UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Greedy or not, here I come: Language production under vocabulary constraints in humans and resource-rational models

Permalink

https://escholarship.org/uc/item/02m6x923

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Clark, Thomas
Chen, Sihan
Nicolae, Laura

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Greedy or not, here I come: Language production under vocabulary constraints in humans and resource-rational models

Thomas Hikaru Clark[1], **Sihan Chen**[1], **Laura Nicolae**[2]
[1]Department of Brain and Cognitive Sciences, MIT [2]Department of Economics, Harvard University
{thclark, sihanc}@mit.edu, lauranicolae@g.harvard.edu

Abstract

Communicating using only a limited vocabulary is a common but challenging cognitive phenomenon, requiring an ideal communicator to plan carefully to optimize for intelligibility while circumventing a constrained lexicon. In this work, we investigate how humans respond to a broad array of questions under variable vocabulary limitations, consisting of only 250 highly frequent words at the most restrictive. We provide theoretically motivated comparisons to greedy and globally optimal sampling algorithms using Sequential Monte Carlo inference with large language models. Humans generally resemble greedy sampling more than globally optimal sampling, though more skilled humans are more likely to backtrack and revise – a non-greedy behavior. An observed human pattern of leaning on semantically light words in high-constraint settings falls out of both greedy and globally optimal sampling. We discuss the results and their broader implications for resource-rational cognition, psycholinguistics, L2 communication, and language impairments.

Keywords: Psycholinguistics; Large Language Models; Explanations; Planning; Resource-Rationality

Introduction

Humans use language to communicate (Fedorenko et al., 2024). Using insights from information theory (Shannon, 1948), researchers model language as communication code between a speaker and a listener: the speaker converts their meaning to utterances, and the listener receives the utterances and recovers the meaning (see Gibson et al., 2019, for a review). The production process is incremental: instead of starting with a full utterance in mind, a person typically plans what to say as they speak or write (e.g. Bock & Levelt, 1994; Ferreira & Swets, 2002; Pickering & Garrod, 2013). At the computational level, Futrell (2023) modeled language production as an action planning problem in which a speaker's incremental production is influenced both by a pressure to say contextually predictable things and a pressure to say things that further the communicative goal. In this work, we specifically investigate how constraints on the available set of vocabulary words affect incremental language production.

Communication under a vocabulary constraint is a common cognitive phenomenon. Many societies contain linguistic minorities with limited knowledge of the majority language, which affects their ability to access basic services or participate in the economy (Bleakley & Chin, 2004; Grogger et al., 2020). Compared to native speakers, learners of a language (L2 learners) tend to use fewer and more frequent words (Laufer, 1991). When L2 learners do not know a specific word, they may employ various strategies to get a meaning across, such as describing key properties of the concept (Poulisse, 2011) or using "all-purpose" or semantically light words as substitutes (Dörnyei & Scott, 1997). This phenomenon suggests that there may be diminishing marginal returns to learning new vocabulary words above a certain point, since a subset of the vocabulary is typically sufficient for communication.

Another common case of vocabulary-constrained communication is when an expert explains a complex phenomenon to a layperson. Past work has considered what makes for a good explanation (e.g. Brewer et al., 2000; Cruz & Lombrozo, 2025; Chandra et al., 2024; McCarthy & Keil, 2023). For example, Sulik et al. (2023) suggests that a good explanation contains functional (i.e., what is something for?) or mechanistic (i.e., how does something work?) information. Besides the type of information, the words used also matter. For example, laypersons find explanations with jargon (terms only understood by a specific group of experts) harder to understand (Bullock et al., 2019; Cruz & Lombrozo, 2025; Keuleers et al., 2015). It remains unclear how exactly the usage of specific words changes in response to vocabulary constraints, and what kind of algorithmic-level (Marr, 1982) model of language generation captures the properties of language generated under vocabulary constraints.

In this work, we define a large space of communicative goals and evaluate the properties of English responses generated by humans under varying vocabulary constraints, comprising only the most common 250 English words at the most restrictive. Recently-developed techniques allow us to sample from a language model subject to custom constraints (Lipkin et al., 2025; Loula et al., 2025), and provide a new opportunity to test algorithmic-level hypotheses about human cognition. In particular, we can manipulate the *greediness* of constrained generation, where a purely greedy algorithm has a bias for locally high-probability continuations, while approximately globally-optimal inference algorithms like particle-based Sequential Monte Carlo (SMC) avoid these biases and are thus more consistent with planning ahead during language production. We investigate whether humans perform constrained language generation in a way that aligns more closely with greedy generation or globally-optimal inference.

To foreshadow our results: for the humans in our study, performance as a function of vocabulary size more closely resembled greedy sampling than SMC-based constrained generation, which handled very small vocabularies better than humans or greedy sampling. Despite this, we observe that top-scoring humans revise valid string prefixes (a non-greedy behavior) significantly more than low-scoring humans. This suggests that human approaches to constrained generation are heterogeneous: some plan ahead or revise their answers while

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

others prioritize simple, greedy, and imperfect responses. We also observe interpretable patterns in the frequency of word use under different vocabulary constraints: semantically "light" words, like *do, thing,* and *people*, are used disproportionately frequently under strict vocabulary constraints, even relative to other allowed words. This pattern is reproduced by constrained generation from language models. We conclude by discussing the implications of these findings for both psycholinguistics and for broader areas of impact.

Methods

Constrained vocabulary definition

We define a constrained vocabulary of size N, consisting of the N most frequent words in English, according to the word list provided by the `wordfreq` Python package (Speer, 2022). For each word in the list, the set of words sharing the same base form (lemma) was also included in the list, using the `lemminflect` Python package, available within `spaCy` (Honnibal et al., 2020). For example, if any of the word forms *drink, drank, drunk, drinks,* or *drinking* were present within the top N words, then all word forms in this set were included. We define seven vocabulary sets, starting at 250 words and doubling up to 16,000 words. For reference, it is estimated that the average native speaker of American English is familiar with 42,000 word lemmas from 11,000 word families (Brysbaert et al., 2016).

Questions dataset

To simulate a variety of real-world situations with diverse communicative goals, we assemble a dataset of 192 questions divided into the following four categories of 48 questions each: Why, How, ExplainSimple, and RedditELI5 (Explain Like I'm 5). The Why questions are sourced from a study by Sulik et al. (2023), who investigated which features make for good explanations. The How and ExplainSimple datasets were created from scratch using fixed sentence patterns ("How is/are/do/does/can/would [BLANK]?" and "Explain [BLANK] in simple terms"), with the goal of covering a wide range of communicative topics, including everyday life, sports, and science. The RedditELI5 questions were sourced from the all-time most popular questions on the Reddit forum "Explain Like I'm 5", lightly edited for length and clarity.

Human behavioral experiment

We created an online behavioral experiment in which participants respond to questions from the above dataset, using an interface which allows them to type only words within a specified vocabulary. While not a fully natural task, the interface simulates a speaker encountering vocabulary limitations, forcing on-line adaptation to the constraint via circumlocution or other strategies. Participants could only type or delete (selecting, replacing, and inserting text was blocked). We recruited 144 English speakers from Prolific, who were each paid $6. The study was conducted under an approved IRB protocol at the authors' institution. Each participant answered 16 questions, beginning with 4 questions with no vocabulary constraints, then 4 questions each for vocabulary sizes of 4000, 1000, and 250. Each question in the dataset was answered by three unique individuals at each vocabulary size. Participants were prompted to move to the next question after 90 seconds without a submission. We record participants' final responses as well as all intermediate keystrokes.

Constrained LLM generation with AWRS

Adaptive Weighted Rejection Sampling (Lipkin et al., 2025) is a Sequential Monte Carlo method for sampling incrementally from a language model, subject to user-defined constraints (implemented as binary functions that can be evaluated on partially generated strings). It does this by maintaining multiple hypotheses in parallel, represented by weighted *particles*. We use AWRS to generate responses to prompts, using a custom potential in the `GenLM-control` library to enforce a hard constraint on the set of words in the vocabulary. Constrained generation was performed using the `Llama-3.2-1B-Instruct` model, with either 16 or 32 particles for SMC inference (denoted AWRS-16 and AWRS-32, respectively). By setting the number of particles to 1, the algorithm is equivalent to local greedy sampling, where only one hypothesis is maintained at each step of generation. A prompt with instructions and two few-shot examples was provided. Example responses from both humans and models, under both a permissive and a restrictive constraint, are shown in Table 1.

Analyses

Automated evaluation of response quality To create an estimate of the quality of responses to each question (from both humans and models), we use a prompted LLM (`Llama-3.1-8B-Instruct`) to assign scores on a 7-point Likert scale, along with accompanying justifications, under an LLM-as-judge paradigm (Gu et al., 2025; Zheng et al., 2023). For each question, we aim to approximate the value $\mathbb{E}[f(X)] = \sum_x f(X)p_X(x)dx$ where X denotes a random variable sampled from the distribution over constrained responses to a question, $f(\cdot)$ denotes the scoring function, and p_X denotes the probability density function of X, approximated by the normalized SMC weights. For computational efficiency, we discard samples from SMC with weights below a chosen threshold of 0.01, which contribute minimally to the sum. We use the same evaluation pipeline on human responses, treating each response as a single sample (all equally weighted). The automated evaluator is an imperfect proxy for response quality that was employed to address the large number of both model and human responses that required annotation, for which it was not practical to collect comprehensive human judgments. While the absolute score assigned to human- vs. model-generated utterances may be biased (e.g. if LLM judges prefer LLM-generated responses), we are primarily concerned with *changes* to the evaluation score as a function of vocabulary size. We conducted a norming study to validate the LLM-as-judge pipeline: one third of questions (64 out of 192) were randomly sampled, and for each of four differ-

ent vocabulary sizes, one response from humans, the greedy model, and the AWRS-32 model was randomly chosen and graded by N=24 human participants on Prolific, who each saw 32 question-answer pairs (users who participated in the constrained generation experiment were excluded). Human graders were given the same instructions provided to the LLM-as-judge, rating the response quality on a 7-point Likert scale. For the subset of questions in the norming study, automated LLM-generated ratings and human ratings had a Spearman ρ of 0.60, suggesting that the automated pipeline captures a substantial amount of variance in human ratings.

We predict that the average response score will decrease monotonically as the allowed vocabulary is reduced, and we evaluate whether scores for human responses decline similarly to the greedy or SMC-based model responses as vocabulary size is restricted.

Shifts in word frequency We compute the frequency of each word in the generated output as a function of vocabulary size. Restricting the vocabulary necessarily removes low-frequency words from the output distribution, but does not *directly* change the relative frequencies of the remaining words. For example, all remaining words might be used slightly more frequently to account for the words that were removed from the vocabulary, but it is not obvious that there would be any change in the usage rank of the remaining words. However, we hypothesize that tightening the vocabulary constraint will systematically change the frequency distribution of used words, even among those that are allowed by the constraint. In particular, words which have a high degree of *substitutability* with low-frequency words will increase in usage rank relative to words with a low degree of substitutability (Varian, 2010). For example, in settings where low-frequency words are not available, we may expect the word "thing" to be substituted instead (Hasselgren, 1994; Klein & Perdue, 1997). Meanwhile, function words, including prepositions, conjunctions, and determiners, may not see a large change in usage.

Revision as an index of greediness vs. planning Under strictly greedy language generation, once a word is output, it can no longer be revised. In language modeling terms, this feature of greedy sampling can lead to so-called "dead ends," where a sequence of high-probability words leads to a context in which the only high-probability continuation is outside of the vocabulary (see Lew et al., 2023, for a discussion). For example, given the sequence "It was a blessing in...", Google N-grams estimates a 90% probability that the next word is "disguise". If this word is outside the vocabulary, a greedy sampler would be forced to choose a different, low-probability continuation, which could rapidly lead to an incoherent sentence being generated. A rational constrained language generator could avoid this problem by either planning ahead (avoiding going down the dead end in the first place) or by revising (backtracking or deleting words once a dead end is detected). It is difficult to directly measure whether a person is planning ahead. However, given the fine-grained typing data from our online task (Figure 1), it is possible to see whether participants ever delete previously written words. For a word to be counted as deleted, it must have first been completed (e.g., by a whitespace or punctuation character) and then erased by repeated presses of the backspace key. We compute the average number of word deletions per response to quantify revisions.

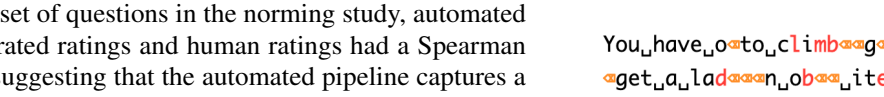

Figure 1: Example keystroke sequence under a 250-word vocabulary, in response to "How do you rescue a cat from a tree?". Red characters denote disallowed keystrokes; the backspace symbol is displayed in orange. The final response was "`You have to get a thing to help you.`".

Results

Table 1: Example responses at different vocabulary sizes.

Question: Explain machine learning in simple terms.

Source (Vocab)	Response
Human (4000)	Machine Learning is when a Machine is taught how to do specific tasks and assist with issues.
Greedy (4000)	Machine learning is a way computers can learn from data without being told what to do. It's like teaching a child to recognize objects by showing them a lot of pictures.
AWRS-32 (4000)	Machine learning is a way computers can learn from data and get better at doing things on their own. It's like having a super smart assistant that can recognize patterns and make decisions automatically.
Human (250)	Over and over it knows.
Greedy (250)	It's a way that a big, very old, very good thinking game helps make things that are like us, like you, not that way. It uses big groups of people giving lots of old, old things into the game ...
AWRS-32 (250)	It's a way of making things good at doing things on their own. You give it some old things it's good at, and it gets better over time.

Effects of vocabulary size reveal similarity between humans and greedy sampling

Figure 2 shows the average evaluation scores of responses generated by humans and the three algorithms (greedy, AWRS-16, and AWRS-32) at each vocabulary size. Exclusion criteria for participants were not pre-defined, but we perform a post-hoc split of participants into the top 50% and bottom 50% by average response score to separate out low-effort responses. In absolute terms, human responses scored lower, on average, than all models at each vocabulary size. This may be influenced by the fact that human responses were generally shorter and

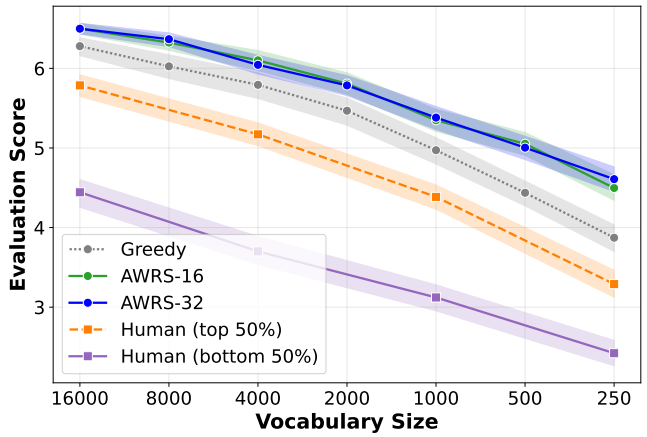

Figure 2: Mean weighted score (automated evaluation on a 1-7 Likert scale) as a function of allowed vocabulary size. The horizontal axis is logarithmic. Shaded bands denote 95% confidence intervals over the scores for all 192 questions. The label '16000' is mapped to the unconstrained condition for humans.

less detailed than model responses. All three algorithms and the human-generated responses exhibit diminishing marginal returns: evaluation scores increase by nearly-constant increments each time the vocabulary size is doubled, suggesting that the average marginal utility of adding a new word to the vocabulary declines as the vocabulary size expands.

Tightening the vocabulary constraint produced similar decreases in the scores of the human responses and the greedy algorithm's responses at all vocabulary sizes. Meanwhile, the AWRS algorithm displays greater robustness to vocabulary constraints compared to humans and the greedy algorithm: the evaluation scores of AWRS-generated responses declined more slowly than those of humans and the greedy algorithm, with generally larger gaps in performance for smaller vocabularies. On the 250-word vocabulary, the AWRS algorithm achieves half a point better (on a 7-point scale) than would be expected if its performance degraded at the same rate as human participants and greedy sampling.

A shift towards semantically light words in constrained generation

Consistent with our prediction, tighter vocabulary constraints induce responses with more frequent use of semantically light content words, such as *thing, do,* or *people*. Both human-generated and model-generated responses used semantically light words more frequently as the vocabulary constraint became stricter (Figure 3). In particular, semantically light content words tended to rise in frequency rank, while function words tended to remain constant or fall in rank. We note that this qualitative pattern occurs for all model results, including AWRS-16 and AWRS-32.

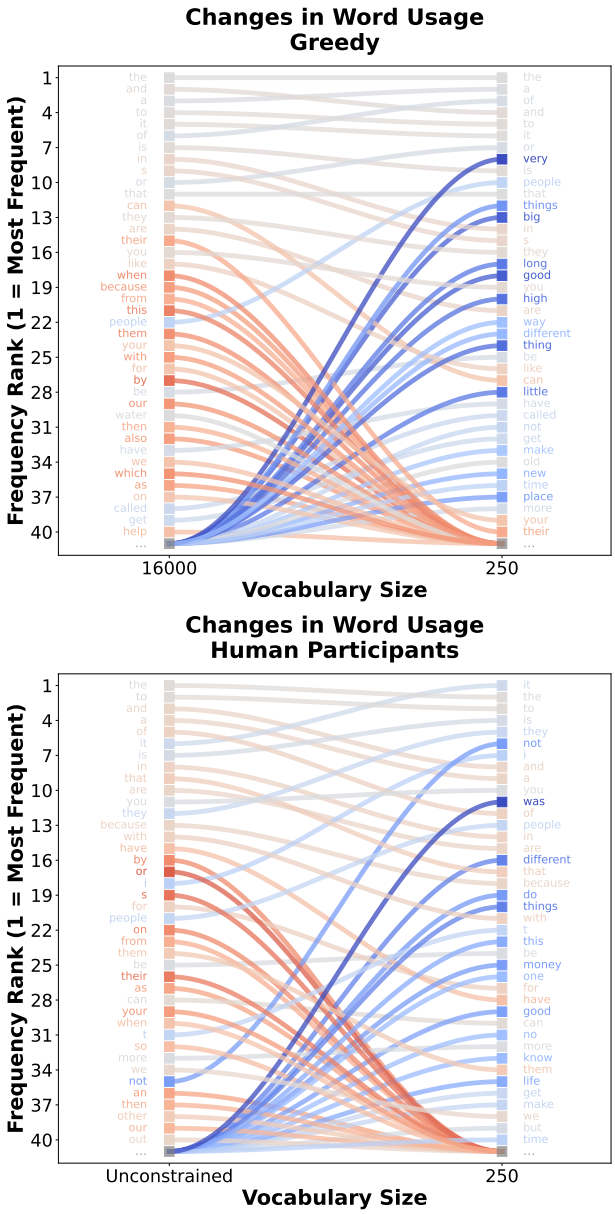

Figure 3: Bump plots for both models and humans showing word frequency rankings at largest and smallest vocabulary sizes. The color scale denotes sign and magnitude of rank change. Semantically light nouns, verbs, and adjectives (e.g., *things*, *make*, etc.) tend to rise in rank as vocabulary size shrinks, while connectives (*or*, *when*) tend to decrease in rank.

Skilled humans revise more under constraint

Figure 4 shows the rate of the human respondents' word deletions per response by vocabulary size and participant skill group (a post-hoc split of the participants into the top and bottom 50% by average score). For participants in the bottom 50%, word deletions remained flat across different constrained vocabulary sizes. In contrast, for participants in the top 50%, word deletions were significantly higher in the constrained vocabularies of 1000 and 250 words than in larger vocabularies.

This much-greater propensity for revision among high-scoring humans suggests an attempt at a non-greedy approach, though these attempts may not be successful, as evidenced by the fact that the responses generated by this group of participants generally scored similarly to the greedy algorithm.

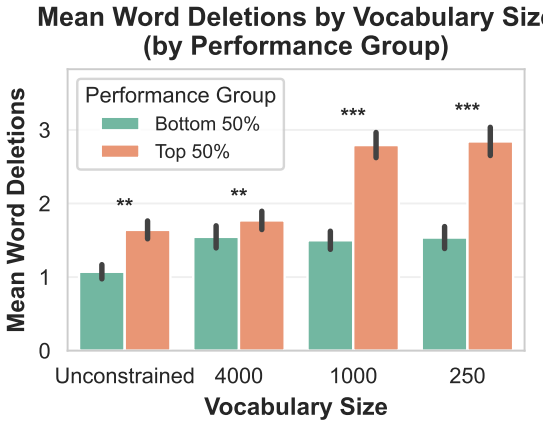

Figure 4: Mean valid word deletions per response, by vocabulary size and post-hoc performance group. Error bars denote 95% confidence intervals. Stars denote p-values of Mann-Whitney U-test with Bonferroni correction (*: <0.05, **: <0.01, ***: <0.001).

Discussion

Resource-rational inference in language production

An extensive body of work in cognitive science and behavioral economics suggests that humans are approximately rational subject to cognitive costs (Kahneman & Tversky, 1979; Kahneman, 2003; Griffiths et al., 2015; Lieder & Griffiths, 2020). In real-time communication, a pressure for communicative utility interacts with constraints on the availability of cognitive resources like time and attention. Evidence also suggests that humans can adapt to *novel* constraints by deploying strategies that make efficient use of available resources: people asked to communicate about novel objects rapidly converge to a set of labels, using shorter phrases for more frequently observed items (Krauss & Weinheimer, 1964). Studies have also shown emergent systematicity in highly constrained communication channels when there was a communicative goal, e.g. whistling to signal a color on a continuum (Chen et al., 2025).

The intersection of language modeling and probabilistic inference provides an opportunity to compare human behavior to algorithmic-level hypotheses about processing. Sequential Monte Carlo has been argued to provide a cognitively plausible algorithm for approximate Bayesian inference for language comprehension (Clark et al., 2025a,b). The AWRS algorithm enables efficient *generation* under a vocabulary constraint. This algorithm is inherently resource-rational: it uses more samples when the constraint is difficult to satisfy than when it is easy, and varying the particle count results in a continuum of behavior from fast, greedy sampling to slower, more exact inference, which here resulted in higher-scoring responses.

Our results indicate that humans presented with the task of constrained generation show behavior that, somewhat surprisingly, more closely resembles greedy sampling than particle-based SMC. Above-average participants, however, show a signature of non-greedy sampling: they are much more likely to revise partially generated sentences. To help explain these results, we note that greedy sampling is one endpoint of a resource-rational trade-off; it minimizes computational effort but maximizes local bias. A plausible hypothesis is that humans are closer to the greedy extreme of this trade-off (possibly due to the high costs of revising or planning ahead), but that highly-motivated individuals may opt to increase effort to counter the dead ends induced by greedy sampling. Future work can consider whether financial incentives, in-person experimentation, or pairing participants together can induce more non-greedy behavior by placing greater weight on communicative success.

Our results also indicate that certain words are used disproportionately when vocabulary is artificially restricted. In the most restricted vocabularies, semantically light content words jump substantially in usage rank. These words have especially high communicative utility in constrained settings, because they can serve as substitutes for low-frequency words (consider the frequent use of the word "way" in Table 1). While intended words may not have perfect substitutes, they can often be replaced by a longer phrase, using relative clauses or prepositional phrases to add detail. This pattern is also seen in both greedy and non-greedy model results, suggesting that it is a basic corollary of the substitutability of these words, rather than a result of active Bayesian inference. These results are consistent with existing theoretical frameworks of built-in redundancy in language providing robustness against failures of communication (Tourtouri et al., 2021; Degen et al., 2020; Mahowald et al., 2023; Leufkens, 2020).

Another property of constrained generation with SMC is the sensitivity of inference quality and efficiency to the proposal distribution used to generate incremental next-word candidates. The greater the divergence between the proposal distribution and the target distribution, the more samples are needed to generate a word that meets the constraint; the algorithm works best when it does not have to reject too many next-word candidates (Lipkin et al., 2025). We speculate that this phenomenon has a natural analog in human cognition. There are two ways that a human can be good at constrained generation. The first is to take many possible "samples"; we expect to see this reflected in slow processing and frequent deletions. The second is to have a good "proposal distribution" that assigns high probability to continuations which already satisfy the constraints; we expect this to be reflected in fluent production with relatively few revisions. Crucially, such a proposal distribution differs from the baseline distribution of language, and thus requires learning via experience. As people gain more familiarity with this task, it may become more natural and require less explicit

inference, in keeping with theories of amortized computation in cognition (Gershman & Goodman, 2014). Future work may consider whether humans with extensive L2 experience are better than monolinguals at this task.

Intrinsic differences in question difficulty

Are some questions simply harder to answer than others? We list the top and bottom questions by average evaluator score for AWRS-32 and for humans, when the vocabulary size was 250:

Top/Bottom AWRS-32 Questions

▲ Explain Christianity in simple terms.
▲ Explain machine learning in simple terms.
▲ Explain consequentialism in simple terms.
▲ Explain the placebo effect in simple terms.
▼ How do I make schnitzel?
▼ Why does it echo if we yell in a cave but not a regular room?
▼ How do you remove salt from water?
▼ How do you prevent a sunburn?

Top/Bottom Human Questions

▲ Explain the placebo effect in simple terms.
▲ How do languages get new words?
▲ How do maggots get into a place like a freezer that's sealed air tight when it loses power?
▲ Why are human babies born so helpless?
▼ How are books printed?
▼ How can I haggle prices at a market effectively?
▼ Explain the International Space Station in simple terms.
▼ What is a hedge fund?

First, we observe that for both the model and humans, most of the lowest-rated questions were in the How category. This may be because How questions typically prompt mechanistic explanations, which are known to be hard to understand, even if the subject of the question is well-known or technically simple (e.g. McCarthy & Keil, 2023; Kelemen et al., 2013; Lombrozo & Carey, 2006; Lombrozo & Wilkenfeld, 2019). Additionally, these questions might involve lexicalized concepts that are difficult to explain via circumlocution. For example, explaining how to make schnitzel is difficult without the word *pork*, which is not among the most common 250 English words. In contrast, the highest-rated questions likely present a communicator with alternative ways of expressing the idea, even if the question appears to be quite technical on the surface. For example, the concept of the placebo effect can be expressed by words such as *feel* and *work*, both of which are among the 250 most frequent words. The highest-scoring questions for AWRS-32 all belong to the ExplainSimple dataset; this is possibly because the questions explicitly prompt the model to explain "in simple terms", which reduces the divergence between the proposal distribution and target constrained distribution. These results suggest that whether a communicative goal is easy or hard to achieve under vocabulary constraints likely depends greatly on the availability of substitute words and well-suited analogies, which varies widely even within a question category, and not necessarily on the "technical" difficulty of the subject matter.

Broader implications

One key application of our work is for language learning. Our finding that semantically light words appear more frequently when the allowed vocabulary is small could provide useful guiding principles for optimizing language learning. For example, our results suggest that a foreign-language learner who expects to learn only 250 – 1000 words in the target language should prioritize learning general-purpose, semantically light words (not necessarily the most frequent words) in order to be able to communicate effectively across a wide variety of everyday situations. In contrast, an intermediate-level speaker who hopes to sound more "native" might consider practicing communication without over-using semantically light words such as *thing, do,* or *people*, which occur less frequently in the unconstrained distribution of language.

Our results on the diminishing marginal returns of expanding one's vocabulary size also suggest that learning the first few hundred or thousand most frequent English words has disproportionately high returns for communicative utility (especially in the presence of a charitable comprehender), despite still being far below the vocabulary size of a native speaker. This could potentially explain why L2 learners often converge to an efficient but limited "Basic Variety" of a language (Klein & Perdue, 1997) or over-use familiar words (Hasselgren, 1994): if one can communicate well using a restricted vocabulary, then there are fewer pressures to expand one's vocabulary.

Additionally, modeling vocabulary-constrained communication has relevance for the study of language disorders. For example, communication for individuals with aphasia may benefit from flexible inference both in language production (choosing which words to produce when production is very difficult) and comprehension (how an interlocutor infers an intended meaning from fragmented utterances) (Beeke, 2012; Fedorenko et al., 2022). Patients with aphasia may be less likely to use low-frequency words like *sailboat* while being more likely to use syntactically rich fragments like *boat that is moved by the wind* to circumvent this limitation (Rezaii et al., 2022). Modeling how people generate language under constraint can shed light on the rapid, online inferences made by conversational partners in the presence of language disorders, and can potentially inform strategies and technologies for easing communication with vulnerable groups (for example, by modifying existing language models to better infer intended messages from atypical inputs).

This study has investigated how vocabulary constraints affect human-generated language, with theoretically-motivated comparisons to greedy and non-greedy sampling from language models using state-of-the-art algorithms. Our results demonstrate both the flexibility and the limitations of human language generated under vocabulary constraints, and provide an approach for computationally characterizing its properties.

References

Beeke, S. (2012). 13. Aphasia: The pragmatics of everyday conversation. In *13. Aphasia: The pragmatics of everyday conversation* (pp. 345–372). De Gruyter Mouton. doi: 10.1515/9783110214215.345

Bleakley, H., & Chin, A. (2004). Language Skills and Earnings: Evidence from Childhood Immigrants*. *The Review of Economics and Statistics*, *86*(2), 481–496. doi: 10.1162/003465304323031067

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In *Handbook of psycholinguistics* (pp. 945–984). San Diego, CA, US: Academic Press.

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (2000). Explanation in scientists and children. In *Explanation and cognition* (pp. 279–298). Cambridge, MA, US: The MIT Press.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*, *7*, 1116. doi: 10.3389/fpsyg.2016.01116

Bullock, O. M., Colón Amill, D., Shulman, H. C., & Dixon, G. N. (2019, October). Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Understanding of Science*, *28*(7), 845–853. Retrieved 2026-04-09, from https://doi.org/10.1177/0963662519865687 doi: 10.1177/0963662519865687

Chandra, K., Chen, T., Li, T.-M., Ragan-Kelley, J., & Tenenbaum, J. (2024). Cooperative explanation as rational communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Chen, A. M., Hofer, M., Poliak, M., Levy, R., & Zaslavsky, N. (2025). Discrete and systematic communication in a continuous signal-meaning space. *Journal of Language Evolution*, *10*(1), lzaf003. doi: 10.1093/jole/lzaf003

Clark, T. H., Vigly, J. H., Gibson, E., & Levy, R. (2025a). A Model of Approximate and Incremental Noisy-Channel Language Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0).

Clark, T. H., Vigly, J. H., Gibson, E., & Levy, R. P. (2025b). Resource-Rational Noisy-Channel Language Processing: Testing the Effect of Algorithmic Constraints on Inferences. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 23659–23672). Suzhou, China: Association for Computational Linguistics.

Cruz, F., & Lombrozo, T. (2025). How laypeople evaluate scientific explanations containing jargon. *Nature Human Behaviour*, *9*(10), 2038–2053. doi: 10.1038/s41562-025-02227-0

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621. doi: 10.1037/rev0000186

Dörnyei, Z., & Scott, M. L. (1997). Communication Strategies in a Second Language: Definitions and Taxonomies. *Language Learning*, *47*(1), 173–210. doi: 10.1111/0023-8333.51997005

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*(8017), 575–586. doi: 10.1038/s41586-024-07522-w

Fedorenko, E., Ryskin, R., & Gibson, E. (2022). Agrammatic output in non-fluent, including Broca's, aphasia as a rational behavior. *Aphasiology*, *0*(0), 1–20. doi: 10.1080/02687038.2022.2143233

Ferreira, F., & Swets, B. (2002, January). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, *46*(1), 57–84. Retrieved 2026-04-09, from https://www.sciencedirect.com/science/article/pii/S0749596X01927974 doi: 10.1006/jmla.2001.2797

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39), e2220593120. doi: 10.1073/pnas.2220593120

Gershman, S., & Goodman, N. (2014). Amortized Inference in Probabilistic Reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. doi: 10.1016/j.tics.2019.02.003

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Grogger, J., Steinmayr, A., & Winter, J. (2020). *The Wage Penalty of Regional Accents* (Working Paper No. 26719). National Bureau of Economic Research. doi: 10.3386/w26719

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., … Guo, J. (2025). *A Survey on LLM-as-a-Judge* (No. arXiv:2411.15594). arXiv. doi: 10.48550/arXiv.2411.15594

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258. doi: 10.1111/j.1473-4192.1994.tb00065.x

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. doi: 10.5281/zenodo.1212303

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475. doi: 10.1257/000282803322655392

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083. doi: 10.1037/a0030399

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*(8), 1665–1692. doi: 10.1080/17470218.2015.1022560

Klein, W., & Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). *Second Language Research*, *13*(4), 301–347. doi: 10.1191/026765897666879396

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1), 113–114. doi: 10.3758/BF03342817

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *Modern Language Journal*, *75*(4), 440–448. doi: 10.2307/329493

Leufkens, S. (2020). A functionalist typology of redundancy. *Revista da ABRALIN*, 79–103. doi: 10.25189/rabralin.v19i3.1722

Lew, A. K., Zhi-Xuan, T., Grand, G., & Mansinghka, V. K. (2023). *Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs* (No. arXiv:2306.03081). arXiv. doi: 10.48550/arXiv.2306.03081

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. doi: 10.1017/S0140525X1900061X

Lipkin, B., LeBrun, B., Vigly, J. H., Loula, J., MacIver, D. R., Du, L., … Vieira, T. (2025). *Fast Controlled Generation from Language Models with Adaptive Weighted Rejection Sampling* (No. arXiv:2504.05410). arXiv. doi: 10.48550/arXiv.2504.05410

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. doi: 10.1016/j.cognition.2004.12.009

Lombrozo, T., & Wilkenfeld, D. A. (2019). Mechanistic versus functional understanding. In S. R. Grimm (Ed.), *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (pp. 209–229). New York, NY: Oxford University Press. (149826 Section: 11)

Loula, J., LeBrun, B., Du, L., Lipkin, B., Pasti, C., Grand, G., … O'Donnell, T. J. (2025). *Syntactic and Semantic Control of Large Language Models via Sequential Monte Carlo* (No. arXiv:2504.13139). arXiv. doi: 10.48550/arXiv.2504.13139

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., & Futrell, R. (2023). Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, *241*, 105543. doi: 10.1016/j.cognition.2023.105543

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc.

McCarthy, A. M., & Keil, F. C. (2023). A right way to explain? Function, mechanism, and the order of explanations. *Cognition*, *238*, 105494. doi: 10.1016/j.cognition.2023.105494

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329–347. doi: 10.1017/S0140525X12001495

Poulisse, N. (2011). A Theoretical Account of Lexical Communication Strategies. In *The Bilingual Lexicon* (pp. 157–190). John Benjamins Publishing Company.

Rezaii, N., Mahowald, K., Ryskin, R., Dickerson, B., & Gibson, E. (2022). A syntax–lexicon trade-off in language production. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(25). doi: 10.1073/pnas.2120203119

Shannon, C. E. (1948). A mathematical theory of communication. , 55.

Speer, R. (2022). *Rspeer/wordfreq: V3.0.* Zenodo. doi: 10.5281/zenodo.7199437

Sulik, J., van Paridon, J., & Lupyan, G. (2023). Explanations in the wild. *Cognition*, *237*, 105464. doi: 10.1016/j.cognition.2023.105464

Tourtouri, E. N., Delogu, F., & Crocker, M. W. (2021). Rational Redundancy in Referring Expressions: Evidence from Event-related Potentials. *Cognitive Science*, *45*(12), e13071. doi: 10.1111/cogs.13071

Varian, H. R. (2010). *Intermediate microeconomics: a modern approach* (8th ed.). New York: W.W. Norton & Co. Retrieved from `http://www.worldcat.org/search?qt=worldcat_org_all&q=0393934241` (varian10 tex.added-at: 2015-02-10T03:45:35.000+0100 tex.interhash: 1075d5839aaa70f12ee85cfd5239e1f5 tex.intrahash: 8c9a8e25755e2c46b53db3505767e471 tex.refid: 317920200 tex.timestamp: 2015-02-10T03:55:59.000+0100)

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* (No. arXiv:2306.05685). arXiv. doi: 10.48550/arXiv.2306.05685