Uniform Information Density

Matthew W. Crocker
Vera Demberg

Block Course – Summer Semester 2015

Surprisal & Psycholinguistics

- The information conveyed by any given linguistic unit (e.g. phoneme, word, utterance) in context is called surprisal:

$$Surprisal(x) = \log_2 \frac{1}{P(x \mid context)}$$

- Surprisal will be high, when *x* has a low conditional probability, and low, when *x* has a high probability.
- Claim: Cognitive effort required to process a word is proportional to its surprisal

Information Theoretic Approaches

- Surprisal offers a (linguistic) theory neutral measure of the information conveyed by linguistic events
- The average surprisal of a word has been shown to correlate with word length, suggesting lexica have "evolved" towards an optimised encoding
    - predictable words (on average) are shorter
- Surprisal also offers a good index of on-line lexical and syntactic processing effort
    - predictable words convey less information, are easier

Rational Communication

- Linguistic forms are being reduced/expanded at all linguistic levels
- Variation enables modulation of the rate and linearization of message transmission
    - Evidence: Word length, speech, reading times
- Rational communication systems:
    - How is information communicated optimally?
    - Are speakers adapted to listeners constraints?

Uniform Information Density Hypothesis

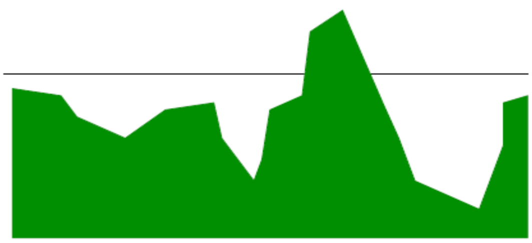
Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).

Jaeger, 2010

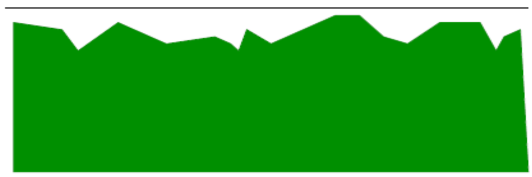
See also:
Entropy Rate Constancy Principle, Genzel & Charniak (2002)
Smooth Signal Redundancy Hypothesis, Aylett & Turk (2004)

UID Hypotheses

- Channel Capacity provides an upper bound on the amount of information
- Language users prefer to distribute information uniformly over a message



bad use of channel
ID very variable



good use of channel
ID uniformly distributed

Information Density

- Uniform Information Density:
  - Maximizes information transmission
  - Minimize comprehender difficulty

$$Information(event) = \log_2 \frac{1}{P(event)}$$

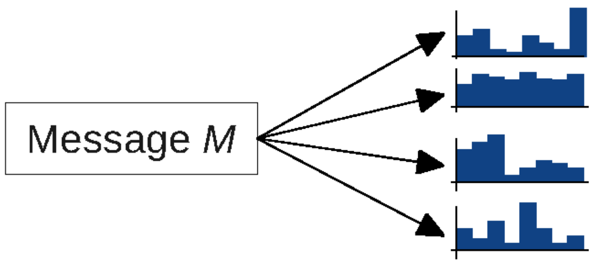
$$= \log_2 \frac{1}{P(w_1)} + \log_2 \frac{1}{P(w_2 \mid w_1)} + \ldots + \log_2 \frac{1}{P(w_n \mid w_1 \ldots w_{n-1})}$$

UID Hypotheses

- Variation in encoding serves to modulate information density
- Uniform information density at all levels of language use: speech to discourse
- Production choices are influenced by predictability:
  - Expansion of informationally dense (high surprisal) expressions
  - Reduction of more predictable expressions
  - Use forms that distribute information peaks over time

Variation and UID

- Within the bounds of the grammar, speakers should adopt the most encoding with greatest uniformity



- Note: assumes the alternatives are sufficiently meaning invariant

Linguistic Levels

- In principle, UID might be expected to be:
  - conditioned by all relevant context
  - relevant to determining encoding as all levels

$$
\begin{aligned}
Surprisal(unit) &= -\log_2 P(unit \mid \text{Context}) \\
&= -\log_2 P(word \mid \text{Script}) \\
&= -\log_2 P(syntactic\_unit \mid \text{Discourse}) \\
&= -\log_2 P(phone \mid \text{Collocation})
\end{aligned}
$$

Scope for variation

- Speech: we can modulate the duration and energy of our vocalisations
- Lexical: we can choose longer and shorter forms
  - *math* versus *mathematics*
- Syntactic reductions, and alternative linearisation
  - *The thief (that was) arrested was guilty.*

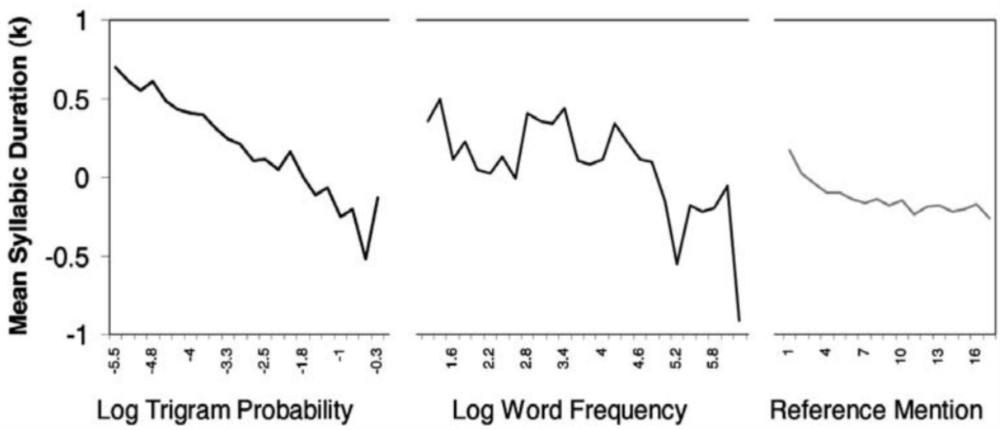
Evidence from Speech

- Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004):

the trade of "robust communication and articulatory effort suggests an inverse relation between redundancy and duration"

Aylett & Turk (2004)

- The SSR hypothesis is similar to UID: expected material is articulated with shorter durations
- Examined a large corpus of spontaneous speech
    - syllables coded with prosodic, durational, and redundancy information
    - redundancy was determined by syllabic **trigrams**, word **frequencies**, and # of previous **mentions**



- a significant effect of prosodic and redundancy factors on duration in a large corpus of spontaneous running speech
- an inverse relationship between redundancy and duration

Constancy Rate Principle

- **Hypothesis**: The entropy rate of generated text should remains constant across that text.
- The accruing context will generally reduce entropy of the text over time.
- **Prediction**: local measure of entropy (ignoring context), should increase with each successive sentence in a text

Two models

- Genzel & Charniak therefore compute sentence level surprisal, across sample texts
- N-gram model:

$$P(S) = P(w_1) \times P(w_2 \mid w_1) \times P(w_3 \mid w_2 w_1) \times \prod_{i=4}^{n} P(w_n \mid w_{n-1} w_{n-2} w_{n-3})$$

- Parsing model:

$$P(S) = \prod_{x \in T} P(x \mid parents(x))$$

Entropy rate

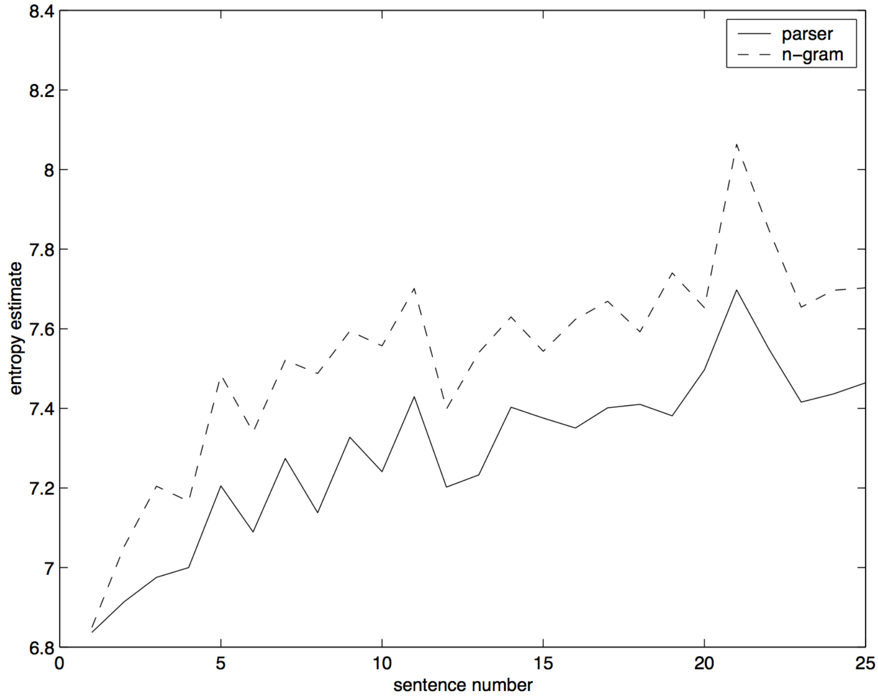


Figure 1: N-gram and parser estimates of entropy (in bits per word)

Syntactic Reduction

- Jaeger (2010 & PhD) tests the UID hypothesis at the syntactic level
- The complementizer "that" is optional in English:

  My boss confirmed (that) I am absolutely crazy.

- UID predicts that *that*-mentioning will be influenced by the surprisal of the complement clause (CC) onset

Example: *that*-omission

- The complementizer "that" is optional in English:

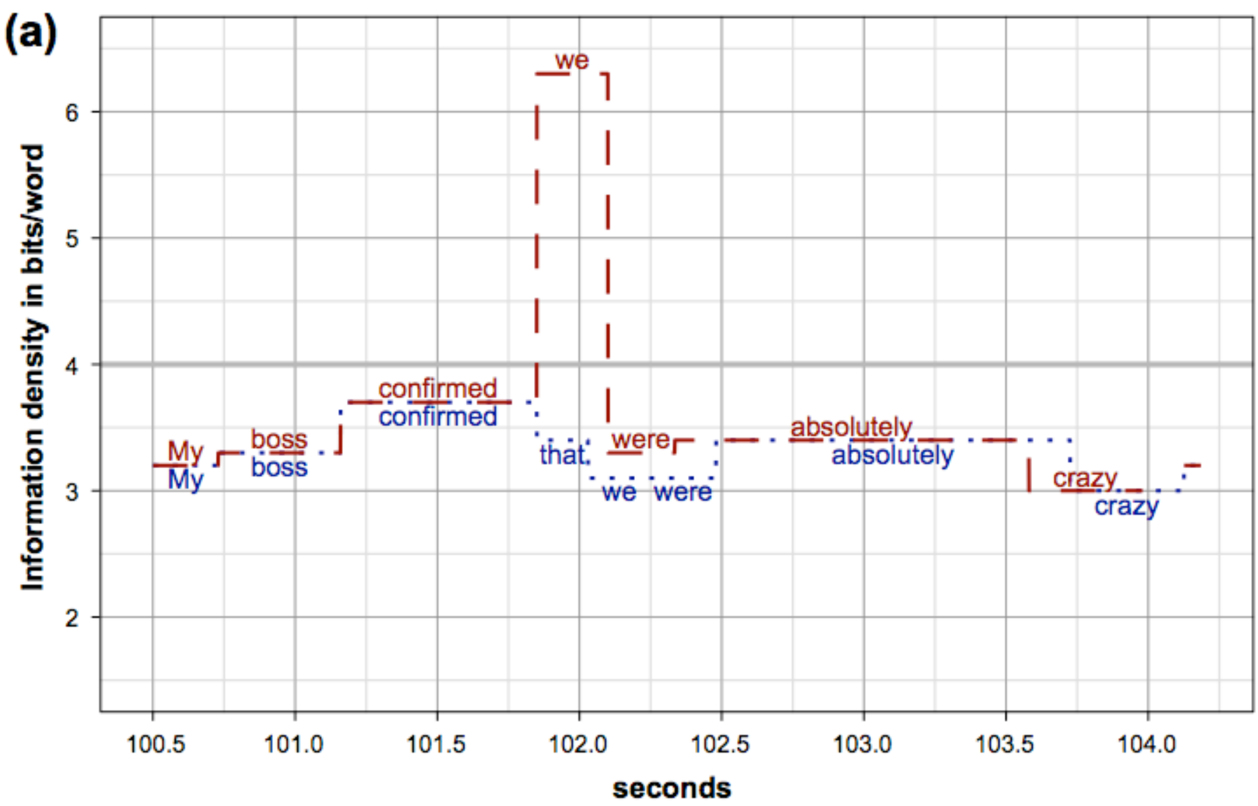
  My boss confirmed (that) I am absolutely crazy.

- Uniform Information Density: Use of overt "that" increases with ID at onset of the CC (i.e. w_1), namely "I ..."

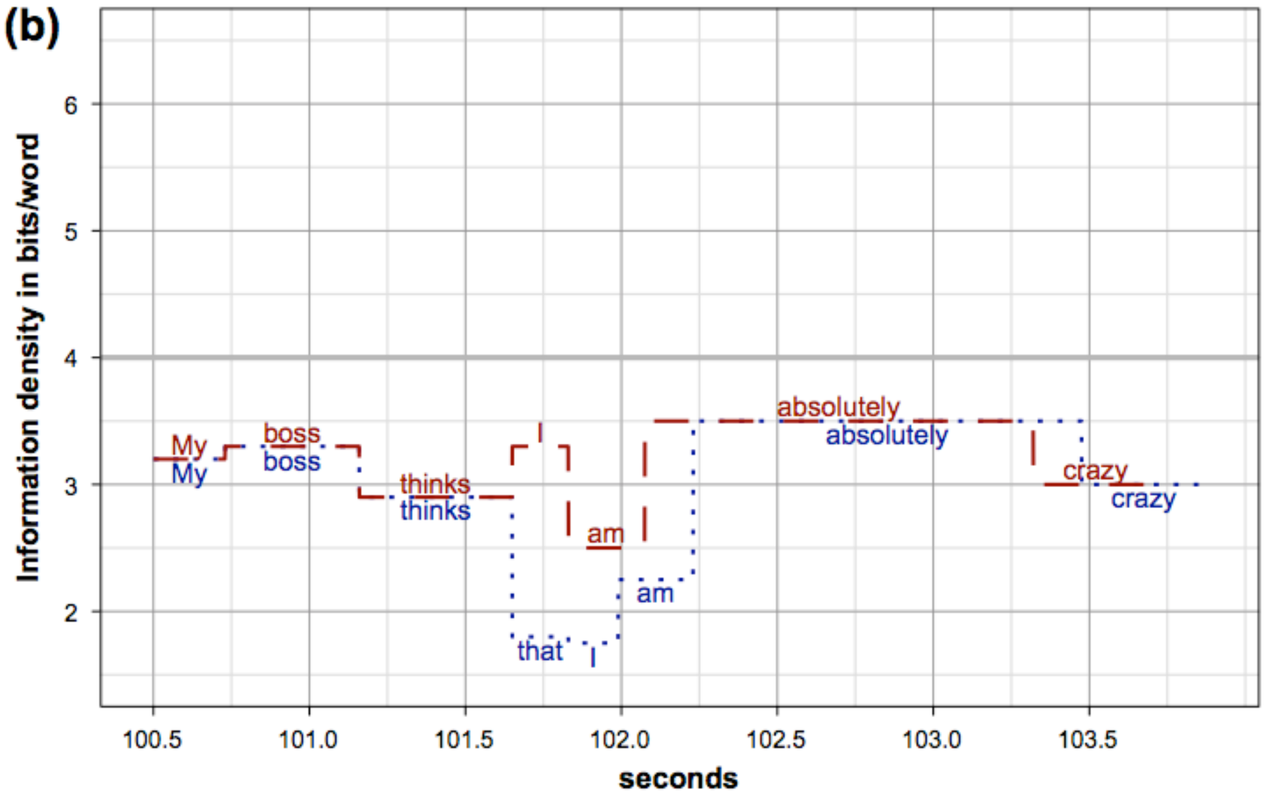
Overt that $= \log_2 \dfrac{1}{P(w_1 \mid CC, that, w_{\ldots -1})}$

Omitted that $= \log_2 \dfrac{1}{P(CC \mid w_{\ldots -1})} + \log_2 \dfrac{1}{P(w_1 \mid CC, w_{\ldots -1})}$

Jaeger, 2010



Jaeger, 2010



Jaeger, 2010

The Study

- A large scale corpus study of complement clause structures in spontaneous speech
    - Switchboard corpus of telephone dialogues
- Compares UID with other theories of *that*-mention
    - availability, ambiguity avoidance, and dependency processing
- Tests the influence of UID above and beyond other known predictors of *that*-mention

Previous accounts

- Availability: this account assumes speakers insert that when they know the following words will be more difficult to retrieve, and want to maintain fluency

- Ambiguity avoidance: *that*-mention occurs when other complements are possible, case doesn't disambiguate:

I know [many of them are doing it].

- Dependency accounts: increasing distance between the matrix verb and the CC correlates with *that*-mention

Predictor	Description	Type (βs) (1)
INTERCEPT		
Dependency length and position of CC		
POSITION(MATRIX VERB)	CC position in the sentence	cont(3)
LENGTH(MATRIX VERB-TO-CC)	Distance of CC from matrix verb	cont(1)
LENGTH(CC ONSET)	Length of CC onset	cont(1)
LENGTH(CC REMAINDER)	Length of remainder of CC	cont(1)
Overt production difficulty at CC onset		
SPEECH RATE	Log and squared log speech rate	cont(2)
PAUSE	Pause immediately preceding CC	cat(1)
DISFLUENCY	Normalized disfluency rate at CC onset	cont(1)
Lexical retrieval at CC onset		
CC SUBJECT	Type of CC subject	cat(3)
SUBJECT IDENTITY	Matrix and CC subject are identical	cat(1)
FREQUENCY(CC SUBJECT HEAD)	Log frequency CC subject head lemma	cont(1)
WORD FORM SIMILARITY	Potential for double *that* sequence	cat(1)
Lexical retrieval before CC onset		
FREQUENCY(MATRIX VERB)	Log frequency of verb lemma	cont(1)
Ambiguity avoidance at CC onset		
AMBIGUOUS CC ONSET	CC onset ambiguous without *that*	cat(1)
Grammaticalization		
MATRIX SUBJECT	Type of matrix subject	cat(3)
Additional controls		
SYNT. PERSISTENCE	Prime (if any) w/ or w/o *that*	cat(2)
MALE SPEAKER	Speaker is male	cat(1)
Total number of control parameters in model plus intercept		25

Predictor	Coef. β	SE(β)	z	p
Intercept	0.12	(0.38)	0.3	>0.7
POSITION(MATRIX VERB)	0.95	(0.14)	6.6	<0.0001
(1st restricted comp.)	−27.94	(5.33)	−5.2	<0.0001
(2nd restricted comp.)	55.43	(10.80)	−5.1	<0.0001
LENGTH(MATRIX VERB-TO-CC)	0.17	(0.065)	2.5	=0.01
LENGTH(CC ONSET)	0.18	(0.014)	12.8	<0.0001
LENGTH(CC REMAINDER)	0.03	(0.006)	4.4	<0.0001
LOG SPEECH RATE	−0.70	(0.13)	−5.5	<0.0001
SQ LOG SPEECH RATE	−0.36	(0.19)	−1.9	<0.06
PAUSE	1.11	(0.11)	10.2	<0.0001
DISFLUENCY	0.39	(0.12)	3.2	<0.002
CC SUBJECT =*it vs. I*	0.04	(0.08)	0.5	>0.6
=*other pro vs. prev. levels*	0.05	(0.03)	1.6	<0.11
=*other NP vs. prev. levels*	0.11	(0.02)	4.9	<0.0001
FREQUENCY(CC SUBJECT HEAD)	−0.02	(0.03)	−0.7	>0.5
SUBJECT IDENTITY	−0.32	(0.17)	−1.9	<0.052
WORD FORM SIMILARITY	−0.31	(0.17)	−1.8	<0.08
FREQUENCY(MATRIX VERB)	−0.23	(0.03)	−7.7	<0.0001
AMBIGUOUS CC ONSET	−0.12	(0.12)	−1.0	>0.2
MATRIX SUBJECT =*you*	0.48	(0.15)	3.1	<0.002
=*other PRO*	0.60	(0.13)	4.8	<0.0001
=*other NP*	0.85	(0.13)	6.7	<0.0001
PERSISTENCE =*no vs. prime w/o that*	0.02	(0.07)	0.3	>0.7
=*prime w/ that vs. prev. levels*	0.06	(0.04)	1.6	<0.11
MALE SPEAKER	−0.15	(0.11)	−1.3	>0.19
Information density	**0.47**	**(0.03)**	**16.9**	**<0.0001**

Support for UID

- ID has a significant influence on *that*-mention, even when all other predictors are controlled
- ID is in fact the stronger predictor in it's contribution to the model's likelihood (15% of model quality due to ID)
- Also support for the availability account (fluency) and dependency accounts, but only very limited support for ambiguity avoidance

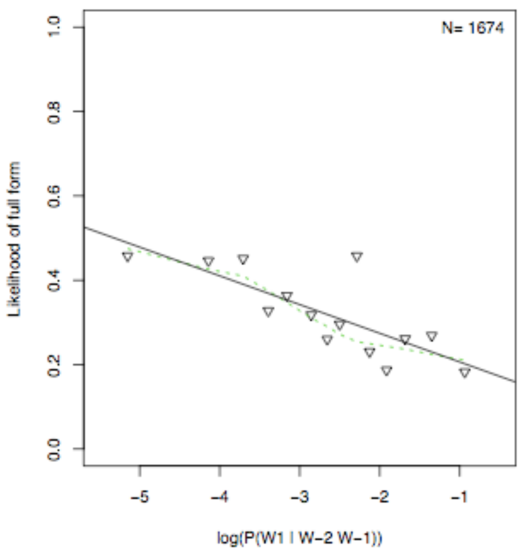
Additional Evidence

- Frank and Jaeger (2008) find evidence that contraction is influence by ID:
  - “I am” vs. “I’m” – “you have” vs. “you’ve” – “did not vs. didn’t”
- for the 4-grams `before host target after`: they compute: $I(t|b,h)$, $I(t|a)$ and $I(a|h,t)$
- ID of the target had consistent influence on reduction, ID of the following word, less so

that-relativiser ommission

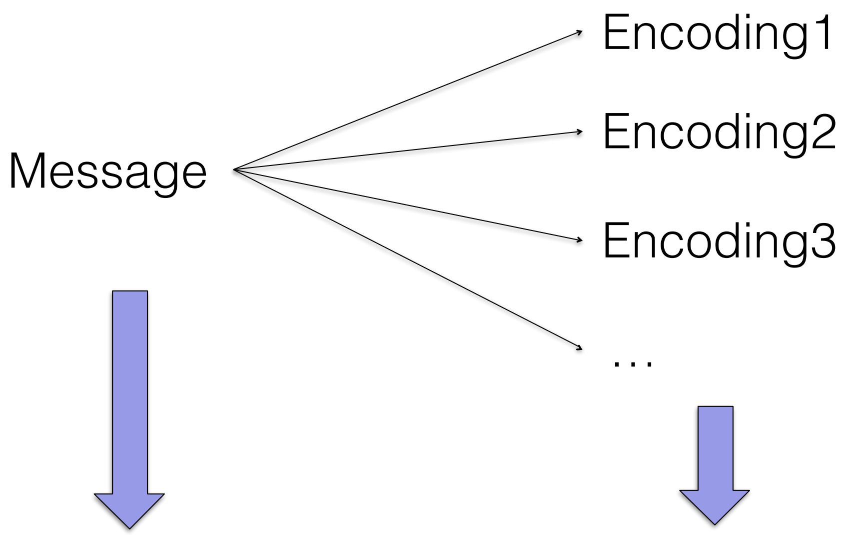
How big is [$_{NP}$ the family$_i$ [$_{RC}$ (that) you cook for $_i$]]

- Similar to that-complementisers, that-mention is relative clauses is optional
- N-gram estimates of ID predicted use of “that”
- Additionally, evidence that purely structural ID also predicts use of “that”



Levy & Jaeger, 2007

Encoding and UID



$$\text{Utterance} = \underset{Enc_i}{\text{argmax}}\, UID(Enc_i)$$

Discussion

- Evidence for uniformity preference …
  - … but not for maximal use of channel capacity
  - … does not claim signal will be uniform
- Is UID really "audience design" or does the speaker just use their own "language model"
  - Does speaker behaviour vary across listeners?
- Omission and contraction are very localised
  - Does UID influence larger encoding choices?