

Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production

Austin F. Frank (afrank@bcs.rochester.edu)
T. Florian Jaeger (fjaeger@bcs.rochester.edu)
Brain and Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14620

Abstract

We provide evidence for a rational account of language production, Uniform Information Density (UID, Jaeger, 2006; Levy & Jaeger, 2007). Under the assumption that communication can usefully be understood as information transmission over a capacity-limited noisy channel, an optimal strategy in language production is to maintain a uniform rate of information transmission close to the channel capacity. This theory predicts that speakers will make strategic use of the flexibility allowed by their languages. Speakers should plan their utterances so that elements with high information are lengthened, and elements with low information are shortened, making the amount of information transmitted per time more uniform (and hence closer to the optimum). In three corpus studies, we show that American English speakers' use of contractions ("you are" → "you're") follows the predictions of UID. We then explore further implications of UID for production planning. **Keywords:** language production; utterance planning; information theory; morphological reduction; contractions

Introduction

Speakers face a large number of choices in the process of going from thought to utterance. Concepts must be mapped to words, words must be arranged into larger linguistic structures, and the articulatory system must be recruited to turn these mental representations into perceptible actions. Moreover, all of this must be accomplished in such a way that the intended recipient of the message will be able to understand it. Faced with this complex task, it seems that speakers should have developed highly efficient strategies for solving the problem of converting thoughts into a serialized stream of articulatory gestures.

Rational approaches to cognition seek to define the optimal performance that could be achieved for a particular task (Chater, Tenenbaum, & Yuille, 2006). This is an example of what Marr called the computational level of analysis (Marr, 1982). Language production, however, is often studied at the algorithmic level, in terms of a series of hypothesized cognitive processing systems (e.g., Dell, 1986; Levelt, 1989). Availability-Based Production (ABP) is an example of this approach to production theories. (Ferreira & Dell, 2000). Under ABP, speakers' utterance planning strategies are, to a large extent, shaped by limitations on the production system's ability to retrieve and integrate upcoming material in the utterance. Algorithmic level theories like ABP help to clarify how speakers execute an utterance. Computational level theories play a complementary role, attempting to explain why the system uses the processes and mechanisms it does.

This paper explores the issue of a rational language production system. As noted above, a rational system performs optimally with respect to some particular goal. Typically, optimal solutions involve the balancing of multiple constraints. For language production, there are at least two pressures that speakers have to balance to achieve efficient communication. On the one hand, speakers want to successfully convey a message (where by message, we do not mean the literal message, but whatever set of directly or indirectly intended effects the speaker wants to achieve). On the other hand, speakers need to produce language efficiently. Pressure for efficient communication may come from several sources, such as limited attentional or memory resources, or other interlocutors who are competing for the ground (i.e. a speaker may be interrupted if information is conveyed too inefficiently). A rational production system, then, is one which maximizes the likelihood of efficient and successful communication, taking into account the limitations imposed by the speaker, listener, and environment.

Previous work has shown that one strategy for language production, Uniform Information Density (UID, Jaeger, 2006; Levy & Jaeger, 2007; building on Aylett & Turk, 2004; Genzel & Charniak, 2002), has several properties that are consistent with the goal of optimizing successful communication. Language use is seen as transmission of information over a bandwidth-limited noisy channel (Shannon, 1948). A unit's Shannon information is determined by its probability ($I(u) = -\log p(u)$). Under these assumptions, it follows from information theory that speakers would optimize the chance of successfully transmitting their message by transmitting a uniform amount of information per transmission (or per time, assuming continuous transmission) close to the Shannon capacity of the channel. UID holds that speakers are indeed approximating optimal production by aiming to produce utterances with uniform information density (within the bounds defined by grammar). Thus UID predicts that the choices speakers have to make when they encode an intended message into an utterance are at least partially determined by information density: if one way to convey a message leads to more uniform information density than another way to convey the same message, the variant with a more uniform distribution of information should be preferred (Jaeger, 2006; Levy & Jaeger, 2007).

Indeed, speakers' productions have been shown to be consistent with a UID strategy at several levels of linguistic representation. Speakers modulate their speech rate so that words with high information content are spread out over a longer

period of time (Aylett & Turk, 2004; Bell et al., 2003) and they produce highly informative phonemes more slowly and with more articulatory detail (van Son & van Santen, 2005). Speakers are also more likely to produce optional function words (such as optional “that”, which in certain types of English complement or relative clauses can be omitted) when the words following them would otherwise be high in information content, thereby avoiding spikes in the rate of information transmission (Jaeger, 2006; Levy & Jaeger, 2007).

We present three corpus studies on speakers’ choices in the use of morphosyntactic contractions in spontaneous American English speech. Specifically, we investigate speakers’ choices between (a) full and reduced BE (e.g. “I am” vs. “I’m”), (b) full and reduced HAVE (e.g. “you have” vs. “you’ve”), and (c) full and reduced NOT (e.g. “did not” vs. “didn’t”). If UID is a general computational strategy that drives language production at all levels of linguistic representation, speakers should prefer a full form whenever the content conveyed by the form is unexpected in its context. Low probability content is high information content. Using a full form spreads this information over a longer time, thereby avoiding a peak in information density. The evidence we present provides further support that speakers employ a UID strategy. The converging evidence for UID across levels of linguistic representation illustrates the power of computational level theories, but it also raises the question of how a general computational principle like UID relates to the cognitive mechanisms that traditional psycholinguistic (algorithmic) models of language production have identified. We address this question in the final part of this paper.

Methods

Corpus and data set

We use the Paraphrase Stanford-Edinburgh LINK SWITCHBOARD corpus (Bresnan et al., 2002; Calhoun, Nissim, Steedman, & Brenier, 2005). The corpus is part of SWITCHBOARD, and consists of over 830,000 words in 642 telephone dialogues between two speakers each (roughly gender-balanced; age range from 20 to 68, $\bar{x} = 38$) on a variety of topics (selected by participants from a pre-determined list). The corpus combines and aligns numerous annotations for SWITCHBOARD (Calhoun et al., 2005), including annotation of disfluencies (Taylor, Marcus, & Santorini, 2003), part-of-speech, syntactic, and grammatical function (Marcus, Santorini, Marcinkiewicz, & Taylor, 1999), and time-aligned orthographic transcriptions.

These annotations make it possible to define syntactic searches to automatically extract only those instances of BE, HAVE, and NOT that can be morphologically reduced,¹ while at the same time extracting syntactic, lexical, phonological, and phonetic information about these words and the contexts they occur in (including speech rate information). We used TGrep2 (Rohde, 2005) to extract all morphologically reducible cases of BE, HAVE, and NOT from the corpus.

¹We use uppercase to refer to all and only reducible forms of a particular lemma. For BE, this includes “am”, “are”, “is”, “m”, “re”, and “s”. For HAVE, “have”, “has”, “had”, “ve”, “s”, and “d”. For NOT, “not” and “n’t”.

Exclusion criteria Not all extracted cases were used in the analysis, for reasons described next. A summary of the amount and composition of the excluded material is presented in Table 1. The first section of the table shows the size of the total data unprocessed data set, and the proportion of utterances that used the full form. The second section of the table shows four different criteria for excluding data. For each source of exclusions, we report the number of cases excluded from the original data set, and the percentage of the excluded cases that occurred in the full form.

Since UID makes predictions about speakers’ choices, we are only interested in cases where speakers actually have a choice between two different realizations of the target lemmas BE, HAVE, and NOT. For example, NOT cannot be reduced if it follows a reduced auxiliary (*“she’sn’t”) or if there is no reduced form (“you aren’t”, but *“I amn’t”). Only auxiliary HAVE, but not possessive or modal *have* is morphologically reducible (*“I’ve a car” and *“I’ve to leave now”). These are all examples of cases where the variation of interest is not available to speakers. We also excluded cases of BE and HAVE that preceded NOT, because in those cases speakers often have two choices (e.g. “we are not/we’re not/we aren’t”). We leave these cases for future study.

Possible inaccuracies in estimating speaking rate led to further exclusions. Short utterances or target words appearing in certain positions of a prosodic phrase can interfere with our estimate of speaking rate. To have reasonably consistent speaking rate estimates, we require that the target element occur in an utterance that has at least five times as many syllables as the element itself, that it not occur in the two initial or final syllables of the utterance, and that the target element does not occur immediately after or before a prosodic break (operationalized as any pause of at least 500 ms). We also exclude cases where the target element was the first or last word in the utterance.

The presence of a filled pause (e.g. “um”, “uh”, “you know”) immediately before a reducible element strongly favors a full form (“It uh is”, but \int “It uh ’s”). We exclude cases immediately following a disfluency, on the grounds that the target could not actually have been contracted in that situation. After excluding all cases with preceding disfluencies, we were left with relatively few cases of following disfluencies. Rather than modeling the effect of disfluent contexts on contraction we therefore decided to limit the current study to fluent contexts.

Finally, when using a corpus to calculate estimates of the probability of an event, the amount of data available can be a limitation. In this case, the events of interest are collocations—sets of neighboring words. Collocation information allows us to estimate the probability of a word appearing in a particular context. When a particular group of words is observed only once in the corpus, it is hard to know whether that count accurately reflects its true rate of occurrence in the language. One convention for dealing with this issue in natural language processing applications is to exclude low-count data. We exclude all utterances where the relevant two-word collocations (*i.e.* bigrams) occur fewer than five times (a standard cut-off, Jurafsky & Martin, 2007, Ch. 4).

Table 1: Exclusion criteria and remaining data

	BE		HAVE		NOT	
Total	20,113	36.70%	4,607	43.54%	10,399	26.92%
Criteria	Excl.	% Full	Excl.	% Full	Excl.	% Full
Irreducible context	17	100%	37	100%	26	100%
More than two options	1494	35.21%	465	97.53%	2458	66.48%
Bad rates	5,026	29.94%	951	35.12%	2,501	24.23%
Disfluency before target	531	100%	10	90%	140	100%
Disfluency after target	1169	49.53%	142	50.70%	274	40.51
Low counts	3,981	66.84%	820	72.32%	1,457	51.41%
Used for analyses	9,379	30.92%	2,411	32.39%	5,034	9.95%

Models

We attempt to determine the extent to which speakers' utterances are consistent with a UID strategy by building statistical models of the influences on speaker choice. Our dependent variable is a binary distinction between the use of a morphologically full or reduced form. Separate models were developed for the three target lemmas BE, HAVE, and NOT. In addition to measures of the information content of the reducible target lemmas, we include several controls in the models that are expected to affect speakers' choices.

Specifically, we use generalized linear models with a logit linking function (Breslow & Clayton, 1993; for an introduction, see Agresti, 2002, Ch. 12) to predict changes in the log-odds of contraction use. Such mixed logit models can be understood as extensions of ordinary logistic regression models that (among other things) allow us to account for random effects due to speakers. Here each speaker is modeled as having a different base rate of contraction use, and the influence of our independent variables is computed relative to each speakers' individual tendencies.² Confidence intervals are calculated for each factor in the model. When the coefficient estimate corresponding to a factor significantly differs from zero, we conclude that that factor influences speakers' decisions about contraction use. While we test the contribution of several factors in each model, no corrections for multiple hypothesis testing are performed. Mixed models perform partial pooling of the variance within each level, resulting in conservative claims about the contribution of each predictor.

For unbalanced corpus data like ours, it is especially important to guard against spurious results due to collinearity. To minimize the impact of collinearity, all factors were centered before being entered into the analysis. For some highly-correlated factors, collinearity remained even after taking these steps. For pairs of highly correlated factors ($r > .5$), we regressed one factor against the other in a linear model and used the residuals from the fit as continuous independent variables in the mixed effects model. These techniques resulted in lower mean and maximum pairwise correlations among predictors in each dataset (maximum pairwise correlations after residualization: BE: 0.43; HAVE: 0.44; NOT: 0.69).

²The analyses presented here were also run using bootstrapping with random cluster replacement to account for random effects of speakers. In this framework, additional significance testing was performed using model comparison (χ^2 tests on changes in log likelihood). The pattern of results was the same.

Factors

Gender We include speaker sex as a factor in the analysis, as male speech has been shown to exhibit a higher degree of reduction than female speech (for phonetic reduction, see Bell et al., 2003; for syntactic reduction, see Jaeger, 2006).

Speech rate Unsurprisingly, speakers are generally more likely to reduce forms during fast speech, although –to the best of our knowledge– this has not been directly tested for morphological contraction (for phonetic reduction, see Bell et al., 2003; for syntactic reduction, see Jaeger, 2006). We expect that faster speech rates will correlate with more contractions. Speech rate was calculated using the time alignment provided by the automatic segmentation process (Calhoun et al., 2005). Each utterance is divided up into a series of “speech windows” characterized by breaths and pauses. The duration of these windows is measured in seconds. The automatic segmentation record is used to provide a count of the number of syllables within a speech window. Speech rate is the number of syllables per second in the speech window containing the reducible element. In an analyses below, we use log-transformed speech rates, which are more normally distributed (Bell et al., 2003). We used the exclusion criteria mentioned above in an attempt to avoid artifacts that could be introduced by phrase boundaries and small sample sizes.

Global naturalness and complexity Each dialogue in the corpus contains information provided by the transcribers to characterize the transcription process. We include the transcribers' ratings of transcription difficulty and conversation naturalness as two measures of the overall perceived complexity and fluency of the conversation.

Phrase length Speakers tend to speed up as the length of their utterances increases. In general, the speech rate at the beginning of an utterance is slower than that at the end. We include two factors to account for these length effects. We measure the length in words of the phrase governing the word before the reducible element, and the length in words of the phrase governing the reducible element.

Frequency Word frequency is known to affect production (Griffin & Bock, 1998). In particular, low-frequency words are harder for speakers to access than high-frequency words. Availability-Based Production (Ferreira & Dell, 2000), mentioned above, maintains that speakers structure their utterances in ways that buy them time to prepare difficult words

and phrases. By this account, the frequency of a word should predict contraction use: speakers should use full forms more before low-frequency words than high-frequency words, to provide themselves more time for lexical access. We include the probability (*i.e.* normalized frequency) of the word preceding the target (the host that the contracted forms encliticize to) and the word following the target in the regression analysis.

Information The information conveyed by a word is closely related to how predictable that word is in its context. Words that are perfectly predictable given their surroundings don't actually provide any new information; conversely, words that are hard to predict are highly informative. We use Shannon information content as a measure of a word's information. Information, again, is defined in terms of its probability, $I(\textit{word}) = -\log p(\textit{word})$ (Shannon, 1948). The probability of a word is approximated as its conditional probability given its immediate context. For each item in our data set, we estimate the conditional probability of a reducible element occurring given the context of its neighboring words. For a string of words *before* *host* *target* *after* (where *target* is the reducible element) we calculate $p(\textit{target}|\textit{before},\textit{host})$, $p(\textit{target}|\textit{after})$, and $p(\textit{after}|\textit{host},\textit{target})$.³ Probability estimates are computed directly from the corpus, and no back-off or smoothing is employed. UID predicts that speakers will tend to use full forms (rather than contractions) when the reducible element is high in information. Uttering a full form takes longer, and so prevents informative elements from causing a non-optimal spike in the rate of information transmission (*cf.* Levy & Jaeger, 2007).

Results

In the following discussion we report the magnitude and significance level of the slope parameters for a series of mixed logit effects models. In this analysis, parameters are measured in units of change in log odds. Positive values for parameters correspond to a change in favor of using a full form, while negative values correspond to a change in favor of using a contraction. Unless otherwise noted, all significant regression parameters are significant at a level of $p < 0.0001$.

BE

Study 1 models reducible BE based on 9,379 instances (31% full forms) from 355 different speakers. The slope parameters for several factors fail to reach significance. Speaker sex, transcription difficulty, speech rate, length of parent phrase are all insignificant ($p > .10$). Higher transcriber ratings for conversation naturalness correlate with a decrease in the log odds of full BE ($\beta = -0.14, p < 0.005$). Speakers' use of full BE correlates with the length of the phrase preceding BE (henceforth host phrase) ($\beta = 0.36$).

Higher log probability of the host correlates with a decrease in full forms ($\log P(h)$ in Figure 1: $\beta = -0.99$), while higher log probability of the word following BE correlates with a small *increase* ($\log P(a) : \beta = 0.09$).

³We use an abbreviated form of this notation when referring to these words and quantities throughout the subsequent analysis.

Finally, all measures of the information load on BE and on the word following BE are significantly correlated with an increase in full forms. When more information is conveyed by BE, speakers are more likely to use a full form, ($I(t|b,h) : \beta = 0.38; I(t|a) : \beta = 0.37$). Additionally, the more full forms are used, the higher the information content on the word following BE ($I(a|h,t) : \beta = 0.16$).

HAVE

Study 2 models reducible HAVE based on 2,411 instances (32% full forms) from 322 different speakers. Speaker sex, conversation naturalness, speech rate, and length of parent phrase all fail to reach significance ($p > 0.10$). Transcription difficulty is marginally significant, with greater difficulty correlating with an increase in full HAVE ($\beta = 0.12, p < 0.10$). Longer host phrases correlate with an increase in full HAVE ($\beta = 0.69$).

The log probability of the preceding and following word have similar effects on the use of the contracted form of HAVE. Speakers are less likely to use full forms in the context of more frequent words ($\log P(h) : \beta = -0.79; \log P(a) : \beta = -0.19$).

The information content of HAVE and of the word after HAVE also affect contraction use. Speakers are more likely to use the full form, as the information content of HAVE increases ($I(t|b,h) : \beta = 0.21; I(t|a) : \beta = 0.65$). Similarly, speakers are more likely to use a full form, the higher the information content of the word following HAVE ($I(a|h,t) : \beta = 0.14$).

NOT

Study 3 models reducible NOT based on 5,034 instances (10% full forms). Speaker sex, conversation naturalness, and length of the parent phrase fail to reach significance ($p > 0.10$). Conversations rated as being more difficult to transcribe have more full forms ($\beta = 0.21, p < 0.005$). Unlike in the first two studies, speech rate reaches significance, and in the expected direction. Speaker are more less likely to use a full form, the faster they speak ($\beta = -0.89, p < 0.005$).

Higher log probability of the host correlates with lower log-odds of the full forms ($\log P(h) : \beta = -0.17$). In contrast, higher log probability of the following word correlates with higher log-odds of the full form ($\log P(a) : \beta = 0.16, p < 0.0005$), as was the case in Study 1 on BE.

Full forms of NOT were more likely to occur as the information borne by NOT increased. This held true both when information was calculated using the preceding context ($I(t|b,h) : \beta = 0.79$) and when it was calculated using the following context ($I(t|a) : \beta = 1.05$). For NOT, the information carried by the following word actually corresponded to a decreased use in full forms ($I(a|h,t) : \beta = -0.11, p < 0.05$). That is, speakers' use of contractions increased when the word following NOT was more informative.

Discussion

Overall, speakers use contractions in a manner that is predicted by UID. As indicated by the information-related parameters shown in the first three panels of Figure 1, full forms tend to be used at points of high information within the utterance, thereby extending the time during which the high

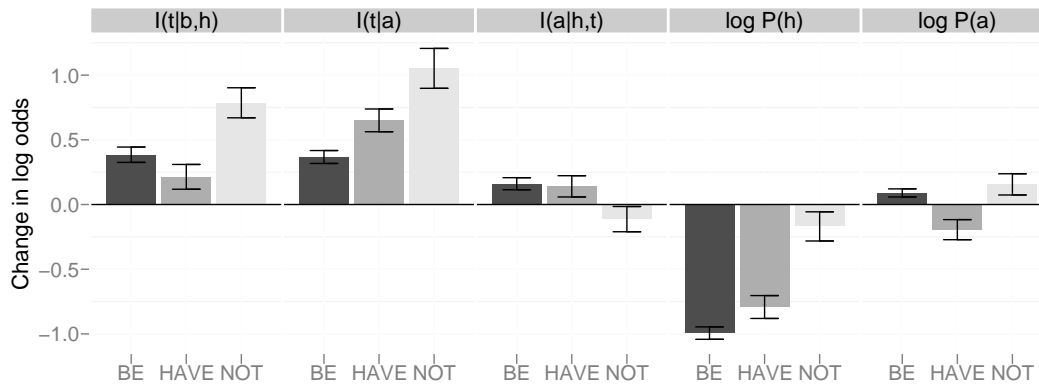


Figure 1: Estimated changes in log-odds (coefficient estimates) associated with predictors in the model

information element is uttered. This result is consistently found on measures of the information of the reducible element ($I(t|b,h)$ and $I(t|a)$). The information of the following word ($I(a|h,t)$), however, has smaller and less consistent effects on reduction. In the cases where the target element encodes a larger amount of information, using a full form is predicted by UID. Use of a full form lowers the rate of information transmission, avoiding a peak and maintaining a more uniform information density. These findings reinforce results reported at multiple levels of linguistic representation.

There are other aspects of these results that are worth noting. Our investigations showed trends in the expected direction for speech rate, but only revealed a significant influence of speech rate on reduction in one data set (NOT). This is troubling, as speaking rate has previously been shown to play an important role in other reduction phenomena at several levels of linguistic representation (Aylett & Turk, 2004; Bell et al., 2003; Jaeger, 2006, *inter alia*). Also, previous studies have found speech rate effects on syntactic reduction using the same automatically extracted speech rate information that we have employed here (Jaeger, 2006). It is possible that transcribers had difficulty distinguishing between full and reduced forms for high speech rates, perhaps reverting to a default, which would make the annotation noisy with regard to speech rate. It may also be the case that BE and HAVE occur in similar positions in prosodic domains, or at least that their use is more similar to each other than to the use of NOT. If this supposition holds, our current methods for estimating speech rate may work well at the positions where NOT is generally found, but poorly for BE and HAVE. It will be important to continue to refine our methods for estimating speech rate as this research goes forward.

Additionally, the role of log probability in driving contraction use merits further attention. We consistently find that higher log frequency of the word preceding the reducible element ($\log P(h)$) is associated with a high probability of using a contraction. While this effect is largely driven by the prevalence of personal pronouns and auxiliary verbs as hosts to contractions, the effect remains when host type is included in the analysis as a categorical factor. High frequency words following a reducible element ($\log P(a)$), on the other hand, have varying effects on reduction across our data sets. This variability is unexpected. If anything, the *a priori* expectation

was that speakers' use of contraction should increase preceding high-frequency words, because the frequency of the following word can be seen as a measure of its availability. Given that previous studies on other reduction phenomena have provided strong evidence for availability-based production accounts that predict more frequent contraction before available words and phrases, and given that we do observe the expected effect for the *conditional* probability of the following word for BE and HAVE (*cf.* the first column in Figure 1), we plan to investigate this puzzling result in future research.

Finally, further evaluation of our statistical models is required. The stability of the confidence intervals presented here will be investigated using Highest Posterior Density intervals. Model comparison will be performed via the Conditional AIC measure in addition to log likelihood tests. Theoretically motivated interactions between predictors will be tested.

General Discussion

Formal descriptions of ideal performance on a cognitive task offer a useful benchmark in the study of cognition. In this paper we provide further evidence for the view that a suitable characterization of ideal performance on a communication task involves transmitting information at a uniform rate close to channel capacity (Genzel & Charniak, 2002), resulting in uniform information density. In three corpus studies of morphosyntactic reduction, we show that speakers make use of the variability licensed by their mental grammar in ways that are consistent with a UID strategy. Our findings extend previous work on the redundancy-sensitivity of segment, word, and syntactic reduction (Aylett & Turk, 2004; Bell et al., 2003; van Son & van Santen, 2005; Jaeger, 2006; Levy & Jaeger, 2007) to the level of morphology.

Studies 1-3 address the most basic predictions of UID, but there are several outstanding issues which we intend to investigate in continuations of this work and novel corpus and behavioral studies. First, while we address the issue of rate of transmission, we have yet to directly investigate channel capacity. Speakers may differ in their channel capacity and these differences may affect speakers' production strategies. Second, it remains to be determined whether speakers can work around channel limitations. For example, in a situation

where channel capacity is highly constrained by the articulatory system, speakers could in theory employ multiple channels. Speakers may use supra-segmental features of language, like gesture and prosody, to modulate the rate of information being transmitted per channel at different points in an utterance.

As mentioned previously, one of the features of rational accounts of cognition is that performance is optimized relative to certain constraints. Perhaps the most important constraint is that these strategies are implemented by human cognitive processes. General computational principles that characterize the human language production system (like UID) must eventually be linked to neural implementation. Similarly, it is important to understand how these computational principles relate to the cognitive mechanisms and representations postulated by algorithmic theories of language production.

Availability-Based Production directly addresses the cognitive mechanisms of language production (Ferreira & Dell, 2000). Recall that in ABP, the form of speakers' utterances is generally determined by limitations on the production system's ability to retrieve and integrate upcoming material in the utterance. This predicts that speakers will put off low-frequency, unpredictable, or highly informative elements of an utterance because those features take more time to prepare. The same cognitive resources that ABP credits for shaping production may also constrain whatever higher-level optimization strategies a speaker is using. For example, the channel capacity assumed in UID may be a speaker-internal construct (as opposed to a limitation that exists between interlocutors). If limits on communication between processing systems determine the availability of mental representations, then availability can be thought of as a measure of how efficiently information is being transmitted within the processing system, and can be directly related to theories like UID.

The study of language production as a rational system stands to provide important insights. In UID, the characterization of communication in information-theoretic terms opens a wide set of issues to be explored. In this paper, one particular prediction of UID was borne out: namely, that speakers use the variability allowed by the morphosyntactic structure of their language to avoid peaks and troughs in the rate of information transmission. Ongoing work tests further predictions of UID in language production, and new efforts will attempt to relate this computational level theory to cognitive mechanisms and representations.

References

- Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.
- Aylett, M., & Turk, A. (2004, March). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language & Speech*, 47, 31-56.
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003, February). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113, 1001-1024.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-24.
- Bresnan, J., Carletta, J., Crouch, R., Nissim, M., Steedman, M., Wasow, T., et al. (2002). *Paraphrase analysis for improved generation*. (In LINK project: HRCR Edinburgh-CLSI Stanford)
- Calhoun, S., Nissim, M., Steedman, M., & Brenier, J. (2005, June). A framework for annotating information structure in discourse. In *Proceedings of frontiers in corpus annotation ii: Pie in the sky, acl2005 conference workshop*. Ann Arbor, MI.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287-291.
- Dell, G. S. (1986, July). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93, 283-321. (PMID: 3749399)
- Ferreira, V. S., & Dell, G. S. (2000). The effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40, 296-340.
- Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. *Proceedings of ACL-2002, Philadelphia*, 199-206.
- Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, 38, 313-338.
- Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.
- Jurafsky, D., & Martin, J. H. (2007). *Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing*. (2nd edition draft)
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), (p. 849-856). Cambridge, MA: MIT Press.
- Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3*.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.
- Rohde, D. L. T. (2005). *Tgrep2 user manual* (version 1.15 ed.).
- Shannon, C. E. (1948). A mathematical theory of communications. *Bell Systems Technical Journal*, 27, 623-656.
- Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An overview. In A. Abeillé (Ed.), (p. 5-22).
- van Son, R., & van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, 47, 100-123.