Language and Cognition (2023), 15: 4, 854–883
doi:10.1017/langcog.2023.29

ARTICLE

Quality, not quantity, impacts the differentiation of near-synonyms

Aja Altenhof[1,2] and Gareth Roberts[2]

[1]Linguistics and English Language, University of Edinburgh, Edinburgh, UK; [2]Department of Linguistics, University of Pennsylvania, Philadelphia, USA
Corresponding author: Gareth Roberts; Email: gareth.roberts@ling.upenn.edu

(Received 24 May 2022; Revised 05 May 2023; Accepted 19 June 2023)

Abstract
How much information do language users need to differentiate potentially absolute synonyms into near-synonyms? How consistent must the information be? We present two simple experiments designed to investigate this. After exposure to two novel verbs, participants generalized them to positive or negative contexts. In Experiment 1, there was a tendency across conditions for the verbs to become differentiated by context, even following inconsistent, random, or neutral information about context during exposure. While a subset of participants matched input probabilities, a high proportion did not. As a consequence, the overall pattern was of growth in differentiation that did not closely track input distributions. Rather, there were two main patterns: When each verb had been presented consistently in a positive or negative context, participants overwhelmingly specialized both verbs in their output. When this was not the case, the verbs tended to become partially differentiated, with one becoming specialized and the other remaining less specialized. Experiment 2 replicated and expanded on Experiment 1 with the addition of a pragmatic judgment task and neutral contexts at test. Its results were consistent with Experiment 1 in supporting the conclusion that quality of input may be more important than quantity in the differentiation of synonyms.

Keywords: synonymy; word learning; semantic differentiation; mutual exclusivity; statistical learning

1. Introduction

Synonymy, where two or more words refer to the same thing, is a common feature of the world's languages. Absolute synonymy, however, whereby synonyms can be substituted for one another in any context with no change to truth value, communicative impact, or connotational meaning, is exceedingly rare (Cruse, 1986; Ullmann, 1962). A typical case is exemplified by the synonym pair *awesome* : *fantastic.*

At first glance, *awesome* and *fantastic* might seem to meet the criteria for absolute synonymy. Swapping them between example sentences (1a) and (1b), for instance, does not substantially alter the intended meaning of either sentence.

(1) a. I had a fantastic time!
b. The show was awesome.

Both words, however, possess meanings aside from 'great'. A hobbit is a fantastic, but hardly awesome, creature; an atomic explosion might be awesome, but it is hardly fantastic. The vast majority of synonym pairs behave like this and can be termed *near-synonyms.* Near-synonyms can differ in both their stylistic and semantic effect, often across multiple dimensions simultaneously (DiMarco et al., 1993).

1.1. Semantic categories of synonymy

Near-synonyms can be broadly classified based on what differentiates them. Following Cruse (1986) and Grove (1973), Edmonds (1999) divided them into four groups: collocational/syntactic, stylistic, denotational, and expressive. Collocational, or syntactic, synonyms differ in how they interact with other words in the sentence (e.g., *die*: *pass away*), while stylistic synonyms differ with respect to such features as dialect or register (e.g., *pissed*: *drunk*: *inebriated*). Denotational synonyms differ in terms of the details of what they denote (e.g., *alligator*: *crocodile*, or *absorb*: *digest*: *assimilate*). Finally, expressive synonyms, the focus of the current study, differ with respect to the emotions, opinions, attitudes, or feelings implied on the part of the speaker. (For a general theory of expressives, see Potts, 2007.) This is distinct from words that are *denotational* synonyms with regard to emotion. For instance, the expressive synonyms in (2) differ with respect to expressive meaning, but not with respect to truth value (Cruse, 1986). That is, they all convey that Sally does not like to spend money, but differ with respect to the speaker's implied attitude toward this fact. In contrast, the denotational synonyms in (3) convey a similar meaning about Sally's emotion, though with subtle denotational differences; they do not imply a difference in the speaker's attitude, with the consequence that *astonished* and *amazed* are not expressive synonyms.

(2) a. Sally is stingy.
b. Sally is thrifty.
c. Sally is cheap.

(3) a. Sally is astonished.
b. Sally is amazed.

Expressive near-synonyms can differ based on their emotive quality in other ways. *Mommy* is more intimate than *mother*, for instance, while being *rejected* implies something more painful than being *declined* (Hayakawa, 1994). As in (2), near-synonyms often reflect differences in speaker attitudes, especially positive, negative, or neutral attitudes. The present study investigates what might cause language users to differentiate potential expressive synonyms.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

1.2. Explanations for differentiation

A potential explanation for the absence (or near absence) of true synonyms comes from work on biases supporting language acquisition. Most obviously, the *mutual exclusivity bias*, according to which people tend to assume that distinct words have distinct referents (Clark, 1992; Lewis et al., 2019; Markman & Wachtel, 1988), might support word learning generally but make true synonyms unlikely to survive. Evidence for this bias comes from experimental work in which children, when presented with one unfamiliar and one familiar object, will apply a new label to the unfamiliar object the majority of the time (Diesendruck, 2005). Adults perform similarly, reliably applying novel terms to novel objects, despite knowing that a familiar object could have more than one label (Golinkoff et al., 1992). This suggests that we should expect people to treat any set of synonyms as non-absolute even if no distinction in meaning is apparent.

Nonetheless, other work shows that the mutual exclusivity bias can be overridden by evidence. Savage and Au (1996) presented young children aged three to five with two novel labels for the same unfamiliar object. Half of the children accepted both labels, suspending mutual exclusivity. Similarly, when adults were given multiple labels for the same objects, they proved highly sensitive to co-occurrence statistics, even for relatively infrequent mappings. Learners were able to maintain multiple hypotheses about the meaning of a word, assigning likelihoods to possible candidates (Vouloumanos, 2008). Likewise, in artificial language paradigms, when faced with probabilistic input, adults often *probability-match*, reproducing input frequencies in their output (Hudson Kam & Newport, 2005), although the precise pattern of behavior depends on properties of the input distribution, with adults *over-matching*, or *regularizing*, in response to complex and highly scattered distributions (Hudson Kam & Newport, 2009). Such statistical-learning patterns are extremely important to language change. If learners closely match input distributions, then there should be relatively little change (although gradual change can occur over generations as a result of small biases in matching behavior; Kirby et al., 2014). Over- and under-matching behavior, by contrast, should lead to change.

This question of initial word distribution is also a particularly important one in the case of synonymy. It is unlikely for any potential synonyms to start off with identical grammatical, stylistic, and social distributions. If potential synonyms enter the language via contact, for instance, then we should expect them to carry social and stylistic associations arising as a result of that contact (Andersen et al., 2017). A well-known example concerns words for farm animals in English after the Norman Conquest of England in 1066. Words of Old English origin (such as ‘pig’, ‘sheep’, and ‘cow’) tend in Modern English to be used for the animals that the English-speaking peasants encountered in the fields, while French-origin equivalents (‘pork’, ‘mutton’, and ‘beef’) are used for the meat that the Anglo-Norman aristocracy had on their plates. Potential synonyms may also have different language-internal associations; for instance, while *awesome* and *fantastic* overlap considerably in meaning, differences between them still owe something to their associations with awe and fantasy, respectively. (This kind of inherited distinction is by no means guaranteed; few English speakers, for instance, are aware of the etymological connection between inspiration and respiration, and it seems to have no obvious effect on the words’ current use.) For these kinds of reasons, potential synonyms are likely to start off with asymmetries built in. If learners reproduce or exaggerate such existing asymmetries, including reanalyzing accidental asymmetries (such as random variation over

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

contexts or speakers) as meaningful, then external distribution is itself a reason to expect potential synonyms to be treated as distinct without the need to appeal to any internal cognitive bias actively pushing them apart. If this is the primary reason for the rareness of absolute synonymy, we should predict that, in cases where words do have equivalent distributions, absolute synonymy should in fact occur. We should also expect patterns of partial synonymy to be related to the historical distributions of the words in question.

This can be thought of as a *quantitative* account of the emergence of synonymy as it implies that the process of differentiation in learning is related to differing statistical distributions in input – probability matching in this case would be a source of non-synonymy. Accounts based on the existence of cognitive biases against synonymy, by contrast, can be thought of as *qualitative* and are associated with regularization in learning. They imply that true synonymy is inherently unstable, regardless of the statistical distribution of the terms involved. And, while patterns of partial synonymy may still reflect historical distributions, the relationship may be rather obscured, as these distributions would not be the only (or perhaps even main) driving force behind modern patterns.

1.3. Patterns of differentiation

In discussing distribution, it is worth laying out the ways in which potential synonym pairs can be distributed relative to each other. Putting the particular *semantic* relationship between potential synonyms (discussed above) to one side, there are three main categories of distributional pattern. First is the rare category of *true synonymy*, where the two words can be used fully interchangeably. The second category, at the other end of the scale, includes potential synonyms that are *fully differentiated* and are never used in quite the same contexts. The pair *easygoing*: *complacent* might serve as an example of this; while the two adjectives could be considered to refer to the same trait, the first has primarily positive connotations, while the second has negative ones. The third clearly defined category of synonymy concerns *partial differentiation* in which one word is used rather generally, while the other is more restricted. For example, *thrifty* and *stingy* both imply that a person is reluctant to spend money, but *stingy* is restricted to a particular (negative) context.

These three possibilities represent the clearest cases, but there are of course intermediate possibilities (such as words that are fully differentiated most but not all of the time). They also represent potential endpoints of statistical learning. If two words are *mostly* differentiated then over-matching could lead them to be absolutely *fully* differentiated. A more interesting case concerns synonyms that are used *mostly interchangeably*. In that case, we might expect learners to make the pattern fully interchangeable (and thus producing true synonyms), but – as discussed – this is rare. Patterns of over- and under-matching, in other words, apparently tend to occur in ways that increase differentiation rather than reduce it. As described above, a key question is whether this tendency is qualitative or quantitative in nature.

1.4. The present study

In terms of this qualitative–quantitative distinction, we can think of our question as concerning whether the different patterns of (full, partial, or non-) differentiation arise directly out of pre-existing distributions in the use of words, represent different outcomes of a cognitive bias against synonymy, or arise from an interaction of the two. The present study investigates this question experimentally. Specifically, if people are exposed to potential synonyms, what effect will distribution – and how *reliably* the words are distinguished by context – have on the extent to which they differentiate them? To examine this, we conducted an experiment in which participants were exposed to two novel synonymous verbs embedded in English sentences, and we manipulated what these sentences implied about the meaning of the words. For simplicity's sake, we focused only on *expressive synonymy*. That is, the containing sentences were varied to imply neutral, positive, or negative shades of meaning. Afterwards, learners were asked to generalize the verbs to new positive, negative (Experiments 1 & 2), and neutral (Experiment 2) contexts. In Experiment 2, participants also completed a sentence judgment task to rate the appropriateness of each word in the three contexts (Experiment 2).

2. Experiment 1

2.1. Overview

In Experiment 1, participants were exposed to two novel verbs that they were told had the same meaning. First, in the *Exposure phase*, each verb was presented multiple times, embedded each time in a different English sentence. Each sentence implied a negative, positive, or neutral meaning for the verb in question, and we manipulated whether implied negative or positive meaning was consistent across sentences, inconsistent, or absent altogether. Afterwards, in the *Generalization phase*, participants were shown new positive and negative sentences with missing verbs; their task was to choose one of the verbs to fill the blank in each sentence. We then measured the extent to which participants differentiated the verbs on the basis of positivity or negativity. We predicted that, if presented with a biasing distribution in the Exposure phase, participants would differentiate the two verbs in the Generalization phase in a way that was related to the biasing distribution. In the absence of such a bias (i.e., if neither verb was biased in its distribution toward positive or negative contexts), different possibilities presented themselves. One possibility was that participants might, in line with the input distribution, treat the verbs as synonymous. Alternatively, they might differentiate them in spite of the input distribution. There would be no quantitative, distributional reason to do so, but a cognitive bias against synonymy (such as a mutual exclusivity bias) might nonetheless motivate a qualitative assumption on the part of participants that the verbs must be distinct, causing them to generate a distinction themselves. A further question concerns the degree of bias in the distribution. Would participants presented with a distributional bias match the distribution in the Generalization phase, or would there be a threshold such that a small bias would (if it passed the threshold) have a similar effect to a large bias? An account in which the mutual exclusivity bias trumped probability matching would lead us to predict such a qualitative effect.

2.2. Method

2.2.1. Participants

A total of 362 participants (225 female, 134 male, 3 non-binary), aged between 18 and 76 (*median* = 30), were recruited from Prolific. All participants were English speakers and were compensated $2 for their time. A total of 58 participants (16%) reported that they spoke more than one language.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

2.2.2. Stimuli

Stimuli consisted of 72 sentences, which were divided equally between the Exposure phase and the Generalization phase. In Exposure, a third of the sentences featured a novel noun, *murp*. Each of the remaining sentences featured one of two novel verbs, *snater* or *fincur*. No sentence contained more than one novel word. Sentences were designed to be Positive (e.g., 'Wow! He got a certificate for _____'.), Negative (e.g., 'I hope he breaks his habit of _____ all the time'), or Neutral (e.g., 'She _____ this morning'.). To ensure that sentences conveyed the intended valence, an independent norming study was conducted. In this norming study, participants were shown each sentence with one of a variety of novel verbs (not including *snater* or *fincur*) and asked to rate how positive or negative they thought the meaning of the verb was on a scale from 'Very negative' (0) to 'Very positive' (100). The mean score for the final sentences of each type in Experiment 1 was as follows: Positive 68.51, Negative 29.18, Neutral 48.34 (Appendix C; Figure 10). A GLM revealed that intended context reliably predicted mean positivity score: $F(2,93) = 585.58$, $\chi^2 = 23203$, $p < 0.001$. A full list of stimuli sentences can be found in Appendix A. Sentences using the noun were not assigned a meaning context but were designed to be relatively neutral (e.g., 'Can you hand me the _____?').

2.2.3. Procedure

Using a Qualtrics survey, participants were taught three novel words: two target verbs (*snater* and *fincur*) and one noun (*murp*). Before being exposed to the words in sentences, participants were informed that *snater* and *fincur* have the same definition and that, at the end of the experiment, they would be asked what they thought this definition was. Participants were further informed that they would see all three new words used in a variety of sentences (*Exposure phase*) before being asked to insert the words into new sentences (*Generalization phase*).

After reading the instructions and before proceeding to the Exposure phase, participants completed an understanding check involving a series of multiple-choice questions to ensure that they recognized the words they would be exposed to, had grasped what parts of speech they were, and understood that the two verbs, *snater* and *fincur*, referred to the same thing (Appendix B). Participants were unable to move on until each question had been answered correctly.

In the Exposure phase, participants viewed each word in twelve unique sentence frames that varied by condition. Sentences were presented one by one and stayed on screen for 10 s before a button appeared allowing participants to move to the next sentence.

There were six conditions, which are laid out in Table 1. In the *Neutral* condition, both *snater* and *fincur* were presented exclusively in neutral contexts (i.e., the sentences did not imply positive or negative connotations for either word). In the *Consistent* condition, one word was used 100% of the time in positive contexts, while the second was used exclusively in negative contexts. In the Random Condition, both verbs were used equally often in positive and negative contexts. In the *75%-Positive* condition, both verbs were presented 75% of the time in positive contexts, and 25% of the time in negative contexts. Similarly, in the *75%-Negative* condition, both verbs were used 75% of the time in negative contexts and 25% of the time in positive contexts. Finally, in the *Overlapping* condition, one verb was shown 75% of the time in positive contexts and 25% of the time in negative contexts, while the reverse was

Table 1. Input distributions across contexts by condition in Experiment 1

	Word 1			Word 2		
Condition	Positive	Negative	Neutral	Positive	Negative	Neutral
Consistent	100%	0%	0%	0%	100%	0%
Random	50%	50%	0%	50%	50%	0%
Neutral	0%	0%	100%	0%	0%	100%
75%-Negative	25%	75%	0%	25%	75%	0%
75%-Positive	75%	25%	0%	75%	25%	0%
Overlapping	75%	25%	0%	25%	75%	0%

true for the other verb. These conditions were designed to allow us to identify if there was a cline in responses such that output distributions followed quantitative patterns in the input distributions.

After exposure, participants proceeded to the Generalization phase. They were shown a series of 36 new sentences (not seen in the Exposure phase), each of which contained a blank. Their task was to insert one of the words into each of the sentences. In all conditions, twelve of the sentences implied positive contexts (e.g., 'I would love to see you _____ sometime!'), twelve implied negative contexts (e.g., 'I can't believe that he _____ in public!'), and twelve required a noun (e.g., 'She lost her _____ yesterday'.). Participants responded via a forced-response, multiple-choice question. All blanks were the same length to ensure this did not influence participants in their choices. Additionally, all words were appropriately inflected for the sentence. For example, given the sentence 'My first time _____ was fantastic', participants were able to select from 'fincuring', 'snatering', or 'murp'.

Including the noun as an option for all Generalization sentences served as an attention check and helped to obscure the real goal of the task. Consistent application of nouns to contexts specifically intended for verbs (or vice versa) might reflect a failure to understand the instructions or a lack of attention. Furthermore, it encouraged participants to think the task might be about parts of speech, reducing demand characteristics.

Measuring differentiation. Synonymy and differentiation can be thought of as a relationship between words and semantic contexts; in particular, they reflect the distribution of one relative to another. In the case of absolute synonyms, two (or more) words are distributed equally across meaning contexts; thought of from another perspective, meaning contexts are distributed equally across the words. In either case, the point is that in all the contexts where one word can occur, the other could occur just as easily; thus, there is no differentiation. At the other end of the scale would be pairs of non-synonymous words that could never replace each other in any of the same contexts. Between these two extremes are non-absolute synonyms, such as *fantastic: awesome,* which are distributed across contexts such that they overlap partly (e.g., in reference to winning the lottery) but not fully (e.g., in reference to hobbits and atomic explosions). Alternatively, we might say that the semantic contexts are unequally distributed across the words.

To measure differentiation, we therefore used *Gini scores* (also known as Gini coefficients). The Gini score is a measure of inequality, often used in economic contexts, in which a Gini score of 0 represents perfect equality and a Gini score of

1 represents perfect inequality (Abounoori & McCloughan, 2003). They are well suited to the question we are asking because they are measures of equality of distribution. In fact, they have been used previously for a similar purpose in linguistic research (see Roberts & Galantucci, 2016). It follows from the above discussion that synonymy and differentiation can be measured in terms of the equality of distribution of context over words. (Alternatively, we could measure this in terms of the distribution of each word over the two contexts. For our purposes, the results are essentially equivalent.) If the two words are equally likely for a given context, then that would result in a Gini score of 0 for that context. Although the highest possible value for a Gini score is 1, the maximum varies according to the number of words available. For our data, in which there were two words the highest possible value was 0.5, indicating that one word occurred in that context but the other did not. This is based on calculating a Gini score without adjusting for population size, as the 'population' (two words) did not vary between conditions so did not need to be normalized. We used the mean score for the two contexts as our measure of differentiation; a mean score of 0.5 would indicate full differentiation. Perfect partial differentiation, in which (e.g.) one word was used 100% of the time in positive contexts, while negative contexts were split equally between both words, would yield a score of 0.33.

2.3. Results

2.3.1. Main patterns of differentiation

Data are available at https://osf.io/qt8xk/. Analyses were performed using R (R Core Team, 2013), with Gini scores calculated using the dineq library (Schulenberg & Schulenberg, 2018). Graphs were created using ggplot2 (Wickham, 2011). Figure 1 shows Gini scores for each condition. Participants who inserted the noun as opposed to a verb in the simple majority of sentences would have been excluded, though none did so. The general pattern is of a high level of differentiation in the Consistent condition and scores more consistent with partial differentiation in all the other conditions. A GLM revealed that condition was a significant predictor of Gini scores,

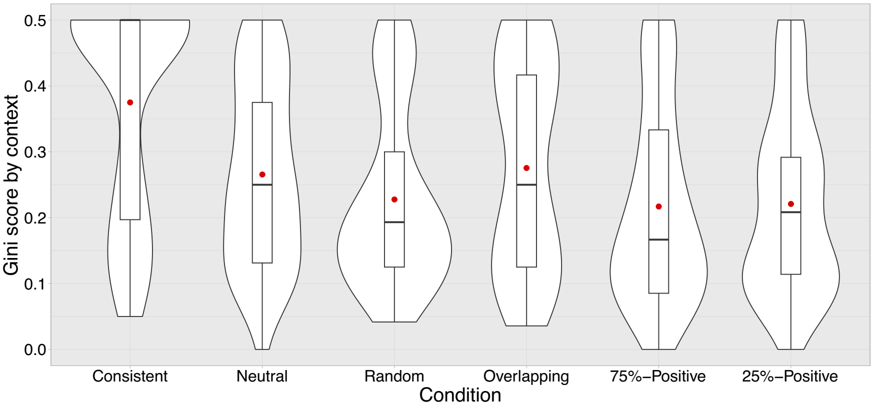

Figure 1. Violin plots of mean participant Gini scores by condition in Experiment 1 overlaid with box and whisker plots. Red dots indicate mean values.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

F(5,356) = 10.03, χ^2 = 1.08, $p < 0.001$. This effect was primarily driven by the Consistent condition (*Est.* = 0.15, *SE* = 0.03, $t = 5.74$, $p < 0.001$), which post-hoc Tukey contrasts revealed differed from all other conditions ($p < 0.001$ for all comparisons except with the Overlapping condition, for which $p = 0.0027$). There were no other significant differences between conditions. All conditions did, however, differ significantly from chance, represented by a mean Gini score of 0 (all $p < 0.001$). In other words, participants do not seem to have been simply inserting verbs at random in the Generalization phase; nor did they tend to treat the verbs as fully differentiated, except in the Consistent condition. The results of the other conditions – for which the mean Gini score was 0.26 – were in line with a pattern of more partial differentiation. More than one pattern could produce this Gini score. One possibility is for both verbs to occur in both contexts, but for each context to be dominated – at about 70% – by a different one of the two verbs, representing not *full*, but still rather high, differentiation. The other possibility is for one 'unmarked' verb to occur in both contexts (and to dominate in one), while the other, 'marked', verb occurs in only one context. Where this pattern of partial differentiation occurred, the marked verb could be positively or negatively marked. Across conditions, the marked verb was equally likely to be positive or negative: $\chi^2(1, N = 263) = 3.19$, $p = 0.07$.

These *overall* mean Gini scores only give part of the story, however. As can be seen clearly in Figure 1, there was quite a lot of variability in our data. While high levels of differentiation were typical of the Consistent condition, the Neutral and Overlapping conditions exhibited a somewhat bimodal pattern, while in the other conditions Gini scores were much more concentrated toward the bottom of the range; the variability was sufficient, however, that mean Gini scores were *overall* consistent with differentiation. Another way of putting this is that patterns of partial differentiation arise not only due to individual language users differentiating in that particular way, but also due to patterns of variability across users exposed to the same input distributions. In the following section, we discuss these patterns.

2.3.2. Further analysis: Statistical-learning patterns

As discussed above, differentiation in output distributions arises through particular responses by participants to input distributions. But what were the statistical-learning patterns involved here? To find this out, we examined each participant's systematicity with regard to their input. Following studies examining the regularization of unpredictable variation, such as that of Hudson Kam and Newport (2009), participants were classified into three categories: *matchers*, *under-matchers*, and *over-matchers*. Matchers matched the probability of the Exposure phase input in their output during the Generalization phase. Under-matchers fell short of the input frequencies, choosing the words less systematically than the input. Over-matchers, in contrast, went beyond their exposure, increasing the systematicity of their input. Participants' groupings were determined using a randomization test with 10,000 replications, comparing participants' Gini scores to the expected one for the given condition.

Figure 2 displays the number of matchers, and under- and over-matchers with their Gini scores by Condition. There was one *systematizer*, who only used a single verb. Participants in the Neutral condition were excluded, as they were given no information in the input regarding positivity or negativity. As can be seen, the distribution of participant types was somewhat similar in the Random, the 75%-Positive, and the 75%-Negative conditions: More than half of participants were

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

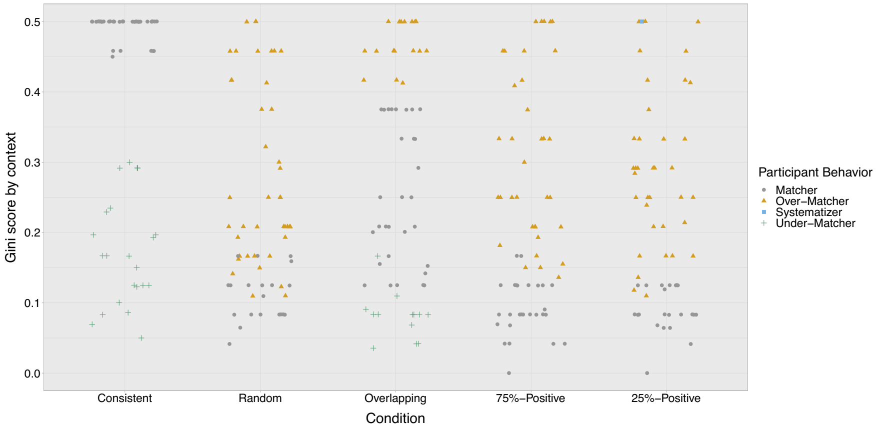

Figure 2. Mean Gini scores for Experiment 1 organized by condition (not including neutral) and colored according to the participant's statistical learning category.

Over-matchers. In the Overlapping condition by contrast, roughly half of participants were Matchers, and a similar number of participants were classified as under-matchers or over-matchers. In the Consistent condition, 63% of participants matched their input, while the rest under-matched. (Overmatching in this condition was of course impossible.) In other words, there was variability in individual participants' responses that was to some extent related to input distributions, but not entirely, with the consequence that (as discussed above) *overall* patterns across conditions were more driven by qualitative than by quantitative features of the input.

We also examined the distribution of *snater* and *fincur* to determine whether the particular verb played any role in motivating participants' decisions. *Snater* and *fincur* were used at about the same rate overall, $Z = 1.64$, $p = 0.1$. Positive uses of *snater* were not more frequent than negative ones: $Z = 0.55$, $p = 0.58$. Likewise, positive uses of *fincur* were not more frequent than negative ones: $Z = -0.57$, $p = 0.57$. In both negative and positive contexts, *snater* and *fincur* were chosen equally often ($Z = 0.682$, $p = 0.49$, and $Z = 1.63$, $p = 0.10$, respectively).

2.4. Discussion of Experiment 1

Experiment 1 investigated the differentiation of expressive near-synonym pairs (like *confident : arrogant*) by exposing adults to two novel verbs with the same definition, *snater* and *fincur*, presented in positive and negative contexts, and then asking them to extend the verbs to new positive and negative sentence frames. We manipulated the distribution of the two verbs across the two contexts in exposure. The overall pattern throughout conditions was of differentiation. That is, the verbs were distinguished by context either fully, with each of the two verbs dominating one context (compare *thrifty : stingy*), or partially, with one verb behaving as an 'unmarked' default and the other being more 'marked', that is, concentrated in one context (compare *young : childish*). Only when input perfectly differentiated the two verbs (in the Consistent condition) was full differentiation in output straightforwardly the most common pattern.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Somewhat unexpectedly, we did not find evidence that participants overall tended to adhere closely to input distributions in differentiating the novel words. To investigate this more closely, we categorized participants based on the extent to which they were matching input distributions and found that, while there were probability matchers in all conditions, probability matching was not the overall pattern for any condition. In fact, input *consistency* again seems to have mattered more than the details of the distribution, which supports a qualitative account of synonymy based on a cognitive bias against synonymy (such as the mutual exclusivity bias) against a quantitative account based on input distributions. That being said, quantity is also relevant to consistency. The Consistent condition involved 100% consistent input. Our results do not allow us to determine if that level of consistency is in fact necessary for the kind of full differentiation observed in this condition. There may in fact be a lower threshold that did not occur in our study. In the experiment, all input distributions outside the Consistent condition involved the same verbs appearing in both positive and negative contexts. Our results suggest that, given the qualitative fact that both verbs are presented as occurring in both contexts, varying the quantitative details of how *often* they occur in each of those two contexts may not make much difference. Our results do not tell us, however, what would happen if each verb occurred most of the time in one context and the rest of the time in *neutral* contexts. This would present a qualitative distinction between the two verbs as in the Consistent condition while also including quantitative variation in how often they occur in the contexts they appear in. Experiment 2 was designed to investigate this.

It should be added that, while the threshold for what we have called *full* differentiation seems to have been somewhat high, the threshold for differentiation of any kind seems to be rather low. In fact it occurred rather often, resulting in mean Gini scores between 0.2 and 0.3 – consistent with differentiation – in all conditions, regardless of the distributions in the input data, and even when the input data provided no information about potential positive or negative meaning. In other words, participants were apparently quite resistant to treating words as exactly synonymous.

It is potentially a little surprising that participants did not generally probability-match, particularly given that this phenomenon has been repeatedly observed in language learning experiments (Hudson Kam & Newport, 2005; Vouloumanos, 2008). It is, however, clear from other work that adults do not *always* match probabilities. For instance, they have been found to over-match when presented with sufficiently complex input. For example, adults increase the frequency of more frequent options when there are three or more alternatives (Gardner, 1957; Weir, 1972). Similarly, in studies of unpredictable variation, adults will regularize – introducing new consistency to the language – when alternative forms are numerous and frequent, in addition to being inconsistent (Hudson Kam & Newport, 2009). Regularization also increases when the task becomes more challenging, as in a study by Wonnacott and Newport (2005) where learners were tested on words that were different from the training stimuli. With only two forms and two contexts, it seems unlikely that our results were due to complexity, although it could be that the various different sentences introduced subtle alternative connotations beyond positive and negative valence, complicating the task. Other work has shown that regularization is encouraged by increased communicative demands. Both Smith and Wonnacott (2010) and Fehér et al. (2016), for instance, found that interaction and contact between learners led to increased systematicity in communication systems. In the

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

present study, the Generalization phase might be seen as a proxy for communicative interaction. Nevertheless, the Generalization sentences were presented in isolation and were not situated within a larger dialogue or explicitly intended for another speaker, so it seems unlikely that they can be interpreted as having presented increased communicative demands.

A better way of thinking about our results may be to consider the relationship between individual behavior and overall patterns in the data. There were undermatchers even in the Consistent condition. And quite a few participants *did* probability-match. Indeed, the extent of variation between participants was itself notable. Differentiation was a feature of some but not all individual participants and at the same time a feature of the overall distributions. An important consequence of this is that it is not necessary for everyone to differentiate synonyms for differentiation to become the overall pattern to which new learners are exposed, which should be expected to lead in not many generations to high levels of differentiation. This could be explored in future work by employing an iterated learning paradigm (Kirby et al., 2014).

Another notable point potentially at odds with existing literature concerns the 'direction' of differentiation. When one of the two verbs in our experiment became specialized or 'marked' (i.e., more concentrated in, or specialized to, one context) in the output, it was equally likely to be positively or negatively marked. This finding conflicts with a body of existing literature on semantic markedness, in which the marked term is more likely to be negative (e.g., Lehrer, 1985). However, this might have been an artifact of how the Generalization phase in our experiment worked. In natural language, unmarked forms are the neutral default forms (Greenberg, 1966; Waugh, 1982; Zwicky, 1978). In our experiment, the Generalization phase included only two contexts, neither of which was neutral. An obvious solution to this issue would be to introduce neutral sentences into the Generalization phase. This was done in Experiment 2, reported below.

Experiment 2 also allowed us to address a related question. This concerns whether the distinction between the two words is more *semantic* or more *pragmatic* in nature. In other words, is the markedness treated as an inherent part of the word's meaning or not? The synonyms *famous : infamous* are an example of more semantic markedness, where 'infamous' necessarily implies some negative valence. The same is not so true of the pair *aggressive : pushy*; while *aggressive* might often imply something more strongly negative than *pushy,* the positive use of 'aggressively' in 'He aggressively advocates for his clients!' seems entirely felicitous. The negative valence, in other words, is not an inherent part of the meaning of *aggressive* but is contingent on context. Experiment 1 did not allow us to make this distinction in our data: Just because a participant chooses one verb over the other for a particular sentence does not tell us if they could in principle have used the other. Experiment 2 remedied this with the introduction of a modified acceptability judgment task, here referred to as the Rating task (cf. Fairchild & Papafragou, 2018; Takimoto, 2009).

3. Experiment 2

3.1. Overview

Experiment 2 replicated Experiment 1 with three changes. First, neutral sentences were included in the Generalization phase. Second, a task was added to the end of the

Generalization phase in which participants were shown all the sentences from the fill-in-the-blank task they had just completed, once with each of the two verbs, and asked to rate their appropriateness. (In what follows we will refer to the fill-in-the-blank task as the *Generalization task* and the sentence rating task as the *Rating task.*) Third, a new condition was introduced: the *Nudge* condition, in which one word was seen 75% of the time in positive contexts, the second was seen 75% of the time in negative contexts, and all remaining uses of both terms were neutral.

The Nudge condition was designed to allow quantitative variation in how often each verb appeared in positive or negative contexts while retaining the qualitative distinction of neither verb appearing in both these contexts. That is, it introduced a new kind of consistency in exposure. In Experiment 1, consistency meant appearing only in positive or only in negative sentences. In the Nudge condition, it was still the case that neither verb ever appeared in both of these contexts; but they no longer occurred so reliably in *only* positive or negative contexts – participants were, in other words, somewhat more *nudged* toward output consistency than given it on a plate. This allowed us an insight into what the threshold might be for full differentiation.

The neutral sentences in the *Generalization phase* were introduced to allow us to better investigate the role of markedness. In Experiment 1, patterns of partial differentiation occurred in some cases in which one ('unmarked') form occurred in both positive and negative contexts, while the other ('marked') form was typically used in only one context. With three contexts, it was possible to have a marked and an unmarked form without the unmarked form dominating all contexts. The addition of the Rating task was intended to shed light on the type of differentiation participants were creating: *semantic* or *pragmatic.* If participants were creating the former, words that were highly differentiated in the output should receive higher rating scores. For example, a verb that takes on a mostly positive meaning in the Generalization task might be consequently rated as most appropriate in positive contexts, and not at all appropriate in negative one (e.g., 'I love people who *fincur*!' might be rated as 'very appropriate' and 'Gross! She's *fincuring*!' as 'very inappropriate'). In contrast, if participants had developed a more flexible, pragmatic distinction, highly positive words might be considered 'okay' in negative contexts, and vice versa (e.g., rating 'I love people who *fincur*!' as 'very appropriate' and 'Gross! She's *fincuring*!' as 'neither appropriate nor inappropriate'). Participants could have preferences about how the words are used, but tolerate variation.

3.2. Method

3.2.1. Participants

241 participants (172 female, 65 male, 3 non-binary), aged between 18 and 86 (*median* = 33), were recruited from Prolific. All participants were English speakers and were compensated $3 for their time; 20 participants reported that they spoke more than one language.

3.2.2. Stimuli

The stimuli used in Experiment 2 were identical to those of Experiment 1, with the addition of 12 new Neutral sentences added to the Generalization phase. As for Experiment 1, we conducted a norming study to ensure that the sentences were conveying what we intended. The mean positivity score for each context type in

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Experiment 2 was as follows: Positive 68.51, Negative 29.19, Neutral 48.64 (Appendix C), and one-way ANOVA revealed that the difference in mean positivity score for each context was indeed significant: F(2,93) = 585.6, $p < 0.001$. A full list of stimuli sentences can be found in Appendix A.

3.2.3. Procedure

The basic procedure of Experiment 2 was nearly identical to that of Experiment 1 except for the following three differences. First, neutral sentences were included in the Generalization task. Second, there was a Rating task after the Generalization task, in which participants were shown, one by one, each of the verb sentences from the Generalization task they had just completed. Each sentence occurred twice, once with each verb, for a total of (12 positive +12 negative +12 neutral) * 2 = 72 sentences. Sentences appeared in a random order. In each case, the participant was asked to 'rate how you feel about the appropriateness of the words' using a 5-point Likert scale, where 1 = 'very inappropriate', 2 = 'somewhat inappropriate', 3 = 'neither appropriate nor inappropriate', 4 = 'somewhat appropriate', and 5 = 'very appropriate'. Third, Experiment 2 included only three of the conditions from Experiment 1: Neutral, Consistent, and Random. Experiment 2 also included a *Nudge* Condition in which one word was used 75% of the time in positive contexts, while the other was used 75% of the time in negative contexts. The remaining 25% of contexts were neutral for both words. The conditions of Experiment 2 are laid out in Table 2.

Scoring. We used mean Gini scores as in Experiment 1 to identify patterns of differentiation. However, the meaning of a given Gini score and of terms like 'full differentiation' has to be interpreted slightly differently in Experiment 2 to take into account the presence of neutral sentences, which increased the range of possibilities. The highest level of differentiation involved one verb being used exclusively in one context, while the other verb was used exclusively in both the remaining contexts. This would yield a mean Gini score of 0.5. A weaker form of differentiation, in which each verb occurred in one and a half contexts (i.e., one verb occurring in all negative sentences and half of the neutral sentences with the other verb occurring in all positive sentences and the remaining neutral sentences), would yield a mean Gini score of 0.44. For the Rating task, mean rating scores were taken by word and by condition for each participant, and Gini scores were computed based on these. An overall Gini score was taken using the mean rating score for both words in all contexts (mean rating of *snater* in positive contexts, mean rating of *fincur* in positive contexts, etc.) to measure appropriateness across words and contexts. Again, higher Gini scores correspond to less spread: in this case, that means values clustered around either 'very inappropriate' (1) or 'very appropriate' (5).

Table 2. Distribution of words across contexts by condition in Experiment 2

	Word 1			Word 2		
Condition	Positive	Negative	Neutral	Positive	Negative	Neutral
Consistent	100%	0%	0%	0%	100%	0%
Random	50%	50%	0%	50%	50%	0%
Neutral	0%	0%	100%	0%	0%	100%
Nudge	75%	0%	25%	0%	75%	25%

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Table 3. Participant category in Experiment 2 based on relationship between behavior in the generalization task and behavior in the rating task

		Generalization task	
		High differentiation	Low differentiation
Rating task	**High differentiation**	Semantic distinction	Inconsistent distinction
	Low differentiation	Pragmatic distinction	Synonymy

We also compared Gini scores between tasks. Based on scores from both sections of the experiment, participants were classified into four different behavior groups (Table 3). People with a high mean Gini score for both tasks were considered to have developed a *semantic* distinction. These participants differentiated the words strongly in the Generalization task and showed less pragmatic tolerance during the Rating task, with valenced words rated as appropriate only in their coordinated context (e.g., the positive word was only rated as 'very appropriate' in positive contexts and as 'very inappropriate' in negative ones). Participants who made a *pragmatic distinction*, in contrast, differentiated the verbs highly in the Generalization task, but proved more flexible in the Rating task (e.g., the positive word was rated as 'very appropriate' in positive contexts and 'neither inappropriate nor appropriate' in negative ones) as evidenced by a lower overall Gini score for the Rating task. Participants with low Gini scores for both tasks (for whom the two verbs were thus only weakly differentiated in use and appropriateness) were categorized as treating the terms as 'synonymous' (though this term is not intended to imply *absolute* synonymy, which would correspond to a Gini score of 0, which was rare). Finally, those with a high Gini score for the Rating task and a low one for the Generalization task were deemed *inconsistent*; their responses to the Rating task implied a distinction not borne out in Generalization. Scores were considered high if they were above the midpoint value for a given index and low if they fell below it.

3.3. Results

3.3.1. Generalization task: main patterns of differentiation

Figure 3 shows participant Gini scores for each condition. Participants who used the noun in the simple majority of verb contexts would have been excluded, though none did so. A GLM revealed an effect of condition on mean Gini score, $F(3,237) = 16.17$, $\chi^2 = 0.71$, $p < 0.001$. The results of post-hoc Tukey comparisons between the different conditions are provided in Table 4. The general pattern was that the Consistent and Nudge conditions did not differ significantly from each other but showed higher levels of differentiation than in the Neutral and Random conditions (which also did not differ from each other). This result mirrors that of Experiment 1, in which participants differentiated the verbs in all conditions but did so most strongly in the Consistent condition. In Experiment 2, the Nudge condition behaved essentially like the Consistent condition, except that (as can be seen in Figure 3) its results were somewhat more bimodally distributed.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

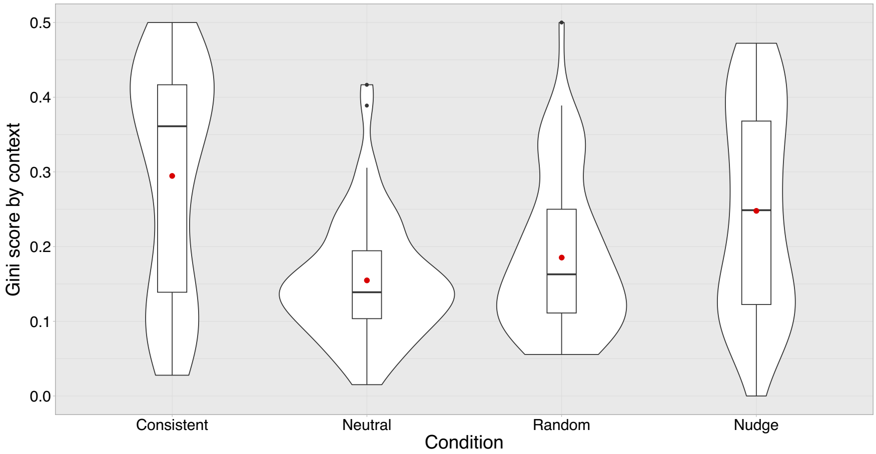

Figure 3. Violin plots of Experiment 2 generalization task results by condition, overlayed with bar and whisker plots. The red dots indicate mean values.

Table 4. Mean, standard deviation, and p-values for post-hoc comparisons of Gini scores in Experiment 2

Condition	M	SD	1 (Consistent)	2 (Nudge)	3 (Neutral)
1. Consistent	0.32	0.17			
2. Nudge	0.27	0.16	0.146		
3. Neutral	0.16	0.09	< 0.001	< 0.001	
4. Random	0.19	0.12	< 0.001	0.024	0.52

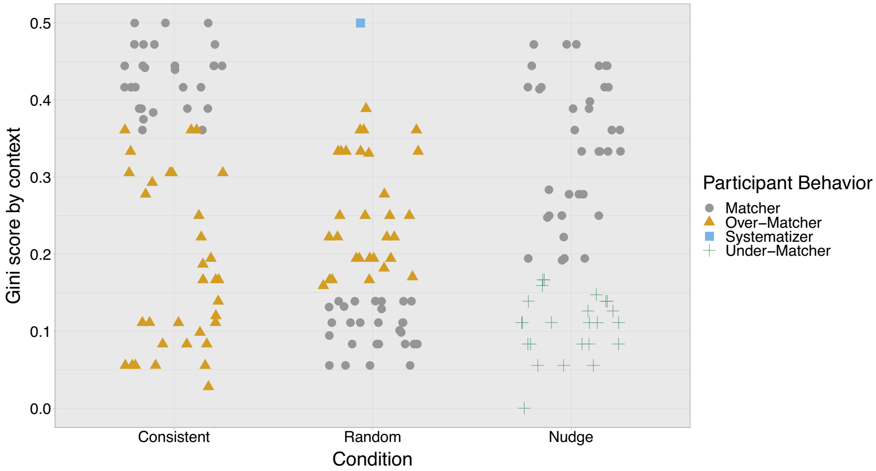

Figure 4. Mean Gini scores for Experiment 2 generalization task by condition (not including Neutral condition); each dot represents a participant, colored according to their statistical learning category.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

3.3.2. Generalization task: further analysis

As in Experiment 1, we classified participants were classified into three categories based on a comparison of their output distribution with the input distribution (*matchers*, *under-matchers*, and *over-matchers*). The Neutral condition is not included as there was no clearly meaningful way to compare output with input distributions. There was also one participant (in the Random condition) who used a single word across all test contexts, who was deemed a *systematizer*. As in Experiment 1, participants in the Neutral condition were removed. Figure 4 displays the distribution of matchers, and under- and over-matchers with their respective mean Gini scores by condition.

We also reviewed the distribution of *snater* and *fincur* themselves by word and by context. As in Experiment 1, *snater* and *fincur* were used equally, $Z = 1.33$, $p = 0.19$.

A Kruskal–Wallis test revealed that the uses of *fincur* did not differ by context type: H(2) = 1.35, $p = 0.51$. Likewise, uses of *snater* were not more frequent in any particular context, H(2) = 1.64, $p = 0.44$. Though *fincur* and *snater* were used at similar rates in positive, H(1) = 0.29, $p = 0.59$, and negative contexts, H(1) = 0.05, $p = 0.83$, *snater* was preferred over *fincur* in neutral ones, H(1) = 11.135, $p < 0.001$.

3.3.3. Rating task: semantic versus pragmatic differentiation

A GLM revealed an effect of condition on rating scores, F(3,237) = 8.76, $\chi^2 = 0.06$, $p < 0.001$. Post-hoc Tukey contrasts suggested that this was driven by differences between the Consistent condition and the neutral and random conditions in particular (*Est.* = 0.04, *SE* = 0.009, $p < 0.001$ in both cases). There was also a small difference between the Nudge and Neutral conditions (*Est.* = 0.02, *SE* = 0.009, $p = 0.038$). As can be seen in Figure 5, this difference is primarily about distribution. Scores for the Consistent condition (and to a lesser extent the Nudge condition) varied more than scores for the other conditions, where most scores were closer to the middle of the range. In this respect, the pattern of results for the Rating task was

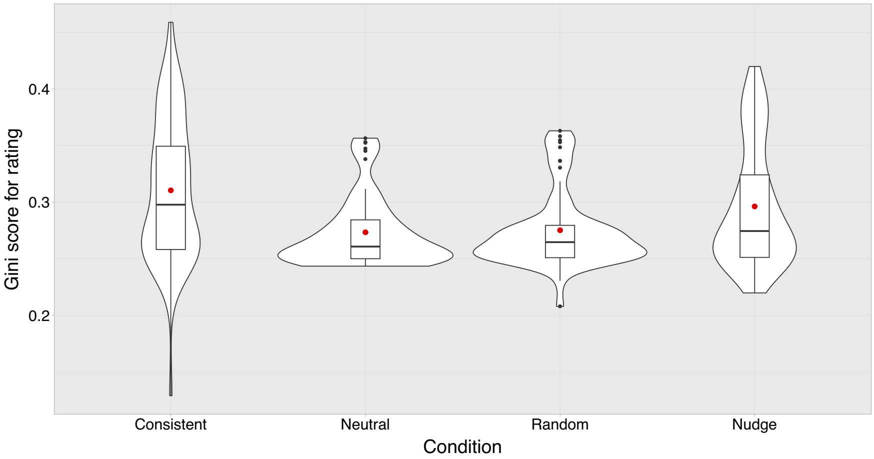

Figure 5. Violin plots of Experiment 2 rating task results by condition, overlayed with bar and whisker plots. The red dots indicate mean values.

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

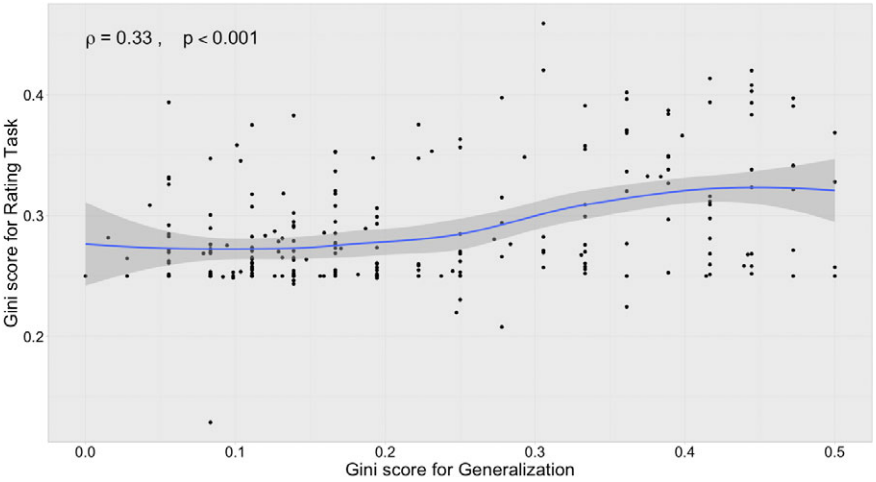

Figure 6. Relationship between Gini score in generalization task and rating task in Experiment 2.

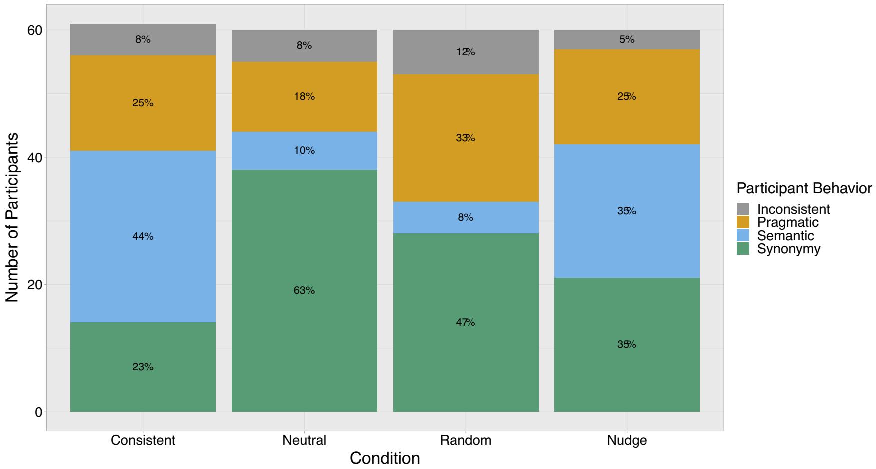

Figure 7. Participant behavior categories by condition.

similar to the pattern for the Generalization task. Indeed, scores for the two tasks were positively correlated overall, $r_s(239) = 0.33$, $p < 0.001$ (Figure 6).

Based on their behavior in the generalization and rating tasks, participants were categorized as Pragmatic, Semantic, Inconsistent, or Synonymous differentiators. as outlined in Table 3. The resulting participant counts per category are displayed in Figure 7. A chi-squared test for independence found a significant difference in behavior group counts by condition, $\chi^2(9, N = 241) = 41.1$, $p < 0.001$. Post-hoc comparisons of participant behaviors across all conditions using the Bonferroni correction confirmed that the number of Inconsistent and Pragmatic responders

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Table 5. P-values for Bonferroni post-hoc comparisons by condition for participant behaviors

		Participant behaviors		
Condition	Inconsistent	Pragmatic	Synonymous	Semantic
Consistent	1.0	1.0	0.008	< .001
Nudge	1.0	1.0	1.0	.46
Neutral	1.0	1.0	0.002	< 0.04
Random	1.0	1.0	1.0	< 0.012

did not vary by condition (Table 5). However, there were more Semantic responders in the Consistent condition and fewer in the Neutral and Random conditions. Additionally, Synonymous differentiators were significantly less common in the Consistent condition and more common in the Neutral condition.

3.3.4. Specialized verb and rating task: further analysis

We next analyzed participant preferences for positivity or negativity when specializing the marked verb. We excluded 31 participants who differentiated the words equally, receiving identical Gini scores for both *snater* and *fincur*. Likewise, 25 participants who more strongly differentiated one form, as evidenced by a higher Gini score, but applied this specialized form equally to positive and negative contexts were not included. For the remaining 185 participants, the marked form was identified as the verb with the higher Gini score in the Generalization task (reflecting increased specialization) and its valence was coded as either positive or negative. A chi-squared goodness-of-fit test revealed that the marked verb was more likely to be used in negatively marked sentences: $\chi^2(1, N = 185) = 6.62$, $p = 0.01$. This pattern held across conditions, $\chi^2(3, N = 185) = 2.19$, $p = 0.53$, suggesting that different information in the Exposure phase did not influence the direction of specialization.

The Gini score for the marked form in the Generalization task was positively correlated with the score for the Rating task, whether that form was *snater* or *fincur*, $r_s(183) = 0.420$, $p < 0.001$ (Figure 8).

Figure 9 shows the distribution of Gini scores for positively and negatively marked forms in the Generalization and Rating Phases. A GLM found a significant effect of valence (i.e., whether the marked form was positive or negative) on the Gini score in the Generalization phase for the relevant context, F(1,183) = 13.09, $\chi^2 = 0.4^{=}$ $p < 0.001$, with positively marked forms receiving lower Gini scores than negatively marked ones. In other words, negatively marked forms were more strongly differentiated in the participants' output than positively marked ones. The same was true for the Rating task: F(1,183) =5.69, $\chi^2 = 0.04$, $p = 0.02$).

3.4. Discussion of Experiment 2

Experiment 2 expanded on Experiment 1 with the incorporation of neutral sentences during exposure, the Nudge condition (which provided a weaker contextual consistency than the Consistent condition), and a Rating task designed to help distinguish semantic differentiation from pragmatic differentiation.

In Experiment 1, we saw two kinds of differentiation: *partial* differentiation – in which one term became specialized as either negative or positive – and *full*

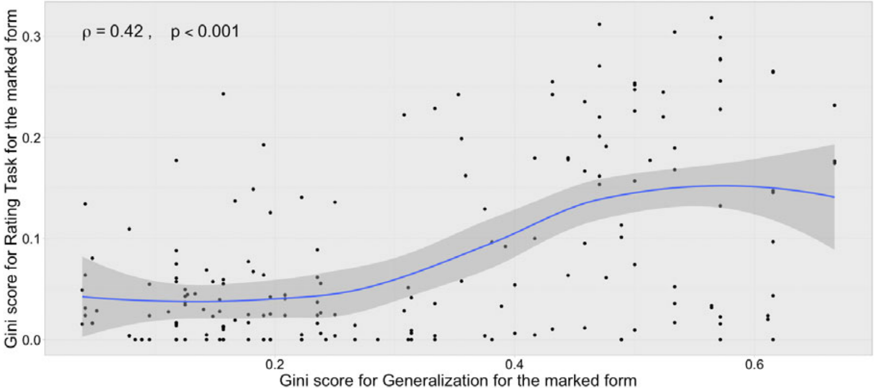

Figure 8. Relationship between Gini score in generalization task and rating task for the marked form only.

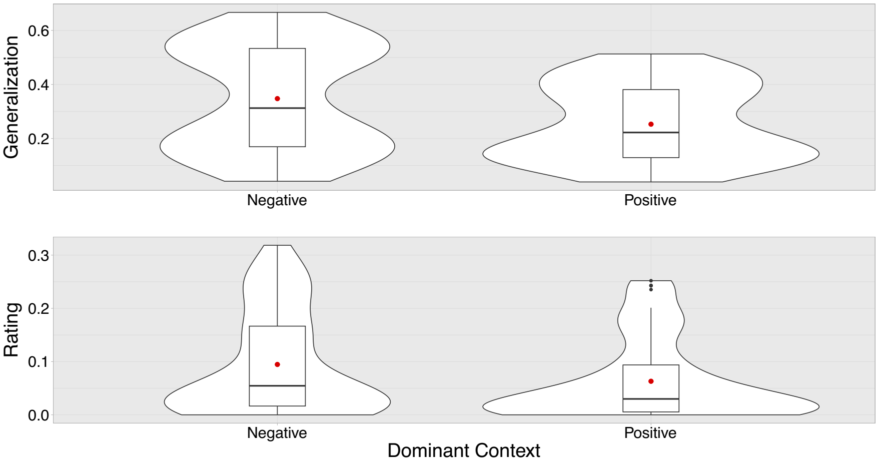

Figure 9. Gini scores for marked forms by context during generalization and rating tasks in Experiment 2. Red dots indicate mean values.

differentiation, in which both terms became specialized. The former occurred in all conditions, while the latter was dominant in the Consistent condition. We found no evidence that responses strongly tracked input probabilities. This result is consistent with an account in which input quality is more important than quantity. But how consistent does input have to be? What if verbs sometimes occur in neutral contexts in input? The Nudge condition in Experiment 2 was designed to investigate this. As in Experiment 1, full differentiation occurred in the Consistent condition. It also occurred in the Nudge condition, though the pattern of results here was more bimodal (Figure 3). As in Experiment 1, the pattern for the Random and Neutral conditions was for lower differentiation, but the verbs still tended to be differentiated in all conditions. The mean Gini score in the Random condition, for instance, was 0.19, close to the value in Experiment 1. This is consistent with a pattern in which both verbs might appear in all three contexts, but each dominates in one of them. In

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

other words, there was again a tendency for participants to introduce greater differentiation than was in the input.

Consistent with the findings of Experiment 1, participants in Experiment 2 did not more strongly differentiate the verbs based on the size of the bias in the input. Participants differentiated just as much in the Random condition, where words were presented equally frequently in both positive and negative sentences, and the Neutral condition, where there was no information about positivity or negativity at all.

Most interestingly, the Consistent condition did not lead to significantly greater differentiation than the Nudge condition (although the distribution was more bimodal in the Nudge condition). The results of the Nudge condition also contrast with the Overlapping condition in Experiment 1, in which participants behaved essentially as in the Random and Neutral conditions. This is interesting, because the Exposure distributions were very similar in Experiment 1's Overlapping condition and Experiment 2's Nudge condition – each word was presented 75% percent of the time in either a negative or positive context. The difference is that, in the Nudge Condition, the remaining 25% of occurrences were neutral. These results strongly suggest, consistently with those from Experiment 1, that *quality* of exposure is the most important factor influencing differentiation. If a word is presented only in a single positively or negatively valenced context, it becomes easier for that word to adopt a specific meaning, even if it does not *only* occur in that context. What matters more is how often it occurs in a directly contradictory context. This is not to say that distribution frequency might not contribute to the differentiation of near-synonyms. The results of the Nudge condition were not quite *identical* to those of the Consistent condition, and it is possible that the 75% rate in the Nudge condition was too high, resulting in a semantic shove instead of a semantic nudge. This raises the question of what the minimum value for differentiation might be: In other words, how much input is enough to optimally distinguish the words? For example, if the frequencies of the Nudge conditions were flipped, and both words appeared in neutral sentences 75% of the time, would participants behave the same? Future experiments can explore any number of probabilistic manipulations, keeping in mind that differentiation seems to be most strongly dependent on the quality of the input signal – whether or not a word appears in both positive and negative contexts.

The fact that Rating task scores were positively correlated with Generalization task scores suggests that participants were consistent in their differentiation behavior across tasks. Most participants rather weakly differentiated the words in both tasks, a behavior that we have labeled 'synonymy' (though this is a relative term, which should not be interpreted as implying absolute synonymy). This is perhaps unsurprising given the varied nature of the input across conditions. Regardless, it remains notable that participants still differentiated the verbs in all conditions, even when there was no bias at all in the input data.

Synonymous responders were, as one might have expected, far less frequent in the Consistent condition than in other conditions. The number of 'semantic' differentiators was also significantly higher in this condition; in other words, participants felt more strongly than in other conditions that each verb had 'right' and 'wrong' contexts. 'Pragmatic' differentiators, who occurred at somewhat similar rates in all conditions perceived things differently. Even if one verb carried a strong negative or strong positive connotation, it was still at least in principle possible in other contexts. Interestingly, Pragmatic differentiators were less common than Semantic ones in general. This is not an obvious result. It is not obvious, that is, that preferences in the

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

forced-choice Generalization task should correspond to strong ratings in the Likert-scale task. It is possible that this is partly a result of having the Rating task after the Generalization task; that is, the Generalization itself might have consolidated participants' attitudes. We chose this order to avoid influence of the reverse kind, but in future work this should be explored. The result may also reflect a broader tendency for semantic distinctions (in the sense we use the term here) to be easier to acquire, or to be more attractive to language users, than pragmatic ones. Storing a usage distribution in terms of a straightforward distinction is simpler than storing a more nuanced rule that a word can be used in any context but that it also tends to be preferred in one context. A similar phenomenon can be observed in the emergence of inaccurate prescriptive rules (such as 'use *between* with two items and *among* with more than two') that do not reflect real usage well but are simpler to recall and to apply than rules that do.

Perceptions of community consensus may also have a role to play in the development of semantic/pragmatic distinctions in natural language. The discrepancy in suitability between the use of a negative word in a positive context (4a) and a positive word in a negative one (4b) is a by-product of community held values. Contrast these with (5) 'I-statements' reflecting the thoughts of individuals (6a–b).

(4) a. Sexism is awesome!
b. World peace sucks!

(5) a. Cilantro is awesome!
b. Pineapple on pizza sucks!

(6) a. I think sexism is awesome!
b. I think world peace sucks!

In cases where our exposure sentences were framed as general statements, as in (4) and (5), they might have been interpreted as reflecting shared beliefs, which could in turn encourage more semantic differentiation; if everyone thinks a term is negative, it should only be acceptable in negative contexts. Sentences framed as personal opinions do not have the same implication. In other words, different kinds of Exposure sentences might have influenced participants differently with regard to semantic or pragmatic differentiation. The current study is not capable of distinguishing these possibilities, but a replication might usefully manipulate the kind of sentence used in a systematic manner.

Finally, we note that – contrary to the results of Experiment 1 – the specialized word in Experiment 2 was more likely to be used in negative contexts; negatively marked forms were also more strongly differentiated during Generalization and received higher Rating scores than their positively marked counterparts. This is consistent with the literature on markedness in natural language (Lehrer, 1985; Sassoon, 2012; Waugh, 1982) and would be interesting to investigate further as a potential driver of differentiation, although (given that we did not find this effect in Experiment 1) it seems unlikely to play a very strong role.

4. General discussion

Across two experiments, we investigated the differentiation of expressive near-synonym pairs, in which at least one term has a positive or negative meaning (e.g., *group* : *clique*). In Experiment 1, participants were exposed to two potentially synonymous novel verbs, *snater* and *fincur*, through sentences where information about positivity and negativity was consistent (Consistent condition), random (Random condition), inconsistent (75%-Positive and 75%-Negative conditions), or absent (Neutral condition). Afterwards, participants were asked to generalize by inserting the verbs into sentences. Participants differentiated the verbs in all conditions. However, full differentiation – with one word acquiring positive connotations and the other acquiring negative connotations (e.g., *satiated* : *crammed*) – was the dominant pattern only when exposure to the terms was consistent. In all other cases, there was a wider range of behavior, with many participants introducing a different kind of distinction between the verbs, with one a positive or negative meaning, while the other was used interchangeably in both contexts. The overall pattern across conditions was for this level of differentiation.

Experiment 2 replicated and built on Experiment 1 with the incorporation of a sentence judgment task, neutral sentences during generalization, and a new condition (Nudge) in which the two verbs overlapped only in neutral sentences. The pattern was essentially as in Experiment 1, with the results of the Nudge condition being similar to those of the Consistent condition.

The inclusion of a rating task in Experiment 2 allowed us to shed further light on the nature of participants' differentiation of the verbs. In the Neutral and Random conditions, participants tended to fall in the 'synonymous' category, meaning that the differentiation was relatively weak and not interpreted as implying that either verb was *unacceptable* in any context. In the Consistent and Nudge conditions, there were fewer such responders, and more responders who treated the verbs as either semantically or pragmatically differentiated.

A limitation of our study concerns the simplicity of the contexts investigated. This study focused – for the sake of simplicity – on expressive synonymy alone. However, dividing all speech situations into simply positive, negative, or neutral sentence contexts is somewhat coarse-grained. There are, as discussed, considerably more dimensions to synonymy. Real-life interactions too are of course much more subtle and complicated than our division might imply, involving complex variation depending on such factors as speaker, prosody, social context, dialect, and so forth We nonetheless consider this a good place to start, opening space for future work to explore a wider and more complex range of dimensions and sentence contexts. It is also worth noting that our study, in leaving the exact semantic content of the verbs an open question, did not touch on the interaction between denotational meaning and the emergence of expressive connotational meaning, and their role in language change. This can include the emergence of opposing expressive meanings in the same lexeme. (See, for instance, Gergel & Kopf-Giammanco, 2021, for a related discussion of change in Austrian German.)

Further work should also include broadening the social contexts involved in language change. Experiments in which participants used the verbs in real communicative interaction with other participants would allow us to better investigate a number of communicative and social factors (Sneller & Roberts, 2018; Stevens & Roberts, 2019; Wade & Roberts, 2020), while an iterated learning paradigm would allow us to explore the role of social transmission (Kirby et al., 2014). Given that the process of iterated learning has been shown to amplify weak biases, it seems likely that this might encourage stronger differentiation and an increased tendency for distinctions to be treated as semantic rather than pragmatic (Kalish et al., 2007).

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

A potential limitation related to the range of contexts investigated concerns the possibility of variation in participants' perception of what is positive, negative, or neutral. This is not a very substantial concern given the results of our norming survey, which – especially taken together with the results of the experiments themselves – give us reason to be confident in our manipulation. This does, however, provide a potential for nuisance variability, which could be further controlled by manipulating participant expectations about what counts as positive and negative, which has been shown to have substantial effects under certain circumstances (Sneller & Roberts, 2018).

A different kind of limitation concerns the number of words involved. Restricting the number of new words to two verbs and one distractor noun had the virtue of simplicity and avoided overtaxing participants. However, it is possible that the task itself amplified the likelihood of differentiation. Future work could investigate this by expanding the number of new words to be learned so that attention is less focused on the target words in particular. It would also be interesting to investigate larger sets of potential synonyms. Real-word near-synonyms only sometimes come in pairs; very often they come in larger sets (e.g., *error : blunder : mistake : accident : oopsie*). This also raises the question of more complex semantic spaces (whether or not the number of potential synonyms is increased).

There is also space for further inquiry concerning lexical category. As mentioned earlier in the discussion of collocational near-synonyms, syntactic context can influence the interpretation of near-synonyms. We used verbs in our study. If we had used adjectives or adverbs, for instance, our results might have looked quite different. On the one hand, the fact that these lexical categories by their nature modify the semantic interpretation of other content words suggests that we might have seen stronger differentiation. On the other hand, this is complicated by the fact that the kind of meanings our contexts provided might be more readily interpreted as part of the core semantic meaning of an adverb or adjective.

There are, in other words, a great many further questions to ask. In the two experiments presented here, we have presented a simple but easily adaptable paradigm that can be used to answer such questions and shed a much broader light on the dynamics of synonymy.

Acknowledgements. We are grateful to Victor Gomes for advice on many aspects of this project, Katie Schuler for assistance in experimental design, Rafael Ventura for suggestions on data analysis, and Julie Legate, Lefteris Paparounas, and Payton Matous for suggestions on early versions of this study.

References

Abounoori, E., & McCloughan, P. (2003). A simple way to calculate the Gini coefficient for grouped as well as ungrouped data. *Applied Economics Letters*, 10(8), 505–509.

Andersen, G., Furiassi, C., & Mišić Ilić, B. (2017). The pragmatic turn in the study of linguistic borrowing. *Journal of Pragmatics*, 113, 71–76.

Clark, E. V. (1992). Conventionality and contrast: Pragmatic principles with lexical consequences. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 171–188). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Diesendruck, G. (2005). The principles of conventionality and contrast in word-learning: An empirical investigation. *Developmental Psychology*, 41, 451–463.

DiMarco, C., Hirst, G., & Stede, M. (1993). The semantic and stylistic differentiation of synonyms and near-synonyms. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*, pp. 114–121.

Edmonds, P. (1999). Semantic Representations of Near-Synonyms for Automatic Lexical Choice. Unpublished Master's Thesis. University of Toronto.

Fairchild, S., & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition*, 181, 80–92. https://doi.org/10.1016/j.cognition.2018.08.010

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, 91, 158–180. https://doi.org/10.1016/j.jml.2016.06.002

Gardner, R. A. (1957). Probability-learning with two and three choices. *American Journal of Psychology*, 70 (2), 174. https://doi.org/10.2307/1419319

Gergel, R., & Kopf-Giammanco, M. (2021). 'Sich ausgehen': On modalizing go constructions in Austrian German. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, 66(2), 141–190.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 28(1), 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Greenberg, J. H. (Ed.). (1966). *Universals of language* (2nd ed.). M.I.T. Press.

Grove, P. B. (Ed.) (1973). *Webster's new dictionary of synonyms: A dictionary of discriminated synonyms with antonyms and analogous and contrasted words*. Springfield, MA: Merriam.

Hayakawa, S. I. (Ed.) (1994). *Choose the rightword: A contemporary guide to selecting the precise word for every situation*. New York: HarperCollins Publishers.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151–195. https://doi.org/10.1207/s15473341lld0102_3

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2), 288–294. https://doi.org/10.3758/BF03194066

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114. https://doi.org/10.1016/j.conb.2014.07.014

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics*, 21(2), 397–429.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2019). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, 198, 104191. https://doi.org/10.31234/osf.io/wsx3a

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121–157.

Potts, C. (2007). *The expressive dimension. Theoretical Linguistics*, 33(2), 165–198. https://doi.org/10.1515/TL.2007.011

R Core Team (2013). R: A language and environment for statistical computing.

Roberts, G., & Galantucci, B. (2016). Investigating meaning in experimental semiotics. *Psychology of Language and Communication*, 20(2), 130–153.

Sassoon, G. W. (2012). A typology of multidimensional adjectives. *Journal of Semantics*, 30(3), 335–380. https://doi.org/10.1093/jos/ffs012

Savage, S. L., & Au, T. K.-F. (1996). What word learners do when input contradicts the mutual exclusivity assumption. *Child Development*, 67(6), 3120. https://doi.org/10.2307/1131770

Schulenberg, R., & Schulenberg, M. R. (2018). Package 'dineq'. Comprehensive R Archive Network. https://cran.r-project.org/package=dineq

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3), 444–449. https://doi.org/10.1016/j.cognition.2010.06.004

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, 170C, 298–311. https://doi.org/10.1016/j.cognition.2017.10.014.

Stevens, J. S., & Roberts, G. (2019). Noise, economy, and the emergence of information structure in a laboratory language. *Cognitive Science*, 43, e12717.

Takimoto, M. (2009). Exploring the effects of input-based treatment and test on the development of learners' pragmatic proficiency. *Journal of Pragmatics*, 41(5), 1029–1046. https://doi.org/10.1016/j.pragma.2008.12.001

Ullmann, S. (1962). *Semantics: An Introduction to the Science of Meaning*. Oxford: Blackwell.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107(2), 729–742. https://doi.org/10.1016/j.cognition.2007.08.00

Wade, L., & Roberts, G. (2020). Linguistic convergence to observed vs. expected behavior in an alien-language map task. *Cognitive Science*, 44(4), e12829. https://doi.org/10.1111/cogs.12829

Waugh, L. R. (1982). Marked and unmarked: A choice between unequals in semiotic structure. *Semiotica*, 38 (3–4), 299–318.

Weir, M. W. (1972). Probability performance: Reinforcement procedure and number of alternatives. *American Journal of Psychology*, 85(2), 261. https://doi.org/10.2307/1420666

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180–185.

Wonnacott, E., & Newport, E. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. R. Clark-Cotton, & S. Ha (Eds.), *Proceedings of the 29th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Zwicky, A. (1978). On markedness in morphology. *Die Sprache* 24. 129–142.

Appendix A

Neutral Sentences – Exposure

Did you _____?
He _____ in the evenings.
They _____ together yesterday.
Didn't they ____?
The child is _____.
I'm done _____. How about you?
Yeah, I think they _____.
People still _____ in some parts of the world.
Dogs cannot _____.
She _____ every day before work.
Are they old enough to ____?
She _____ this morning.
Cannot you see I'm ______.
I've seen him ______.
My dad _____ when he comes home from work.
Do you ___?
My first child sometimes _____.
Did you just _____?
Are you _____?
Will you be ____?
Cat's do not.
She _____ in the afternoons.
They ____ this evening.
Sure, I think he can _____.

Positive Sentences – Exposure

She is great at _____.
Wow! He got a certificate for _____!
I did not know _____ could be so much fun!
I'm proud to be the best at _____!

_____ is the best!
My friends are proud that I've gotten so much better at _____!
_____ is my favorite part of the day.
My friends and family love my _____ skills.
He is amazing at _____!
She got first place for _____.
_____ is his favorite activity!
I enjoy spending my free time _____.
_____ is awesome!
Everyone has a good time when they _____.
Cool! He got an award for _____!
I'm so excited to _____ later!
_____ is so much fun!
They are terrific at _____.

Negative Sentences – Exposure

I'm glad he got arrested for _____!
He will not get my vote after all that _____.
_____ is nasty!
Eww, I overheard that she _____.
They are going to get in trouble if they keep _____!
Yuck! He _____!
_____ is not cool.
_____ is super annoying.
You need to stop _____!
I hope he breaks his habit of _____ all the time.
She thinks _____ is disgusting.
Nice places do not allow _____.
_____ is repulsive.
He deserved to get arrested for _____.
I do not like people who _____.
Do not _____ in front of me!
She will not be friends with guys who _____.
_____ is so irritating.

Noun Sentences – Exposure

The child has a _____.
Look at the _____!
Do you have a _____.
I need to buy some _____.
Can she borrow your _____?
Where is your _____?
Do they sell _____ here?
Can you help me find my _____?
How do you use a _____?
I'm going to give her a _____.
I lent her my _____.
Can I see your _____?

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Neutral Sentences – Generalization

Have you ever seen someone ____?
They did not ____ on Monday.
My daughter ____ every once and a while.
He ___ most days after work.
I just ____ed.
He has never ____ed.
They ___ there.
My son ____.
When did you ____?
When did you first ____?
We ____ last week.
Can you ___?

Positive Sentences – Generalization

Let us _____ today!
I love people who _____!
I would love to see you _____ sometime!
Can you teach me how to _____ like that?
Remember how important it is to practice _____!
He really impressed me by _____.
I wish I could _____ all the time.
People who _____ are so cool.
My first time _____ was fantastic!
I learned to _____ from the best of the best.
Can we _____ together?
My mom taught me how to _____. It's tradition.

Negative Sentences – Generalization

You cannot _____ here!
I cannot believe that he _____ in public!
Do not _____ near me!
I hate people who _____.
Gross! She's _____!
I do not want my child to grow up _____.
I will not hang out with people who _____.
I think she'd be upset if she learned that he _____.
People who _____ are a bad influence.
I cannot believe she is _____ here! It's so rude.
You should not _____.
I'm proud that I do not _____.

Noun Sentences – Generalization

Can you hand me the _____?
I cannot leave without my _____.
She lost her _____ yesterday.
I need the _____.
I got a _____ from the store.
I want a _____ for my birthday.
I hope I did not forget my _____.

Have you seen my _____?
I have a _____.
Let us go shopping for a _____.
He owns a _____.
I used my _____ yesterday.

Appendix B

What three slang words will you be learning? Please select all applicable answers.

- Murp
- Fincur
- Snater
- Monim
- Foncit
- Snooter
- Mop

Which two of the three slang words have the same definition as each other?

- Snater
- Fincur
- Murp

Select the verb(s) from the list below.

- Snater
- Fincur
- Murp

Select the noun(s) from the list below.

- Snater
- Fincur
- Murp

https://doi.org/10.1017/langcog.2023.29 Published online by Cambridge University Press

Appendix C

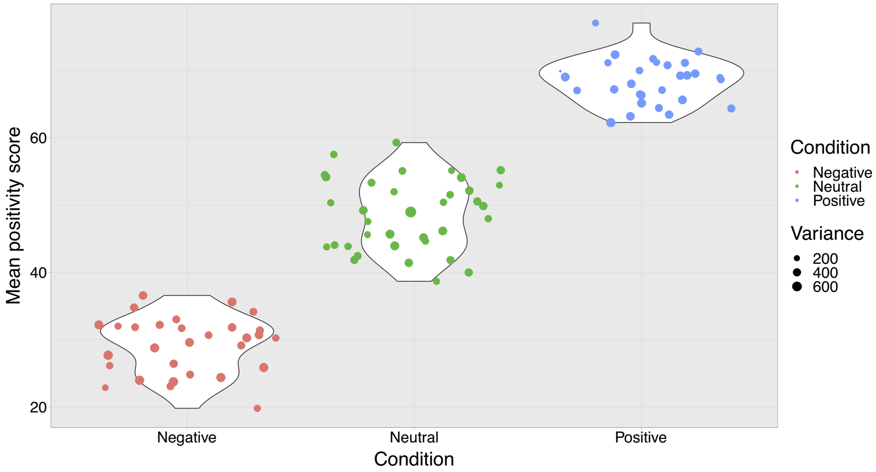

Figure 10. Average positivity scores for stimuli sentences by condition.

Cite this article: Altenhof, A., & Roberts, G. (2023). Quality, not quantity, impacts the differentiation of near-synonyms, *Language and Cognition* 15: 854–883. https://doi.org/10.1017/langcog.2023.29