# Supplemental Figures

These are supplemental figures for Knowles & Lo, 2024, *Calibration and Context in Human Evaluation of Machine Translation*.

In this appendix we provide full figures showing variations in annotator score distributions and how annotators use the 0-100 scale across calibration HITs, for all annotators for each language pair. Figures along the diagonal show histograms of annotator scores, while the off-diagonal figures show a comparison between two annotators' scores, with each point representing a segment (its x-value determined by the score given to it by the annotator for that column and its y-value determined by the score given to it by the annotator for that row).

Note: all figures are scalable; the reader may wish to zoom in for detail.

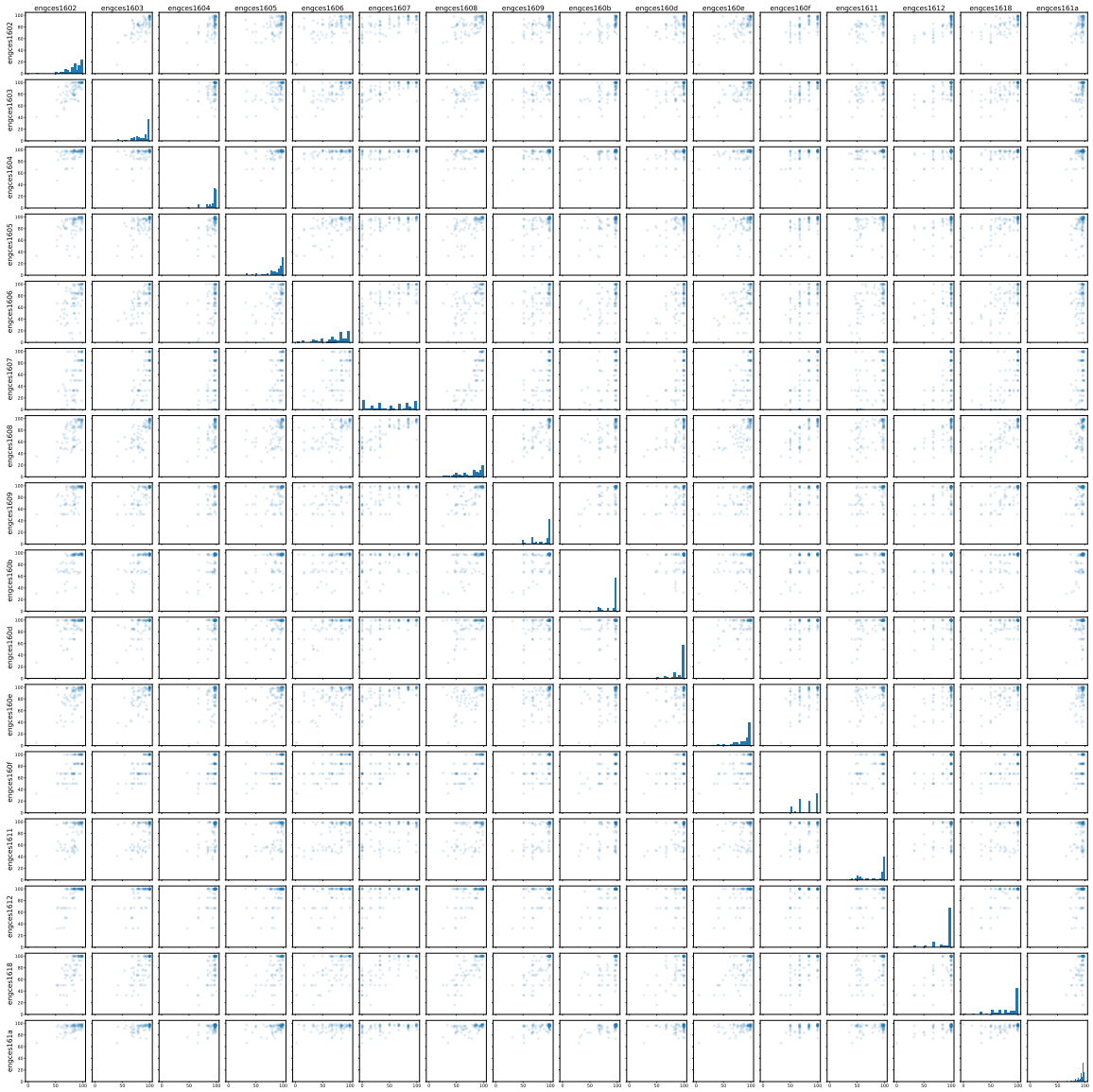


Figure 1: English–Czech

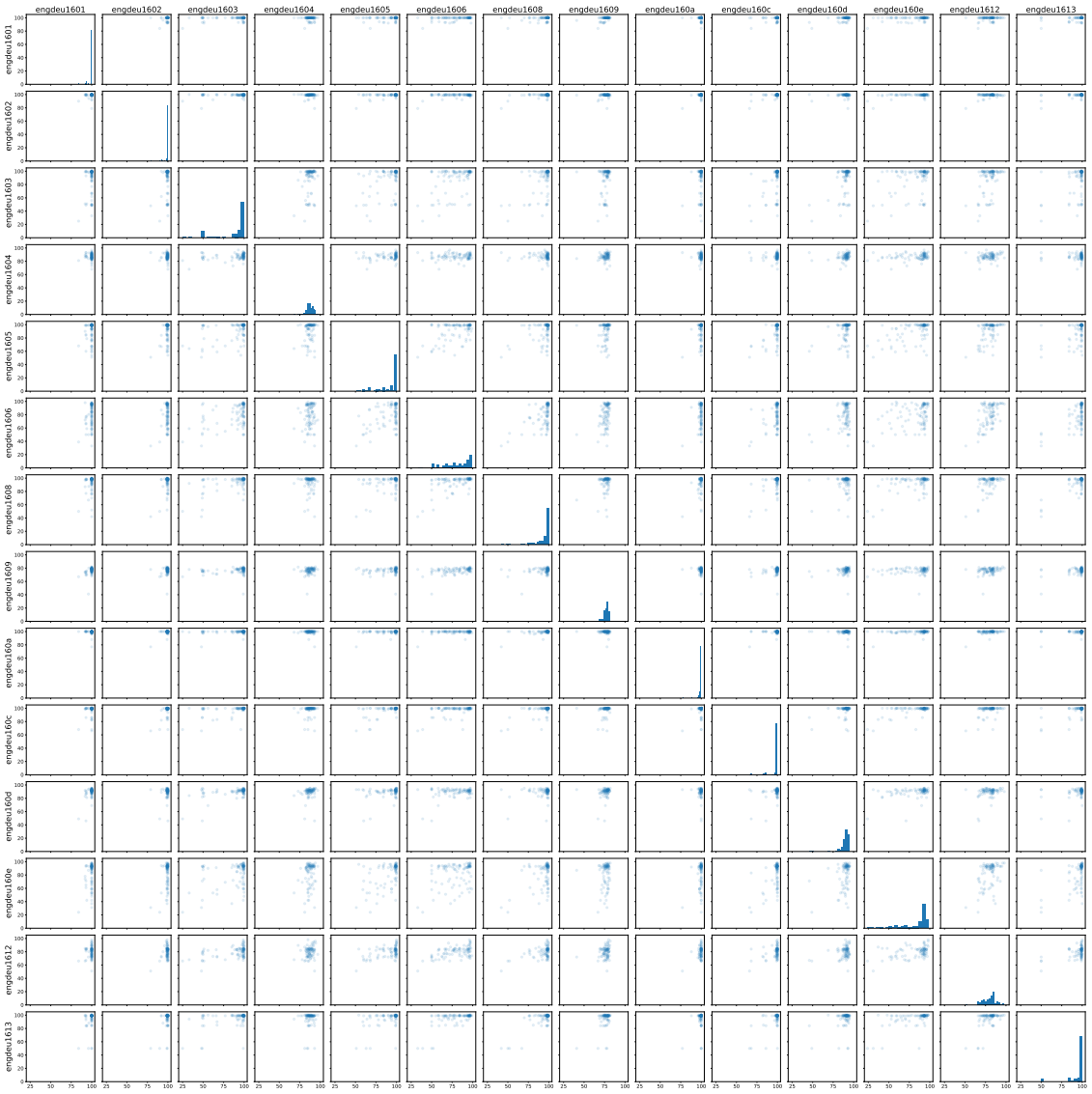


Figure 2: English–German

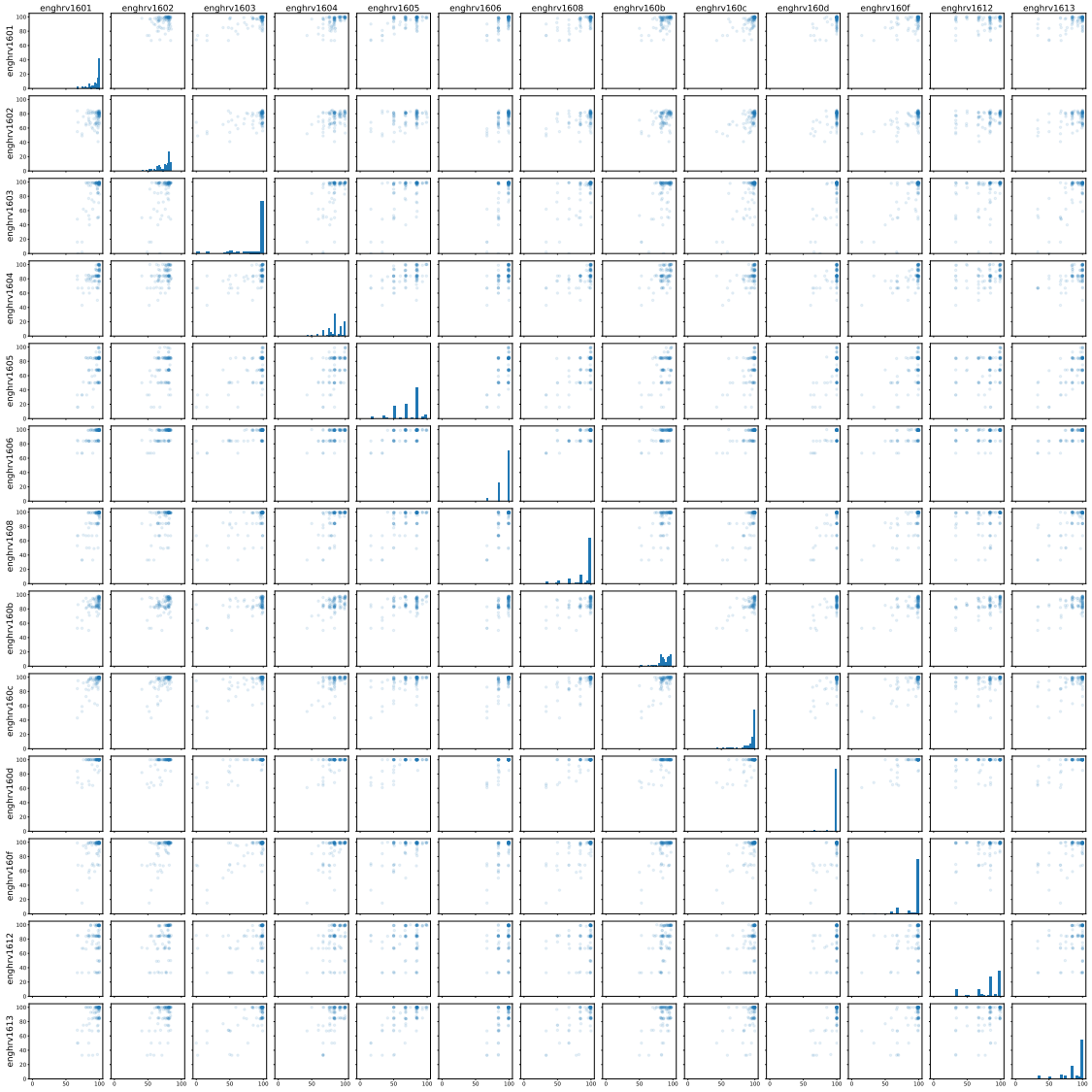


Figure 3: English–Croatian

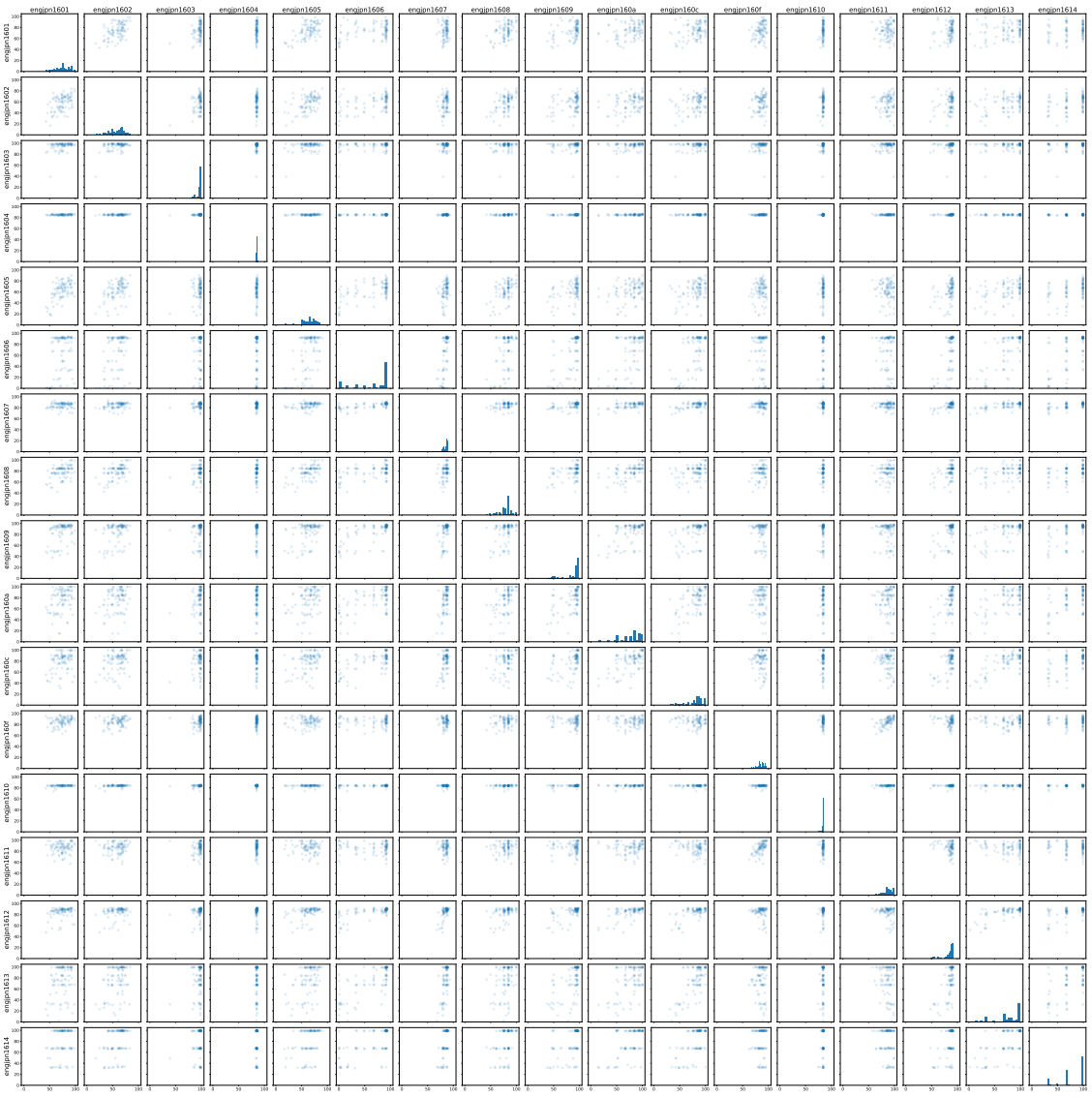


Figure 4: English–Japanese

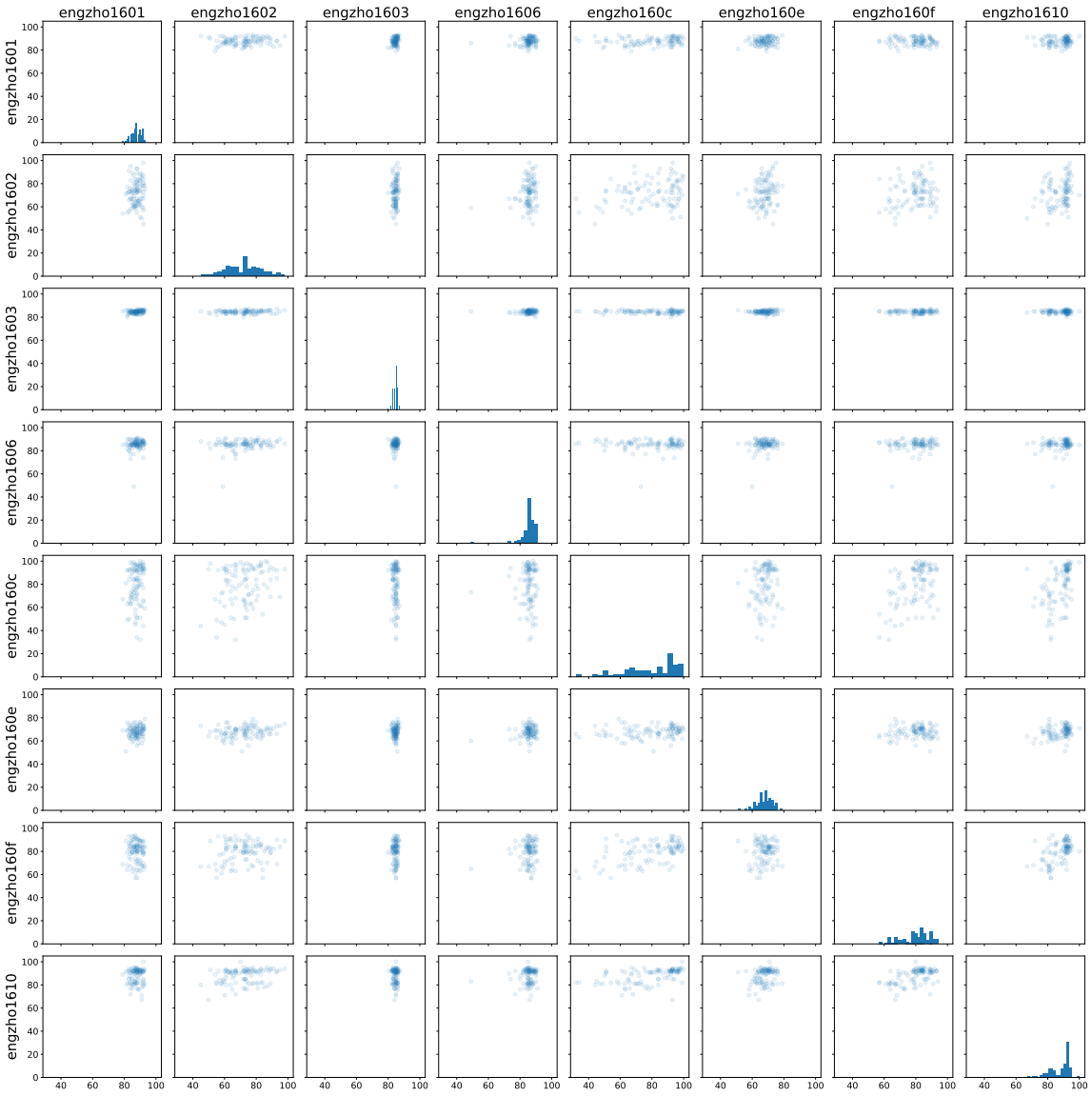


Figure 5: English–Chinese

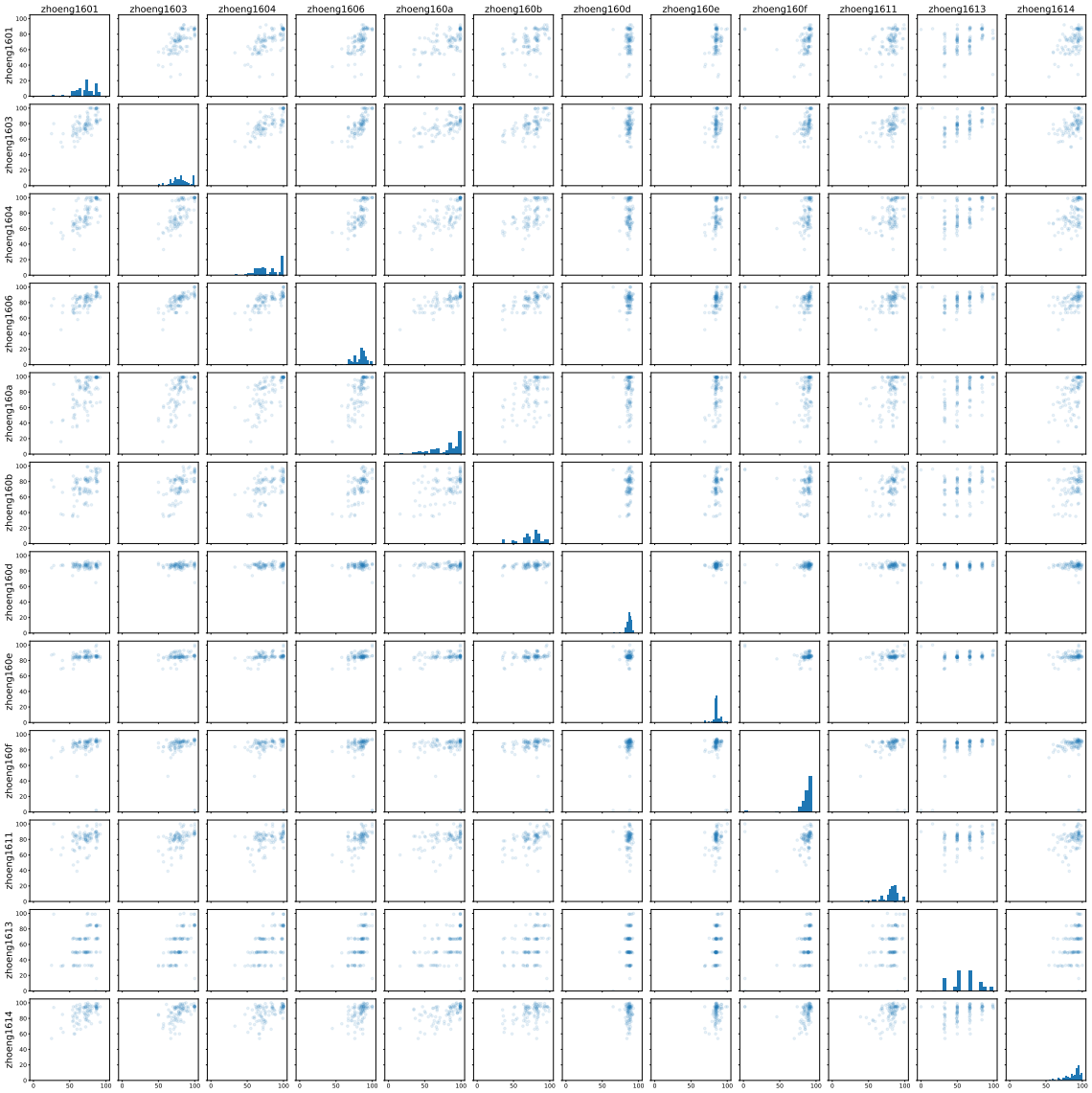


Figure 6: Chinese–English