An information-theoretic account of accessibility effects in language production

The rich literature applying information theory to questions of linguistic processing efficiency has been based primarily on models of comprehension, with less attention paid to the other side of the processing equation: online language production, which is arguably more complex than language comprehension because it involves real-time integration of processes of self-monitoring, planning, grammatical and phonological encoding, and social reasoning, and thus is likely a source of processing bottlenecks that affect the evolution of languages. Here I discuss communicative efficiency within a model of incremental language production based on a combination of information theory and control theory (Todorov, 2009), in which speakers incrementally maximize future-discounted communicative reward while minimizing demands on cognitive control. I show how the model explains effects of **accessibility** in word order in usage preferences (Bock, 1982) and grammar (Bresnan et al., 2001): the well-documented tendency to place words earlier when they are more frequent, discourse-given, animate, definite, etc. The model clarifies what is meant by accessibility, while making successful novel predictions tested in the domain of the dative alternation.

The production model specifies a **policy**: a probability distribution $p(x|g,s)$ on what word x a speaker will produce next given (1) their communicative goal, called g, and (2) what they have produced so far, called the state s. The policy is selected to maximize the average **value** of words produced, where value V of word x for goal g in state s is defined as

(1) $V = \alpha R - C + \gamma \langle V' \rangle$,

where R is the communicative reward associated with a word in context, C is information-processing cost, and $\langle V' \rangle$ is the average value of future words to be produced after the current (terms defined in Figure 1). Scalar parameter α can be interpreted as the channel capacity of cognitive control. Scalar parameter $\gamma < 1$ is the **future-discount** parameter which discounts future value relative to immediate value. Future discounting in this manner is standard in value functions used in economics, robotics, and reinforcement learning (Sutton & Barto, 2018). Future discounting is what drives accessibility effects: because future value is discounted, it is often better produce 'easier' words earlier.

I present two sets of results using the model. In Figure 2, in the domain of the dative alternation, I show that the model predicts a novel and confirmed effect of planning: a phrase (theme or recipient) is more likely to go earlier when it is predictable *given the likely following phrase*. In Figure 3, in the domain of choice of noun classifiers in Mandarin, I show that the model predicts Zhan & Levy's (2018, 2019) results that speakers prefer the generic classifier *ge* over more specific alternatives before low-frequency nouns, a prediction which is contrary to Uniform Information Density but in keeping with accessibility-based production based on the idea that low-frequency nouns (and their corresponding specific classifiers) are hard to access.

Communicative reward: $R(x \mid g,\ s) \ = ln\, p_L(g|x, s)/p_L(g|s)$

Information-processing cost: $C(x \mid g, s) \ = \ ln\, p(x \mid g, s)/p(x \mid s)$

Average future value: $< V' > \ = \ \sum_{x'} p(x' \mid g, s, x)\, V(x' \mid g, s, x)$

Figure 1. Terms used in the value function (1), used to define the incremental language production policy *p*(x | *g*, *s*). The distribution p_L is a simulated listener.

$$p(\text{recipient}) = \sigma\left(\underbrace{\ln\frac{p(\text{recipient}\mid\text{context})}{p(\text{theme}\mid\text{context})}}_{\text{Predictability}} + \gamma\underbrace{\ln\frac{p(\text{theme}\mid\text{context},\text{recipient})}{p(\text{recipient}\mid\text{context},\text{theme},\text{“to”})}}_{\text{Planning}}\right)$$

Predictor	Coefficient	95% Posterior CrI
(Intercept)	-0.68	$[-1.64, 0.23]$
Verb Semantics	-0.36	$[-1.10, 0.38]$
Length	-0.14	$[-0.25, -0.05]$
Definiteness	0.89	$[0.31, 1.61]$
Animacy	1.24	$[0.35, 2.24]$
Predictability	$\mathbf{0.95}$	$\mathbf{[0.75, 1.25]}$
Planning	$\mathbf{0.87}$	$\mathbf{[0.68, 1.15]}$

Figure 2. Dative alternation predictions and results. **Left**, the production model's probability to produce a double-object construction under the assumption that both the DO and PO constructions have equal communicative reward. σ() is the logistic function. "theme" and "recipient" are the words of the theme and recipient phrases; "context" is preceding context. **Right**, results of a logistic regression predicting the English dative alternation in the languageR dataset using the Predictability and Planning terms from the model with standard controls, and using GPT-3 text-davinci-001 as the distribution *p*(*x* | *s*).

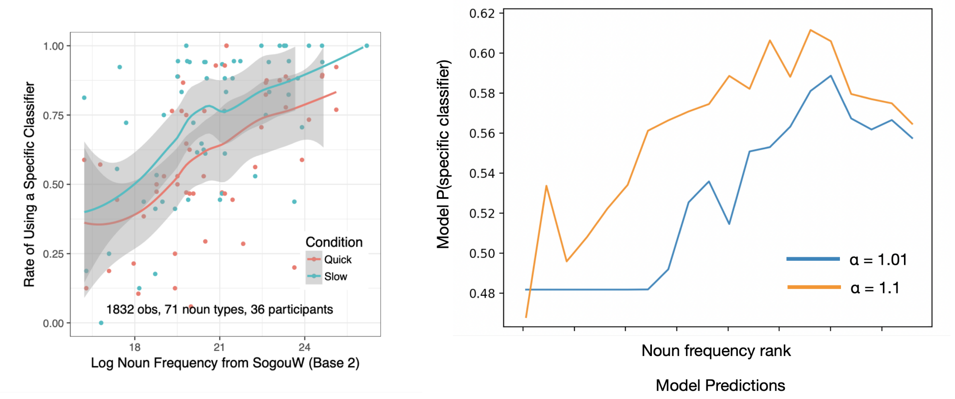

Figure 3. Accessibility effects in Mandarin classifier choice. **Left**, the probability to use a specific classifier rather than the generic classifier *ge* before nouns in a picture naming task, as a function of noun frequency and time pressure. **Right**, production model predictions based on a toy language with 20 nouns, 2 specific classifiers randomly assigned to these nouns, and 1 generic classifier; nouns follow a Zipfian probability distribution. In both cases, low-frequency nouns are more often paired with the generic rather than specific classifiers.