In order to evaluate gender bias in Large Language Models, we define a new bias score and introduce a new dataset of natural contexts this bias score should be applied to. Our dataset aims to be more general than previous works, that were based on restrictive templates and restrictive sources of stereotypes. Combining our bias score with the causal tracing algorithm, we define a word stereotypical score and provide a list of male- and female-stereotyped words. We also use our dataset as finetuning data for the LoRA and DAMA methods: the bias reduction is slight, but similar to more targeted datasets, while preserving language modeling performance. Our code is available at `https://github.com/PaulMouret/master-thesis-gender-bias`.