A Language's Unigram Entropy Distribution Predicts Self-Paced Reading Times

Josef Klafka
jklafka@uchicago.edu
Department of Psychology
University of Chicago

Daniel Yurovsky
yurovsky@uchicago.edu
Department of Psychology
University of Chicago

Abstract

We provide evidence against the Uniform Information Density (Jaeger, 2010) and wrap-up effect hypotheses.

Keywords: Entropy; self-paced reading; information theory; language processing

Uniform Information Density

A major line of research in the current century has asked how speakers and writers distribute information in the linguistic material they produce. This begins with Claude Shannon in the years after World War II. Shannon (1948) defined "information" as a "reduction in uncertainty". Shannon followed by quantifying uncertainty in his measure of *entropy*: the amount of uncertainty on the outcome of a random variable. Shannon proposed that the most efficient method of sending information through a noisy channel is at a constant rate. Genzel & Charniak (2002) apply this principle to human communication: if people communicate information through speech and text optimally, then we should transmit information to one another at a constant rate. Genzel & Charniak created a distribution of estimates for sentence-level entropy, each sentence taken without context, for newspaper articles from the Penn Treebank corpus. Obtaining a roughly monotonically increasing linear distribution, Genzel & Charniak proposed the *Constant Entropy Rate* (CER) principle: the entropy rate in speech and text with context should be constant as sentence number increases. Aylett & Turk (2004) proposed a similar principle for the phonological level.

Jaeger (2010) extends the CER principle to all levels of human communication: the *Uniform Information Density* (UID) hypothesis. Jaeger argues that speakers will try to distribute information more evenly over the course of an utterance, to be as close to channel capacity (in a Shannon sense) as possible. In particular, Jaeger singles out the production of "that" at the beginning of relative clauses in English and the use of contractions such as "he's" and "you're" versus "he is" and "you are" as instances where higher information density in the clause elicits the production of more material by speakers.

The *wrap-up effect* is a popular hypothesis within the eye-tracking community. The hypothesis states that, in written text, sentence-final words are processed more slowly on average then sentence-medial or sentence-initial words, due to readers integrating information from the entire sentence in order to form a final, coherent thought expressed by the sentence, among other reasons. With these two ideas, information production distribution in spoken and written communication, and information integration in reading written texts, we move on to discuss cross-linguistic differences in the distribution of information across sentences.

The UID hypothesis

UID challenges

More recent work has stood out in contrast to the traditional UID perspective. Zhan and Levy (2018) study Mandarin Chinese classifier use, and find that the use of specific classifier, versus general classifiers, appears more often in cases where the production of the corresponding noun is more difficult than when the production is easier, as would be predicted by UID. The UID perspective is challenged in Yu et al. (2016), by performing an analysis of entropy by position in the text portion of the British National Corpus. They use the following formula for each word position X of sentences of fixed length k from the corpus, where each i is a word occurring in position X and pi is the number of times word i occurs in position X divided by the number of total words that occur in position X i.e. the number of sentences of length k.

$$H(X) = \sum_{w} p(w)\log\big(p(w)\big)$$

Yu et al. (2016) refer to their distribution for English as a 'three-step distribution': relatively low entropy at the beginning of a sentence, then a jump, then flat entropy in the middle, a dip before the final position and a jump with the final word. View the figure below for a visual demonstration.

Our Methods

We used the CHILDES TalkBank (Brown 1973; MacWhinney 2000) corpora database of spoken adult-child conversations. We first used the Brown and Providence English corpora from CHILDES. The Brown corpus contains conversations between three young children (over 1.5 years old and under 6 years old) and their families in the home. The Providence corpus contains transcriptions of audio/video files which recorded interactions between children between 1 and 3 years old and their parents in the home.

We found a similar distribution for Spanish and German corpora from CHILDES. The Spanish corpus we used was the XXX and the German corpus we used was the XXX.

Formalities, Footnotes, and Floats

Use standard APA citation format. Citations within the text should include the author's last name and year. If the authors' names are included in the sentence, place only the year in parentheses, as in (1972), but otherwise place the entire reference in parentheses with the authors and year separated by

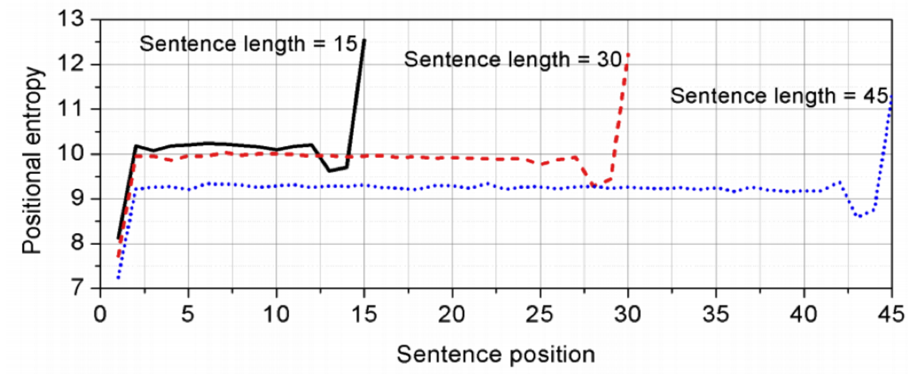

Figure 1: English entropy distribution for three sentence lengths from BNC

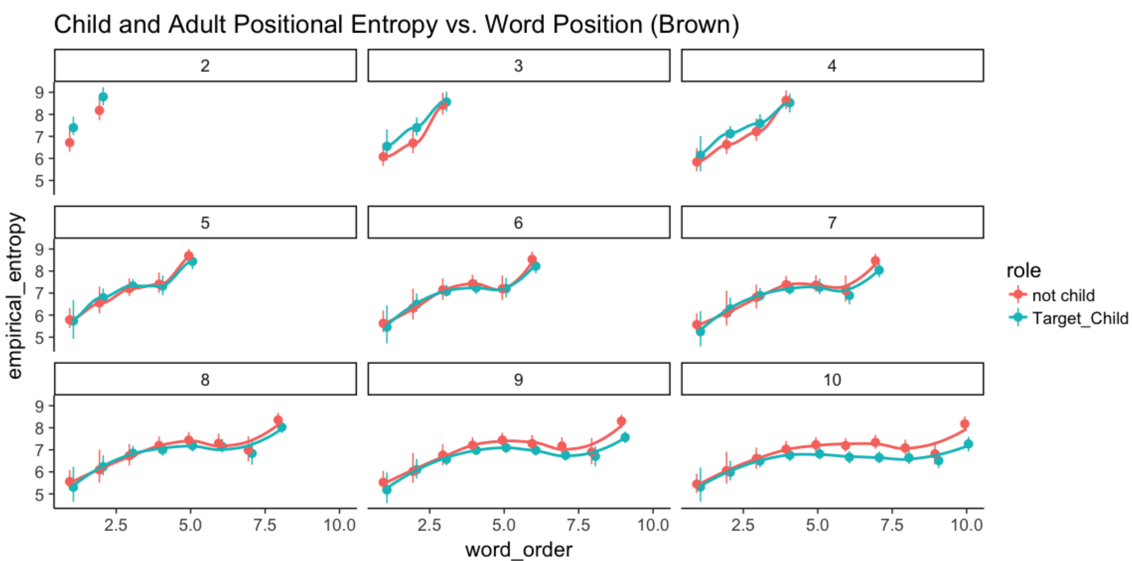

Figure 2: Brown corpus entropy

a comma (Newell & Simon, 1972). List multiple references alphabetically and separate them by semicolons (Chalnick & Billman, 1988; Newell & Simon, 1972). Use the et. al. construction only after listing all the authors to a publication in an earlier reference and for citations with four or more authors.

For more information on citations in RMarkdown, see **here.**

Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes in 9 point type at the bottom of the page on which they appear. Precede the footnote with a horizontal rule.[2] You can also use markdown formatting to include footnotes using this syntax.[3]

[1] Sample of the first footnote.
[2] Sample of the second footnote.
[3] Sample of a markdown footnote.

Figures

All artwork must be very dark for purposes of reproduction and should not be hand drawn. Number figures sequentially, placing the figure number and caption, in 10 point, after the figure with one line space above the caption and one line space below it. If necessary, leave extra white space at the bottom of the page to avoid splitting the figure and figure caption. You may float figures to the top or bottom of a column, or set wide figures across both columns.

Two-column images

You can read local images using png package for example and plot it like a regular plot using grid.raster from the grid package. With this method you have full control of the size of your image. **Note: Image must be in .png file format for the readPNG function to work.**

You might want to display a wide figure across both columns. To do this, you change the `fig.env` chunk option

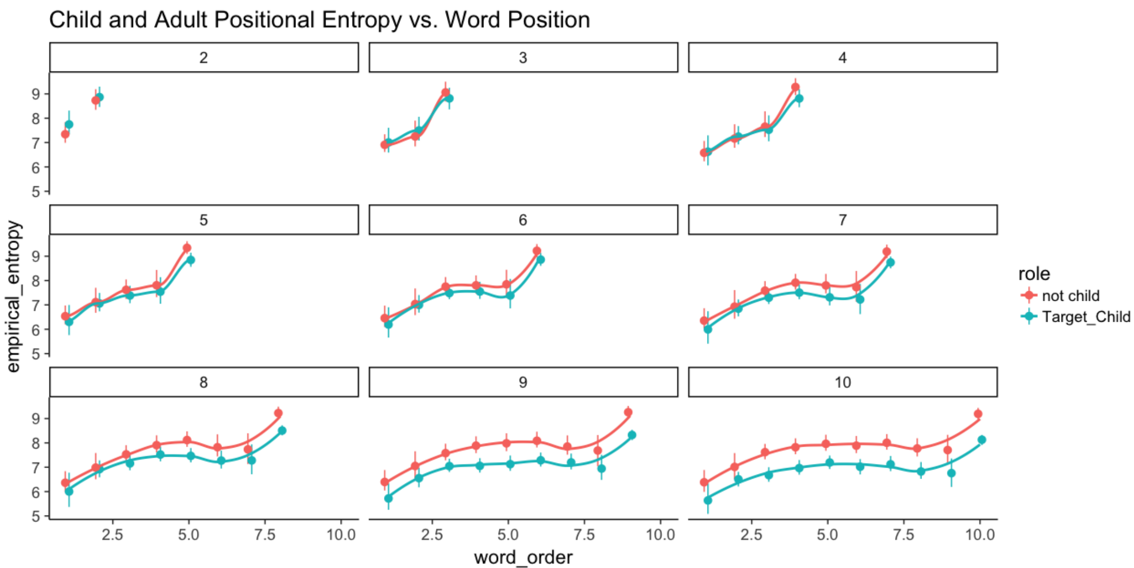

Figure 3: Providence corpus unigram entropy

to `figure*`. To align the image in the center of the page, set `fig.align` option to `center`. To format the width of your caption text, you set the `num.cols.cap` option to `2`.

One-column images

Single column is the default option, but if you want set it explicitly, set `fig.env` to `figure`. Notice that the `num.cols` option for the caption width is set to `1`.

Figure 5: One column image.

R Plots

You can use R chunks directly to plot graphs. And you can use latex floats in the fig.pos chunk option to have more control over the location of your plot on the page. For more information on latex placement specifiers see **here**

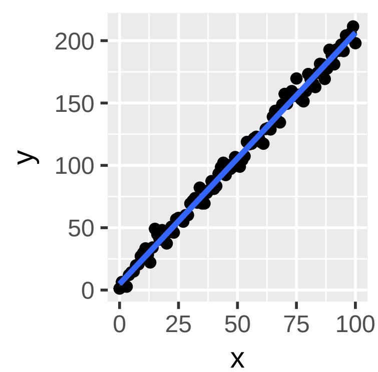

Figure 6: R plot

Tables

Number tables consecutively; place the table number and title (in 10 point) above the table with one line space above the caption and one line space below it, as in Table 1. You may float tables to the top or bottom of a column, set wide tables across both columns.

You can use the xtable function in the xtable package.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.10	0.11	-0.9	0.37
x	1.97	0.12	16.6	0.00

Table 1: This table prints across one column.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

Figure 4: This image spans both columns. And the caption text is limited to 0.8 of the width of the document.

References

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.