Why do objects have many names?
A study on word informativeness in language use and lexical systems

Eleonora Gualdoni*
Universitat Pompeu Fabra
eleonora.gualdoni@upf.edu

Gemma Boleda
Universitat Pompeu Fabra
ICREA
gemma.boleda@upf.edu

Abstract

Human lexicons contain many different words that speakers can use to refer to the same object, e.g., *purple* or *magenta* for the same shade of color. On the one hand, studies on language use have explored how speakers adapt their referring expressions to successfully communicate in context, without focusing on properties of the lexical system. On the other hand, studies in language evolution have discussed how competing pressures for informativeness and simplicity shape lexical systems, without tackling in-context communication. We aim at bridging the gap between these traditions, and explore why a soft mapping between referents and words is a good solution for communication, by taking into account both in-context communication and the structure of the lexicon. We propose a simple measure of informativeness for words and lexical systems, grounded in a visual space, and analyze color naming data for English and Mandarin Chinese. We conclude that optimal lexical systems are those where multiple words can apply to the same referent, conveying different amounts of information. Such systems allow speakers to maximize communication accuracy and minimize the amount of information they convey when communicating about referents in contexts.

1 Introduction

A pervasive property of human lexical systems is that many names can be assigned to the same object. In other words, our semantic system allows for a soft mapping between referents and words (Rosch and Mervis, 1975; Snodgrass and Vanderwart, 1980; Graf et al., 2016; Gualdoni et al., 2023). For instance, speakers can call the same chip *purple* or *magenta* (Monroe et al., 2017), and the same animal *dog* or *Dalmatian* (Graf et al., 2016; Silberer et al., 2020).

At the same time, a large body of literature has claimed that human lexicons are optimized for efficient communication, which implies allowing for accurate communication exchanges while maintaining a compact size (Regier et al., 2015; Xu et al., 2020; Zaslavsky et al., 2018). The existence of a soft mapping, which is not the most compact solution possible, may appear on a first glance at odds with such a pressure for efficiency. In this paper, we ask: is a soft mapping between referents and names an efficient solution? In an analysis of color naming data for English and Mandarin Chinese, we show that, indeed, at least for our domain of interest, a soft mapping is an efficient solution in that it achieves a good a trade-off between the amount of information that speakers have to convey in their *contextual* interactions, on the one hand, and the overall communicative accuracy they can achieve, on the other.

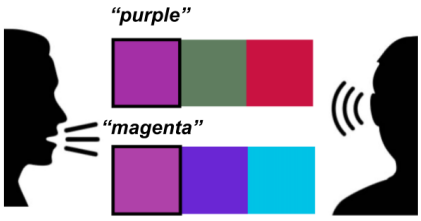


Figure 1: To allow successful identification of a target color chip (in the black frame) within a grid of candidates, a general term like *purple* is sufficient when the context is not challenging (above). A more specific name like *magenta* is needed when the distractors compete more with the target (bottom) —data from Monroe et al. (2017).

Indeed, communication exchanges between interlocutors take place in rich visual contexts. The dynamic nature of our environment and of speakers' goals constrains naming choices. In situations where it's essential to distinguish one item from context objects, some names can be better than others (Graf et al., 2016; Monroe et al., 2017; Mädebach et al., 2022): for instance, when we want our florist to hand us a bouquet of our favourite

*Currently at Apple.

flowers, the word *flowers* does not provide enough information, while the word *daisies* does. Monroe et al. (2017) collected experimental data on the phenomenon, which we analyze in the current study. They asked pairs of speakers and listeners to communicate about target color chips appearing in a grid surrounded by distractor chips —see Figure 1. When the target chip is easily distinguishable from the distractors (top), a general term like *purple* might suffice. However, in a more challenging context, where target and distractors are similar (bottom), a more precise term like *magenta* might be necessary to ensure successful communication. What are the consequences of this on the structure of lexical systems?

Good lexical systems need to be simple, which minimizes cognitive load, and informative, which maximizes communicative effectiveness (Regier et al., 2015). Studies have formalized this principle within an information-theoretical framework, showing that human systems optimize a trade-off between the amount of information provided and system complexity (Regier et al., 2015; Zaslavsky et al., 2018; Xu et al., 2020; Zaslavsky et al., 2021). While these studies often account for flexible semantic mappings (Zaslavsky et al., 2018), they do not study communication as *situated*, with speakers and listeners interacting in an always changing environment.

In this work, we explore why a soft mapping between referents and names is a good solution for in-context communication. Note that most research on the semantic properties of the lexicon has so far focused on a different aspect of the soft mapping between language and its use, complementary to the one we study here: ambiguity and polysemy, or the fact that most words have multiple meanings (Juba et al., 2011; Piantadosi et al., 2012; Regier et al., 2015; O'Connor, 2015). We reverse the question, asking why a referent can be described with different words (Graf et al., 2016). This phenomenon entails that similar, overlapping meanings can be denoted by different words.

Our method introduces a measure of **word informativeness** based on word denotations and grounded in a visual space, which can also be used to measure the information provided by lexical systems as a whole. With it, we analyze the color naming systems of English and Mandarin Chinese, and claim that their structure is key to achieve successful communication *in context*, with interlocutors communicating in differently challenging situations. We first replicate findings from previous studies, showing how speakers adjust their lexical choices to context pressures, leveraging a flexible mapping between referents and words (***I* and Language Use**). We then move to the system level, and show that alternative systems with no such flexible mapping are sub-optimal (***I* and Language Systems**).[1]

2 Related work

Studies modeling language use have explored how speakers adapt their referring expressions and naming choices to the local context in which target referents appear (Graf et al., 2016; Monroe et al., 2017; Degen et al., 2019; Mädebach et al., 2022) or to their communicative goal (Van Der Wege, 2009; Mädebach et al., 2022). These patterns have been formalized in the unified quantitative framework of Rational Speech Act theory (RSA; Frank and Goodman, 2012; Goodman and Frank, 2016; Franke and Jäger, 2016; Graf et al., 2016; Degen et al., 2019).[2]

RSA models focus on the **contextual informativeness** of referring expressions and utterances: the information that a word provides is measured *in context*, factoring in similarities and differences between a target referent and context objects. If a target object, e.g., a dog, appears in a context surrounded by other dogs, the word *dog* will not provide enough information about the target referent, and speakers will avoid it, choosing a more specific expression like *Dalmatian*, in order to help listeners identify the target (Graf et al., 2016). Speaker production choices are also constrained by considerations about utterance cost, often measured in terms of utterance length. Speakers are hypothesized to choose referring expressions to maximize a utility function, trading off the maximization of the contextual informativeness with the minimization of the production costs. The RSA tradition, given this major focus on context-dependent word informativeness, does not discuss properties of lexical systems as a whole.

Cross-linguistic studies on lexical systems have highlighted that, even though different languages partition their semantic space in different ways, this

[1] Scripts are available at `https://osf.io/n3cxh/`.

[2] Another theoretical framework that places emphasis in speaker-hearer interaction mechanisms is Bidirectional Optimality Theory (Blutner et al., 2003; Benz and Mattausch, 2011).

variation is constrained. The structure of lexical systems is believed to derive from the competing communicative principles of informativeness and simplicity, and languages optimize this trade-off —similar in nature to the one discussed for language use— in different ways (Regier et al., 2015; Zaslavsky et al., 2018; Xu et al., 2020; Zaslavsky et al., 2021). In this tradition, rooted in rate-distortion theory, the informativeness of a word is inversely related to the reconstruction error caused in a listener when a speaker uses it to describe a referent. In this sense, word informativeness is **non-contextual**: the informativeness of a word *w* for a target object *t* relates to the word's semantics and the referent's properties, and is not conditional on the *local context* in which *t* appears. This feature is in common with the word informativeness measure we adopt in this study. Of note, Zaslavsky et al. (2020) showed theoretical connections between the objective proposed in the RSA framework and rate-distortion theory, suggesting that similar pressures guide the evolution of lexical systems and their pragmatic use (on a similar topic, see also Brochhagen et al. 2018).

In this work, we propose a new measure of word informativeness that allows us to study speakers' adaptation to context in language use as well as the structure of lexical systems as a whole, bridging a gap between approaches focusing on contextual informativeness and approaches focusing on lexical informativeness —see Section 3.2.

3 Methods

3.1 Dataset

We use Monroe et al. (2017)'s dataset of color chips, including both the English and the Mandarin Chinese data.[3] The English dataset is annotated with more than $53K$ referring expressions collected in a dyadic reference game. In each round, a target color chip is presented to two players in a grid showcasing two other distractor chips, in random order. One player –*speaker*– is tasked to unambiguously describe a target chip, allowing the other player –*listener*– to guess the target chip —see Figure 1 for an illustration. Crucially, the same color chip is annotated multiple times, in differently hard contexts, as defined by the visual distance between target and distractor chips. Such feature of the data enables the analysis of how speakers adapt their referring expressions to the context. Monroe et al. (2017) found that speakers produce longer and more specific referring expressions in harder contexts (specificity was measured via WordNet; Miller 1994). Since we are interested in analyzing properties of the lexicon, after cleaning the data to remove spelling mistakes and noisy annotations (e.g. greetings between annotators), we subset the dataset, considering only rounds that were successfully solved with a single word. This leaves us with 16,168 data points. The Chinese dataset, constructed in the same way, is smaller: it contains around $2K$ referring expressions, and 749 rounds successfully solved with a unique word.

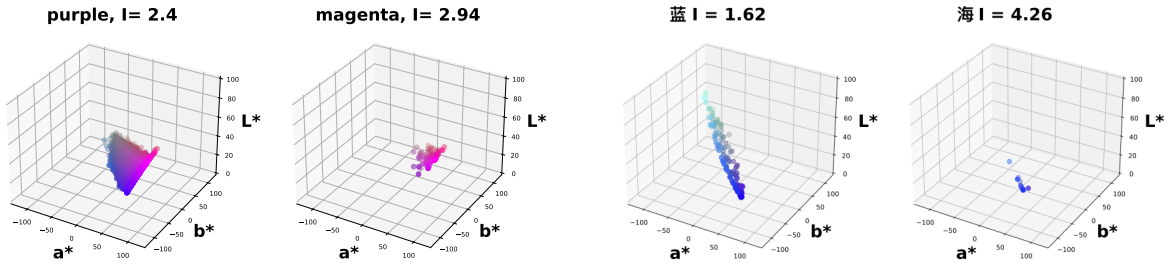
3.2 Word informativeness

We propose a new measure of word informativeness (I).[4] Our measure is inspired in separate traditions in semantics, which have alternatively highlighted the role of things in the world (denotation) or concepts in the mind. First, following the emphasis on reference of formal semantics (Dowty et al., 1981), we ground word meaning in the set of objects that the word denotes —in the case study of this paper, the set of color chips that have been labeled with a given word by the participants. Second, we assume that meanings are convex regions in a meaning space, as in the more cognitively oriented Conceptual Spaces framework (Gärdenfors and Williams, 2001; Gärdenfors, 2014). This way, we approximate the meaning of a color term in terms of a region in the visual space of colors defined by the specific color chips that have been labeled by the color term by a speaker (Erk 2009 is an early example of this kind of approach); and we assume that the region is convex when measuring informativeness, as follows.

The intuition behind I is that smaller volumes in a visual feature space provide more information about a referent than larger volumes: knowing that a referent's visual features are located in a small volume of the space informs a listener about what the referent looks like more than does a large volume. In other words, general words, like *purple* in the case of colors, or *person* in the semantic do-

[3] Distributed under a CC-BY 4.0 license.

[4] We decided to propose this new measure instead of building on the information-theoretic or the RSA traditions because of its simplicity and adequacy for our research question. Note that integrating both system-level and context-dependent informativeness in the aforementioned frameworks is a challenging problem, which implies defining a multi-objective function that interlocutors are believed to optimize, with a well-defined trade-off between task-general and context-dependent pressures; see Gualdoni et al. (2024) for a first attempt.



(a) (b)

Figure 2: Denotation in the CIELAB color space of the words *purple* and *magenta* (a) and 蓝 "blue" and 海 "ocean" (b). Note that there is a difference in numbers of objects, that we control for when computing I. A color chip called 海 "ocean" (b) would not be located in the top and lighter part of the 蓝 "blue" denotation region; more specific names denote objects occupying smaller volumes in a visual feature space. Smaller volumes correspond to more information provided by the word to a listener, and higher utterance costs for a speaker. Best viewed in color.

main of people, are labels for objects that are less similar to each other than the referents of specific words, like *magenta*, or *skier* —see for instance the denotation of *purple vs magenta* in Figure 2a and the denotation of 蓝 and 海 in Figure 2b, and Appendix C for other examples. This results in more specific words being denoted by smaller volumes in a visual feature space (Gualdoni et al., 2023), which we posit corresponds to higher amounts of information conveyed to a listener.

Following our conceptualization, we define the informativeness of a word *w* (I_w) as follows. Given the denotation of w in the space, we compute a measure of the spread of its visual features (S_w), based on the average distance between pairs of objects o that have been referred to by *w*:

$$S_w = \frac{1}{N} \sum_i \sum_{j \neq i} d(o_i, o_j) \quad (1)$$

Then, we define word informativeness as:

$$I_w = \frac{1}{S_w} \quad (2)$$

In this work, $d(o_i, o_j)$ is the Euclidean distance in the CIELAB space (Brainard, 2003) between objects o_i and o_j called by w. The CIELAB space is a color representation model designed to be more perceptually uniform than other accounts, in which Euclidean distances mirror perceptual distance for the human eye (Brainard, 2003). N is the number of object pairs.[5] The same measure could be applied to other metrics and feature spaces as well, modeling nouns from other domains, or other parts of speech as well, e.g. adjectives.[6]

In the English portion of Monroe et al. (2017)'s dataset, we compute I_w for each color name w appearing at least 10 times, and in the Chinese portion we set the threshold at 5 occurrences.[7] We obtain high I_w scores for words like *olive*, *cyan*, or *lavender*, in English, and 灰 "ash" , 橄榄 "olive" , or 海 "ocean" in Chinese; and low I_w scores for words like *blue*, *purple*, and *green*, in English, and 蓝 "blue" , 橙 "orange" , or 红 "red" in Chinese. Speakers generally prefer to refer to objects at their basic level (Rosch and Mervis, 1975; Jolicoeur et al., 1984), such as *blue* or *purple* in the color domain (Collier, 1973), making more specific names like *magenta* or *turquoise* rare options. Rare words come with higher costs, for instance in terms of reading times (Smith and Levy, 2013) or naming latencies (McRae et al., 1990). At least in the set of words we study here, words with higher I_w, not being basic level categories, are expected to be more costly as well.

To connect language use to lexical systems, we define the informativeness of a lexical system L as the average over the I_w of the words uttered to solve N interactions:

[5] Since most I_w scores were of the order of 10^{-2}, we multiply all the scores by 10^2 for readability.

[6] This measure is instead not directly not applicable to other parts of speech denoting relations, such as verbs and adverbs, which cannot be easily reduced to regions in a meaning space (Gärdenfors, 2014).

[7] Since some color names map to more data points than others, to avoid size effects on the value of I_w, we adopt a sampling strategy: if a color name has more than 100 chips associated, we randomly sample N chips for T times, and average the I_w values obtained for each sample. Our results are robust to different sampling sizes and numbers of iterations. We set $N = 100$ and $T = 30$.

$$I_L = \frac{1}{N}\sum_{i=1}^{N} I_w^i \quad (3)$$

4 *I* and Language Use

As discussed in Section 2, models of language use predict that, in harder contexts, speakers will utter longer and more specific referring expressions to achieve successful communication. In our analysis of language use, we aim at replicating these findings with our word informativeness measure and Monroe et al. (2017)'s data. Easier contexts are those where targets and distractors share fewer properties (Graf et al., 2016; Degen et al., 2019) or, in the case of color chips, target and distractor chips are further away in the color space (Monroe et al., 2017). Our expectation for I_w is that in harder contexts higher I_w will be needed to reach successful communication —Figure 1, bottom. Recall that we are especially interested in exploring what happens *to the same target* in different contexts. Thus, we subset the data to keep chips that appear at least twice in the dataset (see Appendix B for models fitted on all the data). This leaves us with 5491 data points across 2524 target chips, for English, and 60 data points across 29 target chips, for Chinese.

Models We use the distance between the target chip and the hardest distractor as a measure of **context ease**: the larger the distance, the easier the task.[8] We build a linear mixed-effects model, predicting I_w based on context ease. We add random intercepts and random slopes for the target chips (for English and Chinese) and for the worker ids (for English only, since they are not available for the Chinese data). Our hypothesis, based on previous work, is that easier visual contexts will be characterized by a decrease in I_w.

Results We replicate findings from previous literature with our I measure and data for English (Table 1, first row): for the same target, when the context is easier, lower values of I_w allow for communication success. We do not find an effect for Chinese (Table 1, second row). Of note, as mentioned above, if we consider only the chips that appear at least twice in the dataset, we only have 29 possible targets to fit the model on Chinese, which is probably too little to identify an effect (see Appendix B for results on the whole data).

		Estimate	Std. Error
English	Intercept	3.51***	0.05
	Ctx ease	-0.01***	0.00
Chinese	Intercept	2.54***	0.36
	Ctx ease	0.00	0.01

Table 1: Fixed effects of the linear mixed-effects model fitted on the English and Chinese data subset of repeated chips. Asterisks express p values: *** = $p < 0.001$.

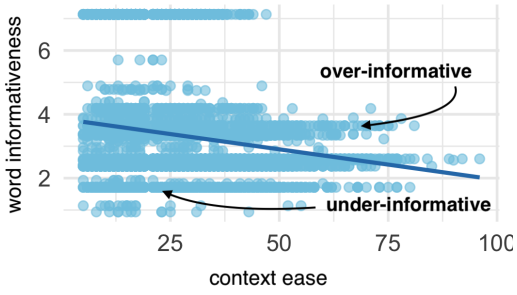


Figure 3: Relationship between context ease and word informativeness (I_w) in the portion of Monroe et al. (2017)'s English dataset considered in Table 1. Communication in easier contexts can be successful with less informative words.

Considering our results for English, we can see that, on average, for the same chip, an increase of 50 in context ease leads to a decrease of 0.5 in name I_w. How to interpret this? For I_w, this means moving, approximately, from *magenta* to *purple* ($I_w = 2.93$; $I_w = 2.30$) or from *grass* to *green* ($I_w = 3.07$; $I_w = 2.59$). As for context ease, in Figure 1 - top, the distance between the purple target and the green distractor (middle) is 54, while in Figure 1 - bottom, the distance between the magenta target and the purple distractor (middle) is 17. Therefore, moving from the bottom case to the top case means increasing context ease of 37.[9]

The relationship between context ease and I_w for the English data subset is illustrated in Figure 3. Even if the general trend follows our hypothesis, there are some data points that indicate an over-informative (center-right) or under-informative (bottom-left) behavior by speakers, which we analyze next.

Analysis of mismatches A qualitative inspection of the mismatches yields two trends. First, some

[8] The distance to the other distractor, in our data, is quite highly correlated with the distance to the closest one, which is supposed to compete more with the target (r=0.58, p<0.001). Therefore, we only consider the latter.

[9] Here we are considering, for the sake of the example, the chips in the middle as the only distractors.

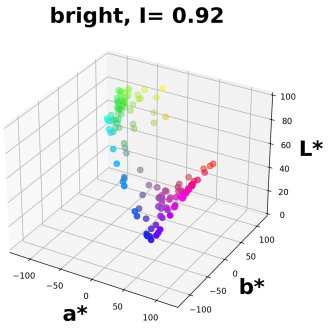


Figure 4: Words like “bright” or “dark” are denoted by non-convex regions, resulting in low informativeness (I_w) scores.

(a) *mint*: I_w = 3.34; context ease: 51

(b) *blue*: I_w = 1.71; context ease: 6

Figure 5: Typicality effects in language production. A word with high *I* like *mint* (panel a) can be used when the context is not hard, if the target is very typical for that word. A word with low *I* like *blue* (panel b) can solve the ambiguity in a very hard contexts, if the target is much more typical for the color compared to the distractors.

interactions in very hard contexts are unintuitively solved with low-informativeness words. The majority of these cases comes from the use of the words *bright*, *dark*, and *light*; or 亮 “bright”, 暗 ‘dark”, or 浅 “pale” . These words are characterized by a non-convex shape in the visual feature space: adjectives like *dark* and *bright* can apply to many different chips that are far from each other in the space —see Figure 4. Our I measure, which is based on the assumption of convexity, results in very low informativeness scores for them.[10]

We also find pragmatic effects related to object prototypicality. Psycholinguistic studies have analyzed the effects of prototypicality in descriptive naming tasks (Snodgrass and Vanderwart, 1980; Brodeur et al., 2010; Liu et al., 2011; Tsaparina et al., 2011; Gualdoni et al., 2023), showing that the probability of producing a given object name increases with the object’s typicality for the name. Graf et al. (2016) found this effect also for in-context communication. They found that speakers deviate from frequent words to use a more costly, specific name when the target is very typical for it, even if a less costly name would suffice to identify the target. In this sense, typicality modulates the cost of the word. We find a similar pattern: when the target is very typical for a name, speakers can be over-informative, producing words with high informativeness, thus more costly, even in cases of easy disambiguation —see, for instance, the mint chip in Figure 5a, a context in which “green" would suffice.

As for under-informativeness, speakers can produce words with low informativeness in hard contexts, successfully solving the ambiguity anyway —see, for instance, the blue chip in Figure 5b. We interpret this as the result of interlocutors’ pragmatic iterative reasoning about word interpretations (Goodman and Frank, 2016; Graf et al., 2016; Degen et al., 2019): in a hard context with multiple chips that could be called *blue*, if the speaker uttered *blue*, it is likely that their intention was to refer to the most prototypical blue. Object prototypicality and interlocutors’ reasoning expand the information provided by words beyond denotation, or more generally the semantics of words. In other words, pragmatics enriches word meanings.

5 *I* and Language Systems

We have seen that words with different informativeness values are used by speakers in differently hard contexts. What are the consequences of this on the structure of the lexical system? We argue that, to communicate successfully across differently hard interactions, we need a lexical system where multiple entries providing different amounts of information map to the same referent, allowing for a dynamic adaptation in lexical choice. We first formalize this idea in a simulation, and then run an empirical test to confirm it.

Simulation Table 2 exemplifies a lexical system with a soft mapping between referents and names, listing 6 referents with 2 possible names each. Note

[10] Of note, given the setup of the referential game these words could be abbreviations for longer and syntactically more complex referring expressions like *dark green* or *the dark one*: different measures may need to be designed to assess the information provided by more complex constructions.

that, since general names denote larger volumes in feature spaces, it is more likely for objects to share general names (e.g. *blue*) rather than specific ones (e.g. *teal*). However, since lexical systems are complex and not perfectly organized hierarchically, it is also possible to encounter pairs like *referent 2* and *referent 3* that share a specific name (*teal*), but not the general one (*blue* vs. *green*). This may be due to prototypicality: *referent 2* may be more typical for the color blue, and *referent 3* for the color green.

referent id	general name	specific name
referent 1	blue	turquoise
referent 2	blue	teal
referent 3	green	teal
referent 4	purple	magenta
referent 5	purple	mauve
referent 6	purple	mauve

Table 2: Naming system for 6 hypothetical referents, with a soft mapping between referents and words.

A lexical system like the one just described is not the most compact option: it lists 7 words for 6 referents, while, for instance, keeping only the general names would result in 3 words, and keeping only the specific names would result in 4 words. However, as we will show, this kind of system is more efficient: given that referents appear in context, the system can maintain high accuracy in communication, allowing a listener to identify the referent, while minimizing the overall information provided by speakers with their utterances.

How is this achieved? Imagine that each referent can appear in a visual context with another referent, with uniform probabilities (e.g., {*referent 1-referent 2*}, {*referent 1-referent 3*}, and so on). Assume furthermore that a listener's accuracy in guessing the target is at chance (50%) if the name uttered to describe the target applies to the distractor as well. Then, a speaker-listener pair could achieve very high accuracy by leveraging the system structure —as we have shown in Section 4— uttering the general name (low informativeness) when the two referents do not share it (easy context), and the specific name (high informativeness) when they do (i.e. in a harder context where the objects share more properties).

We run a simulation with this setup, using the color data of Monroe et al. (2017). In particular, we use the target chips that were annotated with at least two different names, and the I_w values of their corresponding names; and we generate from these data all the possible target-distractor pairs. Results are reported in Table 3, first two columns.

The simulation confirms our hypothesis: for both English and Chinese data, the best accuracy-cost trade-off is achieved by the actual system. The actual naming system achieves the highest accuracy (98% English / 99% Chinese) with quite low informativeness I_L ($2.78/1.99$). The only cases where communication success is at chance are those where referents share both general and specific names, akin to the case of {*referent 5 - referent 6*} in Table 2. The other systems are sub-optimal. The hypothetical system keeping only the general name of each referent has a lower I_L (2.56 / 1.83) but achieves a lower accuracy as well (93% in both cases). The hypothetical system keeping only the specific name of each referent achieves an accuracy of 96% / 98%, comparable to the one of the actual system (if slightly lower), but exhibits a much higher I_L of 3.99 / 3.13. Mistakes occur in cases where the referents share both specific and general names (as was the case for the actual system), or to the cases where referents share the same specific name, but not the general one (as in the *teal* example above).

Overall, thus, a system with a soft mapping between referents and names is an optimal solution, maximizing communicative accuracy with lower overall I_L.

Empirical test We collect human data to complement our simulation. We sample 100 target-distractor1-distractor2 datapoints from the English dataset, uniformly distributed with respect to their context ease, and consider the name that the target received in the sampled triplet as a reference name. To generate lexical systems alternative to the actual one, we simulate for each target a more general name (lower informativeness) and/or a more specific name (higher informativeness). We do so by taking the name with highest or lowest I_w that the same chip received across contexts. This way, we make sure that the word is adequate for the chip.[11] In order to measure the accuracies achieved by the different resulting lexical systems, we asked 3 English native speakers (unrelated to this study) to act as listeners, guessing the target based on the word we provide –see Appendix A for further details.

[11] Given that only a few chips were annotated more than twice in different contexts with a single word, in the majority of the cases we either simulate the general name, or the specific name.

	English - sim		Chinese - sim		English - emp.	
	Acc	I_L	**Acc**	I_L	**Acc**	I_L
Actual	98%	2.78	99%	1.99	96%	3.33
General	93%	2.56	93%	1.83	81%	2.36
Specific	96%	3.99	98%	3.13	89%	4.24

Table 3: Results of accuracy and I_L for actual *vs* hypothetical lexical systems (general words only and specific words only). Column 1 and 2: results of simulation; column 3: empirical data.

Results are reported in Table 3, column 3. The relationship between the listeners' performance in the 3 conditions mirror what we found in our previous simulation. The actual naming system achieves the highest accuracy (96%), with an intermediate I_L value (3.33). This makes it the best system: the general system we simulated comes with lower costs ($I_L = 2.36$), but is not accurate (81%), while the specific system we simulated is more costly (4.24), without a gain in accuracy (89%).

Note that the scores in our empirical test are generally lower, which is due to the sampled contexts being harder than in the simulation: in the simulation, we created all the possible target-distractor pairs, thus automatically generating a larger number of easier cases, given that each chip has a high similarity only with a few chips, and is visually very different from the majority of the other chips. To have a better grasp on our listeners' behavior, we next dive deeper into their mistakes.

Analysis The mistakes that the listeners made are in line with those that arose in the simulation. Figure 6 shows an example. The target chip in the black frame, called *pink* in the shown context, was assigned the name *mauve* as a simulated specific name. However, the specific name *mauve* can apply to the rightmost distractor as well, leading to a case where two referents appearing in the same context share the same specific name (as in the *teal* example), and in this case the listener failed to identify the target.[12]

Cases of this nature, which can result in mistakes in the annotation, are actually good examples of how pragmatics is again at play, expanding word meanings beyond denotational semantics. For instance, both *pink* and *mauve* could describe the target in Figure 6 in isolation. The word *mauve* is *per se* more informative than the word *pink*, but it is not *contextually* informative, since the distractor on the right may also be called *mauve*. Given that the target is more prototypical for *pink* than the distractor, a listener may expect the word *pink* –and not *mauve*– to be used to describe it, even if the context is hard and the word *pink* is less specific. Moving from *pink* to *mauve* increases word informativeness but does not factor in pragmatics, which in this case leads to unsuccessful communication (note that Figure 5b, discussed in Section 4, constitutes a successful case of the same type of pragmatic reasoning).

Figure 6: Target chip, called *pink* in this context, for which we simulated *mauve*, here misleading, as the rightmost chip was chosen instead.

6 Discussion

In this work, we have studied why a lexical system where multiple names map to the same referent is a good solution for human communication. We have done so by proposing a measure of word informativeness grounded in a visual feature space and based on word denotations.

Previous studies on the optimality of lexical systems often consider the number of lexical entries in a system as a measure of system complexity (Regier et al., 2015; Xu et al., 2020), with smaller lexicons preferable over large ones due to cognitive constraints. These approaches would fail at capturing how an increased lexicon size can become advantageous when we factor in communication in context, allowing for the minimization of the overall amount of information transmitted in language use. Our study, drawing a connection between lan-

[12] Recall that for some chips we could not simulate a name more specific / general than the actual one (already specific / general). As a sanity check, we report that our annotator's accuracy on this portion of data is in line with that of the actual system (95% for simulated specific data; 98% for the simulated general data), while it decreases for the portion of data with simulated names (81% for simulated specific; 69% for the simulated general).

guage use in context and the consequent structure of an efficient lexical system, bridges this gap.

Connecting properties of language production in context with properties of the lexicon is much in the spirit of previous work connecting language production and properties of grammars regarding language universals (Hawkins, 2004; Franzon and Zanini, 2023). Future work should explore parallelisms between the lexicon and the grammar in this respect. This is also intimately connected to diachronic dynamics, and the causes and consequences of semantic change. Adopting an evolutionary perspective, Gualdoni et al. (2024) study how a human-like semantics can emerge from contextually-rich, pragmatic interactions; and Kobrock et al. (2024) compare lexicons emerging in artificial agents that have access to context information to those of context-agnostic ones. Future work should delve deeper into how word informativeness and reference to objects in context interact to produce a given lexicon. A related question is how system learnability affects communication accuracy and shapes language evolution (on the topic see, for instance, Carlsson et al., 2024; Gyevnar et al., 2022; Tucker et al., 2022).

Our work also resonates with previous research on why the lexicon presents pervasive ambiguity, and specifically the fact that most words have multiple meanings (Juba et al., 2011; O'Connor, 2015; Fortuny and Corominas-Murtra, 2015; Piantadosi et al., 2012). As mentioned in the Introduction, lexical ambiguity corresponds to one-to-many relationships between words and meanings, while we have focused on many-to-one relationships between words and referents. We believe that both phenomena are different consequences of efficiency constraints acting on the lexicon. Our findings complement previous research (Piantadosi et al., 2012) by showing that, in general, many-to-many mappings between words and meanings can be characterized as efficient solutions.

An advantage of our specific approach lies in the simplicity and the flexibility of our informativeness measure, that could be adapted to other kinds of distributed representations, allowing us to study different phenomena besides referential language in language and vision. For instance, when the subject of a sentence is unexpected for a listener, a more informative word like *president* instead of *man* may be preferred (Aina et al., 2021). Our measure (which could be derived from language models considering the distances between contextualized embeddings of the same word) could allow for joint analyses of discourse and vision data under the same light, taking a step towards a unified view of human referential acts (see, for instance, Franzon and Zanini, 2023, for the analysis of a related phenomenon in morphology).

That being said, our measure only accounts for the semantic meaning of words, suffering from a limitation: in every interaction, pragmatics enriches word meanings and modulates the information provided by words in context, or their cost for speakers. A denotational measure of word informativeness cannot capture the full range of phenomena characterizing language production in context. Moreover, our formulation of the measure cannot accurately describe words denoted by non-convex regions in meaning spaces. We leave it to future work to define alternative measures that take into account more complex shapes (see Figure 4), while still being general enough.

Finally, in our analyses we have made the simplified assumption that speakers want to provide the right amount of information, avoiding over-informative utterances. There is literature showing that this is not always the case (Engelhardt et al., 2006; Koolen et al., 2011), and that the production of redundant and over-informative referring expressions can fall in the set of behaviors that maximize efficient communication (Rubio-Fernandez, 2016; Degen et al., 2019). Moreover, considerations beyond informativeness affect speakers' naming choices, e.g. speakers could choose *professor* instead of *woman* to highlight aspects of the referent contingently relevant to them (Silberer et al., 2020), even without any explicit pressure for discrimination. A more comprehensive analysis of language use should account for these factors as well.

7 Conclusion

Objects have many names. In this work, we have analyzed human color naming data exploring why some degree of soft mapping is a feature of an optimal lexical system, bridging the gap between analyses of lexical systems and language use. We conclude that systems where multiple words conveying different amounts of information can be used to describe the same referent are optimal, in that they maximize communicative accuracy while minimizing the amount of information conveyed.

8 Limitations

Our study aims at analyzing the structure of human lexical systems. However, given the scarce availability of cross-linguistic datasets studying situated communication, we limit ourselves to English and Mandarin Chinese; the latter, with less coverage than the former. This constitutes a limitation of our work, which would benefit from the analysis of a more diverse set of lexical systems. Along the same lines, our analysis is limited to the color semantic domain: validating the word informativeness measure on a different system of categories, as well as running the analyses on richer semantic domains, would strengthen our conclusions (see Gualdoni et al. 2023 for preliminary evidence in the domain of people).

It is also worth noticing that our analysis of the sub-optimality of hypothetical lexical systems is limited to two alternative systems that we could simulate with the data available to us. The overall considerations on the optimality of human-like lexical systems would benefit from the analysis of more, and more diverse, hypothetical alternatives.

Finally, our simulation in Section 5 relies on simplified assumptions, such as referents co-occurring in contexts with uniform probabilities. Estimating the real probabilities of referents co-occurring constitutes a big challenge, but would also constitute a great improvement for all studies interested in understanding the pressures that shape the human lexical system.

Acknowledgements

This research is partially an output of grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033, funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain) and has received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715154). We thank Thomas Brochhagen, Francesca Franzon and Louise McNally for feedback on an earlier version of the paper.

References

Laura Aina, Xixian Liao, Gemma Boleda, and Matthijs Westera. 2021. Does referent predictability affect the choice of referential form? a computational approach using masked coreference resolution. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 454–469, Online. Association for Computational Linguistics.

Anton Benz and Jason Mattausch, editors. 2011. *Bidirectional Optimality Theory*. John Benjamins.

Reinhard Blutner, Anne Bezuidenhout, Richard Breheny, Sam Glucksberg, and Francesca Happé. 2003. *Optimality theory and pragmatics*. Springer.

David Brainard. 2003. *Color Appearance and Color Difference Specification*, pages 191–216.

Thomas Brochhagen, Michael Franke, and Robert van Rooij. 2018. Coevolution of lexical meaning and pragmatic use. *Cognitive Science*, 42(8):2757–2789.

Mathieu Brodeur, Emmanuelle Dionne-Dostie, Tina Montreuil, and Martin Lepage. 2010. The bank of standardized stimuli (boss), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, 5:e10773.

Emil Carlsson, Devdatt Dubhashi, and Terry Regier. 2024. Cultural evolution via iterated learning and communication explains efficient color naming systems.

George A. Collier. 1973. *Language*, 49(1):245–248.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2019. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological review*.

David R. Dowty, Robert Eugene Wall, and Stanley Peters. 1981. *Introduction to Montague Semantics*. Springer.

Paul Engelhardt, Karl Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54:554–573.

Katrin Erk. 2009. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 57–65.

Jordi Fortuny and Bernat Corominas-Murtra. 2015. Introduction. on the locus of ambiguity and the design of language. *The Linguistic Review*, 32(1):1–4.

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

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35.

Francesca Franzon and Chiara Zanini. 2023. The entropy of morphological systems in natural languages is modulated by functional and semantic properties. *Journal of Quantitative Linguistics*, 30(1):42–66.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Caroline Graf, Judith Degen, Robert X D Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 2261–2266, Austin, TX. Cognitive Science Society.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2023. What's in a name? A large-scale computational study on how competition between names affects naming variation. *Journal of Memory and Language*, 133:104459.

Eleonora Gualdoni, Mycal Tucker, Roger P. Levy, and Noga Zaslavsky. 2024. Bridging semantics and pragmatics in information-theoretic emergent communication. In *Advances in Neural Information Processing Systems*. To appear.

Balint Gyevnar, Gautier Dagan, Coleman Haley, Shangmin Guo, and Frank Mollica. 2022. Communicative efficiency or iconic learning: Do acquisition and communicative pressures interact to shape colour- naming systems? *Entropy*, 24(11).

Peter Gärdenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT.

Peter Gärdenfors and Mary-Anne Williams. 2001. Reasoning about categories in conceptual spaces. In *Proceedings of the IJCAI*, pages 385–392.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.

Pierre Jolicoeur, Mark A. Gluck, and Stephen M. Kosslyn. 1984. Pictures and names: Making the connection. *Cognitive Psychology*, 16(2):243–275.

Brendan Juba, Adam Tauman Kalai, Sanjeev Khanna, and Madhu Sudan. 2011. Compression without a common prior: an information-theoretic justification for ambiguity in language. In *Proceedings of the Innovations in Computer Science*, Tsinghua University, China.

Kristina Kobrock, Xenia Isabel Ohmer, Elia Bruni, and Nicole Gotzner. 2024. Context shapes emergent communication about concepts at different levels of abstraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3831–3848, Torino, Italia. ELRA and ICCL.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Youyi Liu, Meiling Hao, Ping li, and Hua Shu. 2011. Timed picture naming norms for mandarin chinese. *PloS one*, 6:e16505.

Ken McRae, Debra Jared, and Mark S. Seidenberg. 1990. On the roles of frequency and lexical access in word naming. *Journal of Memory and Language*, 29(1):43–65.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Andreas Mädebach, Torubarova Ekaterina, Eleonora Gualdoni, and Gemma Boleda. 2022. Effects of task and visual context on referring expressions using natural scenes. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Cailin O'Connor. 2015. Ambiguity is kinda good sometimes. *Philosophy of Science*, 82(1):110–121.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cogn.*, 122(3):280–291.

Terry Regier, Charles Kemp, and Paul Kay. 2015. *Word Meanings across Languages Support Efficient Communication*, pages 237–263.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Paula Rubio-Fernandez. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020. Object naming in language and vision: A survey and a new dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5792–5801, Marseille, France. European Language Resources Association.

Nathaniel J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Joan Gay Snodgrass and Mary Vanderwart. 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, 6 2:174–215.

Diana Tsaparina, Patrick Bonin, and Alain Méot. 2011. Russian norms for name agreement, image agreement for the colorized version of the Snodgrass and Vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for modal object names. *Behavior Research Methods*, 43(4):1085–1099.

Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. 2022. Trading off utility, informativeness, and complexity in emergent communication. In *Advances in Neural Information Processing Systems*.

Mija M. Van Der Wege. 2009. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60(4):448–463.

Yang Xu, Emmy Liu, and Terry Regier. 2020. Numeral systems across languages support efficient communication. *Open Mind*, 4:1–14.

Noga Zaslavsky, Jennifer Hu, and Roger Levy. 2020. A Rate–Distortion view of human pragmatic reasoning.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *PNAS*, 115(31):7937–7942.

Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. 2021. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *43st Annual Meeting of the Cognitive Science Society*.

A Details on the data collection

Each annotator was presented with the same set of target chips to annotate, but with different names: one annotator received the actual system, one received the simulated general system, and one received the simulated specific system. Each block of questions contained 5 randomly placed controls, designed to ensure that annotators were paying attention to the task. These cases were made intentionally very simple. The data collection routine was written in Psychopy (Peirce et al., 2019). There was no time limit for completing the study. Instructions for annotators: "Welcome! In this study, we ask you to identify a target color chip in a set of 3 chips, based on a word. You will always see 3 color chips. Above them, there will be a word describing the target. We ask you to click on the target. Sometimes you will not be sure about your answer. Please make your best guess. Reply with what you think is the most plausible answer".

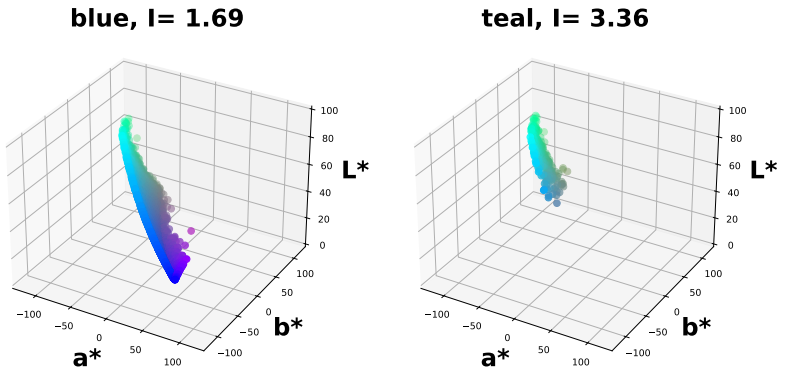
B Models fitted on all the data

Table 4 shows effects when modeling the whole data. For both English and Mandarin Chinese, we identify the expected trend (recall from Section 4, this effect disappears when we subset the data to keep only chips annotated at least twice across different contexts, which reduces the total of the target chips available to fit the model to 29).

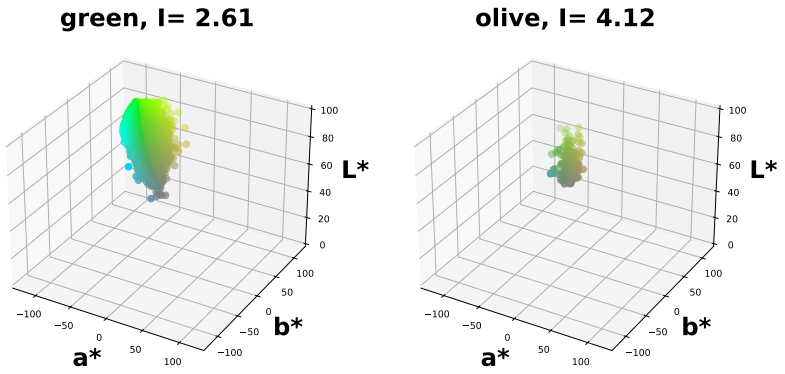
		Estimate	Std. Error
English	Intercept	3.80***	0.04
	Ctx ease	-0.02***	0.00
Chinese	Intercept	3.27***	0.14
	Ctx ease	-0.01*	0.00

Table 4: Fixed effects of the linear mixed-effects model fitted on the English data (random intercepts and random slopes for worker-ids) and of the linear model fitted on the Chinese data —without subsetting the data to include only chips annotated at least twice. Asterisks express p values: *** = $p < 0.001$; * = $p < 0.05$.

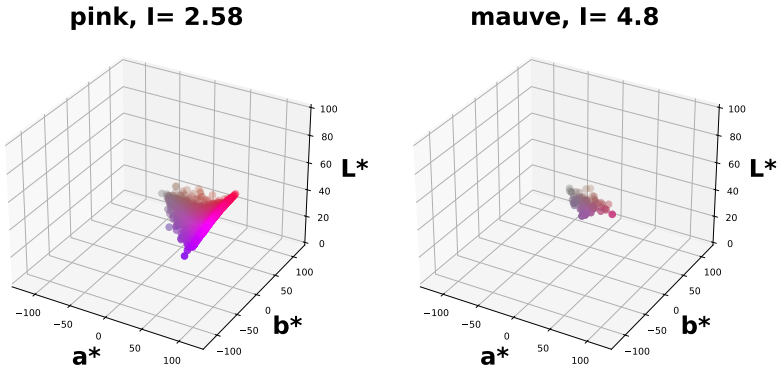
C Color denotation in visual space



(a)

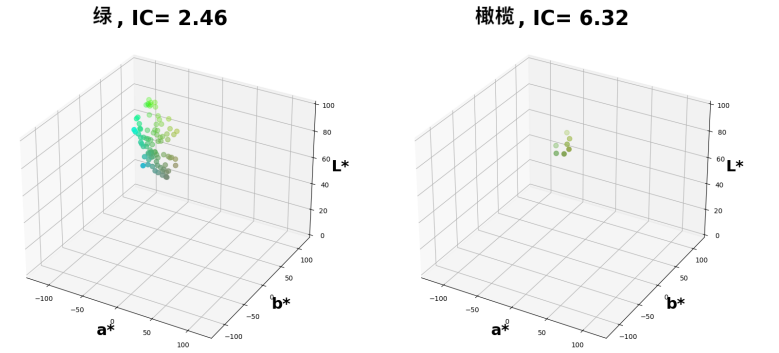


(b)

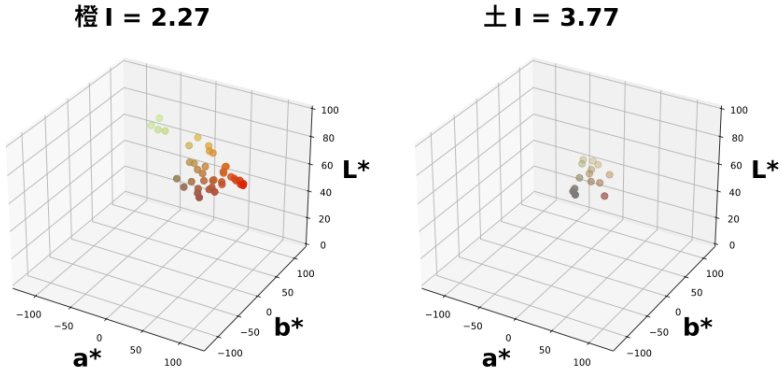


(c)

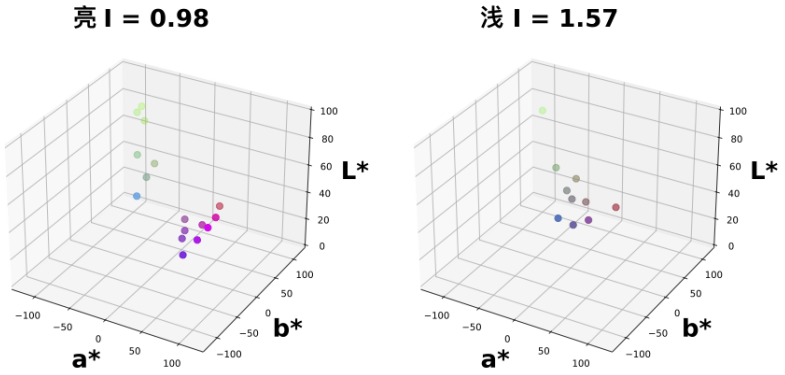
Figure 7: Denotation in the CIELAB color space for *blue* and *turquoise* (panel a), *green* and *olive* (panel b), *pink* and *blood* (panel c).



(a)



(b)



(c)

Figure 8: Denotation in the CIELAB color space for 绿 and 橄榄 (panel a), 橙 and 土 (panel b), 亮 and 浅 (panel c). Translations, in order: *green*, *olive*, *orange*, *soil*, *bright*, *pale*.