Incremental Generation Drives "Efficient" Language Production

Spencer Caplan (University of Pennsylvania)
spcaplan@sas.upenn.edu

A major testing ground for mechanistic accounts of language production (De Smedt, 1990) is the study of "syntactic optionality"; i.e. given multiple potential syntactic encodings for equivalent semantic sentences, what factors govern the use of one form rather than another (Ferreira and Dell, 2000).

A prominent previous account, the "Uniform Information Density" hypothesis (UID) (Jaeger, 2010), proposes that such syntactic optionality is driven by a speaker's implicit managing of computable information content to maximize communicative efficiency. On this model, conditional probability serves as a proxy for optimized information content. Previous corpus modeling of optional 'that'-omission has supported UID (Jaeger, 2010). However, such evidence is potentially problematic. In addition to syntactic confounds (Grimshaw 2009), rates of 'that'-omission show a great deal of variability by genre, ranging from 1% in formal writing to 85% in conversational speech (Biber 1999). With so much variance attributable to sociolinguistic register, it is not clear what we might learn about the cognitive architecture of the production system.

We propose the English verb-particle (VP) construction (e.g. 'John picked up the book' vs. 'John picked the book up') as a better case to evaluate theories of optionality. We extracted a large database of VP alternations from COCA, a balanced corpus of modern English (Davies, 2009). From these we can compare a UID account with the general framework of incremental generation (IG) (Bock and Levelt, 2002). Under IG, the architecture of sentence generation requires several components: retrieve lemmas from memory, assign such elements their proper functional roles, as well as assign linear order in adherence with syntactic restrictions. If we assume such modules operate incrementally and in parallel, then variations in the order in which information is delivered from one component to the next can readily affect the linear order elements appear in speech. So long as the system does not intentionally hold retrieved lemmas back in a buffer, any factors which speed up lexical access will also be proxies for spoken linear order (see Rayner, 1998 for a review of lexical access factors). While both UID and IG make convergent predictions regarding conditional probability, only IG predicts that the factors of frequency, definiteness, and constituent length, etc. should all predict linear order in optional constructions. This is because such factors correlate with lexical retrieval times yet are orthogonal to "information density".

Following Jaeger's original study, a multilevel logit model was used to evaluate predictions of an IG account of sentence production compared with UID. The dependent variable was the binary outcome of linear order (particle-first rather than object-first) in VP sentences. Evaluating over the entire database, we see a strong correlation between all IG related factors and output order (Blue in Table). This includes the effect of conditional probability posited by both UID and IG. However, when limited to evaluation over even moderately long objects (at least four words), then the predictions of UID are not borne out while the other instantiations of IG remain significant (Green in Table). To whatever degree we can characterize the output of the language production system as "efficient" in information ordering, this is an emergent property of a simple, incremental generation system.

	All VPs (58,619 cases)				Longer object VPs (5,679 cases)			
Factor	**Estimate**	**Std. Err**	**Z-Value**	**P**	**Estimate**	**Std. Err**	**Z-Value**	**P**
Freq(obj)	-282.8	38.12	-7.42	~0	-502	77.9	-2.84	0.01
Info(obj|verb)	0.039	0.01	8.6	~0	0.02	0.03	0.69	0.49
NP-length	1.0	0.03	40.37	~0	0.63	0.12	5.18	~0
Definite-Obj	-0.59	0.02	-26.9	~0	-1.13	0.15	-7.33	~0