Edinburgh Research Explorer

Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation

OPEN ACCESS

Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation

Paula Rubio-Fernández[1,2]
Francis Mollica[3]
Julian Jara-Ettinger[4,5]

[1]Department of Philosophy, University of Oslo
[2]Department of Brain and Cognitive Sciences, MIT
[3]School of Psychological Sciences, University of Melbourne
[4]Department of Psychology Yale, University
[5]Department of Computer Science, Yale University

Abstract

Pragmatic theories and computational models of reference must account for people's frequent use of redundant color adjectives (e.g., referring to a single triangle as 'the blue triangle'). The standard pragmatic view holds that the informativity of a referential expression depends on pragmatic contrast: color adjectives should be used to contrast competitors of the same kind to preempt an ambiguity (e.g., between two triangles of different colors), otherwise they are redundant. Here we propose an alternative to the standard view, the incremental efficiency hypothesis, according to which the efficiency of a referential expression must be calculated incrementally, over the entire visual context. This is the first theoretical account of referential efficiency that is sensitive to the incrementality of language processing, making different cross-linguistic predictions depending on word order. Experiment 1 confirmed that English speakers produced more redundant color adjectives (e.g., 'the blue triangle') than Spanish speakers (e.g., 'el triángulo azul'), but both language groups used more redundant color adjectives in denser displays where it would be more efficient. In Experiments 2a and 2b, we used eye tracking to show that pragmatic contrast is not a processing constraint. Instead, incrementality and efficiency determine that English listeners establish color contrast across categories (BLUE SHAPES > TRIANGULAR ONE), whereas Spanish listeners establish color contrast within a category (TRIANGLES > BLUE ONE). Spanish listeners, however, reversed their visual search strategy when tested in English immediately after. Our results show that speakers and listeners of different languages exploit word order to increase communicative efficiency.

In foundational work, Zipf (1949) proposed that language is shaped by the competing pressures of minimizing the production costs for the speaker and the comprehension costs for the listener. In the last two decades, Zipf's Principle of Least Effort has been formalized in information-theoretic terms and evaluated positively against the structure of linguistic representations (see also Givón, 1979), with research in cognitive science confirming that language is shaped by a pressure to minimize complexity and communicate efficiently (Kemp, Xu, & Regier, 2018; Gibson et al., 2019).

The idea that language is optimized for fast, easy and reliable information transmission has been highly influential in linguistics, explaining language structure at a phonetic (Aylett & Turk, 2004), morphological (Piantadosi, Tily, & Gibson, 2011), semantic (Kemp & Regier, 2012) and syntactic level (Gibson et al., 2013). Beyond structural properties, this information-theoretic approach also explains language use, providing normative and descriptive accounts for why speakers ought to convey meaning efficiently (Jaeger, 2010; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013), and how listeners, by holding speakers accountable to this expectation, can go beyond literal meanings and infer things that were left unsaid (Grice, 1975). It follows from this view that if someone asked you 'Could you pass me the blue bowl?', you would assume that there is more than one bowl to choose from, and that not all bowls are blue. Yet if there was only one bowl, the adjective 'blue' would be redundant (as it would not preempt an ambiguity between several bowls) and the speaker would have failed to communicate efficiently.
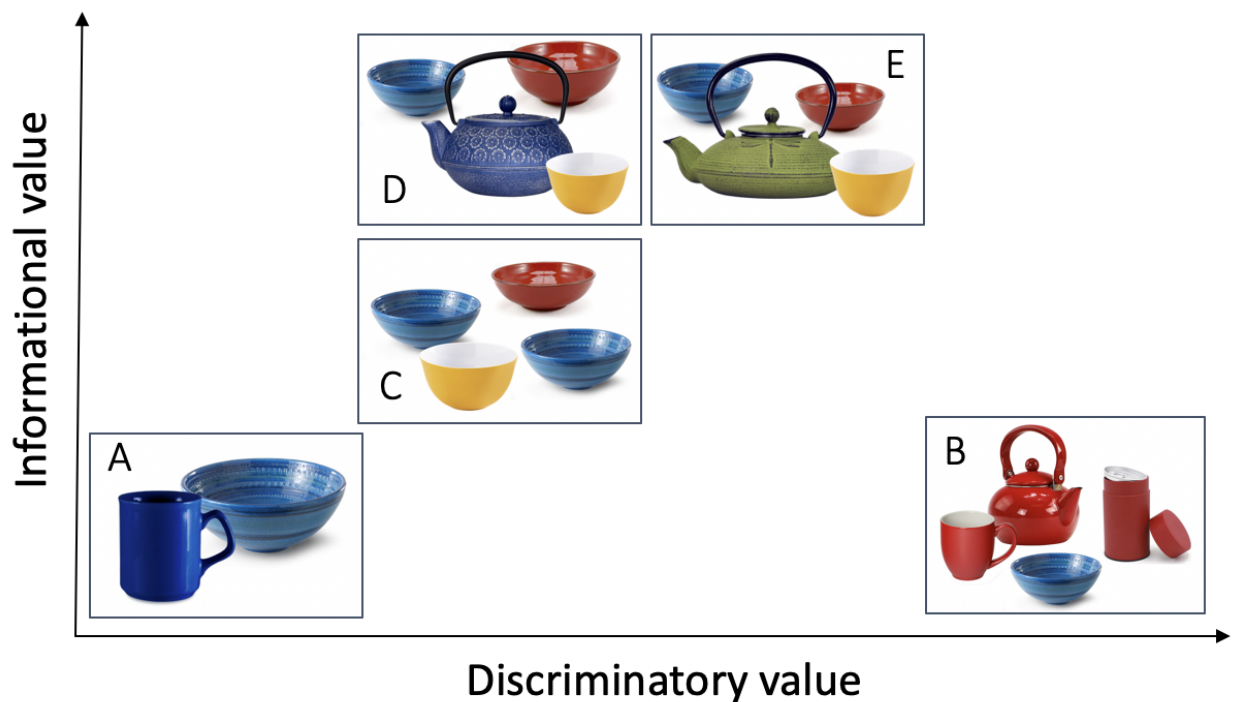
Figure 1: Visualization of the informational vs. discriminatory value of the color adjective 'blue' in the description 'the blue bowl'. In Displays A and B, 'blue' has no informational value because there is only one bowl (i.e. color does not preempt an ambiguity). In Display C, 'blue' has more informational value because there are several bowls, but not enough to secure unique reference amongst the competitors. In Displays D and E, 'blue' has the highest informational value because it resolves reference amongst all bowls. Regarding discriminatory value, 'blue' has none in Display A because both objects are blue, whereas it has more in Displays C and D because two of the four objects are blue. In Display E, 'blue' has higher discriminatory value because there is a single blue object, and it has even more in Display B because it relies on a pop-out effect (i.e. the uniform color of the other objects makes the color blue stand out).

This information-theoretic account has great explanatory power, yet it does not always map onto the experimental record. A large number of studies challenge in fact its most basic predictions: speakers are often redundant or over-specific, using descriptive color adjectives that do not add necessary information (Sedivy, 2003, 2005; Maes, Arts, & Noordman, 2004; Engelhardt, Bailey, & Ferreira, 2006; Van der Sluis & Krahmer, 2007; Arts, Maes, Noordman, & Jansen, 2011a, 2011b; Engelhardt & Ferreira, 2014; Rubio-Fernández, 2016; Rubio-Fernandez, 2019). These unexpected results are normally taken to challenge the Gricean Maxim of Quantity, whereby speakers should not provide more information than is necessary for the purposes of the exchange (Grice, 1975). In response to this puzzle, some have argued that speakers use color adjectives without checking if they are necessary in the context (Pechmann, 1989; Belke & Meyer, 2002; Belke, 2006; Koolen, Goudbeek, & Krahmer, 2013), or alternatively, that they do so strategically to preempt a possible ambiguity (Frank & Goodman, 2012; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Hawkins, Gweon, & Goodman, 2018). Here we adopt the view that redundant color adjectives can facilitate the listener's visual search for a referent, making their use rational and efficient.

In face-to-face interaction, a cooperative speaker should ensure that the listener not only identifies the referent, but also does so rapidly and easily in the visual context (Rubio-Fernández, 2016; Rubio-Fernandez, 2019). In this view, referential expressions ought to be analyzed in terms of efficiency, rather than purely in terms of informativity (see also Ramscar & Port, 2016). A referential expression's efficiency depends not only on the likelihood of communicative success, but also on the processing effort it requires. Therefore, a color adjective may have informational value if it preempts an ambiguity between several competitors of the same kind (e.g., various bowls of different colors), but it may also have discriminatory value if it facilitates
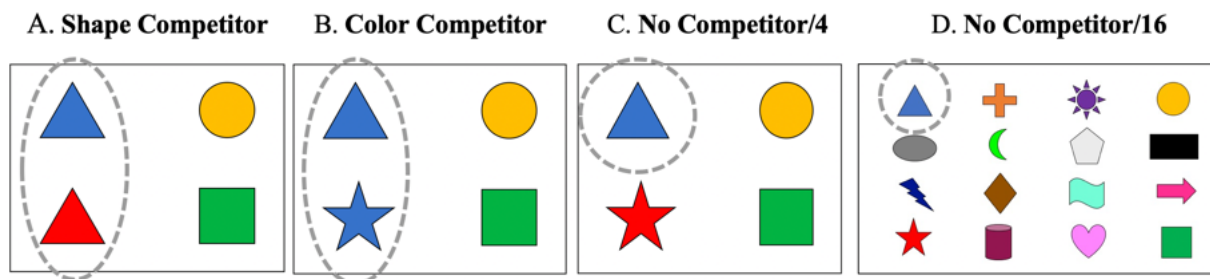
Figure 2: Sample displays from the critical conditions in the study. The target was the blue triangle in all displays. Experiment 1 used the No Competitor/4 and No Competitor/16 conditions in a language-production task, and Experiments 2a and 2b used the Shape Competitor and Color Competitor conditions in two eye-tracking tasks.

the listener's visual search for the referent (e.g., if the bowl is the only blue object in the display; see Fig. 1). Following the standard pragmatic view, we calculate informational value in relation to a referent's competitors (as the degree to which the adjective resolves reference amongst objects of the same kind), whereas we calculate discriminatory value over the entire visual context (regardless of the category of the objects). Previous work has shown that discriminatory value can explain the production of redundant color adjectives (Rubio-Fernández, 2016; Rubio-Fernandez, 2019). Here we expand on this work to develop a theoretical account of referential efficiency that builds on informativity and visual search, and accounts for the incrementality of language processing.

If cooperative speakers aim to produce efficient referential expressions, they should be sensitive to the incremental nature of language processing, as it determines the order in which information becomes available to the listener. This is particularly important when considering cross-linguistic variation: in a language like English, for example, adjectives are encoded before the noun (e.g., 'blue triangle'), whereas in languages like Spanish, they are encoded after the noun (e.g., 'triángulo azul'). That means that, in processing a color description, an English listener would search for the referent guided by the adjective (e.g., by color), whereas a Spanish listener would do so guided by the noun (e.g., by shape). Thus, in the same visual context, equivalent referential expressions in different languages can vary in their efficiency, depending on the visual search procedures they instantiate.

Our new account of communicative efficiency treats reference as a collaborative process (Zipf, 1949; Clark & Marshall, 1981; Clark & Schaefer, 1989, for a review of the Rational Speech Act model of pragmatic reasoning, see ; Franke & Jäger, 2016; Goodman & Frank, 2016), and predicts cross-linguistic differences depending on word order. In the remainder of the paper, we review how the notion of referential contrast—a key linguistic construct in analyses of over-specification—changes when considering the incrementality of language processing. We then turn to our theory, and describe its tenets and cross-linguistic implications. Finally, we make an empirical contribution by testing our theory: we report a language-production experiment investigating how English and Spanish speakers use color adjectives depending on their word order and on the efficiency pressures of the task at hand, and two eye-tracking experiments investigating how adjective position affects the way listeners establish color contrast during language processing.

Reconceptualizing the notion of contrast during language processing

According to the standard pragmatic view, the purpose of color adjectives is to help listeners distinguish between objects that belong to the same category (e.g., when referring to 'the blue triangle' in Fig. 2A, the function of 'blue' is to distinguish, or contrast, the blue triangle from the red one). This contrast is called a pragmatic contrast because it requires that speakers engage in audience design and use a color adjective to preempt an ambiguity. When there are no objects of the same category as the target, a color adjective is therefore considered redundant or non-contrastive (as in Figs. 2B-2C Sedivy, 2003, 2005). However, from an incremental point of view, contrast can be established in two ways: within members of a category or across members of different categories (Rubio-Fernández, 2016).

When English listeners process a description such as 'the blue triangle', their visual search for the referent is guided by color and refined by shape (i.e. they should look for a blue shape that is triangular), whereas

when Spanish listeners process the mirror phrase 'el triángulo azul', their visual search is guided by shape and refined by color (i.e. they should look for a triangular shape that is blue). It follows from this basic difference that adjective position affects how listeners establish color contrast during processing: in hearing 'the blue triangle' in Figure 2A, English listeners would first contrast blue vs. non-blue items, regardless of their shape, while in hearing 'el triángulo azul', Spanish listeners would first contrast triangular vs. non-triangular items, regardless of their color. Thus, given their word order, Spanish listeners are more likely to establish color contrast within a category (TRIANGLES > BLUE ONE), while English listeners would do so across categories (BLUE SHAPES > TRIANGULAR ONE)[1].

The incrementality of language production and comprehension has been documented in psycholinguistic studies using real-time eye tracking (e.g., Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Sedivy, 2003, 2005; Brown-Schmidt & Konopka, 2008). Some of these studies investigated the effect of pragmatic contrast on real-time adjective interpretation by presenting English listeners with two objects of the same color (e.g., a blue triangle and a blue square), only one of which had a contrast object in the display (e.g., a red triangle, but no other square; see Supplemental Materials). In this visual context, English listeners were sometimes able to anticipate the noun by deriving a contrastive inference: they anticipated that 'blue' was intended to preempt an ambiguity between the two triangles, rather than referring to the square (Sedivy, 2003, 2005; Aparicio, Xiang, & Kennedy, 2016; Rubio-Fernandez, 2020). In this paradigm, English listeners reveal sensitivity to pragmatic contrast during language processing, rather than establishing color contrast across categories, as we predict. However, pragmatic contrast does not trump incrementality altogether: listeners' initial search is nonetheless guided by the color adjective (i.e. they first identify the two blue shapes when hearing 'blue'; Rubio-Fernandez & Jara-Ettinger, 2020).

Whereas the incrementality of language production and comprehension have been amply documented, the implications of incremental processing for referential contrast have been overlooked in both pragmatics and psycholinguistics, despite their theoretical import: from an incremental processing perspective, redundant color adjectives in prenominal position are normally contrastive, but they establish contrast across categories (e.g., blue vs. non-blue shapes in Figs. 2B-2C). Under this view, prenominal color adjectives are non-contrastive only if they have no discriminatory value in the visual display (i.e. if they do not distinguish the target referent from any other object; see Fig. 1). The incremental view therefore reconceptualizes what counts as a contrastive adjective.

Adopting an incremental perspective on color contrast results in a graded notion of efficiency that is sensitive to the discriminatory value of a color word in the entire visual space. This view of color contrast is different from the standard notion of pragmatic contrast, which applies within categories and treats all redundant uses as non-contrastive. One of the aims of the present study is to use eye tracking to evaluate the incremental view of color contrast against the canonical notion of pragmatic contrast during real-time language processing.

The graded notion of efficiency that we propose is also different from the standard notion of redundancy or over-informativity, which is based on pragmatic contrast. Adopting an incremental view of efficiency results in different and more nuanced empirical predictions for the production of color adjectives than the standard notion of over-informativity. Thus, whereas psycholinguistics research has long established the incremental nature of language production and comprehension, our account is the first to consider the implications of incrementality for referential contrast and efficiency.

The incremental efficiency hypothesis

Since speech unfolds linearly, listeners interpret language incrementally (Eberhard et al., 1995; Spivey et al., 2001). We propose that speakers aim to produce referential expressions that are incrementally efficient for listeners. This is what we call the incremental efficiency hypothesis. This hypothesis has implications for both speakers and listeners. On the production side, the incremental efficiency hypothesis predicts that speakers will be sensitive to adjective position in so far as it affects the efficiency of the color cue for the listener's visual search. On the comprehension side, the same hypothesis predicts that listeners will interpret

---

[1]The incremental view predicts that when English listeners process the color adjective in a postnominal relative clause (e.g., 'The triangle that is blue'), they would establish color contrast within a category.

language incrementally and efficiently, using color information as it becomes available during processing rather than necessarily establishing a pragmatic contrast between competitors of the same kind.

This incremental analysis makes different predictions from the standard global analysis (Sedivy, 2003, 2005; Engelhardt et al., 2006; Koolen et al., 2013, cf. ; Degen et al., 2020). From a global standpoint, the word 'blue' is optimally informative in Figure 2A because it is necessary to distinguish the two triangles, but it is over-informative in Figures 2B-2D because there is only one triangle in each display. From an incremental standpoint, however, the relative efficiency of 'blue' should be evaluated based on the incomplete phrase 'the blue . . . ' rather than on the full phrase 'the blue triangle'. Under this analysis, 'blue' is now efficient in Figures 2A, 2C and 2D because it secures unique reference for the listener, whereas it would be inefficient in 2B because it would create a temporary ambiguity between the blue triangle and the blue star.

From an incremental point of view, mentioning the color of the target triangle in displays 2A, 2C and 2D would be more efficient than in 2B, yet this does not explain why referring to 'the blue triangle' may be more efficient than referring to 'the triangle' in displays 2C and 2D, given that both expressions secure unique reference incrementally, and the unmodified expression is shorter. Here we appeal to the relative efficiency of different visual search procedures. An extensive literature on visual cognition has shown that color is a highly salient cue for visual search, which plays a key role in object recognition (for a review and meta-analysis, see Bramão, Reis, Petersson, & Faísca, 2011). Psycholinguistic studies have revealed that properties such as color, size and spatial position can reduce target identification times when used redundantly (Sonnenschein & Whitehurst, 1982; Mangold & Pobel, 1988; Paraboni, Van Deemter, & Masthoff, 2007; Arts et al., 2011a; Paraboni & van Deemter, 2014; Tourtouri, Delogu, Sikos, & Crocker, 2019). Rubio-Fernandez and Jara-Ettinger (2020) showed that searching by color in displays like 2A, 2C and 2D is faster than searching by shape; in other words, participants were faster to identify 'the blue star' than 'the star'. However, as predicted by the incremental efficiency hypothesis, color does not always lead to more efficient visual search: in visual displays where color was not distinctive of the target (as in display 2B), participants were faster to process the shorter description because color created a temporary ambiguity that delayed target identification.

Language production studies also offer support to the incremental efficiency hypothesis. In polychrome displays where the color of the target is distinctive, participants tend to use redundant color adjectives; whereas in monochrome displays where color is a highly inefficient visual cue, they prefer minimal descriptions (Belke, 2006; Koolen et al., 2013; Rubio-Fernández, 2016; Rubio-Fernandez, 2019; Long, Moore, Mollica, & Rubio-Fernandez, 2020). We take the parallel results observed in reference production and comprehension studies as evidence that reference is a collaborative process (Clark & Marshall, 1981; Clark & Schaefer, 1989).

In summary, psycholinguistic studies have previously suggested that redundant adjectives can facilitate the listener's visual search for a referent. Likewise, eye-tracking studies have also shown that language is interpreted incrementally. However, ours is the first pragmatic account to explain overspecification as resulting from efficiency pressures that are shaped by incrementality (see also Rubio-Fernandez & Jara-Ettinger, 2020). Another innovation with our account is that it predicts cross-linguistic differences in overspecification depending on adjective position. Current pragmatic theories and computational models of reference production have only been tested in languages like English or Dutch, which have prenominal adjectives (e.g., Sedivy, 2003, 2005; van Gompel, van Deemter, Gatt, Snoeren, & Krahmer, 2019; Degen et al., 2020) and do not make different predictions for other languages. However, as we aim to show in this study, adjective position can affect the relative efficiency of a color modifier for visual search, affecting their production as a result.

Cross-linguistic implications

Since language is processed incrementally, the relative efficiency of color for visual search depends on the position of the color adjective relative to the noun it accompanies. Here we aimed to test the incremental efficiency hypothesis by comparing English and Spanish, which have reverse adjective-noun orders and should have reverse strategies to increase the efficiency of color for visual search.

We started by investigating the effect of adjective position on the use of redundant color adjectives. According to the incremental efficiency hypothesis, in languages like English, speakers often use redundant color adjectives to facilitate the listener's visual search, allowing them to quickly identify the color-matching referent, even before they hear the noun (Rubio-Fernandez, 2019). However, in languages like Spanish,

redundant color adjectives are less efficient because they are processed after the listener has begun searching for the noun. The incremental efficiency account therefore predicts that Spanish speakers should produce fewer redundant color adjectives than English speakers.

However, if this cross-linguistic difference is driven by efficiency pressures, the effect of adjective position should be modulated by the density of the display, with redundant color adjectives being generally more efficient in denser displays (e.g., Paraboni et al., 2007; Clarke, Elsner, & Rohde, 2013; Koolen et al., 2013; Koolen, Krahmer, & Swerts, 2016; Gatt, Krahmer, Van Deemter, & van Gompel, 2017; Rubio-Fernandez, 2019). Thus, in Figure 2C, a Spanish listener should be able to identify the target in hearing 'triángulo' (making the ensuing adjective 'azul' inefficient). However, in Figure 2D, the same listener would probably benefit from learning the color of the target since they are more likely to still be searching for it at the end of the noun. It follows from the incremental efficiency hypothesis that the difference in the production of redundant color adjectives between English and Spanish speakers should be reduced in denser displays. This twofold hypothesis was tested with native speakers of English and Spanish in Experiment 1.

We also investigated the effect of adjective position on visual search in order to establish the differential efficiency of prenominal and postnominal color adjectives. The incremental efficiency hypothesis is based on the incremental nature of language processing and the communicative pressures affecting interpretation. However, its predictions are not in line with the canonical notion of pragmatic contrast. From the standard pragmatic view, a cooperative speaker should use color adjectives to distinguish the target referent from other competitors of the same category, and that way preempt an ambiguity. However, whereas speakers of all languages may contrast a target referent with its category competitors in order to produce a sufficiently informative description, the incremental efficiency hypothesis predicts that in languages with prenominal adjectives, listeners will nonetheless establish color contrast across categories. That means that an English speaker may contrast the two triangles in Figure 2A to produce the unambiguous description 'the blue triangle', but in processing this description, an English listener would contrast the blue shape against all the others, not necessarily establishing color contrast between the two triangles. On the other hand, the incremental efficiency hypothesis predicts that, in the same display, Spanish listeners will establish the intended pragmatic contrast between the two triangles, using the postnominal color adjective to disambiguate the description. In Experiment 2a, we used eye tracking to test the predictions of the incremental efficiency hypothesis against the canonical notion of pragmatic contrast .

To further investigate whether the pragmatic contrast intended by the speaker affects listeners' processing of color adjectives, the Spanish listeners participated a second time in Experiment 2b, now completing the English version of the eye-tracking task. If the predicted cross-linguistic differences result from a pressure to communicate efficiently, they should be flexible, with Spanish listeners reversing their visual search strategy in English. However, if pragmatic contrast drives language interpretation (the same way it drives language production; see Brown-Schmidt & Tanenhaus, 2006; Davies & Kreysa, 2017), Spanish listeners should continue to suffer interference from the shape competitor when tested in English. That is, in processing the shape noun in 'The blue triangle', they should consider both triangles in Figure 2A. Given that the Spanish word order supports the canonical pragmatic contrast established within members of the same category (e.g., blue triangle vs red triangle), it is possible that the frequency of this interpretation in Spanish introduces a bias when processing English as a second language. Such a bias, however, would not be efficient for visual search and would run counter to the incremental efficiency hypothesis.

Overall, empirical support for the above hypotheses would confirm that the redundant use of color adjectives is modulated by efficiency pressures on both speakers and listeners, rather than being pragmatically infelicitous (cf. Engelhardt et al., 2006; Engelhardt, Demiral, & Ferreira, 2011). Moreover, the predicted results would support the incremental efficiency hypothesis, according to which a referential expression's efficiency should be calculated incrementally in relation to the entire visual context, rather than on the informativity of the full message.
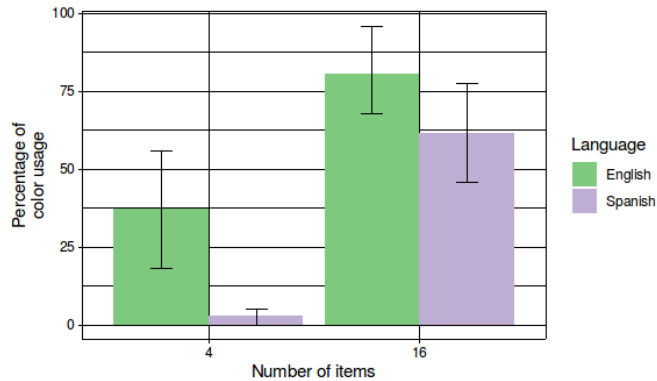
Figure 3: Average percentage of times speakers used redundant color adjectives in each language and condition. Vertical bars show 95% bootstrapped confidence intervals.

Experiment 1

Methods

Participants

A group of 22 undergraduates from University College London and 22 undergraduates from the Universidad de las Islas Baleares (Spain) took part in Experiment 1, after IRB approval had been obtained from each university to conduct the study. All participants in the study signed an informed consent form at the start of the session, and received debriefing at the end. The UCL undergraduates were native speakers of English and the UIB undergraduates were native speakers of Spanish. Both groups participated for monetary compensation. All participants reported having normal color vision.

Materials and procedure

Two types of displays were created, one for the Experimenter (consisting of 20 displays of shapes) and another one for the participant (consisting of empty grids with a cross marking the position of the target in the Experimenter's display). The target shapes were the following: circle, cross, diamond, heart, oval, rectangle, square, triangle, star and sun; and came in the following colors: black, blue, brown, green, grey, orange, pink, purple, red and yellow. Target position was counterbalanced across trials. The first block of trials consisted of ten displays from the No Competitor/4 condition (NC/4; Fig. 2C) and the second block consisted of ten displays from the No Competitor/16 condition (NC/16; Fig. 2D), presented in the same random order.

The displays were shown on a computer monitor placed in front of the Experimenter. The participant sat beside and behind the Experimenter and their task was to ask the Experimenter to click on the target shape in each trial. In order to determine which shape was the target, participants were given printouts of 20 empty grids with a cross indicating the position of the target in the Experimenter's display. The instructions stressed that the Experimenter did not know which shape was the target in each trial, and that all shapes in the displays were different. It followed from this description that color adjectives would be redundant in all trials.

Participants were told that their responses would serve as control data in a study originally designed for children. This was done in order to avoid that participants may become self-conscious and start producing unnatural responses because of the simplicity of the task (this was observed in a pilot study where participants described the shapes in great detail). Participants' requests were recorded and later coded as redundant or not redundant by two blind coders. Only referential expressions including both an adjective and a noun (e.g., 'The blue triangle') were coded as redundant. The task lasted less than 10 minutes.

Results

Participants produced either minimal or modified descriptions (i.e. 'The triangle' or 'The blue triangle'). This type of response consistency is often observed in referential communication studies where participants tend to adopt different referential strategies (Tarenskeen, Broersma, & Geurts, 2015; Rubio-Fernandez, 2019). Figure 3 shows the percentage of times participants used color adjectives as a function of the number of shapes in the display. When there were only four shapes, English speakers used color adjectives 37.3% of the time (95% CI:18.2-55.9) whereas Spanish speakers used color adjectives only 2.73% of the time (95% CI: 0-5.0), with these rates being reliably[2] different (difference = 34.57%; 95% CI: 14.09-53.64). In the display with 16 shapes, English speakers now produced color adjectives 80.5% of time (95% CI:67.7-95.9), a reliably higher rate than their production in the four-shape condition (difference = 43.18%; 95% CI:19.55-68.18). Spanish speakers also produced more color adjectives in the 16-shape condition at 61.4% (95% CI: 45.9-77.7), which was reliably higher than their production in the four-shape condition (difference = 58.64%; 95% CI:42.72-75.45) but was no longer lower than the English speakers' rate (difference = 19.09; 95% CI: -2.27-40.9).

We found similar results through a mixed-effects logistic regression predicting participant's use of redundant color words as a function of language (dummy-coded, with English coded as 0 and Spanish as 1) and number of items in the display (4 vs. 16), with random intercepts and slopes (as a function of number of items) per subject (see Supplemental Materials for details). Consistent with our main analyses, our regression showed that participants were more likely to use redundant color words as a function of the number of items in the visual display ($\beta = 0.57; p < 0.01$), and that Spanish speakers were less likely to use redundant color words ($\beta = -10.65; p < 0.05$). However, we did not find an interaction between language and number of items ($\beta = 0.47; p = 0.15$). This lack of a significant interaction shows that our model did not find evidence that the tendency to use more redundant color words in denser displays was stronger among Spanish speakers relative to English speakers. Note, however, that, in this experimental design, our theoretical predictions did not hinge on the existence or absence of an interaction. The overall pattern of results from Experiment 1 was also visible at the subject-level. 50% (n=11) of English speakers used more color adjectives in the 16-shape condition than in the four-shape condition, 50% (n=11) used the same amount (8 of these participants used color adjectives in every trial of the four-shape block, making it impossible for them to use more color adjectives in the 16-shape block), and 0% used fewer color adjectives. Among Spanish speakers, 82% (n=18) used more color adjectives in the 16-shape condition than in the four-shape condition, 18% (n=4) used an equal amount, and 0% used fewer color adjectives (for data visualizations, see Supplemental Materials).

The results of Experiment 1 offered support to the incremental efficiency hypothesis, with both English and Spanish speakers producing redundant color adjectives to the extent that it would be efficient for the listener's visual search. However, the question remains as to whether the cross-linguistic differences observed in language production are related to differences in language processing. Such a relation would support the view that reference is a collaborative process between speakers and listeners (Clark & Marshall, 1981; Rubio-Fernández, 2016; Rubio-Fernandez, 2019). The aim of Experiments 2a and 2b was to test the incremental efficiency hypothesis using eye tracking during language processing.

Experiment 2a

Methods

Participants

A new group of 25 undergraduates from each university took part in Experiment 2a. The UCL undergraduates were native speakers of English and the UIB undergraduates were native speakers of Spanish. Both

---

[2]Due to conceptual and practical challenges associated with null-hypothesis significance testing (Cohen, 1994; Cumming, 2014), our analyses focused on estimating the magnitude of effect sizes through 95% bootstrapped confidence intervals. Throughout, we refer to effects as reliable whenever all values within a confidence interval are consistent with our theory, and as unreliable whenever there is at least one value within a confidence interval that is inconsistent with our theory. As suggested by reviewers, we supplement these main analyses with null-hypothesis significance tests throughout the study.

groups participated for monetary compensation. All participants reported having normal color vision. Sample size was determined by the time available to collect data at UCL and UIB. A post-hoc power analysis revealed power > 0.9 for our key predictions (see Supplemental Materials).

Materials and procedure

Materials for all three experiments are available at OSF https://osf.io/9hw68/. The visual materials consisted of 72 displays of four geometrical shapes, including 12 critical items from the Shape Competitor condition (SC; e.g., Fig. 2A), 12 from the Color Competitor condition (CC; e.g., Fig. 2B), and 48 fillers. Filler trials also included 4 geometrical shapes, but the target had both a shape and a color competitor (for a sample item from the Two Competitors condition, see Supplementary Materials). This kind of displays has been used to investigate the derivation of contrastive inferences using long preview windows (Sedivy, 2003, 2005; Aparicio et al., 2016; Rubio-Fernandez & Jara-Ettinger, 2020; Rubio-Fernandez, 2020). Here, the Two Competitors condition was intended to add variability to the types of displays used in the study, rather than being a test of our main hypotheses (for analysis and discussion of the filler trials, see Supplementary Materials).

The target shapes (2 per critical condition) were: circle, diamond, rectangle, square, star and triangle; and the colors of the shapes were: blue (x2), brown, green (x2), orange (x2), pink (x2), purple, red and yellow. The position of the target and competitor shapes were counterbalanced across trials, and trials were randomized individually for each participant.

The displays were on the screen for 400 ms before the instructions started. This is a relatively short preview window, especially considering that launching a saccade takes 200 ms. This short preview window was intended to prevent participants from conceptualizing all the color shapes in the display prior to the start of the instruction, which would have allowed them to identify the target artificially fast.

All instructions were of the form 'Click on the [COLOR ADJECTIVE + SHAPE NOUN]' or 'Haz click en el [SHAPE NOUN + COLOR ADJECTIVE]'. The instructions were recorded by male native speakers of British English and Castilian Spanish, respectively, who did not stress the adjectives contrastively. For data analysis, a critical time window was calculated for each instruction from the onset of the adjective until the offset of the noun in English, and from the onset of the noun until the offset of the adjective in Spanish. The structure of the time window resulted in an earlier disambiguation point in English than in Spanish in the Shape Competitor condition (because the adjective disambiguated the description), whereas it resulted in an earlier disambiguation point in Spanish than in English in the Color Competitor condition (because the noun disambiguated the description). The mean duration of the critical time window was 743 ms in English (range: 630-891) and 996 ms in Spanish (range: 841-1219). We decided on this time window because its structure is parallel in the two languages and would allow observing an early effect of adjective position. We did not use RTs to calculate individual time windows because it introduced the issue of having longer time-windows in conditions with an interfering competitor.

Participants were told that they were going to listen to a series of instructions to click on different geometrical shapes and their task was to click on the target as fast and accurately as possible. All participants were first presented with four displays that served as warm-up trials, followed by the remaining 68 trials in a random order. In between trials, participants had to click on a cross in the center of the screen to move on to the next trial. This ensured that participants' gaze and the mouse cursor were always in the center of the screen at the start of each trial.

Predictions

According to the standard pragmatic view, color adjectives should distinguish objects of the same kind. The incremental efficiency hypothesis predicts that Spanish listeners will indeed establish color contrast within a category, contrasting both triangles in Figure 2A when searching for 'el triángulo azul', for example. However, in the same display, English listeners should identify the target in hearing 'blue', not considering the second triangle when processing the noun. Likewise, because English listeners establish color contrast across categories, when they search for 'the blue triangle' in Figure 2B, the blue star should interfere with their visual search. In the same display, however, Spanish listeners should identify the target in hearing 'triángulo', disregarding the other blue shape when processing the color adjective.
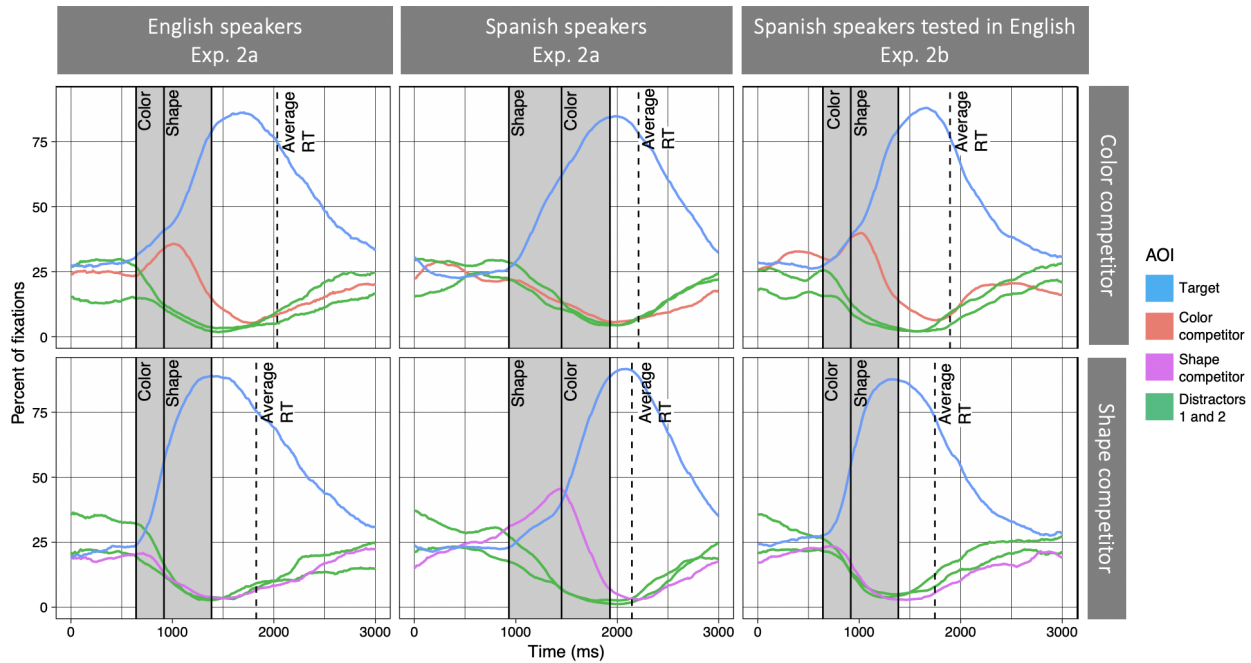
Figure 4: Percentage of eye fixations on the four shapes during the processing of instructions including color and shape words (e.g., 'Click on the blue triangle'). At each time point, the curve represents the percentage of fixations in the previous 200ms. The grey region represents the duration of the critical phrase and the dashed lines represent participants' average reaction times. Experiment 2a tested English and Spanish speakers, and Experiment 2b retested the Spanish speakers in English.

Results

Data and analysis code for all three experiments are available at OSF https://osf.io/9hw68/. Eye-tracking data were standardly corrected by +200 ms in both experiments in order to account for the time it takes to launch a saccade. Figure 4 shows the percentage of fixations over time. As predicted, English listeners fixated more on the target when there was a shape competitor (53.9%; 95% CI: 49.2-58.6) than when there was a color competitor (42.1%; 95% CI:39.2-45.1), with a reliable difference between the two conditions (Target advantage in CC = 11.8%; 95% CI: 6.05-17.43). By contrast, Spanish listeners fixated less on the target when there was a shape competitor (40.5%; 95% CI: 37.4-43.6) than when there was a color competitor (52.5%; 95% CI: 48.8-56.0), also revealing a significant difference (Target advantage in SC = 12%; 95% CI: 7.07-16.63). Consistent with this, English listeners fixated more on the target than Spanish listeners did when there was a shape competitor (Difference = 13.4%; 95% CI: 7.58-19.07), whereas Spanish listeners fixated more on the target than English listeners did when there was a color competitor (Difference = 10.4%; 95% CI: 5.59-15.05).

We further confirmed these results through a mixed-effects logistic regression predicting participant fixations to the target as a function of time, language, and condition (with all their interactions). Time was centered and divided by two times the standard deviation (Gelman & Hill, 2006), and language and condition were sum-coded (English coded as 0.5 and Spanish as -0.5; Shape Competitor condition coded as 0.5 and Color Competitor condition as -0.5). We also included random intercepts as a function of participant and random intercepts and time slopes as a function of item (the maximal model that converged; see Supplemental Materials for full details). Consistent with our hypothesis, we found a significant interaction between condition and language ($\beta = 1.18; p < 2e - 16$), and a significant interaction between condition, language, and time ($\beta = 1.78; p < 2e - 16$), showing that English speakers in the Shape Competitor condition and Spanish speakers in the Color Competitor condition were faster to identify the target relative to English speakers in the Color Competitor condition and Spanish speakers in the Shape Competitor condition (see Supplemental Materials for full regression tables).
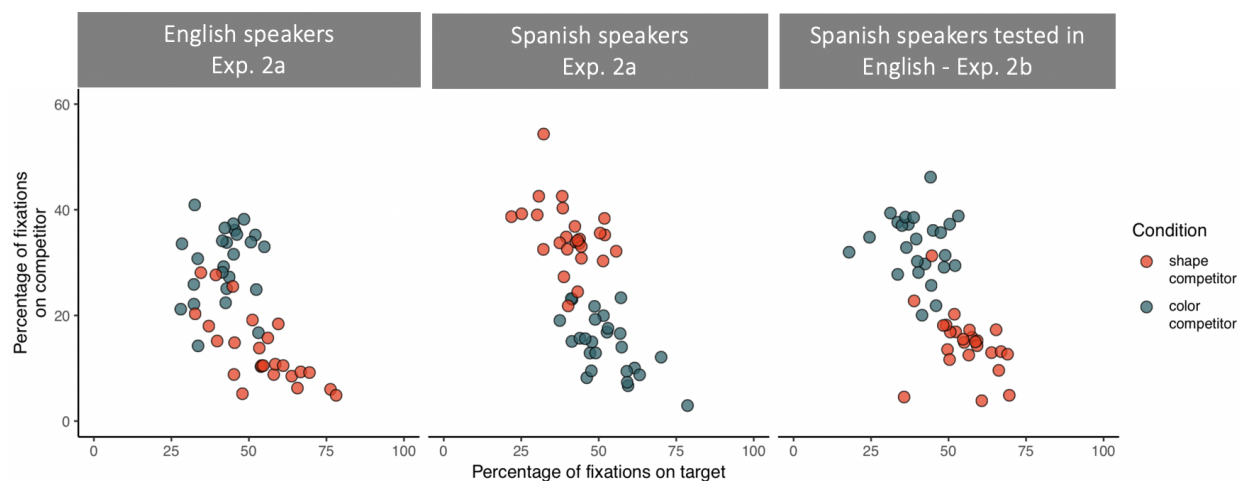
Figure 5: Percentage of fixations on the target and the competitor during the critical window. Each dot represents a participant, with two dots per participant in each plot, one corresponding to their average fixations on the Shape Competitor condition, and the other corresponding to their average fixations on the Color Competitor condition. The x axis shows the percentage of fixations on the target, and the y axis shows the percentage of fixations on the relevant competitor (shape or color, depending on the condition). Supporting the incremental efficiency hypothesis, English listeners and Spanish listeners tested in English showed higher rates of target fixations in the shape competitor trials relative to the color competitor trials. By contrast, Spanish listeners showed higher rates of target fixations in the color competitor trials relative to the shape competitor trials.

Fixations on the competitor also supported the incremental efficiency hypothesis (for data visualizations, see Supplemental Materials), with Spanish listeners suffering more interference from the shape competitor relative to English listeners (Interference difference = 21.7%; 95% CI: 18.13-25.23). By contrast, English listeners suffered more interference from the color competitor relative to Spanish listeners (Interference difference = 15.4%; 95% CI: 12.14-18.86).

The interaction between language and competitor type in the looking patterns also appeared at the subject-level (see Fig. 5). 88% (n=22) of English listeners fixated more on the target when there was a shape competitor than when there was a color competitor, whereas 88.5% (n=23) of Spanish listeners fixated more on the target when there was a color competitor than when there was a shape competitor. The results of competitor interference were consistent with this. 92% (n=23) of English listeners fixated more on the competitor when it was a color match than when it was a shape match, whereas 100% (n=26) of Spanish listeners fixated more on the competitor when it was a shape match than when it was a color match.

The results of Experiment 2a replicate previous findings in English and extend them in a cross-linguistic comparison: both English and Spanish listeners identified the target as soon as they had sufficient information to do so, not suffering interference from competitors past the disambiguation point in the instructions. This confirms that participants in both language groups interpreted language incrementally and efficiently (see Eberhard et al., 1995; Spivey et al., 2001). In addition, the results support the incremental efficiency hypothesis, with English listeners establishing color contrast across categories, while Spanish listeners do so within a category.

The results of Experiment 2a confirm that pragmatic contrast, which is supposed to be established between category competitors, does not drive language processing, at least not the way it drives language production (with speakers relying on pragmatic contrast to produce sufficiently informative descriptions; see Brown-Schmidt & Tanenhaus, 2006; Davies & Kreysa, 2017).

Experiment 2b

Methods

Participants

The same group of 25 undergraduates from UIB who took part in the Spanish version of Experiment 2a were tested again in the English version immediately after. Participants had not been told about this second task before performing the first. All participants in Experiment 2b reported having an intermediate level of English and none reported speaking English natively. One of the participants did not perform the second task because they did not feel comfortable being tested in English.

Materials and procedure

The same English materials and procedure used in Experiment 2a were used again in Experiment 2b.

Predictions

According to the standard pragmatic view, color adjectives should be used to preempt an ambiguity between potential referents of the same kind, being otherwise redundant. If the canonical notion of pragmatic contrast drives not only language production, but also language comprehension, Spanish listeners should consider both competitors in the Shape Competitor condition when tested in English. However, according to the incremental efficiency hypothesis, listeners should identify a referent as soon as they have enough information to do so. This means that Spanish listeners should reverse their visual search when tested in English and establish color contrast across categories.

Results

In contrast to the results of Experiment 2a, Spanish listeners' looking pattern was reversed when tested in English (see Fig. 3). Participants in Experiment 2b fixated more on the target when there was a shape competitor (55.6%; 95% CI: 52.4-59.1) relative to when there was a color competitor (40.3%; 95% CI:37.2-43.6), with a reliable difference between conditions (Target advantage in SC = 15.3%; 95% CI: 10.75-19.98). This pattern was reliably different from their looking pattern when tested in Spanish, both when there was a shape competitor (Spanish tested in Spanish vs in English: Difference in SC = 15.1%; 95% CI: 10.58-19.63) and a color competitor (Spanish tested in Spanish vs in English: Difference in CC = 12.2%; 95% CI: 7.29-16.94). Consistent with this, participants' looking pattern was no longer different from that of English listeners when there was a shape competitor (Spanish tested in English vs English listeners: Difference in SC = 1.7%; 95% CI: -3.97-7.78) or when there was a color competitor (Spanish tested in English vs English listeners: Difference in CC= -1.8%; 95% CI: -6.14-2.74). Similarly, Spanish listeners' interference was reversed when tested in English (for data visualizations, see Supplemental Materials). Participants now showed increased interference from the color competitor (Interference increase in CC = 18.5%; 95% CI: 15.51-21.81) and reduced interference from the color competitor (Interference decrease in SC = 20.3%; 95% CI: 17.08-23.57).

These results were consistent with those from a mixed-effects logistic regression where we compared Spanish speaker's fixations to the target when tested in English and in Spanish. In this regression we predicted participant fixations to the target as a function of time, language, and condition (with all their interactions). All variables were coded in the same way as the regression analysis in Experiment 2a. We also included random intercepts, time slopes, and condition slopes as a function of participants, and random intercepts and time slopes as a function of item (the maximal model that converged; see Supplemental Materials for full details). We found a significant interaction between condition and language ($\beta = 1.35; p < 2e-16$), and a significant interaction between condition, language, and time ($\beta = 1.55; p < 2e-16$), showing that Spanish speakers were faster in the Shape Competitor condition when tested in English, and faster in the Color Competitor condition when tested in Spanish (see Supplemental Materials for full regression tables).

The interaction between language and competitor type in the looking patterns also appeared at the subject-level (see Fig. 5). 96% (n=24) of Spanish listeners tested in English fixated more on the target when there was a shape competitor than when there was a color competitor. The results of competitor interference

were consistent with this: 100% (n=25) of Spanish listeners tested in English fixated more on the competitor when it was a color match than when it was a shape match.

The results of Experiment 2b confirm that pragmatic contrast does not drive the processing of prenominal color adjectives, not even when listeners are used to processing color adjectives in postnominal position and establish color contrast within a category in their mother tongue.

General Discussion

Here we presented and tested an incremental account of communicative efficiency, whereby speakers of different languages exploit word order to coordinate more efficiently. Supporting the incremental efficiency hypothesis, English speakers produced more redundant color adjectives than Spanish speakers when referring to shapes in sparse displays (where prenominal color adjectives are more efficient than postnominal color adjectives), but both groups did so more frequently and to comparable rates in denser displays (where the listener's search for the referent was harder and both types of adjectives would be efficient).

Further supporting the incremental efficiency hypothesis, our eye-tracking results revealed that the visual search for a referent was guided by color and refined by shape in English, whereas the search for the same referent was guided by shape and refined by color in Spanish. Importantly, the Spanish listeners reversed their visual search strategy when retested in English immediately after, showing efficient incremental processing, rather than a native bias to establish a pragmatic contrast between entities of the same kind when processing color adjectives.

It must be noted, however, that the cross-linguistic differences observed in both the production and comprehension of color adjectives do not support linguistic determinism (Whorf in Carroll, 1957). Spanish speakers did not produce more redundant color adjectives than English speakers across all visual contexts, nor did they establish color contrast within a category regardless of the language they were processing. We therefore conclude that English and Spanish speakers did not reveal 'differently structured minds' (Pinker, 2007), but flexible effects of word order driven by efficiency pressures.

Our results inform psycholinguistic analyses of the use and comprehension of color adjectives. Continuous eye-tracking measures revealed that Spanish listeners established color contrast within a category (TRIANGLES > BLUE ONE), whereas English listeners did so across categories (BLUE SHAPES > TRIANGULAR ONE). This difference highlights the importance of adopting an incremental perspective when studying communicative efficiency. Thus, the idea that redundant color adjectives exploit color contrast across categories challenges the general view that they are non-contrastive (Sedivy, 2003, 2005) and explains why redundant color adjectives can facilitate the listener's visual search for a referent (Sonnenschein & Whitehurst, 1982; Mangold & Pobel, 1988; Arts et al., 2011a; Paraboni et al., 2007; Paraboni & van Deemter, 2014; Rubio-Fernandez, 2020) and be efficient (Rubio-Fernández, 2016; Rubio-Fernandez, 2019), rather than being pragmatically infelicitous (Engelhardt et al., 2006, 2011).

Our results are also relevant to computational models of reference generation, which so far have only been developed to account for data from languages with prenominal modification (e.g., Dale & Reiter, 1995; Frank & Goodman, 2012; van Gompel et al., 2019; Deemter, Gatt, Sluis, & Power, 2012; Degen et al., 2020). However, languages where adjectives precede nouns are a minority, with most world languages positioning their adjectives after their nouns (Dryer, 2013). Like pragmatic theories, computational models may not have looked at reference from a cross-linguistic perspective because pragmatics is supposed to apply above and across all languages and communicative situations. However, our research shows that, like other language components, pragmatics is affected by incrementality constraints and efficiency pressures, also requiring a cross-linguistic investigation (see also Rubio-Fernandez & Jara-Ettinger, 2020).

Importantly, incrementality is not the only factor that affects the production of redundant color adjectives. The discriminability of the referent and the density of the display have been shown to affect color overspecification (Rubio-Fernandez, 2019), as well as the typicality of the color (with atypical colors being used redundantly more often than typical colors; Sedivy, 2003; Westerbeek, Koolen, & Maes, 2015; Rubio-Fernández, 2016) and the lexical category of the noun (with clothes eliciting higher rates of redundant color adjectives than geometrical shapes; Rubio-Fernández, 2016; Rubio-Fernandez, 2019). In addition, not all adjectives are equally efficient in prenominal position: scalar adjectives are more context-dependent than color adjectives because of their relational semantics (Kennedy & McNally, 2005; Kennedy, 2007; Kennedy

& McNally, 1999), resulting in higher processing costs that may not justify their overspecification (for eye-tracking evidence, see Aparicio et al., 2016; Rubio-Fernandez, 2020). Likewise, material adjectives are not as visually salient as color adjectives, being less efficient when used redundantly (Jara-Ettinger & Rubio-Fernandez, 2020). We therefore see overspecification as a multifaceted phenomenon determined not only by informativity considerations, but by the complex interaction of pragmatic, semantic and perceptual factors during speaker-listener coordination.

Regarding informativity, our results challenge the standard notion of redundancy or over-informativity, which is based on pragmatic contrast. According to the standard view, color adjectives are used contrastively to distinguish objects of the same kind, being used redundantly otherwise. While speakers of all languages may indeed contrast objects of the same kind in order to produce sufficiently informative descriptions of an intended referent (see Brown-Schmidt & Tanenhaus, 2006; Davies & Kreysa, 2017), our eye-tracking results show that pragmatic contrast does not drive language processing (although in some visual contexts, listeners can show sensitivity to pragmatic contrast; see Supplementary Materials). Instead, our findings confirm that incrementality and efficiency determine how listeners establish color contrast during processing (i.e. within a category or across categories), whereas the canonical notion of pragmatic contrast does not always constrain processing.

Our results are relevant for pragmatic theories of referential communication since it has been argued that the redundant use of color adjectives results from speakers' failure to take the listener's perspective (for discussion, see Arnold, 2008; Davies & Arnold, 2018). However, if listeners establish color contrast incrementally and assign reference to an expression as soon as they have enough information to do so (rather than necessarily establishing a pragmatic contrast between category competitors), then redundant color adjectives are not 'egocentric' or 'non-contrastive', but may in fact be rational, cooperative and efficient (Grice, 1975).

To conclude, our results confirm that speaker-listener coordination is subject to efficiency pressures that affect the production and processing of redundant color adjectives, supporting an incremental view of communicative efficiency. Moreover, these findings highlight the importance of cross-linguistic research for developing nuanced pragmatic theories and computational models that explain how language users maximize communicative efficiency given the constraints and affordances of their languages.

Context of the Research

This project emerged from a 'cross-linguistic' intuition that Paula Rubio-Fernandez had during her postdoctoral studies with Sam Glucksberg in Princeton: whereas it sounded natural to her to use color adjectives redundantly in English, she felt that she would not do so as frequently in Spanish. This intuition led to a set of empirical studies looking at the role of efficiency in referential communication, culminating in the theoretical perspective presented here. This research question, in turn, is part of a broader research program looking at the interaction between perception, Theory of Mind, and communication. Paula Rubio-Fernandez and Francis Mollica are currently studying the interaction between perception and communication, and Paula Rubio-Fernandez and Julian Jara-Ettinger are studying the interaction between communication and Theory of Mind.

Acknowledgements

References

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. In *Semantics and linguistic theory* (Vol. 25, pp. 413–432).

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, *23*(4), 495–527.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011a). Overspecification facilitates object identification. *Journal of pragmatics*, *43*(1), 361–374.

Arts, A., Maes, A., Noordman, L. G., & Jansen, C. (2011b). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31–56.

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, *14*(3), 261–294.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different "decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

Bramão, I., Reis, A., Petersson, K. M., & Faísca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta psychologica*, *138*(1), 244–253.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of english and spanish. *Cognition*, *109*(2), 274–280.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Carroll, J. B. (1957). *Language, thought, and reality: Selected writings of benjamin lee whorf*. MIT Press.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259–294.

Clarke, A. D. F., Elsner, M., & Rohde, H. (2013). Where's wally: the influence of visual salience on referring expression generation. *Frontiers in psychology*, *4*, 329.

Cohen, J. (1994). The earth is round (p¡. 05). *American psychologist*, *49*(12), 997.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263.

Davies, C., & Arnold, J. (2018). Reference and informativeness: How context shapes referential choice. *Handbook of Experimental Semantics and Pragmatics. Oxford: OUP*.

Davies, C., & Kreysa, H. (2017). Looking at a contrast object before speaking boosts referential informativeness, but is not essential. *Acta Psychologica*, *178*, 87–99.

Deemter, K. v., Gatt, A., Sluis, I. v. d., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, *36*(5), 799–836.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*.

Dryer, S. M. (2013). Order of adjective and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures*. Max Planck Institute for Evolutionary Anthropology at Leipzig.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, *24*(6), 409–436.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of memory and language*, *54*(4), 554–573.

Engelhardt, P. E., Demiral, Ş. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An erp study. *Brain and cognition*, *77*(2), 304–314.

Engelhardt, P. E., & Ferreira, F. (2014). Do speakers articulate over-described modifiers differently from modifiers that are required by context? implications for models of reference production. *Language, Cognition and Neuroscience*, *29*(8), 975–985.

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

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

Gatt, A., Krahmer, E., Van Deemter, K., & van Gompel, R. P. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive science*, *41*, 1457–1492.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. P. (2019). How efficiency shapes human language, tics 2019. *Trends in Cognitive Science*.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological science*, *24*(7), 1079–1088.

Givón, T. (1979). *On understanding grammar*. Academic Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts.* Academic Press.

Hawkins, R. X., Gweon, H., & Goodman, N. D. (2018). Speakers account for asymmetries in visual perspective so listeners don't have to. *arXiv preprint arXiv:1807.09000*.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2020). Speakers construct reference based on listeners' expected visual search.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, *30*(1), 1–45.

Kennedy, C., & McNally, L. (1999). From event structure to scale structure: Degree modification in deverbal adjectives. In *Semantics and linguistic theory* (Vol. 9, pp. 163–180).

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*(2), 395–411.

Koolen, R., Krahmer, E., & Swerts, M. (2016). How distractor objects trigger referential overspecification: testing the effects of visual clutter and distractor distance. *Cognitive science*, *40*(7), 1617–1647.

Long, M., Moore, I., Mollica, F., & Rubio-Fernandez, P. (2020). Contrast perception as a visual heuristic in the formulation of referential expressions.

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, *37*(2), 117–144.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Mangold, R., & Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, *7*(3-4), 181–191.

Paraboni, I., & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, *29*(8), 1002–1017.

Paraboni, I., Van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational linguistics*, *33*(2), 229–254.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, *53*, 58–74.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, *7*, 153.

Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive science*, *43*(11), e12797.

Rubio-Fernandez, P. (2020). Redundant color words are more efficient than shorter descriptions.

Rubio-Fernandez, P., & Jara-Ettinger, J. (2020). Incrementality and efficiency shape pragmatics across languages. *Proceedings of the National Academy of Sciences*.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3–23.

Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for cricean mechanisms in online language interpretation. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, 345.

Sonnenschein, S., & Whitehurst, G. J. (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research*, *11*(2), 115–125.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological science*, *12*(4), 282–286.

Tarenskeen, S., Broersma, M., & Geurts, B. (2015). Overspecification of color, pattern, and size: salience, absoluteness, and consistency. *Frontiers in Psychology*, *6*, 1703.

Tourtouri, E. N., Delogu, F., Sikos, L., & Crocker, M. W. (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science*, *3*(2), 175–202.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*(3), 145–174.

van Gompel, R. P., van Deemter, K., Gatt, A., Snoeren, R., & Krahmer, E. J. (2019). Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological review*, *126*(3), 345.

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in psychology*, *6*, 935.

Zipf, G. K. (1949). Human behaviour and the principle of least-effort. cambridge ma edn. *Reading: Addison-Wesley*.

Supplemental materials

1. Post-hoc power analyses

We determined the power of our studies by bootstrapping our experimental data for each key prediction. All code for these post-hoc power analyses is available at https://osf.io/9hw68/. All our key predictions (see below) relied on differences between conditions. We analyzed these differences by computing bootstrapped 95% confidence intervals over the difference in effect sizes and seeing if the interval crossed 0 (suggesting there may be no difference in effect sizes) or not.

The bootstrap process followed two stages. First, we bootstrapped the data to obtain a replicate of the two target experimental conditions (depending on prediction; see below). Next, for each experimental replicate, we conducted the analysis from the manuscript: bootstrapping the results to obtain a confidence interval over the difference in effect sizes. We ran 5000 experimental replicates and within each replicate we bootstrapped the difference in effect sizes with 1000 replicates.

Experiment 1

Prediction 1. Percentage of redundant color words should be different in the four-item displays across languages.

In 94.9% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different and that our experiment's power is 0.949.

Prediction 2. Percentage of redundant color words in English should be higher in the 16-item displays relative to the four-item displays.

In 92.0% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different and that our experiment's power is 0.92.

Prediction 3. Percentage of redundant color words in Spanish should be higher in the 16-item displays relative to the four-item displays.

In 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different. Naturally, this does not imply that our power is 1, as it is likely that at least one replicate would eventually not produce a difference. Note, however, that the difference across conditions was striking in Spanish (see Figure 1S below), with almost no Spanish speaker producing color words in the four-item display, and the majority of them producing color words more than half of the time in the 16-item display. Given that we found no replicates that went against our predictions in the 5000 bootstrap samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/500 = 1e-04.

Experiment 2a

Prediction 1. English listeners should show increased target fixations during the adjusted NP in the shape competitor trials relative to the color competitor trials.

In 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects (fixations on target in shape competitor trials, and fixation on target in color competitor trials) were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

Prediction 2. Spanish listeners tested in Spanish should show increased target fixations during the adjusted NP in the color competitor trials relative to the shape competitor trials.

Similar to Prediction 1, in 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

Experiment 2b

Prediction 1. Spanish listeners tested in English should show increased target fixations during the adjusted NP in the shape competitor trials relative to the color competitor trials.

In 100% of the bootstrapped replicates, the corresponding confidence interval once again did not cross 0, suggesting that the two effects (fixations on target in shape competitor trials, and fixation on target in color competitor trials) were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

2. Supplemental figures for main analyses
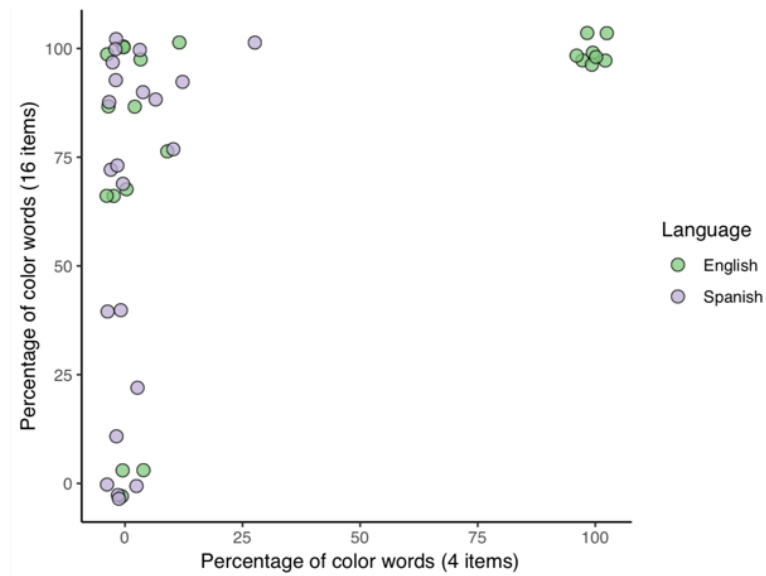

Figure 1: Percentage of redundant color adjectives used by each subject (represented by a dot). The x axis shows the percentage of color words that subjects produced in 4-shape displays and the y axis shows the percentage of color words that they produced in 16-shape displays. The color of the dot indicates the speaker's language. Data have been minimally jittered to avoid overplotting.
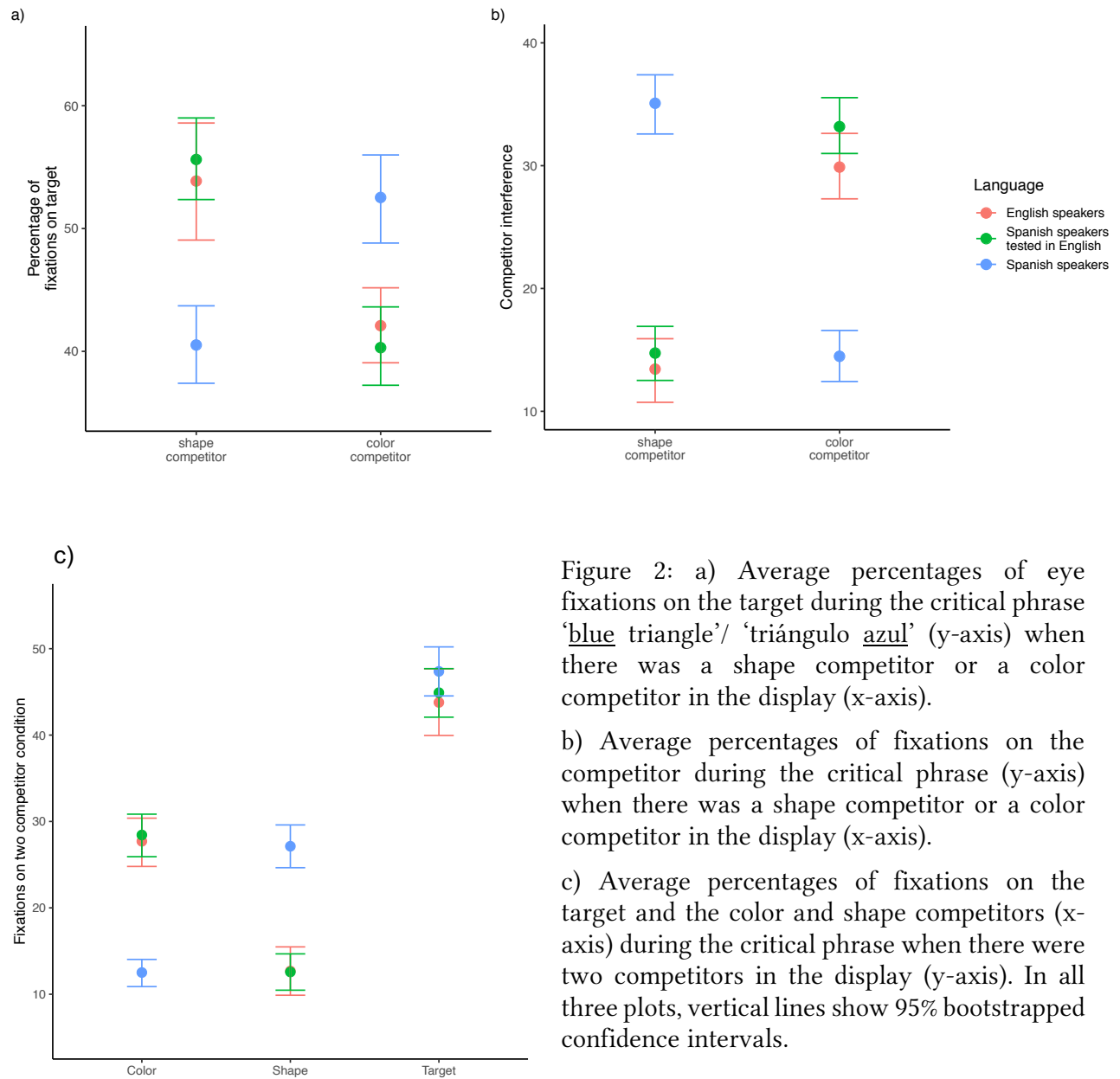
Figure 2: a) Average percentages of eye fixations on the target during the critical phrase 'blue triangle'/ 'triángulo azul' (y-axis) when there was a shape competitor or a color competitor in the display (x-axis).

b) Average percentages of fixations on the competitor during the critical phrase (y-axis) when there was a shape competitor or a color competitor in the display (x-axis).

c) Average percentages of fixations on the target and the color and shape competitors (x-axis) during the critical phrase when there were two competitors in the display (y-axis). In all three plots, vertical lines show 95% bootstrapped confidence intervals.

3 Regression analyses

3.1 Experiment 1

The tables below present the result of a logistic mixed-effects model [1, 2] predicting participant's use of redundant color words as a function of number of items in the display (coded numerically, either four or sixteen) and language (dummy-coded, with English set to 0 and Spanish to 1). We also included the maximal random effects structure:

$$Target \sim Items * Language+$$
$$(1 + Items|Participant)$$

Note that we do not include language-based slopes per participant because language varies across participants.

Scaled residuals:				
Min	1Q	Median	3Q	Max
-3.5538	-0.0524	-0.0001	0.1037	3.5441

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Subject	(Intercept)	118.7392	10.8968	
	Items	0.2717	0.5213	-0.95

Fixed effects:				
	Estimate	Std. Error	z value	Pr($> |z|$)
(Intercept)	-5.2552	3.1689	-1.658	0.09724 .
Items	0.5742	0.1768	3.247	0.00116 **
Language(Sp)	-10.6457	5.4160	-1.966	0.04934 *
Items:Language(Sp)	0.4653	0.3210	1.450	0.14713

3.2 Experiment 2

The following tables show the results of two logistic mixed-effects models [1, 2] comparing English vs. Spanish listeners (Experiment 2a), and comparing Spanish listeners tested in English vs Spanish listeners tested in Spanish (Experiment 2b). In each case we identified the largest maximal model that converged through the buildmer package [3].

In each trial we considered only the NP window (adjusted by 200 milliseconds), and we excluded trials where participants failed to select the correct target. Time was standardized by subtracting the mean and dividing by two times the standard deviations in each trial. Language and condition were both sum-coded (see individual regressions for details).

3.2.1 Experiment 2a

We first considered the maximal logical model:

$$Target \sim Time * Language * Cond+$$
$$(1 + Time + Cond|Participant)+$$
$$(1 + Time|Item)$$

In the language variable, English was coded as 0.5 and Spanish as -0.5. In the condition variable, the Shape Competitor condition was coded as 0.5 and the Color Competitor condition as -0.5 [1]. The largest model that converged was

$$Target \sim Time * Language * Cond+$$
$$(1|Participant)+$$
$$(1 + Time|Item)$$

Our predictions focused on the speed of identifying the target and we thus focused on how language type and condition interacted with time.

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.9940	-0.7506	-0.3079	0.7679	6.1587

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.1578	0.3973	
Item	(Intercept)	0.2040	0.4516	
	Time	0.3344	0.5783	-0.29

Fixed effects:

	Estimate	Std. Error	z value	$Pr(>|z|)$	$Pr(>|t|)$
(Intercept)	-0.132697	0.107819	-1.230735	0.218	0.21842
Time	1.850445	0.119071	15.540725	0.000	$< 2e-16$ ***
Cond	-0.009818	0.184884	-0.053103	0.958	0.95765
Language	0.091396	0.112124	0.815133	0.415	0.41500
Time:Cond	0.720336	0.237137	3.037636	0.002	0.00238 **
Cond:Language	1.178885	0.026640	44.253191	0.000	$< 2e-16$ ***
Time:Language	-0.010657	0.028990	-0.367623	0.713	0.71315
Time:Cond:Language	1.783963	0.057598	30.972843	0.000	$< 2e-16$ ***

The significant positive effect of the interaction between condition and language ($\hat{\beta}_{C:L} = 1.18$), and time, condition and language ($\hat{\beta}_{T:C:L} = 1.78$), show

evidence for our hypothesis. This means that English listeners in the Shape Competitor condition (both coded as 0.5, creating interactions of $0.25 * \hat{\beta}_{C:L}$ and $0.25 * \hat{\beta}_{T:C:L}$) and Spanish listeners in the Color Competitor condition (both coded as -0.5, creating the same interaction effects) identified the target significantly faster than English listeners in the Color Competitor condition and Spanish listeners in the Shape Competitor condition (interaction effects = $-0.25 * \hat{\beta}_{C:L}$ and $-0.25 * \hat{\beta}_{T:C:L}$ in both cases).

3.2.2 Experiment 2b

In the regression for Experiment 2b we compared Spanish listeners tested in Spanish and English.

We first considered the maximal logical model:

$$
\begin{aligned}
Target \sim &Time * Language * Cond+ \\
&(1 + Time + Cond + Language|Participant)+ \\
&(1 + Time|Item)
\end{aligned}
$$

Notice that this maximal model now included random language slopes per participant. This was now possible because the same participants completed the task in both languages. In the language variable, English was coded as 0.5 and Spanish as -0.5. In the condition variable, the Shape Competitor condition was coded as 0.5 and the Color Competitor condition as -0.5 [1].

The largest model that converged was

$$
\begin{aligned}
Target \sim &Time * Language * Cond+ \\
&(1 + Time + Cond|Participant)+ \\
&(1 + Time|Item)
\end{aligned}
$$

Our predictions focused on the speed of identifying the target and we thus focused on how language type and condition interacted with time.

Scaled residuals:				
Min	1Q	Median	3Q	Max
-5.1131	-0.7543	-0.3568	0.7809	5.7258

The significant positive effect of the interaction between condition and language ($\hat{\beta}_{C:L} = 1.35$), and time, condition and language ($\hat{\beta}_{T:C:L} = 1.55$), show evidence for our hypothesis. This means that when tested in English in the Shape Competitor condition (both coded as 0.5, creating interactions of $0.25 * \hat{\beta}_{C:L}$ and $0.25 * \hat{\beta}_{T:C:L}$) and in Spanish in the Color Competitor condition (both coded as -0.5, creating the same interaction effects), participants identified the target significantly faster than when tested in English in the Color Competitor condition and in Spanish in the Shape Competitor condition (interaction effects = $-0.25 * \hat{\beta}_{C:L}$ and $-0.25 * \hat{\beta}_{T:C:L}$ in both cases).

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.04886	0.2210	
	Cond	0.12844	0.3584	-0.29
	Time	0.04669	0.2161	0.48 -0.02
Item	(Intercept)	0.14438	0.3800	
	Time	0.28916	0.5377	-0.11

Fixed effects:					
	Estimate	Std. Error	z value	$Pr(>|z|)$	$Pr(>|t|)$
(Intercept)	-0.12984	0.08901	-1.45875	0.145	0.1446
Time	1.78146	0.11839	15.04720	0.000	$< 2e-16$ ***
Language	0.08839	0.01320	6.69488	0.000	$2.16e-11$ ***
Cond	0.09969	0.17007	0.58616	0.558	0.5578
Time:Language	-0.05508	0.02848	-1.93438	0.053	0.0531 .
Language:Cond	1.35165	0.02641	51.17822	0.000	$< 2e-16$ ***
Time:Cond	0.54240	0.22056	2.45924	0.014	0.0139 *
Time:Language:Cond	1.54676	0.05659	27.33337	0.000	$< 2e-16$ ***

3.3   Analysis of the Two Competitors vs Color Competitor conditions

The Two Competitors condition was used as filler trials intended to add variability to the types of displays used in the study. However, previous studies have compared the results of this condition with those of the Color Competitor condition in order to investigate the derivation of contrastive inferences in English (Sedivy, 2003, 2004; Aparicio et al., 2016; Rubio-Fernandez et al., under review). We report additional analyses of these two conditions for completeness, although they do not directly test our key hypotheses.

Previous studies comparing the Two Competitors vs Color Competitor conditions used long preview windows that allowed English listeners to anticipate the noun in the Two Competitors condition through the derivation of a contrastive inference (see Rubio-Fernandez et al. (under review) for discussion). Since we used short preview windows of 400ms in all our conditions, we expected weaker results than previous studies. That is, we suspected that English listeners in Experiment 2a and Spanish listeners tested in English in Experiment 2b may not have sufficient preview time to derive an anticipatory inference in the Two Competitors condition.

We analyzed English listeners and Spanish listeners tested in English on the Two Competitors and the Color Competitor conditions. We considered the maximal logical model:

$$Target \sim Time * Language * Cond+$$
$$(1 + Time|Participant)+$$
$$(1 + Time|Item)$$

In the language variable, English listeners were coded as -0.5 and Spanish listeners tested in English were coded as 0.5. In the condition variable, the filler trials were coded as -0.5 and the Color Competitor condition as 0.5 [1]. The largest model that converged was

$$Target \sim Time * Language * Cond+$$
$$(1|Participant)+$$
$$(1 + Time|Item)$$

Scaled residuals:				
Min	1Q	Median	3Q	Max
-2.9615	-0.8065	-0.4493	0.9054	6.5398

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.1184	0.3441	
Item	(Intercept)	0.1685	0.4105	
	Time	0.4117	0.6416	-0.28

Fixed effects:					
	Estimate	Std. Error	z value	Pr($> |z|$)	Pr($> |t|$)
(Intercept)	-0.36207	0.09712	-3.72817	0.000	0.000193 ***
Time	1.21279	0.13194	9.19220	0.000	$< 2e-16$ ***
Cond	0.14696	0.16828	0.87334	0.382	0.382476
Language	0.01158	0.09826	0.11788	0.906	0.906164
Time:Cond	0.25449	0.26487	0.96078	0.337	0.336661
Cond:Language	-0.13876	0.02718	-5.10531	0.000	$3.30e-07$ ***
Time:Language	0.08668	0.02829	3.06369	0.002	0.002186 **
Time:Cond:Language	0.37164	0.05644	6.58502	0.000	$4.55e-11$ ***

The additional analyses of the Two Competitors vs Color Competitor conditions showed that English listeners did not reveal sensitivity to pragmatic contrast in the Two Competitors condition. We explain this lack of an effect as a result of the short preview window used in our study.

Contrary to what we observed with English listeners, Spanish listeners tested in English revealed a target advantage in the Two Competitors condition relative to the Color Contrast condition. This pattern of results confirms that

pragmatic contrast can affect real-time language processing in some visual contexts (see Sedivy, 2003, 2004; Aparicio et al., 2016; Rubio-Fernandez et al., under review). However, we interpret the target advantage observed in Experiment 2b as a familiarity effect since Spanish participants had interpreted color adjectives contrastively when first tested in Spanish in Experiment 2a (see Fig. 4 and eye-tracking analyses in the main text). We suppose that this first experience with the task may have allowed Spanish listeners to derive contrastive inferences in Experiment 2b, despite the short preview window.

In summary, the results of the Two Competitors vs Color Competitor conditions seem to suggest that, without a sufficiently long preview window or some experience with the task, pragmatic contrast may not affect real-time processing in this paradigm. Future studies should further investigate these open questions.
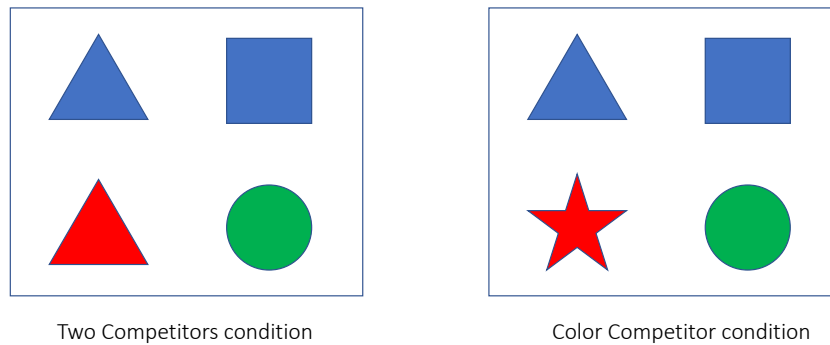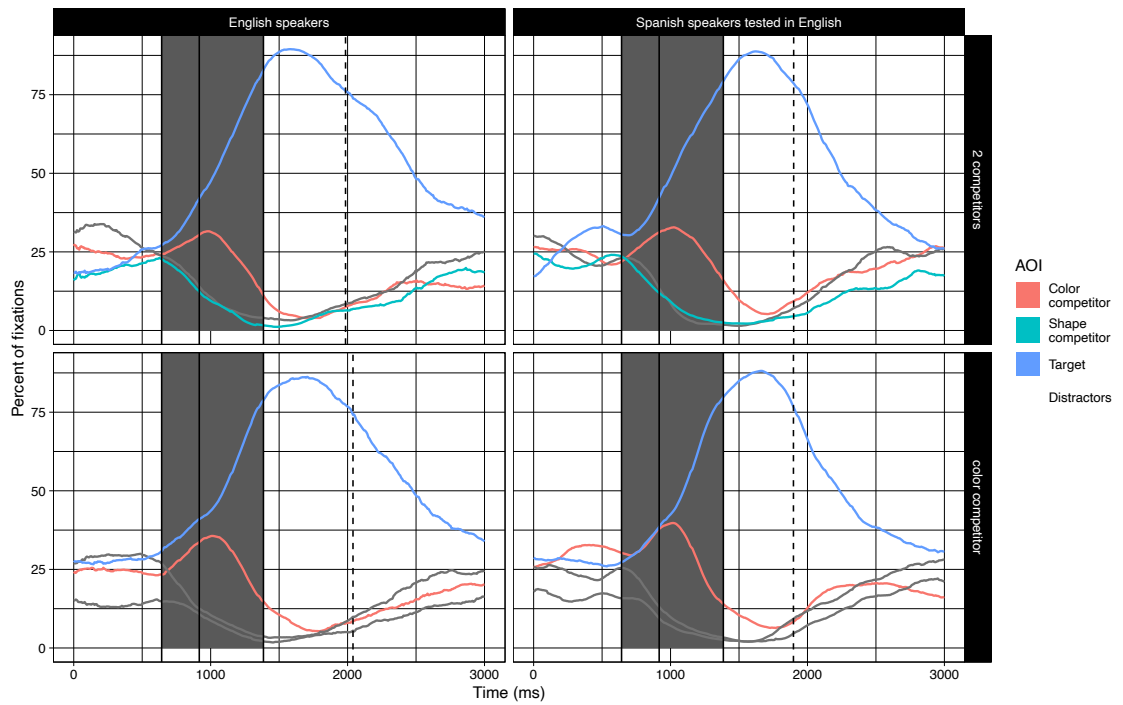

Figure 3: Sample displays.

Figure 4: Supplemental looking plots.

References

[1] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press, 2006.

[2] T Florian Jaeger. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446, 2008.

[3] Cesko C. Voeten. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*, 2020. R package version 1.5.