportion, and therefore

---

6In fact, the original syntactic criterion could also be used if we take ‘more than half’ to be ‘more than one half’.

---
only produces utterances implying at most a certain level of granularity. Moreover, the communicative aims generally do not require transmission of precise proportions. The idea that the notion of alternative is graded and depends on the complexity of the concept, developed in Buccola et al. 2018, could also account for why ‘five sixth’ seems to compete with simpler fractions such as ‘two thirds’, while the opposite is not true; the difference would depend on the different conceptual complexity of different fractions.

The second assumption we make is that the quantifiers constructed by substitution satisfy the conservativity, extensionality, and isomorphism-closure invariance properties as discussed in Peters and Westerståhl 2006. This is equivalent to assuming that the only predicates to be considered correspond to combinations of $|A - B|$ and $A \cap B$.

The third assumption we make is that ‘most’ and ‘more than half’ are conceptually structured as proposed by Hackl 2009, i.e. as in equations 1 and 2 above. Crucially, the conceptual structure of ‘more than half’ contains a fraction $1/2$, the numerator and denominator of which can be substituted with other simple integers. Therefore, the main consequence of this assumption is that ‘two thirds’ and structurally equivalent expressions are alternatives to ‘half’ according to the criterion in equation 3, while ‘most’, which lacks the fraction in its conceptual structure, has a smaller set of conceptual alternatives. Under the two assumptions just discussed, the criterion defined in equation 3 has the correct consequences for the cases at hand. Namely, $A_{str}$(‘most’) contains ‘all’ and does not contain ‘more than three quarters’. On the other hand, $A_{str}$(‘(one) half’) contains e.g. ‘three quarters’.

In this section, we have presented two mechanisms that play a role in the way quantifiers are interpreted. These two mechanisms have already been discussed and supported in the literature in other contexts [Franke, 2014, Jäger et al., 2011, Katzir, 2007]. The main contribution of this paper is therefore to show how these two mechanisms can, together with the analysis of conceptual structure of ‘most’ and ‘more than half’ in Hackl 2009, explain the difference between ‘most’ and ‘more than half’ with respect to the proportions that the two expressions typically convey. In particular, our account does not need to introduce a difference in scale structure presented in Solt 2016.

In the next section, we model the mechanisms we discussed. We consider

---

7 For a discussion of the role of granularity in scalar language, see e.g. Cummins et al. 2012.

8 The lexicalization of the fractional concept $1/2$ with the omission of ‘one’ in ‘one half’ has a pragmatic justification: ‘one’ (or ‘a’) is generally superfluous when combined with ‘half’, since except in very rare occasions ‘two halves’ would not be uttered, given the simpler available option ‘one’. This is opposed to every other denominator, which can informatively combine with more than one numerator in a way that is not reducible to other fractions.
Figure 1: (a) Simple RSA model with three possible utterances $u$ (y-axis) and three states $s$ (x-axis). $L_1$ calculates a scalar implicature for utterances $u_1$ and $u_2$ ($\alpha = 4$). The left, central, and right plots correspond to $L_0$, $S_1$, and $L_1$ respectively. Note that the color indicates the probability of guessing a state given a signal for $L_0$ and $L_1$, and the probability of producing a signal given a state for $S_1$. (b) RSA model with a distance-minimizing $L_1$. The model displayed in the plot uses a language with three utterances and 20 states. The listener $L_1$ does not simply guess the signal observed by the speaker by sampling their posterior, but rather attempt to minimize the expected distance between their guess and the speaker’s observation ($\alpha = 4, \rho = 0.1$). See figure 1a for more detail.

A pragmatic speaker who picks ‘most’ or ‘more than half’ not simply as a function of their extension on the scale of proportions, but instead also implicitly selecting the set of alternatives that will allow a pragmatic listener to choose a proportion that is as close as possible to the speaker’s observation. Since a pragmatic listener guesses points closer to 0.5 for a rich alternative set such as the one induced by ‘more than half’, the speaker selects ‘more than half’ for such proportions. On the other hand, since the listener will guess alternatives close to 0.75 for ‘most’, the speaker produces ‘most’ for such proportions. In order to formalize this intuition, in the next section we present the RSA modelling framework for pragmatic language use.

4 An RSA model of the two mechanisms

4.1 Basic RSA model

The RSA framework is meant to model the process of recursive mindreading that lies behind the pragmatic interpretation or production of utterances [Goodman and Stuhlmüller, 2013, Franke, 2014, Frank, 2017]. RSA models usually start with a pragmatic listener who interprets utterances based on the simulated behaviour of a pragmatic speaker. The pragmatic speaker in turn given an observation tends to choose the most useful utterance for a literal listener who interprets it based solely on its literal meaning. We will first explain the simplest type of RSA model, and then a modification that will be useful to model numerals.
The simplest RSA model starts with a set of utterances $u$ and a set of possible states $s$. The meaning of each utterance can be encoded as the set of those states that verify the utterance. The pragmatic listener $L_1$ receives an utterance $u$ and calculates a posterior over states by Bayesian update, combining their prior over states with the probability that the pragmatic speaker $S_1$ would have produced the utterance given each state:

$$p_{L_1}(s|u) \propto p_{L_1}(s)p_{S_1}(u|s)$$  \hspace{1cm} (4)

The pragmatic speaker in turn observes a state and produces an utterance that aims at optimizing the utility $U(u|s)$ for a literal listener $L_0$ given the state, while minimizing the utterance cost $c(u)$:

$$p_{S_1}(u|s) \propto \exp(\alpha U(u|s) - c(u))$$  \hspace{1cm} (5)

The utility $U(u|s)$ is the negative surprisal of the state given the utterance, so that the speaker favours utterances that make the state less surprising for the literal listener:

$$U(u|s) = \log(p_{L_0}(s|u))$$  \hspace{1cm} (6)

Finally, the probability that literal listener $L_0$ attributes to each state given an utterance is simply 0 if the utterance is not verified by the state, and proportional to the prior for the state otherwise:

$$p_{L_0}(s|u) \propto \begin{cases} p(s) & \text{if } s \text{ verifies } u \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (7)

Figure 1a shows $L_0$, $S_1$, and $L_1$ in this simple RSA model. The crucial phenomenon that can be observed in figure 1a is that $L_1$ calculates a scalar implicature: although utterance $u_1$ is, in its literal sense, compatible with both $s_1$ and $s_2$, $S_1$ tends to produce $u_1$ mostly for $s_2$, because when $s_1$ is observed $S_1$ tends to use the more useful signal $u_1$. Therefore, when hearing $u_1$ $L_1$ is more likely to guess $s_2$.

### 4.2 Distance based listeners

In the simple RSA models above, the success of communication is binary, solely a function of whether the listener’s guess coincides with the speaker’s observed state. This is plausible in cases where the set of states has no internal structure. However, as discussed above in the case where a notion of distance is well-defined on the set of states, the listener might not be simply trying to guess the speaker’s observation, but rather might strive to minimize the (expected) distance between the state they select and the speaker’s observation.

---

9 Cf previous work where the effects of a distance structure affects the speaker but not the listener, such as [Franke 2014](#). It is worth noticing that our model does not have more
In order to model the effects of a well-defined distance $D$ on the set of states, we modify the listener $L_1$ so that instead of selecting a state by sampling from their posterior distribution given the signal, they try to minimize the expected distance between their selection $s$ and the true state. Therefore, we define the choice probability for listeners as follows:

$$p_{CL}(s|u) \propto \exp \left( -\rho \sum_{i \in \text{states}} p_{L_1}(i|u)D(i, s) \right)$$

where $\rho$ is the parameter of a softmax function which determines how strongly the listener tends to minimize the expected distance and $p_{L_1}$ is defined as above in equation 4. The listener described in equation 8 tends therefore to minimize the expected linear distance function. Figure 1b shows the effects of this modification of the model for 20 states, when $D(s_n, s_m) = |n - m|$. The right plot shows that in this modified RSA model, $L_1$ tends to guess points that are located centrally in the category, after the category has been restricted by scalar implicature.

4.3 Varying sets of alternatives

The modification to the basic RSA model above is an implementation of the first mechanism discussed in section 3. The second mechanism concerns the way that the comparison set depends on the speaker’s utterance.

In the basic RSA framework, the set of possible utterances considered by the pragmatic speaker and the pragmatic listener that the speaker models are identical. However, according to the structural account of alternatives discussed above the set of utterances considered by the listener depends on the actual utterance that is picked by the speaker. For instance, if the speaker utters ‘101’, the listener will consider all alternative utterances that are at most at a similar level of granularity as 101, such as 91 and 100. However, if the speaker utters ‘100’, the listener in the model considers an alternatives set containing e.g. only 90 and 100, but not 101.

In order to model this in the RSA model, we introduce a speaker $S_2$. $S_2$, much like $S_1$, tends to select the signal that minimizes the listener’s surprise for the real state given the signal. However, the set of alternative utterances considered by $L_1$ is not independent of the signal received by $L_1$. Instead, the set of alternative utterances considered by $L_1$ (and therefore by the lower levels $S_1$ and $L_0$) depends on the actual utterance of $S_2$, as described in the

---

10\text{We apply this modification only to } L_1, \text{assumming that the attempt to minimize distance is something above and beyond the literal reading of the signals. We leave to future work an investigation of the effects of modifying both listeners.}
Consider now for illustration the case of ‘some’, ‘all’, and ‘some but not all’ discussed above. Figure 2 shows the reasoning of speaker $S_2$ as they decide which signal to produce given that they observed a 100% state or a state < 100% (and > 0%). Being a rational speaker, $S_2$ produces signals that tend to maximize the probability that $L_1$ attributes to the true state. Since $L_1$ is a rational listener themselves, the probability attributed to each state given a signal depends on the set of alternatives to that signal. According to the structural account of alternatives, the set of alternatives is itself a function of the received signal. So if $S_2$ sends ‘some but not all’ (SBNA), $L_1$ will run pragmatic reasoning on the set of utterances \{‘All’, ‘Some’, SBNA\} (bottom row of plots in figure 2). Note that in this set of alternatives, corresponding to the symmetric case of the symmetry problem, ‘some’ does not implicate SBNA. On the other hand, if $S_2$ utters ‘some’, $L_1$ will reason only with \{‘All’, ‘Some’ \} according to the structural account of alternatives, and therefore calculate the implicature from ‘some’ to SBNA (top row of plots in figure 2). In sum, given a state $S_2$ will tend to produce the utterance that is most useful for a hypothetical $L_1$ who reasons about a set of alternatives which itself depends on $S_2$’s utterance.
Figure 3: Structure of modified RSA model. The set of alternative utterances considered by $L_1$ is not fixed, but rather depends on the received utterance. Moreover, $L_0$ and $L_1$ do not simply guess a state based on their posterior probability given the received signal, but rather tend to guess a state that is expected to be close to the speaker’s observation.

This picture of alternatives is in many respects a simplification. For instance, it is likely that from the point of the listener there is uncertainty as to the set of alternatives that ought to be considered in the context. More complex discussions of issues related to granularity and alternatives can be found in the literature, see e.g. Bastiaanse [2011] for numerals. However, these more complex models are not needed to explain the issue at hand, and therefore we leave investigation of the subtleties to future work.

In sum, the only requirements for the model presented in this section to apply are (1) that the listener is trying to minimise the distance between their guess and the speaker’s observation, and (2) that different terms induce different sets of alternatives. Crucially, the model applies even if two expressions with different alternatives sets are truth-conditionally equivalent.

In this section, we have formalized the two mechanisms discussed in section 3 within the RSA framework. The resulting model is summarized in natural language in figure 3. In the resulting model, structurally different expressions induce the pragmatic listener to consider different sets of alternative utterances. Moreover, the listener does not simply guess uniformly from the enriched part of the parameter space, but rather tends to guess points that are central in the pragmatically enriched category. Therefore, even intensionally equivalent expressions will be used differently, as long as they are structurally different. In the next section, we show how this model applies to the specific case of the contrast between ‘most’ and ‘more than half’.
<table>
<thead>
<tr>
<th>Utterance</th>
<th>Structure</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>All ((\forall)) (</td>
<td>A \cap B</td>
<td>\geq</td>
</tr>
<tr>
<td>Most ((-\exists)) (</td>
<td>A \cap B</td>
<td>&gt;</td>
</tr>
<tr>
<td>None ((\exists)) (</td>
<td>A - B</td>
<td>\geq</td>
</tr>
<tr>
<td>MT a half ((&gt; 1/2)) (</td>
<td>A \cap B</td>
<td>&gt; \frac{1}{2}</td>
</tr>
<tr>
<td>MT one third ((&gt; 1/3)) (</td>
<td>A \cap B</td>
<td>&gt; \frac{1}{3}</td>
</tr>
<tr>
<td>MT two thirds ((&gt; 2/3)) (</td>
<td>A \cap B</td>
<td>&gt; \frac{2}{3}</td>
</tr>
<tr>
<td>LT a half (&lt; 1/2) (</td>
<td>A \cap B</td>
<td>&lt; \frac{1}{2}</td>
</tr>
<tr>
<td>LT one third (&lt; 1/3) (</td>
<td>A \cap B</td>
<td>&lt; \frac{1}{3}</td>
</tr>
<tr>
<td>LT two thirds (&lt; 2/3) (</td>
<td>A \cap B</td>
<td>&lt; \frac{2}{3}</td>
</tr>
</tbody>
</table>

Table 1: Meaning of each signal in the model. MT=‘more than’, LT=‘less than’.

5 An alternatives account of ‘most’ vs ‘most than half’

5.1 An RSA model of the contrast

In the following, we will model communication with quantifiers by applying the RSA model described above to the following simple referential communication task, modelled after Pezzelle et al. [2018] where communication was set up similarly in a production task. A speaker observes two sets, \(A\) and \(B\), and attempts to communicate to a listener which proportion of \(A\) is also in \(B\) in the way modelled by the modified RSA model introduced above. As possible signals, we have chosen the Aristotelian quantifiers and some minimal set of alternatives for ‘more than half’. The literal meaning of each quantifier in the model corresponds to a portion of the scale of proportions (see [table 1]). The set of structural alternatives to ‘more than half’ is closed under substitution of ‘more’ by ‘less’, and by (semantically meaningful) substitutions of one, two, and three (both their cardinal and ordinal versions) with each other.

As in the modified RSA model presented above, the alternatives considered by the pragmatic listener depend on the speaker’s utterance. For instance, if the speaker uttered ‘some’ the listener would consider a set of alternatives containing ‘all’ but not ‘more than two thirds’, while if the speaker uttered ‘more than one third’ both ‘all’ and ‘more than two thirds’ would be

\[\text{Note that the meanings of the Aristotelian quantifiers can be obtained by exchanging } >, \geq, <, \text{ and } \leq \text{ with each other, and } A \cap B, A, \text{ and } A - B \text{ with each other. We exclude the meanings of ‘not all’ and non-conservative quantifiers, obtained by adding } B \text{ to the substitution source, as they are never lexicalized [Horn 1989, Barwise and Cooper 1981, Szymanik 2016], indicating that for reasons presently not fully understood they might not be valid conceptual alternatives.}\]
possible options for the listener. In the present case, the utterances above can be divided in two groups, the first containing ‘all’, ‘most’, ‘none’, and ‘some’, and the second containing the remaining utterances. Each utterance in the first group contains all other utterances in that group as alternatives, and none of the utterances in the second group. Each of the utterances in the second group contains all utterances in its set of alternatives.12

In order to isolate the effects of the account of alternatives discussed above from the consequences of utterance cost, we assume that signals have no cost. Moreover, to keep the results as simple as possible $S_2$ can only produce ‘all’, ‘most’, ‘none’, ‘some’, ‘more than a half’, and ‘less than a half’, rather than the full set of alternatives in table 1. In order for the speaker to be able to calculate a distribution over utterances given any state, there has to be at least one utterance to refer to each state in each set of alternative utterances.

The results of the model are shown in figure 4a. $L_0$ guesses uniformly within the categories expressed by each signal considered by $S_2$. $L_0$ treats ‘most’ and ‘more than half’ identically, guessing uniformly among the states between 51 and 100. Finally, $S_0$ selects the maximum for ‘every’ and the minimum for ‘none’.

With $S_1$, the set of alternatives for each signal matters (second plot from top in figure 4a). More specifically, while the lower bound for both ‘most’ and ‘more than half’ are similar for $S_1$, their upper bounds are different as a consequence of the different ways that the respective set of alternatives cover the scale. ‘More than half’ implicates less than two thirds, and therefore tends to not be used for proportions higher than two thirds, while most only implicates ‘not all’. Note that while the six signals are plotted together in figure 4a, the distribution for each signal is computed independently with a possibly different set of alternatives utterances. Therefore, $S_1$ does not suffice to explain the difference between ‘most’ and ‘more than half’.

$L_1$ tends to pick the central point in the categories as produced by $S_1$ (third plot from top in figure 4a). Therefore, $L_1$ tends to guess points closer to the middle of the scale for ‘more than half’ than for ‘most’, because the former is produced by $S_1$ for a range of proportions closer to the scale’s midpoint. Finally, the pragmatic speaker $S_2$ tends to pick ‘more than half’ for signals closer to the midpoint of the scale than ‘most’ (bottom plot in figure 4a).

The results in figure 4a, while qualitatively correct, are quantitatively surprising in that the upper bound of ‘more than half’ goes higher than in the data presented by Solt [2016]. However, the positions of the involved bounds are sensitive to the parameter values. For instance, figure 4b shows a

---

12Previous work has argued that utterances with different monotonicity profiles do not appear in the same set of alternatives [e.g. Horn 1989]. This observation would be contradicted by the set of sets of alternatives we consider. However, Katzir [2007] has argued the structural theory of alternatives can lift this restriction.
Figure 4: (a) The plots shows the results with $|A| = 100$, $\alpha = 3$ and $\rho = 1$. Each plot shows the behaviour of a different agent in the RSA model. Note that while the lines for $L_0$, $S_1$, and $L_1$ are shown on the same plot, they implicitly have different comparison sets. (b) The plots shows the results with $|A| = 100$, $\alpha = 4$ and $\rho = 3$. $L_0$ is not shown as it is identical to figure 4a.

parameter setting that makes predictions closer to Solt’s data. Moreover, as more proportions are included in the set of alternatives to ‘more than half’ the alternatives will divide the scale with a higher granularity, moving the upper bound of ‘more than half’ strictly closer to 50%.

5.2 Ignoring the mechanisms

In order to see what role each of the two mechanisms play in the predictions of the model above, it is instructive to observe the consequences of ignoring each of the two mechanisms.

When both mechanisms are ignored, the model reduces to the simple model presented in section 4.1. Bergen et al. [2016] has proposed RSA modelling as an alternative solution to the symmetry problem which does not need the structural account of alternatives. Specifically, it solved the symmetry problem by introducing different costs for different signals, using neither of the mechanisms we discussed in section 3. This raises the question of whether it is possible to account for the contrast between ‘most’ and ‘more than half’ only by assuming different costs for the different signals, without implementing the two mechanisms.

In order to study whether costs suffice to explain the difference between ‘most’ and ‘more than half’, we implement a simple RSA model as described in 4.1. The results are shown in figure 5. When both mechanisms are ignored and production costs are implemented, speaker $S_2$ does not introduce
any substantial innovation over the listener $L_1$, and therefore we stop the computation at the level of $L_1$. The conclusion is that costs are not enough to solve the problem at hand. The effect of cost in this setting is simply to make $S_1$’s production probability for ‘more than half’ uniformly lower than the one of ‘most’ for any given state. However, the difference in cost cannot be exploited by $L_1$ to draw an inference about the state observed by $S_1$. In sum, a difference in cost alone cannot be exploited by a pragmatic speaker to convey different information two truth-conditionally equivalent signals.

When only the structural account of alternatives is ignored, all utterances compete with each other, and therefore $S_2$ does not introduce interesting results. Therefore, figure 6a shows the result of such a change up to $L_1$. Again, since the symmetry is not broken by the different set of alternatives, ‘most’ and ‘more than half’ end up conveying identical information to $S_1$.

When only the distance-minimizing listener is ignored, we still obtain the crucial result that ‘more than half’ is generally used to convey proportions closer to the scale’s midpoint than ‘half’. However, the results differ from section 5.1 in two crucial ways. First, the speaker $S_2$ ends up producing each signal with uniform probability within the pragmatically enriched category, as shown in figure 6b. For instance, ‘more than half’ is produced with uniform probability for proportions above $1/2$ and below $1/3$. Second, the model without distance-minimizing $L_2$ predicts that the speaker would use ‘all’ and
Figure 6: (a) Results without the first mechanism. Since ‘most’ and ‘more than half’ are truth-conditionally equivalent, if they compete with the same set of alternatives they will use used interchangeably. (b) Results without the second mechanism: when the distance-minimizing listener is substituted with the pragmatic listener of equation 4, speaker $S_2$ has a uniform probability of producing ‘more than half’ across its whole pragmatically enriched domain. Only $L_1$ and $S_2$ are plotted as $L_0$ and $S_1$ are essentially the same as in plot 4a.

‘none’ exclusively for the maximum and minimum of the scale respectively.\textsuperscript{13} These two consequences contradict both Solt’s data and the data in Pezzelle et al. [2018]. In the data, the production probabilities for ‘more than half’ resembles a Gaussian distribution rather than a uniform distribution, ‘all’ and ‘none’ are used for signals close to the scale’s extremes rather than exclusively to the extremes.

In this section, we have shown that the two mechanisms discussed in section 3 are not only independently motivated, but are also both needed to make sense of the difference in typical sets between ‘most’ and ‘more than half’.

6 Experiment

In the previous sections, we presented an explanation for the difference in the typical proportions conveyed by ‘most’ and ‘more than half’. We implemented the proposed account in an RSA model of production, and showed that this model can qualitatively produce the observed effect. In this section, we present the results of a quantifier production experiment and analyse how well the RSA can fit them quantitatively.

\textsuperscript{13}Note that noise cannot play the same role as distance minimization in this context. Under the assumption that the effect of noise is independent of the true category, noise alone would not skew the distribution towards the points that truly belong to the category, but rather uniformly on the whole scale.
<table>
<thead>
<tr>
<th></th>
<th>(a) resp</th>
<th>(b) % targ</th>
<th>(c) n targ</th>
<th>(d) n non-targ</th>
<th>(e) n total</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1186</td>
<td>0.06 (0.2)</td>
<td>0.65 (2.53)</td>
<td>10.88 (5.27)</td>
<td>11.53 (4.92)</td>
</tr>
<tr>
<td>Few</td>
<td>3784</td>
<td>0.23 (0.14)</td>
<td>2.41 (1.75)</td>
<td>9.2 (5.0)</td>
<td>11.62 (5.46)</td>
</tr>
<tr>
<td>Less than half</td>
<td>2647</td>
<td>0.36 (0.13)</td>
<td>4.2 (2.35)</td>
<td>7.72 (4.0)</td>
<td>11.92 (5.12)</td>
</tr>
<tr>
<td>Some</td>
<td>1459</td>
<td>0.37 (0.19)</td>
<td>4.56 (2.79)</td>
<td>8.5 (4.56)</td>
<td>13.06 (5.07)</td>
</tr>
<tr>
<td>Half</td>
<td>1421</td>
<td>0.49 (0.1)</td>
<td>5.69 (2.7)</td>
<td>5.97 (2.92)</td>
<td>11.67 (4.96)</td>
</tr>
<tr>
<td>More than half</td>
<td>2762</td>
<td>0.63 (0.13)</td>
<td>7.27 (3.82)</td>
<td>4.13 (2.48)</td>
<td>11.4 (5.0)</td>
</tr>
<tr>
<td>Many</td>
<td>1449</td>
<td>0.7 (0.17)</td>
<td>9.82 (4.24)</td>
<td>4.22 (2.87)</td>
<td>14.04 (4.66)</td>
</tr>
<tr>
<td>Most</td>
<td>3517</td>
<td>0.77 (0.14)</td>
<td>9.67 (4.81)</td>
<td>2.81 (2.28)</td>
<td>12.48 (5.43)</td>
</tr>
<tr>
<td>All</td>
<td>1155</td>
<td>0.95 (0.19)</td>
<td>10.92 (5.33)</td>
<td>0.6 (2.6)</td>
<td>11.52 (5.02)</td>
</tr>
</tbody>
</table>

Table 2: Descriptive statistics for each signal in the experiment. This table matches Table 1 in Pezzelle et al. [2018]. Values in brackets refer to SD.

6.1 Task

The experiment is an almost exact replication of the ‘grounded task’ in Pezzelle et al. [2018] with a slightly different set of quantifiers. The original experiment was conducted in Italian, whereas our experiment was in English. The data was gathered on the Prolific\(^{14}\) platform and successfully obtained for 57 participants (43 females, 14 males), while 8 participants were excluded as they did not finish the experiment. 340 judgments were obtained for each participant, for a total of \((340*57=)\) 19380 data points. The experiment was coded in PsychoPy 3.2.4 [Peirce et al., 2019]. Since the experiment is described in detail in Pezzelle et al. [2018], we only report here the main design choices.

Each participant completed 340 rounds, each round consisting of three screens. The first screen, which lasted 500ms, only contained a fixation cross. The second screen, which lasted one second, showed objects arranged in a grid, with some possibly empty slots. The objects were a mixture of one type of animal and one type of artifact, the exact types varying across pictures. Each image contained between 3 and 20 (inclusive) objects. Finally, the third screen showed a grid of nine quantifiers: ‘most’, ‘more than half’, ‘all’, ‘half’, ‘many’, ‘none’, ‘less than half’, ‘few’, ‘some’ (the choice of quantifiers is the only difference in design to Pezzelle et al. [2018]).

6.2 Results and discussion

Table 2 reports summary statistics for each quantifier, aggregated across all participants. The results are comparable to table 1 of Pezzelle et al. [2018]. The order of the quantifiers that appeared both in Pezzelle et al. [2018], and in our experiment is exactly the same, namely ‘None’ < ‘Few’ < ‘Some’ < ‘Many’ < ‘Most’ < ‘All’. The percentages for which ‘None’ and ‘All’ were used are less extreme in our results (respectively 0.06 and 0.95)

\(^{14}\)https://www.prolific.co/
than in Pezzelle et al. [2018] (respectively 0.01 and 0.99). The reason for this difference is presumably that, for reasons explained below, we did not exclude any participant from the experiment, and therefore the production data is noisier. The proportions are close for the remaining quantifiers, especially ‘Few’ (0.23 in our vs 0.26 in Pezzelle et al. [2018]) and ‘Many’ (0.7 vs 0.64). In the case of ‘Some’ (0.37 vs 0.44), the average in our experiment is lower, indicating that ‘almost none’ in Pezzelle et al. [2018] moved ‘Some’ higher. Similarly, ‘Most’ (0.77 vs 0.69) is higher in our data, indicating that ‘almost all’ in Pezzelle et al. [2018] moved ‘Most’ lower.

Figure 7 shows the data for each quantifier aggregated across participants, for some subsets of the data. Figure 7a shows the results for all participants, which as discussed above are similar to Pezzelle et al. [2018]. Figure 7b shows the aggregated data with more than 3 animals and more than 3 artifacts. The distributions for the signals in 7b is close to 7a except for ‘None’ and ‘All’, which shows random behaviour. This is expected, as the data in 7b excludes all the stimuli where ‘None’ and ‘All’ apply and therefore production of those signals comes from noise. Figure 7c shows the data with fewer than 4 targets (animals). While ‘None’ is produced correctly, ‘All’ is as expected noisy. A similar effect, although to a smaller degree, is seen for ‘Many’ and ‘Most’. In figure 7d, the reversed pattern is seen for ‘None’, which is as expected noisy, and to a lesser extent for ‘Some’. Overall, our results are similar to the results in Pezzelle et al. [2018] for the signals shared by the two experiments.

The quantifiers which were not in Pezzelle et al. [2018] show the expected behaviour. ‘Few’ is lower than ‘Less than half’, and they are respectively close to ‘Almost none’ and ‘The smaller part’ in Pezzelle et al. [2018]. ‘Half’ is, as expected, centered around the midpoint of the scale.

The data is shown in greater detail in figure 8, which shows the number of times each quantifier was used for each stimulus, aggregated across participants. The y-axis of the figure represents the total number of objects, the x-axis the number of target objects. This way of representing quantifiers in a triangle originates from van Benthem [van Benthem, 1986, 1987]. A cardinal quantifier used to express ‘between \(a\) and \(b\)’ (with \(a, b \in \mathbb{N}\)) would appear in the plot as a group of light squares between the vertical lines Target = \(a\) and Target = \(b\). On the other hand, a proportional quantifiers used to express ‘between \(a\) and \(b\)’ (with \(a, b \in [0, 1]\)) appears as a group of red squares between the lines Target = \(a \times \text{Total}\) and Target = \(b \times \text{Total}\). All the quantifiers’ lower and upper bounds are roughly straight lines in the plot. This shows that the quantifiers were interpreted proportionally, i.e. did not depend on the absolute number of objects on the screen but only on the proportion between the total number and the number of target objects. This is important mainly in the case of ‘Few’ and ‘Many’, which have been argued to be ambiguous between a cardinal and a proportional interpretation.

The crucial result is that the difference between ‘most’ and ‘more than
Figure 7: Kernel density estimation of data aggregated across participants for various subsets of the data. (a) All data. (b) Data where both the number of targets (animals) in the picture nor the number of non-targets (artifacts) was greater than 3. (c) Data with 3 or fewer target objects (animals). Y-axis for ‘None’ is scaled down for visualization purposes. (d) Data with 3 or fewer non-target objects (artifacts).

half’ observed by Solt [2016] in the corpus is reproduced in our experiment. More precisely, three of Solt [2016]’s predictions were verified. First, the approximate lower bound of ‘More than half’ is right above 0.5, as observed by Solt [2016]. Second, the upper bound of ‘More than half’ roughly corresponds to the lower bound of ‘Most’. Third, the upper bound of ‘Most’ is close to 100%. However, the upper bound of ‘More than half’ is higher than the one observed by Solt [2016] in the corpus. We return to this point in more detail below.

In this section, we presented a partial replication of Pezzelle et al. [2018] with a different set of quantifiers modified to study the question in this paper. The main insight of Solt [2016] was confirmed, although the exact position of the thresholds of ‘most’ and ‘more than half’ are different. In the next section, we connect the experimental data and the RSA model described in section 4 in a Bayesian hierarchical model.
Figure 8: Graded van Benthem triangle. The plot shows, for each combination of total number of objects (targets+non-targets) and number of target objects, the counts of usages of each quantifier aggregated across participants. The completely white squares are combinations that none of the participants saw. The black squares are observations for which the quantifier was never produced, and squares become lighter with increasing number of uses. A specific proportion can be imagined as a straight line starting from (0,0) (outside of each plot). The bottom right plot is an illustration of the proportion 0.5. As expected for proportional quantifiers, the upper and lower bounds for each quantifier roughly follow straight lines with various inclinations.

7 Model fitting

7.1 Extending the production model

The production model presented above consisted of an RSA model with 10 signals, 6 of which could be produced by the speaker and 4 of which were only implicitly considered as alternatives. In order to use the RSA model developed in section 4 to fit the experimental data, the language has to be slightly enriched to include the signals in the experiment. The signals included in the model are the ones in table 1 plus the ones in table 3. In the experimental production model, we include fourths, effectively increasing the granularity of the set of alternatives to ‘more than half’. In our opinion, this makes the model more realistic, but in the future experimental work can try to directly elicit cognitively plausible alternatives.

Based on the conceptual structure arguments discussed above, we retain two groups of alternatives as follows: ‘All’, ‘Most’, ‘None’, ‘Some’, ‘Half’, ‘Many’, and ‘Few’ are all alternatives of each other, and do not have any of the other signals as alternatives. The remaining signals are all alternatives of each other, as well as all the signals in the first group. For instance, using the notation of Katzir [2007] introduced in equation 3, ‘most’/‘all’/...
Table 3: Meaning of signals added to the ones in table 1 for the experimental data fitting. MT='more than', LT='less than'.

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Structure</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half (1/2)</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>Many</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>Few ($\neg \exists$)</td>
<td>$</td>
<td>A - B</td>
</tr>
<tr>
<td>MT one fourth ($&gt; 1/4$)</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>MT three fourths ($&gt; 3/4$)</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>LT one fourth ($&lt; 1/4$)</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>LT three fourths ($&lt; 3/4$)</td>
<td>$</td>
<td>A \cap B</td>
</tr>
</tbody>
</table>

'Throughout, we refer to the following alternatives as 'more than half'/'less than three quarters'/

This way of structuring the set of alternatives assumes that 'half' is analysed as $|A \cap B| = |A - B|$ rather than $|A \cap B| = \frac{1}{2}|A|$, since the latter conceptual structure would imply that the alternatives to 'half' include 'more than half' etc. However, simulations show that this choice does not have a substantial influence on the resulting production behaviour. Moreover, the setup assumes that 'many' and 'few' do not conceptually encode a comparison with precise proportions, which could conceptually compete with concepts such as 'three quarters'. We leave to future work an analysis of the implications of this choice.

In addition to specifying the alternatives set for each quantifier, the model requires a specification of their literal meanings. While most of the quantifiers we included in the experiment have a default interpretation in the literature, 'few' and 'many' are generally taken to be vague and to lack a precise threshold. In order to keep things simple, we attribute 'few' a hard threshold at 0.2 and 'many' a hard threshold at 0.4. We leave to future work to more precisely determine the position of the thresholds for these two quantifiers, or to implement participant-wise estimation of their thresholds from the data. The meaning of 'half' is also not trivial. In order to have 'half' refer to at least one state in every picture, we encode 'half' as being verified in a situation with a total number of objects $n$ by the number(s) of target objects $i$ such that $i/n$ is at least as close to 0.5 as any other number:

$$[\text{half}] (A)(B) \iff |A \cap B| = \arg \min_i \left| \frac{i}{|A|} - 0.5 \right|$$

For instance, if there are 4 objects in total, 'half' applies iff there are 2 animals. On the other hand, if there is a total of 5 objects, 'half' applies if there are either 2 or 3 objects.\footnote{While this analysis of 'half' has a substantial impact on the usage for small proportions, the effect becomes negligible for enough total number of objects.}

Up to this point, we have considered the production behaviour of a
rational RSA agent. However, behaviour of real participants will not perfectly conform to the RSA model. First, there will be systematic error from aspects of quantifier usage that is not captured by the RSA model. Second, there will be noise coming from participants pressing the wrong button or not paying attention. In order to account mainly for the latter kind of error, we add production noise in the model. The production noise introduces a third production noise parameter $\epsilon$ for each participants, in addition to the RSA $\alpha$ and $\rho$ parameters. We give the following generative characterization of production noise. Assume the participant has made an observation and has calculated a posterior distribution over quantifiers given the observation. Then, the participant’s response is sampled from the RSA posterior distribution with probability $1 - \epsilon$, and sampled from a uniform distribution over quantifiers with probability $\epsilon$. Therefore, the production noise is a mixture with the RSA distribution and a uniform distribution as components, and $\epsilon - 1$ and $\epsilon$ as the mixture weights respectively:

$$f_E(S_2, \epsilon) = (1 - \epsilon)S_2 + \epsilon \frac{1}{9} I_9$$

where $S_2$ is the vector of production probabilities according to the RSA model and $I_9$ is a vector of ones of length 9, i.e. the number of signals in the model. Intuitively, the greater the value of $\epsilon$ the noisier the behaviour of the participant, i.e. the less the participant’s decision depends on the observed state. When $\epsilon = 0$, the participant’s behaviour is fully determined by the RSA predictions. When $\epsilon = 1$, the participant selects a completely random quantifier by clicking a random button.

In sum, the behaviour of a participant $p$ producing a judgment about a stimulus $s$ is modelled by two functions, namely an RSA function and a noise function $f_E$. The RSA function takes four parameters: (1) the participant’s alpha parameter $\alpha_p$, (2) the participant’s distance-minimization parameter $\rho_p$, (3) a total picture size $a_s$, and (4) the number of target objects $b_s$. The output of RSA, $\text{RSA}(\alpha_p, \rho_p, a_s, b_s)$, is the input of the error function which calculates the mixture probability of each quantifier $f_E(\text{RSA}(\alpha_p, \rho_p, a_s, b_s), \epsilon_p)$. Thus, the probability of the participant producing each quantifier is determined by the output of the error function.

7.2 RSA based Bayesian models

The hierarchical model, displayed as a Bayesian directed acyclic graph in figure[9] has three nested levels. The bottom level is the level of specific participants’ judgments for specific stimuli. The participant at index $i$ is indicated by $P[i]$ and the stimulus at index $j$ for participant $i$ is indicated by $S[i, j]$. Index $i$ takes 57 values, and for each participant $j$ takes 229 values. The bottom level has therefore a total of $57 \times 229 = 13053$ values. The quantifier produced
Figure 9: DAG for the Bayesian model. See section 7.1 for definitions of functions RSA and \( f_E \). We use the following parameterizations: the normal distribution \( \mathcal{N} \) takes a mean \( \mu \) parameter and a \( \sigma \) parameter, the halfnormal distribution \( \mathcal{HN} \) only a \( \sigma \) parameter, truncated normal distribution \( \mathcal{T\mathcal{N}} \) a \( \mu, \sigma \), and lower bound parameters in this order, the beta distribution \( B \) parameters \( \alpha \) and \( \beta \) in this order, and the categorical distribution a probability vector.

by participant \( P[i] \) for stimulus \( S[i, j] \) is sampled from a categorical distribution with the production probability vector parameter calculated, in the way described in the previous section, as \( f_E(RSA(\alpha_{P[i]}, \rho_{P[i]}, a_{S[i,j]}, b_{S[i,j]}, \epsilon_{P[i]})) \).

The bottom level also includes the two properties of the stimuli that are relevant to production, namely the total number of objects \( a_{S[i,j]} \) and the number of target objects \( b_{S[i,j]} \) in the \( j \)th round for participant \( i \).

The middle level of the hierarchy is the level of the individual participants \( P[i] \), where \( i \) takes on 57 values. Three parameters are associated with the participant, namely the two RSA parameters \( \alpha_{P[i]} \) and \( \rho_{P[i]} \) and the error parameter \( \epsilon_{P[i]} \). These three parameters control the predicted behaviour at the bottom level, and are themselves sampled from population-level distributions at the top level.

The top level is the population-level, from which the participant-level parameters are sampled. The top level contains six distributions: a distribution over the (1) \( \mu \) and (2) \( \sigma \) parameters for the distribution of each participant’s RSA \( \alpha \) parameter, a distribution over the (3) \( \mu \) and (4) \( \sigma \) parameter of
the distribution of each participant’s RSA $\rho$ parameter, and a distribution
over the $\alpha$ and $\beta$ parameters for the distribution of each participant’s error
parameter $\epsilon$.

Overall, the generative model is as follows. First, an $\alpha$, $\rho$, and $\epsilon$ para-
meters are drawn for each participant. Then, the RSA production probabilities
of each signal are calculated for each stimulus observed by each participant,
taking into account the participant’s parameters as well as the stimulus’
properties (number of target objects and total number of objects). Then, the
RSA production probabilities are disturbed by adding noise. Finally, each
participant samples a quantifier for each stimulus with the noisy production
probabilities.

The prior values for the population-level distributions can be seen in
figure 9, and are visualized in figure 10a. We chose weakly regularizing priors
which included a variety of possible behaviours. The prior predictions at
the level of single participant’s quantifier production behaviour are shown in
figure 10.

In addition to the model with structural account of conceptual alternatives
described above, we fit a model without the structural account of conceptual
alternatives, where each signal has all and only the other signals seen by the
participants as alternatives. The 95% HPD intervals for the marginalized
prior production probabilities are shown in figures 10d and 10e. The model
without structural alternatives has the same hyperprior parameters as the
model presented above. This model differs from the model with structural
alternatives in the predicted production probabilities. First, ‘most’ and ‘more
than half’ are used in exactly the same way in the model without structural
alternatives. Given the difference in usage between the two expressions that
can be seen in the raw data in figure 7, a lack of predicted difference will
diminish the fit to the data. We perform model comparison to quantify the
difference in fit between the two models. The second main difference between
the two models is in the predicted behaviour of ‘some’. Since ‘more than half’
is now used for higher proportions, and ‘some’ also applies to proportions
higher than half, the prediction of the model without structural alternatives
is that ‘some’ will be used for proportions slightly higher than half. Since
the peak of ‘half’ is very stable across prior samples, even in the marginal
distribution ‘some’ has a dip at 0.5.

It is worth noting the way that the noise mechanism affects the estimation
for both models. If the noise parameter for a participant is high, the partici-
plant’s behaviour will depend less on their $\alpha$ and $\rho$ parameters. Therefore,
the participant’s data will give less information about those parameters,
influencing less the population-level estimates. On the other hand, since the
behaviour and the underlying parameters will be less tied to each other for
a noisy participant, the hierarchical model’s estimation of the participant’s
RSA parameters will depend more on the population-level distributions. The
population-level RSA distributions will therefore be impacted less by data
Figure 10: Prior predictive simulations. (a) Marginal distribution of prior samples for various model parameters. (b) 95% HPD interval for $S_2$'s behaviour for each signal. (c) 95% HPD interval for predicted participant behaviour, i.e. RSA agent with added noise. The main effect of adding noise is an increased probability of producing signals outside of the usual range of usage of the quantifier. Note that prior predictions of noisy production behaviour includes nearly uniform languages, which appear to be close to 0, because uniform production probabilities are small for each state. Therefore, despite (b) and (c) looking similar, they include substantially different predicted production behaviours. (d) and (e) plot the same information as (b) and (c) respectively, for the model without the structural account of conceptual information. Crucially, in (d) and (e) ‘most’ and ‘more than half’ are used identically.
of participants estimated to be noisy, and in turn will play a bigger role in estimating the individual-level RSA parameters for such participants. As well as being an intuitive way to deal with noisy participants, this mechanism eliminates the need to define arbitrary criteria for data exclusion.

In this section, we described how we embed the RSA model within a hierarchical Bayesian model whose hidden parameters can be fitted to experimental data. We discussed two models which can be compared, with and without the structural account of alternatives. In the next section, we present the results of fitting for the two models and their comparison.

7.3 Model fitting and results

We fit the models with the Python library PyMC3 [Salvatier et al., 2016], which implements a NUTS sampler to fit hierarchical Bayesian models. In order to reduce the effect of subitation while still including production data of ‘all’ and ‘none’, as well as reducing computation time, we only fit the model on responses for stimuli where the total number of objects in the picture was greater than or equal to 10. As a result, we have 229 responses for each participant, for a total of 13053 responses. We fit two chains for each of the two model in order to perform convergence checks. We drew for each chain 3000 NUTS tuning samples and 2000 non-tuning samples.

For both of the models, the posterior distributions of all the population-level parameters are substantially more precise than the prior distributions, indicating that the data contained substantial information about the underlying parameters (fig 11). Plot of the joint distributions does not indicate any strong correlation between population-level distributions.

The \( \hat{r} \) convergence statistics is close to 1 for all the population-level estimates of both the model with both mechanisms and the model without the structural account of alternatives (‘w/o sa’), indicating that the sampling converged to the posterior:

<table>
<thead>
<tr>
<th></th>
<th>( \mu_\alpha )</th>
<th>( \sigma_\alpha )</th>
<th>( \alpha_\epsilon )</th>
<th>( \beta_\epsilon )</th>
<th>( \mu_\rho )</th>
<th>( \sigma_\rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both</td>
<td>1.000229</td>
<td>0.999937</td>
<td>0.999665</td>
<td>1.000097</td>
<td>0.999755</td>
<td>1.000751</td>
</tr>
<tr>
<td>w/o sa</td>
<td>1.000297</td>
<td>0.999968</td>
<td>1.000156</td>
<td>1.000085</td>
<td>1.000280</td>
<td>0.999982</td>
</tr>
</tbody>
</table>

We compare the models with and without the structural account of alternatives with the Watanabe-Akaike Information Criterion (WAIC) [Watanabe, 2013; Gelman et al., 2014] (fig 12), an information criterion that considers the deviance for all posterior samples and is apt for Bayesian analyses. The result of the comparison is that the model with both mechanisms is estimated to have a higher out-of-sample predictive accuracy (WAIC \( \approx 35584 \), \( SE \approx 235 \)) than the model without the structural account of alternatives (WAIC \( \approx 36803 \), \( SE \approx 211 \)), with a difference in WAIC of...
approximately 1218 (SE ≈ 83). This indicates that the structural account of alternatives plays a crucial role in explaining the difference in usage between ‘most’ and ‘more than half’. For both model, the posterior variance of the log predictive density exceeded 0.4 for fewer than 30 of the 13279 observations, and within those was mostly close to 0.4 and never greater than 1.

A crucial question about the two models is how well they predict the participants’ behaviour with respect to each of the signals. Figure 13 shows, for each signal, the posterior distribution of the difference between the deviance of the two models across all data points. Predictably, the model including structural alternatives performs much better than the other for ‘most’ and ‘more than half’. However, the in-sample predictive accuracy of the two models differs starkly also for the other signals. This might be a consequence of the model without structural alternatives compromising the fit of the other signals while trying to find parameter values appropriate for ‘most’ and ‘more than half’.

Figure 14 shows the posterior predictive simulations. Behaviour for a new participant is predicted by sampling a set of individual-level parameters from the population-level distributions of each posterior sample. For both models, the predictions for a new participant are more precise than the predictions from the prior shown in figure 10b. In the case of the model with the structural account of alternatives (fig 14a), the difference is particularly stark for ‘more than half’, ‘most’, and ‘many’. For the model without the structural account of alternatives, the biggest difference from the prior to the
Figure 12: Comparison of the model with both mechanisms (‘Both’) and the model without structural alternatives (‘W/o sa’) with the WAIC. The plot is on the deviance scale, where smaller values indicate a better fit. The empty circles are the WAIC values for the two models, with their associated standard deviations shown as black bars. The black dots show the in-sample deviance of each model. Finally, the triangle and its error bar show the standard error of the difference between the WAIC of the top ranked model, i.e. the one with both mechanisms, and the WAIC of the model without structural alternatives. The main result in the plot is that the model with both mechanisms performs much better than the model without the structural account of alternatives.

Figure 13: Distribution of differences of posterior signal-wise deviance, marginalized across participants. Positive values indicate better in-sample fit for the model with both mechanisms. The model with the structural account of alternatives performed better for all signals except ‘Many’ and ‘Half’.

posterior is in the distributions of ‘some’ and ‘many’. Moreover, since the marginal distribution of the production error tends more towards 0 in the posterior, adding the production error has a smaller impact for the posterior than the prior.

In this section, we have presented two Bayesian hierarchical models encoding two minimally different pictures of how quantifiers are produces, with and without the structural account of alternatives. In sum, model comparison lends strong support to the model which includes the structural account of conceptual alternatives over a minimally different model without it. Moreover, the model with structural alternatives has a closer fit to the data not only for the signals we have discussed focused on—‘most’ and ‘more than half’—but also for most of the other signals, a consequence of the fact that in an RSA model the production distributions for the signals are
Figure 14: Posterior predictive simulations (without production error) for a new participant, for 20 total number of objects. We predict a new participant in this plot by sampling a set of individual-level parameters from the population-level distributions defined by each posterior sample and then calculating $S_2$. The left column of plots shows the 95% HPD interval of production probabilities, for the model with both mechanisms and without the structural account of alternatives. Adding production error does not make a substantial visual difference in the plots, as the predicted production error is generally low. The right plots show the distribution of the means of the population-level distributions for the plots on the right.

In the next section, we compare our model to Solt [2016]'s account.

8 Comparison with Solt [2016]'s account

In the previous section, we discussed and compared two models with respect to their ability to fit the experimental data, namely a model with and without the structural account of alternatives. It would have been desirable to directly compare the two models with the account proposed in Solt [2016].
However, it is unclear how the latter could be implemented in a generative Bayesian cognitive model in something like the production task presented above. The fundamental problem is that Solt’s account relies on the difference between information on a ratio scale and information on a semi-ordered scale obtained through the approximate number system (henceforth ANS) [Dehaene, 1999]. However, in our experiment almost all the data is approximated via the ANS (except for observations in the subitation range to which we return below), making it unclear what predictions Solt would make.

More specifically, in Solt’s account the signals that the speaker can produce partially depend on the type of available data. In particular, in a literal reading of Solt’s account ‘more than half’ could not be used for observations on a scale with less structure than a ratio scale. This contradicts the data gathered in our experiment, where the expression is used by participants even for stimuli that we know require the ANS. It is unclear what Solt’s theory predicts in such cases. A more sophisticated interpretation of Solt’s account could be developed to allow a computational implementation, but at the moment it is unclear how to do so.

This problem was reduced by Solt’s original corpus data, where the two quantifiers always occurred together with a percentage—Solt’s inclusion criteria for determining the typical sets of the two expressions—effectively weakening the reliance on approximate knowledge. However, the dual problem, namely why ‘most’ is used so often when precise proportions are available, is relevant for Solt’s data. Occurrences with proportions are precisely those situations where amounts are known precisely, excluding those scales which according to Solt are most typical of ‘most’, namely non-ratio scales. Solt argues that ‘most’ is used even in these cases as if the scale was semi-ordered, since the approximate meaning of ‘most’ becomes an R-implicature. However, this move further detaches the input to the perceptual system and the predicted output, making it harder to extract a production model from Solt’s account.

While the ANS plays a crucial role in Solt’s account, we did not include it in our model. The reason for this is that the ANS was not necessary for our analysis, i.e. the advantage of including the structural account of alternatives was not conditional on including the ANS, allowing us to keep the model simple. It is prima facie tempting to think that the ANS could play the same role as the distance-minimizing listeners, essentially allowing the speaker to produce a signal for states that are not compatible with its literal meaning but close to it. However, this substitution is not as simple as it might first seem, because the ANS can only perform the function of distance-minimization for stimuli outside the subitation zone.

For stimuli within the subitation zone, e.g. those stimuli where ‘None’ applies, the ANS predicts that participants would only produce signals that literally applied to the stimulus. In contrast to the ANS, the distance-
minimizing mechanism applies to every proportion, and predicts correctly that participants sometimes use ‘All’ and ‘None’ for non-extreme, albeit close to extreme, proportions. In order to substitute the distance-minimizing listeners with the ANS, an additional mechanism could be added to the generative model so that participants might miss some of the target stimuli when many non-target stimuli are shown (or viceversa), allowing e.g. for perceptual confusion between 20/20 (target/total) and 19/20. This would allow for misapplication of the literal meaning within the substitution zone, while however further complicating the model. In conclusion, while the distance-minimizing listeners do not play as crucial a role as the structural account of alternatives and alternatives could be developed for the former, we leave such developments for future work.

Another point where our and Solt [2016]’s account differ is on the difference between ‘more than half’ and ‘less than half’. In Solt’s account, the two are predicted to be symmetric. Solt argues that pragmatic competition between ‘most’ and ‘more than half’ is irrelevant to explain the difference between them: all that is required is that ‘more than half’ is upper bounded by other proportions. In a similar way, ‘less than half’ is bounded below by smaller proportions even in the absence of a dual to ‘most’. Therefore, the case of ‘less than half’ is symmetric to the case of ‘more than half’. On the other hand, in our model two factors contribute to the production behaviour of ‘most’ and ‘more than half’. First, the pragmatic competition with a quantifier’s own set of alternatives for $S_1$ and $L_1$. Second, the competition of each quantifier with the other real alternatives at the level of $S_2$. While implicit alternatives like ‘more than 3/4’ might lower the upper bound of ‘more than half’ compared to ‘most’, if ‘most’ was not an available option ‘more than half’ would be used by $S_2$ for higher proportions for lack of a better signal. Our model therefore implies that the two expressions pragmatically compete with each other: ‘more than half’ is upper bounded by the lower bound of ‘most’, and viceversa. Since there is no equivalent of ‘most’ for points below 0.5, a prima facie consequence of our account is that ‘more than half’ and ‘less than half’ should not be symmetric. However, the situation is further complicated by ‘some’ competing with ‘less than half’ at the level of $S_2$ in a way similar to how ‘more than half’ competes with ‘most’. As can be seen in figure 13, ‘some’ is better explained by the model with structural alternatives, and albeit to a lesser extent this is also true of ‘less than half’.

Overall, our model offers various advantages over Solt’s account. First, it is a quantitative rather than a verbal model, and can be used directly to analyse or predict behaviour as we have done with experimental data above. Second, it only uses independently motivated mechanisms, whereas Solt’s account introduces a novel claim about the way scale structures affect quantifiers usage. Third, it offers a unified explanation for bounds of both ‘most’ and ‘more than half’, in contrast to Solt’s disjunctive account. Fourth, our model is not tailored specifically for the opposition between ‘most’ and
‘more than half’, but can rather fit quantifier production behaviour more generally as demonstrated above.

Despite these advantages of our model, we have discussed only one of the features analysed by Solt [2016], namely the proportions for which the quantifiers are used. Solt’s account makes sense of other aspects of the distribution of ‘most’ and ‘more than half’ in the corpus. For instance, Solt finds difference in the way the two phrases behave with respect to generics, noun phrase structure, kind vs. group nominals, and vagueness. Future research will investigate how much of these further differences can be accounted for by the model presented here.

9 Conclusions

‘Most’ and ‘more than half’, while traditionally assumed to be truth conditionally equivalent, are typically associated with different proportions. In the most developed explanation of this difference, Solt [2016] introduces a difference between the structures of the scales used by the two expressions. In contrast, in this paper we proposed a novel account of the difference that is based on independently motivated mechanisms and does not rely on different scale structures. Moreover, we analysed the predictions of the account by implementing it in a popular computational model of pragmatic reasoning, the RSA model. Finally, we presented a replication of Pezzelle et al. [2018], and fitted our model to the quantifier production data. We found that our model explained the data better than a minimally similar model without the structural account of alternatives.

The RSA model we presented can be extended in various possible directions. First, a similar model could be used to account for the usage of modified numerals, since a similar contrast to the one discussed here can be found e.g. between ‘more than 100’ and ‘more than 101’, where the typical guessed number for the former utterance is higher than the for the latter. Another possible development would look at whether the model predicts that the quantifiers’ thresholds stay the same even in downward entailing contexts, as suggested by the experimental data in Denić and Szymanik [2020]. Lastly, in the models presented in this paper we only considered a small set of alternative proportions for ‘half’. However, it would be valuable to study the predictions of the model when more alternative utterances, containing more complex proportions, are included. The hierarchical Bayesian model can also be extended in various ways, e.g. by implementing more alternative accounts of the difference between ‘most’ and ‘more than half’ and comparing them to our account. We leave all these exciting possible developments to future work.
Acknowledgments

The authors have received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ.
References


Mark Davies. Corpus of Contemporary American English (COCA), September 2017.


Milica Denić and Jakub Szymanik. Differences in thresholds between ‘most’ and ‘more than half’: Semantics or pragmatics? In *Sinn Und Bedeutung 25*, London, July 2020. OSF.


