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

Authors: Yizhi Tang, Lucy Yu-Chuan Chiang, and Lisa Levinson

Aims: Broadening the linguistic diversity of research focus has proved difficult in the domain of psycholinguistics, where English dominates and English-based experimental resources perpetuate this dominance. In this project we extend the recently innovated A-maze[1] experiment generation tools to work with Mandarin (the most orthographically challenging) and a wide range of other languages, and have created an open source software package that will be easily accessible to researchers of different technical skill levels around the world.

Challenge: Psycholinguistics has especially struggled to escape English dominance due to the unavailability of important resources not only for minoritized languages, but for any languages other than English. These resources historically have included large corpora and descriptive lexical statistics, but even as such resources grow for other languages, the recent shift towards using large language models in computational psycholinguistics has shifted the density of work again back towards English. This not only limits cross-linguistic work on topics like predictive processing, but also experimental tasks. One such task is the Maze task[2,3], which is a sentence reading paradigm measuring reaction times as participants choose the best sentence continuation from a pair of words (as in Figure 1). The task is applicable in similar contexts as self-paced reading, but requires fuller incremental parsing and does not induce spillover effects.

Original A-maze: Due to the difficulty of stimuli creation, the Maze task was not widely used until the development of the A-maze software package[1], which automates the generation of the alternative words using predictive language models. After downloading and installing Python, packages, scripts, and the language model files, researchers can input a list of stimuli as a text file and run the relevant scripts from the command line. The maze alternatives will be output in another text file, optionally pre-formatted for the original Ibex[4] experiment platform. Rather than use the language model to predict high probability continuations, the algorithm selects for continuations that are significantly less probable than the stimuli continuation (according to a user-determined threshold). The software also automates lexical frequency and word-length matching between stimuli and alternatives, as well as providing an option to match alternatives across multiple conditions within one item. 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.

Multilingual Maze: In this work we re-implemented the algorithms designed for the original A-maze software to extend its use in three dimensions. (1) Our software works with a broad range of languages, as probabilities are extracted using Hugging Face models and transformers library with options for BERT[5] and GPT-2[6] models, including multilingual BERT which includes 104 languages. (2) Our implementation can generate maze alternatives for Mandarin Chinese, the only highly-resourced language not available in multilingual BERT due to its orthography. Most Mandarin language models are tokenized by characters due to the lack of word spacing in text. This complicates the task of extracting multi-character words within a specified range of word (vs character) probabilities. Matching word-level probabilities and frequencies is necessitated by the prevalence of words such as complex verbs where the word frequency is not closely estimated by the joint probabilities of characters within the word. (3) Use of the open Hugging Face models helped to implement the scripts via well-documented Google Colab Jupyter notebooks that can be run in a web browser with no local installation. This makes the software much more accessible to researchers with limited programming or Python background.

Data Quality: To test the quality of the Multilingual Maze alternatives, we conducted a study in simplified Mandarin and evaluated participant accuracy. For the maze task to work well, participants must choose the “grammatical” or higher probability most of the time, as incorrect responses are discarded and (optionally) end sentence presentation. Our stimuli contained three Mandarin sentence types that varied in length and verb types due to the challenge that Mandarin verbs pose for the frequency and matching algorithms (Table 1). 25 native Mandarin speakers currently living in the US completed an online experiment using the PCIbex[7] platform. We calculated the percent of correct responses for each of the first 5 words (first 3 for 20 shorter sentences), as error rates vary by sentence position. Error rates ranged from only .3 to 2.3% incorrect, which verifies that the Multilingual A-maze is generating alternatives which are sufficiently distinct from the stimuli words. These rates are also much lower than reported for words 1-5 in the original online A-maze studies (e.g. as high as 13% incorrect for word 2 with Amazon Mechanical Turk participants), and similar to those for their in-lab traditional maze[1]. We also compared these results to accuracy in previously-conducted English study using original A-maze and Prolific participants. As can be seen in Figure 2, accuracy rates across the packages (across different sentence types and languages), are within the same range.

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

References:

[1] V. Boyce, R. Futrell, and R. P. Levy (2020) *J. Mem. Lang.*, vol. 111.[2] K. I. Forster *et al*, (2009) *Behav. Res. Methods*, vol. 41, no. 1. [3] N. Witzel, J. Witzel, and K. Forster, (2012) *J. Psycholinguist. Res.*, vol. 41, no. 2 [4] A. Drummond, (2013) *Ibex farm*. [5] J. Devlin *et al* (2018) *CoRR*, vol. bs/1810.04805. [6] A. Radford *et al.*, (2019) *OpenAI Blog*, vol. 1, no. 8. [7] F. Schwarz and J. Zehr,(2021) *Proc. Annu. Meet. Cogn. Sci. Soc.*, vol. 43.

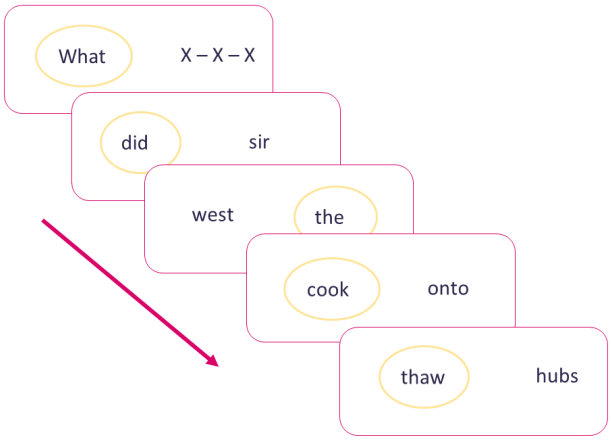

Figure 1: Maze task illustration.

Sentence Type	Mandarin (/ = word break)	English translation	items
Simple transitive verb with temporal adverbial (for sentence length)	星期天 / 中午 / 小明 / 吃了 / 面包	Xiaoming ate bread at noon on Sunday.	50
Compound resultative verb	小明 / 擦干了 / 眼泪	Xiaoming wiped away his tears.	20
Serial verbs	李静 / 爬了 / 果树 / 摘 / 水果	Li Jing climbed the fruit tree to pick the fruit.	20

Table 1: Mandarin stimuli

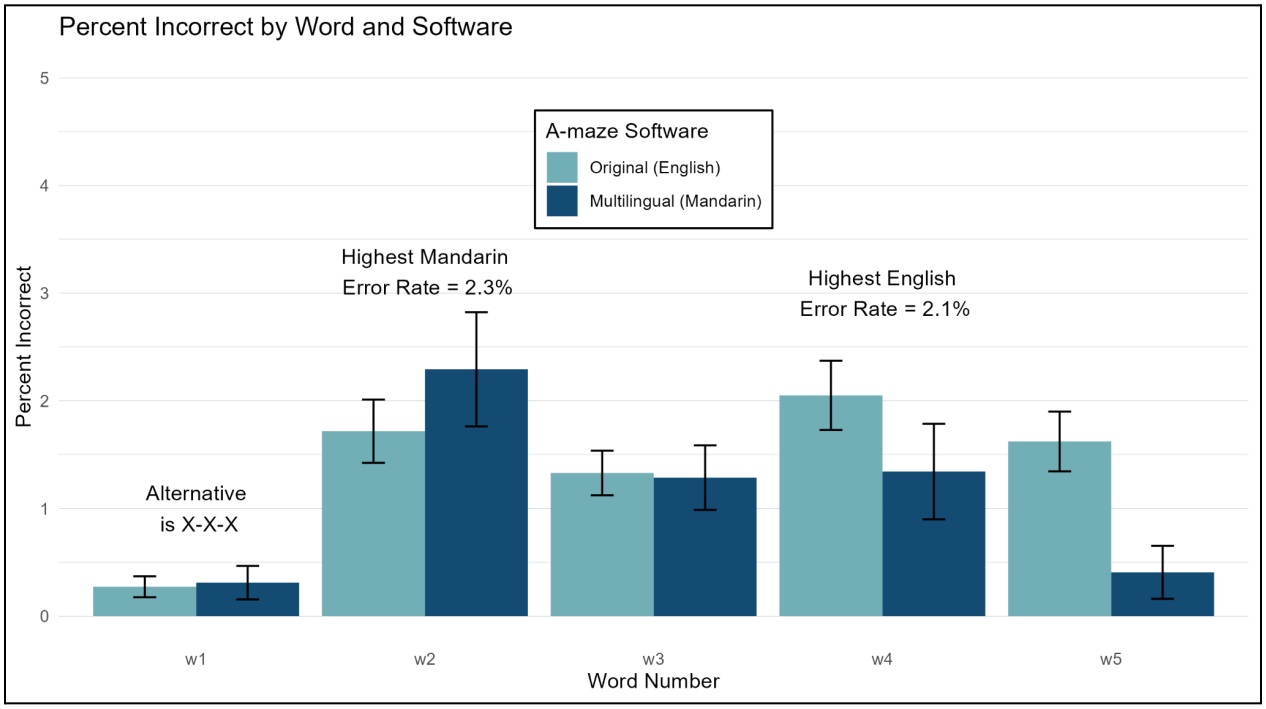

Figure 2: Word-by-Word comparison of original English A-maze participant accuracy (n=60) and Mandarin (Multilingual) A-maze (n=25). Words vary in category within and across languages.