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

The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading Abstract should be 10 point, bold, centered, with one line space below it. This one-paragraph abstract section is required only for standard spoken papers and standard posters (i.e., those presentations that will be represented by six page papers in the Proceedings).

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

Uniform Information Density

The average distribution of information in English sentences- what is it? Genzel & Charniak (2002) looked at the log-probability of words occurring in sentences as you proceed through a paragraph or article and found that less common/less likely words occurred more frequently as you moved through the text. The relationship they found was roughly an affine linear function, leading them to propose the *constant entropy rate* (CER) principle: the entropy of words in a sentence, i.e. the uncertainty about what words will appear in a sentence, will increase at a constant, linear rate through a paragraph. This idea was used by Aylett & Turk (2004), Levy & Jaeger (2007) and Frank & Jaeger (2008) among others. The CER principle was adapted by Jaeger (2008) into the *uniform information density* (UID) principle: uniformation is evenly spread throughout a sentence. From a theoretical standpoint, this makes sense as the optimal distribution for information throughout a sentence: if you miss any word I've said, the rest of the information in the sentence is intact, and you lose no more than if you miss any other word I'd have said. Description of the UID perspective and what it predicts for entropy distributions and what that means.

Yu challenge

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.

This diverges from the UID account, which would predict a simple affine function. Our replication with CHILDES. Conditional entropy and mutual information. Link with Thiessen & Onnis (to appear). Mention Ferrer-i-Cancho (2017), which provides a theoretical basis for the end of the sentence being the center of information. Cross-linguistic variation. Our Wikipedia analysis Link with Kuperman et al. (2010). Consequences for eye-tracking. Talk about Zhan and Levy (2018) and how current work in eye-tracking and language processing disputes the UID account. Our eye-tracking experiment.

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 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]

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.

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

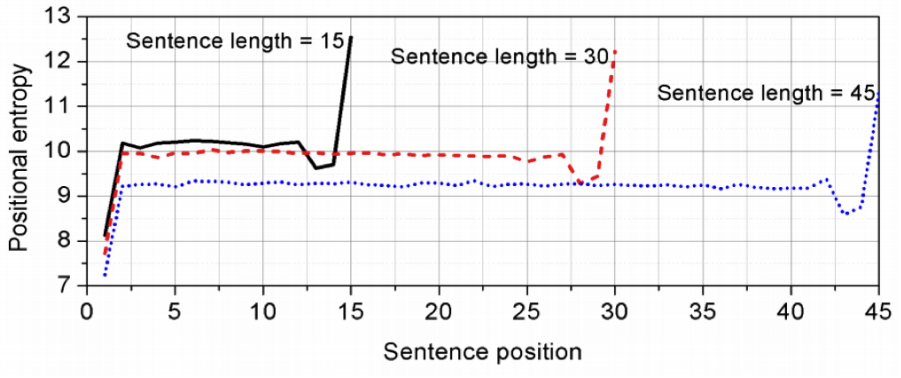

Figure 1: English entropy distribution for three sentence lengths from BNC

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 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`.

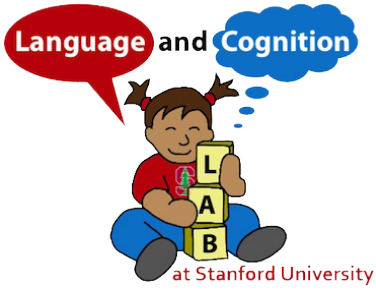

Figure 3: One column image.

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**

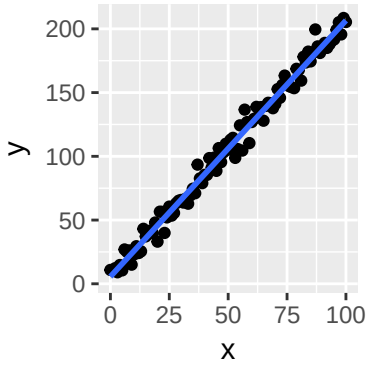

Figure 4: R plot

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)	-0.02	0.10	-0.2	0.82
x	1.95	0.10	19.2	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.

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*.

Figure 2: This image spans both columns. And the caption text is limited to 0.8 of the width of the document.

Englewood Cliffs, NJ: Prentice-Hall.