Multilingual A-maze: Generating Maze Experiments in Mandarin and Beyond

Lisa Levinson[1], Yizhi Tang[2], Lucy Yu-Chuan Chiang[1], Wei-Jie Zhou[1], Sohee Chung[1] ([1]University of Michigan, [2]Columbia University)

Aims

- Linguistically diversify the accessibility of automated alternative generation (A-maze) (Boyce, Futrell, and Levy 2020) for **maze task** experiments
- Especially provide generation for **Mandarin**, a challenging case
- Increase **user accessibility** to automated alternative generation tools

Unfamiliar with the maze task? See the rightmost panel!

Background: Original A-maze

Rather than use a language model to predict high probability continuations, the original A-maze algorithm (Boyce, Futrell, and Levy 2020) selects for **continuations that are significantly less probable** than the stimuli continuation (according to a user-determined threshold).

While this package has rapidly accelerated the adoption of the highly compelling Maze task, it only includes scripts that work with pre-trained LSTMs for English and French and requires non-trivial setup of a local Python environment.

Multilingual A-maze

We have re-implemented a subset of the algorithms designed for the original A-maze to adapt it in 3 dimensions.

1. **Broad range of languages**, using Hugging Face models including BERT (currently) and (soon) GPT-2, with multilingual BERT (104 languages).
2. **Mandarin version**. Highly-resourced language, but orthography requires different tokenization (and thus a need for multi-character alternative matching algorithm).
3. **Web interface**. Scripts shared as Google Colab notebooks that can be run via browser, without a local Python environment. This makes the software accessible to researchers with limited Python experience. Users can upload csv-format stimuli, specify parameters, run code cells, and download outputs. For multilingual maze, upload of a csv with frequency data also required.

https://github.com/UMWordLab/multilingual_amaze

Alternative Quality

Poor quality alternatives lead to participant selection of the alternative rather than target, and thus **experimental data loss**. This can be measured by assessing error (alternative selection) rates over **uncorrected a-maze outputs**.

So far we have tested the most recent version of the **uncorrected outputs of the Mandarin implementation** and compared to the same analysis conducted by Boyce, Futrell, and Levy (2020).

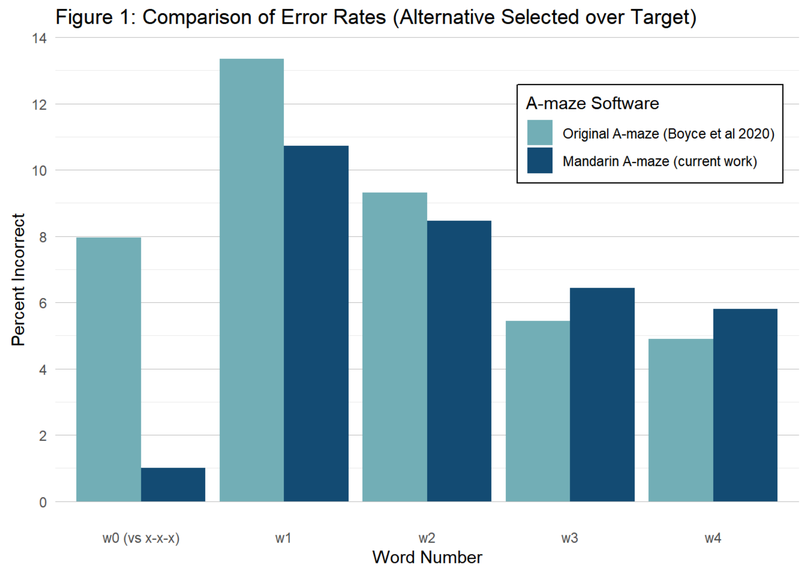
Here we are only concerned with accuracy, but the Mandarin stimuli were from Jäger et al. (2015), a study testing for subject-relative advantage.

23 (thus far) native Mandarin speakers recruited from the UM community completed an online experiment on a lab server based on the PCIbex (Zehr and Schwarz 2018) controller.

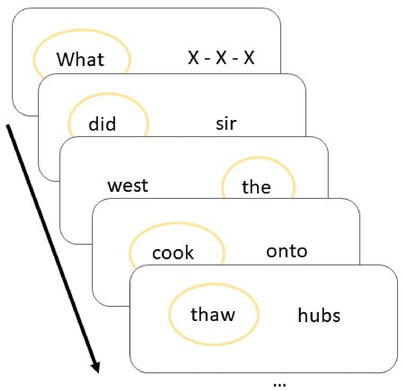
We calculated the percent of correct responses for each of the first 5 words, shown in Figure 1.

Results

Figure 1: Comparison of Error Rates (Alternative Selected over Target)



Maze Task

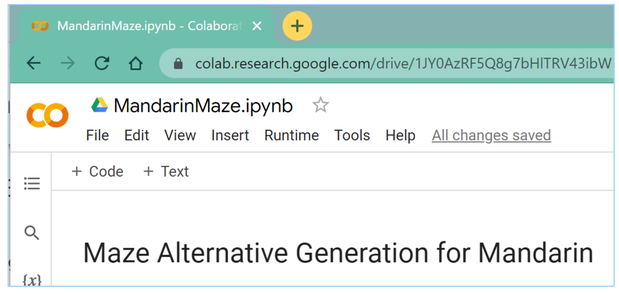


Choice of sentence continuation - target vs. alternative.

Highly incremental, focalized task. Long RTs, no spillover. (Forster, Guerrera, and Elliot 2009; Witzel, Witzel, and Forster 2012)

A-maze is a variant of "grammatical maze", where alternatives are valid words but poor fits for the sentence context (low cloze/high surprisal).

Platform



Conclusion

The accuracy results suggest that the Mandarin implementation of Multilingual Maze **performs comparably** to the original A-maze software, while also **expanding application** to a broader range of languages and researchers.

References

Boyce, Veronica, Richard Futrell, and Roger P. Levy. 2020. *Journal of Memory and Language* 111 (April): 104082. https://doi.org/10.1016/j.jml.2019.104082.

Forster, Kenneth I., Christine Guerrera, and Lisa Elliot. 2009. *Behavior Research Methods* 41 (1): 163–71. https://doi.org/10.3758/BRM.41.1.163.

Jäger, Lena, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth. 2015. *Journal of Memory and Language* 79-80 (February): 97–120. https://doi.org/10.1016/j.jml.2014.10.005.

Witzel, Naoko, Jeffrey Witzel, and Kenneth Forster. 2012. *Journal of Psycholinguistic Research* 41 (2): 105–28. https://doi.org/10.1007/s10936-011-9179-x.

Zehr, Jérémy, and Florian Schwarz. 2018. *PennController for Internet Based Experiments (IBEX).* https://osf.io/md832/.