Soft production preferences emerge from a bottleneck on memory

Neil Rathi
Stanford University
rathi@stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Soft production preferences are a key feature of incremental language production, yet they lack a well-defined unified explanatory theory. Here, we propose an information-theoretic theory of availability effects grounded in the notion of lossy-context working memory, which takes the form of a cost function that can be applied to any computational-level model of language production. We show that production policies that minimize this cost function naturally give rise to key soft preferences observed in empirical data, including frequency bias, heavy-NP shift, and agreement attraction. We then show a novel prediction made by the model regarding the entropy of arguments' thematic roles, and show that this effect holds in corpus data.

Keywords: language production; working memory; availability; information theory

Introduction

Language production is thought to be largely incremental, meaning that speakers choose their words as they produce them, in real time (Bock, 1982; Levelt, 1989; Griffin, 2001; Brown-Schmidt & Hanna, 2011). Across languages, this incrementality results in speakers routinely expressing strong preferences between semantically equivalent (or nearly equivalent) utterances. Take, for example,

(1) a. I gave my friend a beautiful red Italian car.
b. I gave a beautiful red Italian car to my friend.

Speakers almost universally produce utterances like (1a) over those like (1b), where the shorter constituent ('my friend') comes before the longer one ('a beautiful red Italian car').

Here, we take a broad view of such effects, which we term **soft production preferences** (Bresnan, Dingare, & Manning, 2001). These are not statements about grammaticality, but rather strong behavioral preferences that inform our understanding of computational models of production. This terminology is inclusive of ordering effects like heavy-NP shift as in Example 1 (Stallings, MacDonald, & O'Seaghdha, 1998; Arnold, Losongco, Wasow, & Ginstrom, 2000), and availability effects, where speakers prefer to use words that are 'easier' to produce, and will therefore produce them earlier in an utterance, or over less 'available' alternatives (Koranda, Zettersten, & MacDonald, 2021). It also includes systematic speech errors, such as agreement attraction. In language data, these effects manifest as preferences to produce items with certain features earlier: for example, short constituents are produced before long ones, frequent before infrequent, animate before inanimate, and so on. Yet most computational-level models of production do not explain why these features in particular are associated with availability and other soft production preferences (e.g. Degen, 2023; Cohn-Gordon, Goodman, & Potts, 2019; Ferreira & Dell, 2000; Bock & Warren, 1985).

Here, we propose an explanatory theory of soft production preferences. Previous work has hypothesized that some of these preferences arise from a constraint on working memory (e.g. Chang, 2009; Slevc, 2011; Arnold, Kaiser, Kahn, & Kim, 2013; Futrell, 2019; Rathi, Waldon, & Degen, 2024). In line with this idea, we show that a variety of soft production preferences emerge if we assume that speakers act in a way that minimizes an information-theoretic quantity called **predictive information**, which quantifies the amount of information that must be extracted and stored about the past of an utterance needed to specify its future. Our model is best thought of as a universal cost function compatible with many concrete models of production. We hold that soft production preferences will emerge from any policy that minimizes predictive information, whether implicitly or explicitly.

In what follows, we first formally characterize predictive information. We present three simulations that show how minimizing predictive information gives rise to three key empirical soft production preferences: frequency bias in constituent order, heavy-NP shift, and agreement attraction. We then show that our theory predicts a novel ordering effect as a function of the entropy of the thematic roles of the arguments, and verify that this effect appears in corpus data.

Model

Our goal is to characterize how production effects emerge from a speaker's **production policy**, a distribution over utterances conditional on the world state, which includes the linguistic context of what the speaker has said so far. We do so by means of a cost function building on the notion of **predictive information** (Futrell & Hahn, 2024), a measure of the complexity of sequential prediction.

Predictive information measures the amount of information about the past of a stochastic process that an agent needs to extract and store in order to predict or generate its future (Bialek, Nemenman, & Tishby, 2001; Crutchfield & Feldman, 2003; Palmer, Marre, Berry, & Bialek, 2015). Thus, it implicitly formalizes the notion of memory cost in terms

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

of a concrete information-theoretic quantity. If speakers act under a resource constraint on memory, then they should prefer policies that have low predictive information, and thus require less memory resources. The idea of minimizing predictive information is a simpler form of the memory–surprisal tradeoff from previous work (Hahn, Degen, & Futrell, 2021; Hahn, Mathew, & Degen, 2021; Futrell & Hahn, 2022; Rathi, Hahn, & Futrell, 2022): predictive information is the amount of information that must be extracted from context in order to achieve the highest possible predictability on average.

Memory constraints often appear in models of online language *comprehension* (Lewis & Vasishth, 2005; Futrell, Gibson, & Levy, 2020; Hahn, Futrell, Levy, & Gibson, 2022; Kuribayashi, Oseki, Brassard, & Inui, 2022). Here we extend this concept to production. The core idea is that during incremental language production—indeed, during the production of any sequence of actions (Lai & Gershman, 2021)—it is easier and more automatic to produce units that are incrementally predictable (Goldman-Eisler, 1957; Futrell, 2023). However, these incremental predictions must be made based on constrained memory. Production preferences emerge from policies that minimize the amount of memory that must be used to achieve optimal prediction.

Predictive Information

Given a stationary stochastic process generating symbols labeled $\ldots, X_{t-1}, X_t, X_{t+1}, \ldots$, predictive information is defined as the mutual information between the entire past of the process and the entire future of the process:

$$E := \mathbf{I}[X_{<t} : X_{\geq t}]. \tag{1}$$

This quantity can be calculated from the ***n*-gram entropy rate**, which is the average entropy of a symbol given the previous $n-1$ symbols as context. The n-gram entropy rate is the amount of information that the size $n-1$ context window gives us when we produce X_t. Formally, for a sequence $\{X_i\}_{i=1}^n$, at time t, the n-gram entropy rate is

$$h_n := \mathbf{H}[X_t \mid X_{t-n+1}, \ldots, X_{t-1}] \tag{2}$$

where $\mathbf{H}[X \mid Y]$ is the entropy of X conditional on Y: $\mathbf{H}[X \mid Y] = \mathbb{E}[-\log p(x \mid y)]$. If $\{X_i\}_{i=1}^n$ is a stationary stochastic process, h_n will converge asymptotically to the **entropy rate** h. Predictive information can be calculated as the rate of convergence to the entropy rate (Crutchfield & Feldman, 2003):

$$E = \sum_{n=1}^{\infty} (h_n - h). \tag{3}$$

Intuitively, this means that if X_t can be predicted with a small context window, the process has *low* predictive information, since h_n converges to h quickly, but if X_t resolves long-distance dependencies, a small context window will not be informative and thus $h_n - h$ will be large for small n; in other words, the predictive information will be high.[1]

[1]Note crucially that h_n is monotonically decreasing in n. This reflects the intuitive fact that, with more context, the next symbol is made easier to predict.

Discounting

To explore the relative importance of near vs. distant context, we apply **discounting** to predictive information. For a discount factor $\gamma \in [0,1]$, we define the γ-discounted predictive information as

$$E_\gamma = \sum_{n=1}^{\infty} \gamma^{n-1} (h_n - h), \tag{4}$$

which recovers pure predictive information E when the discount factor $\gamma = 1$. While predictive information has been applied to analyze natural language in previous work (Li, 1989; Ebeling & Pöschel, 1994; Debowski, 2011; Futrell & Hahn, 2024), this notion of discounted predictive information is novel to our knowledge.

The discount factor γ determines how much weight is given to near versus distant contexts. When γ is small, only the most local context matters, and when it is large, more and more context is taken into account. We may also take discounted predictive information as a cognitive cost function, reflecting the idea that memory for distant contexts is simply lost or becomes unavailable. In that case, this lost information does not contribute to memory cost, although it presumably manifests as decreased predictability for symbols.

Measuring Predictive Information

To calculate predictive information for a given text training set, we first fit n-gram models for $n = 1, \ldots, N$ for some large N. The n-gram probability of producing a symbol x_t given context $x_1, \ldots, x_{t-1}$ is given by

$$p(x_t \mid x_1, \ldots, x_{t-1}) = p_n(x_t \mid x_{t-n+1}, \ldots, x_{t-1}), \tag{5}$$

where $p_n(\cdot)$ is the distribution from the n-gram model. These n-gram probabilities are then used to calculate n-gram entropy rates and then (discounted) predictive information using the formulas above. We estimate the n-gram probabilities using modifier Kneser-Ney smoothing (Heafield, Pouzyrevsky, Clark, & Koehn, 2013).

Study 1: Frequency Effects in Ordering

One key soft production preference is a bias towards producing more frequently occurring units earlier in a sentence when a construction allows for flexible ordering, as they are more 'accessible' to speakers (Jescheniak & Levelt, 1994; Bock, 1982; Koranda et al., 2021). For example:

(2) a. I gave the girl the aardvark.
b. I gave the aardvark to the girl.

A speaker would more likely produce (2a), since 'girl' is a more frequent lexeme than 'aardvark.'

We show that policies which have this bias result in lower predictive information. The core intuition is that, in producing more a frequent constituent first, the second constituent is made more predictable with more local information.

We do so by first fitting a predictive information model to a toy corpus with no frequency bias, and then computing the

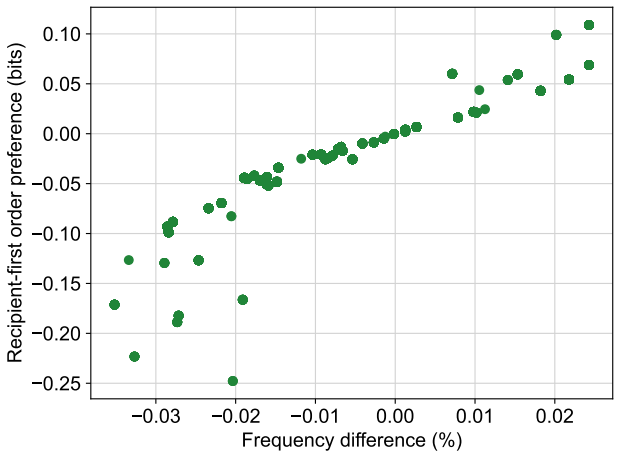


Figure 1: Preference for recipient-first order (predictive information difference, discount $\gamma = .5$) as a function of normalized frequency difference between recipient and theme. Recipient-first order is preferred when the recipient is more frequent.

predictive information of frequent-first vs. infrequent-first constructions. This allows us to model the extent to which predictive information is minimized even under a speaker policy that has only ever been exposed to a balanced corpus.

Simulation Setup

The corpus consists of multiple sets of 2 monotransitive and 2 ditransitive sentences, with one possible verb for each sentence type ('kept' or 'gave'). To construct a set of sentences, we draw two nouns and generate both possible ditransitives and monotransitives:

```
I gave the noun1 the noun2
I gave the noun2 the noun1
I kept the noun1
I kept the noun2.
```

We generate a corpus with no explicit frequency effects, where each noun appears equally many times in both the first and second argument of the distransitive verb. We do so by randomly assigning each noun a frequency weight, and then using a linear program to produce pairs of sentences satisfying these weight constraints. The resulting corpus consists of 11200 sentences (5600 ditransitives) based on a vocabulary of 50 nouns. We show results with discount $\gamma = 0.5$.

After we fit n-gram models to the training corpus, we measure frequency bias by calculating the predictive information over each ditransitive sentence. These quantities give us a sense of the memory usage for generating such sentences, for a speaker who has only ever been exposed to a balanced corpus. We compute the quantity **recipient-first order preference** as E_γ(theme-before-recipient) − E_γ(recipient-before-theme), which is positive when the recipient-before-theme order has lower predictive information and is thus preferred. We compare this to the difference in (normalized) frequency between the recipient and theme.

Results

Results are shown in Figure 1. We find that when the recipient is more frequent than the theme, the recipient-first order has lower predictive information, and the opposite when the theme is more frequent. In other words, production policies with a frequent-first bias reduce predictive information.

Study 2: Modeling Heavy-NP Shift

Heavy-NP shift is a phenomenon in which speakers delay long ('heavy') constituents in constructions that allow for flexible constituent ordering. For example, (1a) above is an example of heavy-NP shift: the short constituent 'my friend' is produced before the long constituent 'a beautiful red Italian car,' even though the opposite order (theme before recipient) is the canonical baseline (Stallings et al., 1998; Stallings & MacDonald, 2011). In contrast, in head-final languages like Japanese, heavy-NP shift acts in the opposite direction: longer constituents shift to be produced *before* shorter ones (Yamashita & Chang, 2001).

Here we show that heavy-NP shift decreases predictive information, in both head-initial languages (with short-before-long shift) and head-final languages (with long-before-short shift). The intuition, in line with previous work, is that heavy-NP shift decreases the dependency length between the verb and the head of each of constituent (Temperley & Gildea, 2018; Futrell, Levy, & Gibson, 2020).

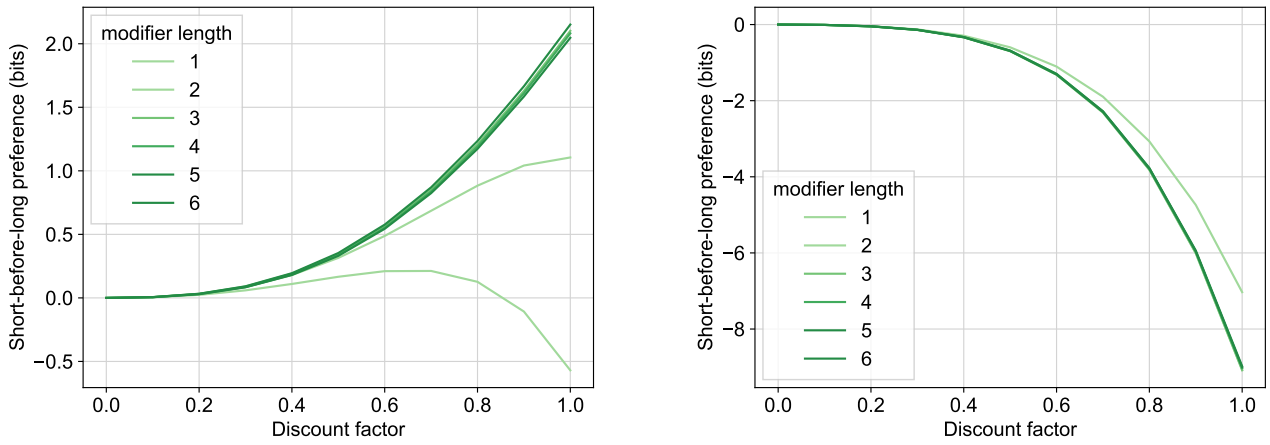
Simulation Setup

Similarly to Study 1, we compute predictive information from toy corpora which do not show any length-based bias. We construct such corpora for two types of toy languages: one with head-initial syntax and one with head-final syntax.

Each corpus consists of multiple sets of 4 monotransitive and 8 ditransitive sentences, again with one possible verb for each sentence type. To construct a set of sentences, we uniformly draw a theme noun and modifier and a recipient noun and modifier. We generate all 8 possible ditransitives, expressing all permutations of order (theme / recipient first) and length (i.e. including / not including the modifier), as well as all 4 possible monotransitives (i.e. with each noun as the object, with or without the modifier):

```
I gave the noun1 the noun2
I gave the noun2 the noun1
I gave the adj1 noun1 the noun2
I gave the noun2 the adj1 noun1
...
I kept the noun1
I kept the adj1 noun1
...
```

We experiment with a variety of modifier lengths in this simulated corpus, ranging from 1 to 6, generating a new cor-



(a) head-initial (English-like) toy language

(b) head-final (Japanese-like) toy language

Figure 2: Difference in predictive information for long-before-short minus short-before-long orders, as a function of the discount factor. A positive value indicates that the short-before-long order is preferred, that is, it has lower predictive information than the long-before-short order. In the head-initial toy language, we find a consistent preference for the short-before-long order except for very short modifiers at discount factor near 1; in the head-final toy language, we find a consistent preference for the long-before-short order.

pus for each of these lengths. Each dataset consisted of 6000 total sentences. For each dataset, we fit several models, varying the discount factor γ.

We measure heavy-NP shift preference again by computing predictive information, here comparing sets of sentences instantiating a long-before-short or short-before-long order. Since we are interested in the extent to which short-before-long constructions create less cost, we compute the quantity **short-before-long preference** as $E_\gamma(\text{long-before-short}) - E_\gamma(\text{short-before-long})$, which is positive when the short-before-long order has lower predictive information (that is, consistent with heavy-NP shift in head-initial languages).

Results

Results are in Figure 2. We see that in the head-initial toy corpora, short-before-long constructions have lower predictive information than long-before-short constructions, and that the opposite is true in the head-final toy corpora. In other words, heavy-NP shift indeed leads to production policies with lower predictive information.

Study 3: Agreement Attraction

Agreement attraction (Bock & Miller, 1991) is a type of production error in which a speaker produces a verb that agrees with a noun (the 'attractor') other than its subject:

(3) The key to the cabinets *are* missing.

That agreement attraction might be the result of memory constraints is intuitive: if a speaker upweights local information, they will be susceptible to making errors based on this information, rather than using distant information.

A key feature of agreement attraction is that plural nouns tend to be stronger attractors than singular ones (Bock & Cutting, 1992). Most previous memory-based approaches require this bias to be built into the model in some capacity (Badecker & Kuminiak, 2007; Wagers, Lau, & Phillips, 2009; Keshev, Cartner, Meltzer-Asscher, & Dillon, 2024). However, we hypothesize that this is due to a simple distributional difference: nouns are plural less frequently than singular.

We show that sentences exhibiting agreement attraction errors have lower predictive information. We also show that attraction to less frequent noun forms reduces predictive information even more than attraction to frequent ones.

Simulation Setup

We construct a toy corpus consisting of sets of 2 simple and 4 'complex' embedded RC sentences. Each set consists of a subject, an attractor, and a preposition. We generate all four possible embedded RC constructions following the structure of Example 2, with nouns marked for number using a postpositional token (SG or PL).

```
the noun1 S prep the noun2 S is adj
the noun1 S prep the noun2 P is adj
the noun1 P prep the noun2 S are adj
the noun1 P prep the noun2 P are adj
the noun1 S is adj
the noun1 P are adj
```

We 'padded' these sets with multiple copies of sentences using only singular nouns, such that the final distribution of number was 3:1 singular to plural.[2] Each sentence had a 10%

[2]This approximately matches the distribution of singular and plu-

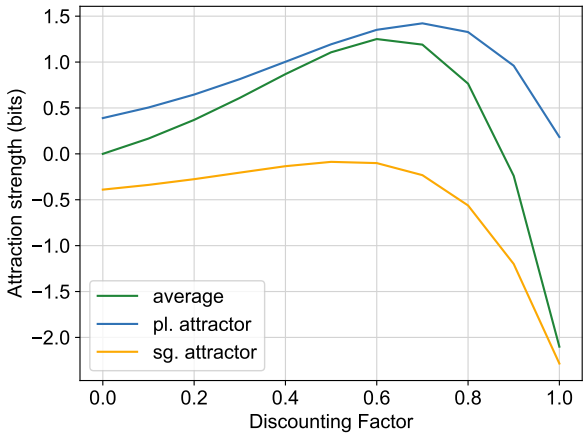


Figure 3: Difference in predictive information for unattracted minus attracted sentences, as a function of the discount factor γ. A positive value indicates that the attracted sentence has lower predictive information than its unattracted (grammatical) version. At lower discounting factors, attraction reduces predictive information, while for gamma near 1, attraction increases predictive information. Plural nouns (less frequent in the training dataset) are stronger attractors than singular ones.

probability of exhibiting an attraction error.[3] The training set comprised of 37500 total sentences.

We are interested in the predictive information of sentences exhibiting attraction errors, compared to their grammatical counterparts. We thus measure the **attraction strength** $E_\gamma(\text{unattracted}) - E_\gamma(\text{attracted})$, which will be high if the attraction error does indeed decrease predictive information.

Results

Results are in Figure 3. We observe a clear attraction effect at low discounting factors, when distant contexts are downweighted. For high discounting factors, as the discounting factor approaches 1 (i.e. as E_γ utilizes more context), attraction errors begin to *increase* predictive information. We also note that the attraction strength of plural attractors (which were uncommon in the training dataset) is significantly greater than that of singular attractors. In this sense, predictive information recovers the number asymmetry with no additional machinery.

Study 4: Thematic Role Entropy

The framework of predictive information minimization can also inform novel predictions about soft production preferences. We show that if two constituents have highly certain thematic roles (i.e. they are highly 'theme-like' or 'recipient-like'), then their length plays a smaller role in predicting their order. We verify that this prediction holds in natural language data. Our intuition here is that when one argument has a highly certain thematic role, it is highly predictive about the thematic role of the *next* argument. So the predictive information required to produce the second argument is thus made lower.

Model Prediction

As above, we train a predictive information model on an order-balanced corpus. However, here we use a single corpus with constituents of lengths between 1 and 12, and, rather than each noun appearing equally frequently as a theme or recipient, we give each noun a probabilistic thematic role bias, which we draw uniformly.

We first formalize the notion of thematic role certainty. We consider the **binary entropy** of the thematic roles of both constituents. Formally, for a sentence with two arguments c_1, c_2, the role entropy is given by

$$\text{Role Entropy} = -\sum_i \bar{p}(c_i = t) \log \bar{p}(c_i = t),$$

where $\bar{p}(c_i = t)$ is the probability that c_i is a theme:

$$\bar{p}(c_i = t) := \frac{p(c_i = t)}{p(c_1 = t) + p(c_2 = t)}. \quad (6)$$

We then fit a logistic regression in the simulated order-balanced data to predict preferred order (theme- or recipient-first) as defined by minimizing predictive information, with role entropy, difference in length (recipient length minus theme length in words), and their interaction as predictors. We find that role entropy does lead to attenuated heavy-NP shift, as shown by a negative interaction in Table 1 and visualized in Figure 4.

Corpus Preparation

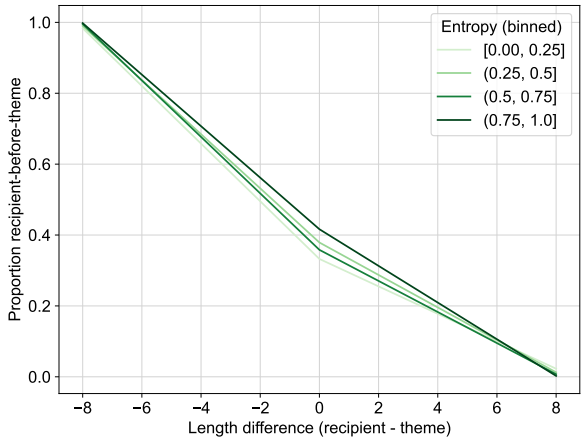
To verify this effect in natural language data, we use Bresnan, Cueni, Nikitina, and Baayen's (2007) parsed set of 2360 datives from the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992). To compute role entropy, we extract probabilities from Llama-3.1-8B, an open-source large language model (Dubey et al., 2024). For a given argument (in context) c, we measure $p(\text{to} \mid c)$. If this probability is high, it means that the constituent is highly likely to be a *theme*, since the language model expects a recipient to follow. In other words, this approximates $p(c = t)$ above.

Similarly to the simulation, we fit a logistic regression to predict order from binary thematic role entropy and length difference. Here, we also include control variables for animacy and definiteness, which were hand-annotated by Bresnan et al. (2007), and probability in context, which was computed by Futrell (2023).[4] We also include random intercepts for speaker ID and corpus.

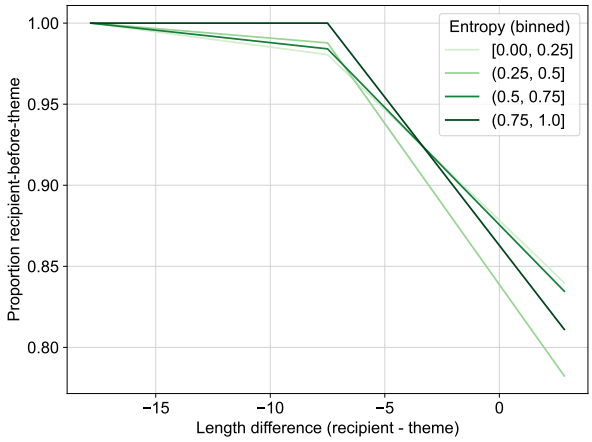
ral nouns in English data, which we measured using the parsed English Web Treebank (Bies, Mott, Warner, & Kulick, 2012).

[3] Without this error probability, the predictive information model would be infinite for sentences with attraction errors, as the attracted verb would have probability 0.

[4] The probability in context was computed by using GPT-3 (`text-davinci-001`) (Brown et al., 2020).



(a) Proportion of constructions where predictive information favors recipient-first order, in simulated dataset.



(b) Proportion of recipient-first constructions from Switchboard.

Figure 4: Preference for recipient-first form in the dative alternation, as predicted by a logistic regression as an effect of length difference and role entropy. In all plots, as length difference increases, the proportion of recipient-first decreases, and the slope of each line gets approximately more negative when entropy is high.

	Coefficient	p Value	**Coefficient**	p Value
Intercept	-0.92	***	1.90	***
Role Entropy	0.51	***	-1.00	*
Length Difference	-0.144	***	-1.93	***
Role Entropy $\times$ Length Difference	-1.99	***	-1.886	**

Table 1: Logistic regression fit to simulated (left) and natural (right) corpus. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. The natural corpus regression included controls for animacy, definiteness, and probability in-context; these all came out significant ($p < 0.001$) with coefficients with the expected sign (theme negative, recipient positive). Crucially, both regressions show a significant negative interaction between entropy and length difference, indicating that greater certainty dampens heavy-NP shift.

Results

Regression coefficients for entropy and length difference from both corpora are in Table 1. Figure 4 plots the proportion of theme- and recipient-first orders against length difference at different role entropy levels. In both settings, we find a significant negative interaction between role entropy and length difference. This indicates that, as predicted, increased certainty about thematic role (lower entropy) has a dampening effect on length difference.

Conclusion

Prior work on soft production preferences do not typically reach a concrete formalization of the underlying *causes* of these effects, and instead model them with empirically identified features such as animacy, frequency, and length. We have proposed a general explanatory theory for some of these effects based on the predictive information bottleneck.

Our model of production preferences is highly general in its form as a cost function, and as such, it can be integrated into a variety of computational-level theories of production. For example, the Rational Speech Acts (RSA) model of interactive language use describes a speaker's production policy as a distribution subject to an unspecified cost function (Frank & Goodman, 2012; Cohn-Gordon et al., 2019; Degen, 2023). Similarly, Futrell's (2023) Rate-Distortion theory of Control (RDC) model operationalizes cost in terms of deviation from an 'automatic policy,' which may be constrained by predictive information. Our theory also builds on a growing body of work that models language use as process of using finite, lossy memory resources to make predictions about future material (Futrell, 2019; Futrell, Gibson, & Levy, 2020; Hahn et al., 2022). It remains to be seen the extent to which predictive information gives the whole picture of soft production preferences, or whether additional machinery is required.

Acknowledgements

We thank for discussion Judith Degen, Brandon Waldon, Emily Goodwin, Aryaman Arora, Michael Hahn, Roger Levy, the Stanford ALPS Lab and NLP Group, and audiences at CogSci 2024, CAMP, and BlackboxNLP 2023.

References

Arnold, J. E., Kaiser, E., Kahn, J. M., & Kim, L. K. (2013). Information structure: Linguistic, cognitive, and processing approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 403–413.

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*(1), 28–55.

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of memory and language*, *56*(1), 65–85.

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, *13*(11), 2409–2463.

Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). *English Web Treebank* (Tech. Rep. No. LDC2012T13). Philadelphia, PA: Linguistic Data Consortium.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1.

Bock, J. K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of memory and language*, *31*(1), 99–127.

Bock, J. K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In *Cognitive Foundations of Interpration* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J., Dingare, S., & Manning, C. D. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 01 Conference* (pp. 13–32). CSLI Publications.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue & Discourse*, *2*(1), 11–33.

Chang, F. (2009). Learning to order words: A connectionist model of Heavy NP Shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, *61*, 374–397.

Cohn-Gordon, R., Goodman, N., & Potts, C. (2019). An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019* (pp. 81–90). Retrieved from `https://aclanthology.org/W19-0109` doi: 10.7275/cprc-8x17

Crutchfield, J. P., & Feldman, D. P. (2003). Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *13*(1), 25–54.

Debowski, Ł. (2011). Excess entropy in natural language: Present state and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*(3), 037105.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... others (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ebeling, W., & Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters*, *26*(4), 241.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340.

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

Futrell, R. (2019). Information-theoretic locality properties of natural language. In X. Chen & R. Ferrer-i-Cancho (Eds.), *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)* (pp. 2–15). Paris, France: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W19-7902`

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

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.

Futrell, R., & Hahn, M. (2022). Information theory as a bridge between language function and language form. *Frontiers in Communication*, *7*, 657725.

Futrell, R., & Hahn, M. (2024). Linguistic structure from a bottleneck on sequential information processing. *arXiv preprint arXiv:2405.12109*.

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–413.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)* (Vol. 1, pp. 517–520).

Goldman-Eisler, F. (1957). Speech production and language statistics. *Nature*, *180*(4600), 1497–1497.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*(1), B1–14.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient

tradeoff of memory and surprisal. *Psychological Review*, *128*(4), 726-756.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119.

Hahn, M., Mathew, R., & Degen, J. (2021). Morpheme ordering across languages reflects optimization for processing efficiency. *Open Mind*, *5*, 208–232.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Keshev, M., Cartner, M., Meltzer-Asscher, A., & Dillon, B. (2024). A working memory model of sentence processing as binding morphemes to syntactic positions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Koranda, M. J., Zettersten, M., & MacDonald, M. C. (2021). Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science*, 09567976221089603.

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022, December). Context limitations make neural language models more human-like. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10421–10436). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.712/ doi: 10.18653/v1/2022.emnlp-main.712

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of learning and motivation* (Vol. 74, pp. 195–232). Elsevier.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Li, W. (1989). *Mutual information functions of natural language texts* (Tech. Rep.). Santa Fe Institute Working Paper #1989-10-008.

Palmer, S. E., Marre, O., Berry, M. J., & Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, *112*(22), 6908–6913.

Rathi, N., Hahn, M., & Futrell, R. (2022). Explaining patterns of fusion in morphological paradigms using the memory-surprisal tradeoff. In J. Culbertson, H. Rabagliati, V. C. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society, CogSci 2022*. Toronto. Retrieved from https://escholarship.org/uc/item/0v03z6xb

Rathi, N., Waldon, B., & Degen, J. (2024). Informativity and accessibility in incremental production of the dative alternation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1503.

Stallings, L. M., & MacDonald, M. C. (2011). It's not just the "heavy NP"': Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research*, *40*, 177–187.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, *39*(3), 392–417.

Temperley, D., & Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, *4*, 1–15.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*(2), B45–B55.