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], and Sohee Chung[1]
[1]University of Michigan, [2]Columbia University

Aims: Broadening the linguistic diversity of research focus has proved difficult in the domain of sentence processing. 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 reduced the Python experience required to make this tool more accessible to researchers of different technical skill levels around the world.

Challenge: Psycholinguistic work has especially struggled to escape English dominance due to the unavailability of important resources for other languages. This resource limitation extends to tasks such as 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 Fig. 1).

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 and optionally produces pre-formatted stimuli 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). 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: We have re-implemented a subset of the algorithms designed for the original A-maze to extend its use in 3 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, the only highly-resourced language not available in multilingual BERT due to its orthography and tokenization. (3) We use open Hugging Face models to implement the scripts via well-documented Google Colab notebooks that can be run in a browser. This makes the software accessible to researchers with limited programming experience.

Data Quality: We evaluated participant accuracy for a study in simplified Mandarin. For the maze task to work well, participants must choose the higher probability alternative most of the time. Our stimuli contained 3 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 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); error rates ranged from only .3 to 2.3% incorrect, which verifies that alternatives are sufficiently distinct from the target words. These rates are 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. Accuracy rates across the packages (across different sentence types and languages), are within the same range (Figure 2). We are currently testing timing data and a new version of the software with additional features.

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.

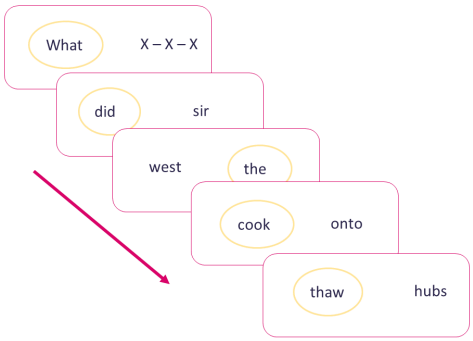

Figure 1: Maze task illustration.

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

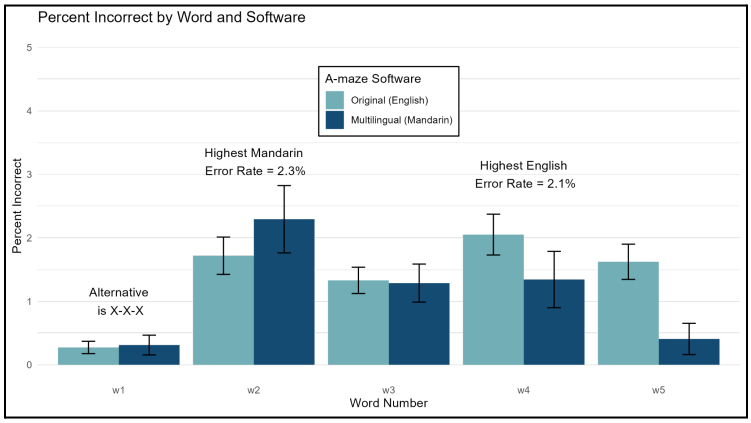

Figure 2: Word-by-Word comparison of original English A-maze participant accuracy (n=60) and Mandarin (Multilingual) A-maze (n=25). Words vary in category within and across languages.