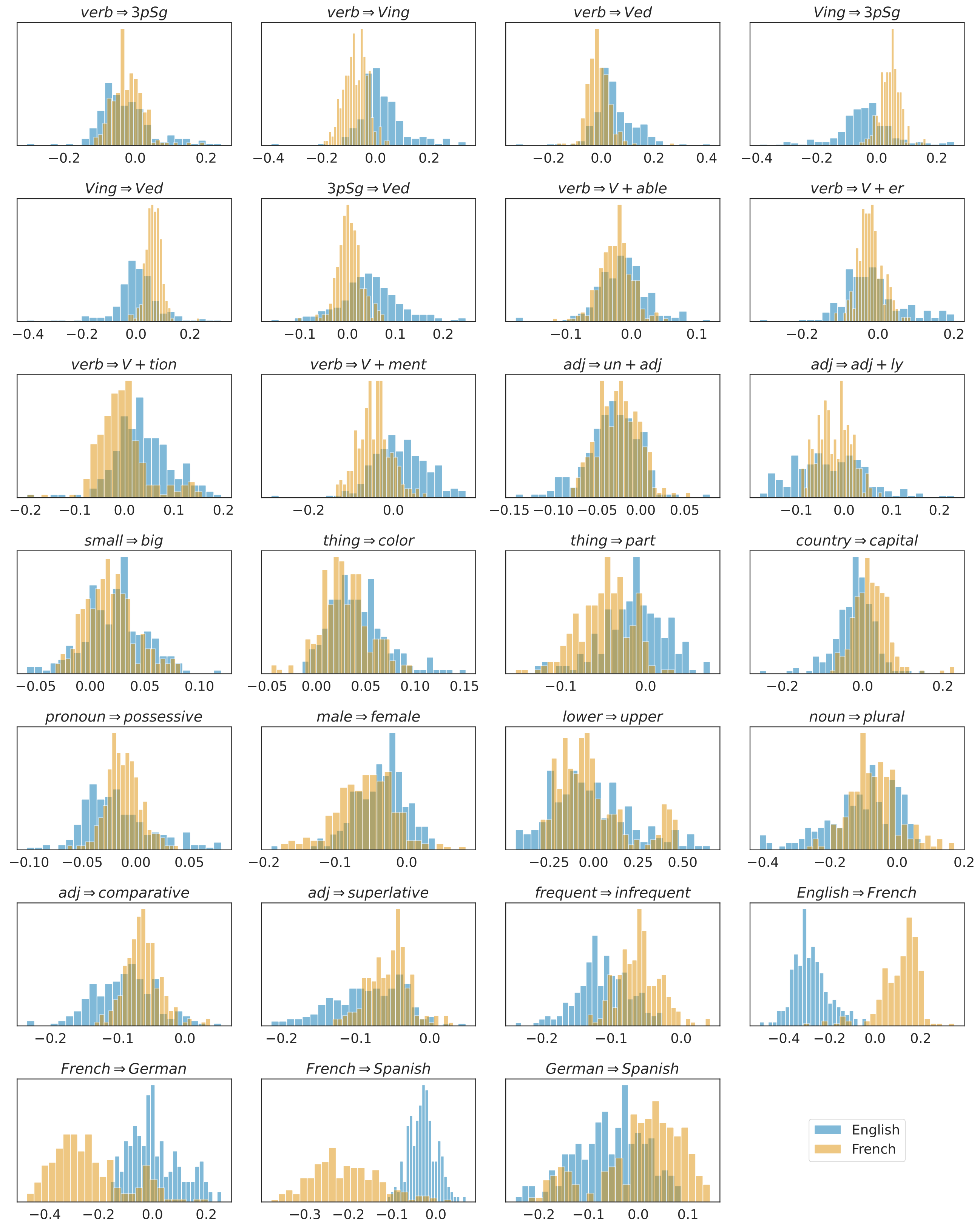

Figure 11. Histogram of $\bar{\gamma}_C^\top \lambda(x_j^{\texttt{en}})$ vs $\bar{\gamma}_C^\top \lambda(x_j^{\texttt{fr}})$ for all concepts C, where $\{x_j^{\texttt{en}}\}$ are random contexts from English Wikipedia, and $\{x_j^{\texttt{fr}}\}$ are random contexts from French Wikipedia.

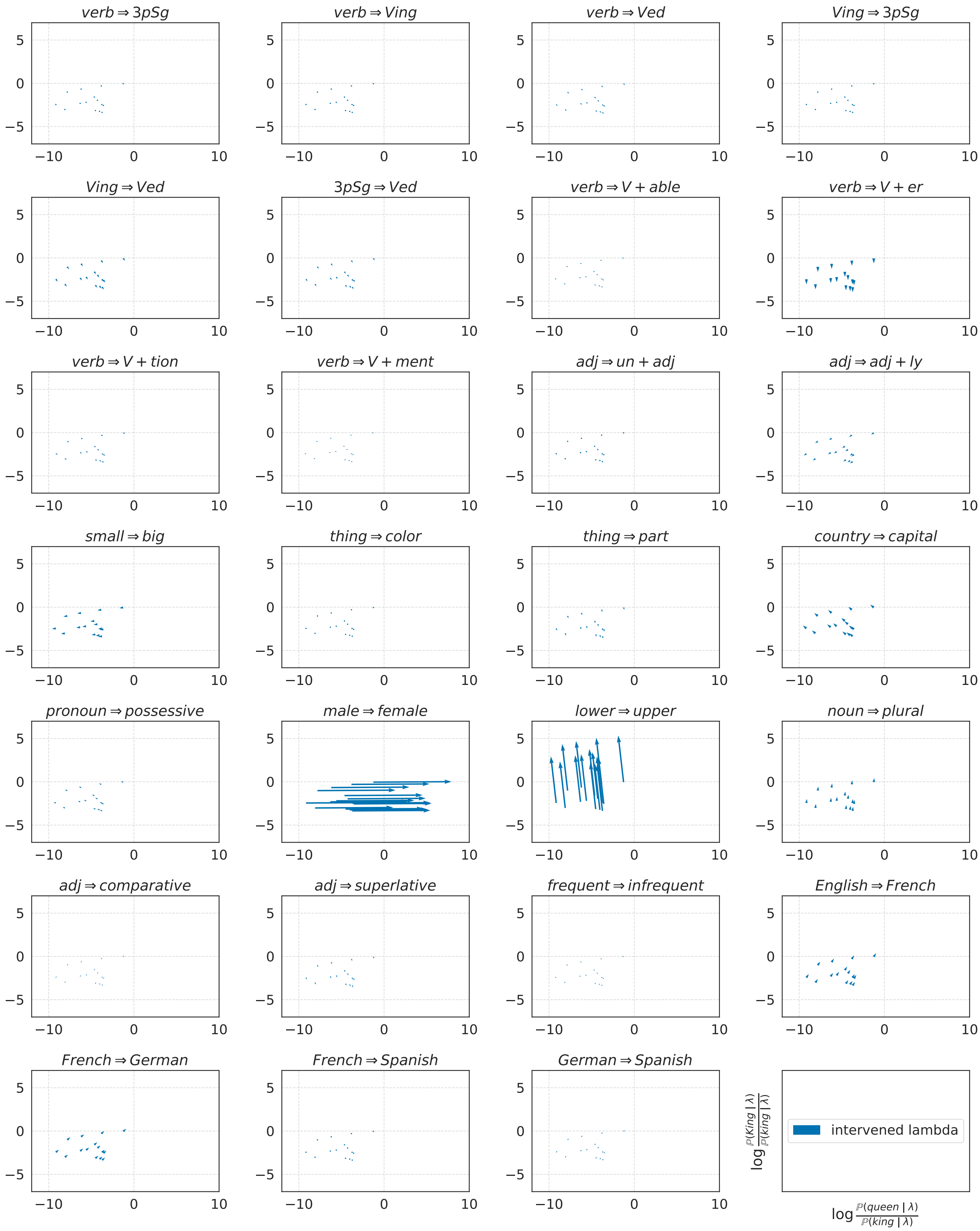

Figure 12. Change in $\log(\mathbb{P}(\text{“queen”} \mid x)/\mathbb{P}(\text{“king”} \mid x))$ and $\log(\mathbb{P}(\text{“King”} \mid x)/\mathbb{P}(\text{“king”} \mid x))$, after changing $\lambda(x_j)$ to $\lambda_{C,\alpha}(x_j)$ for $\alpha \in [0, 0.4]$ and any concept C. The starting point and ending point of each arrow correspond to the $\lambda(x_j)$ and $\lambda_{C,0.4}(x_j)$, respectively.

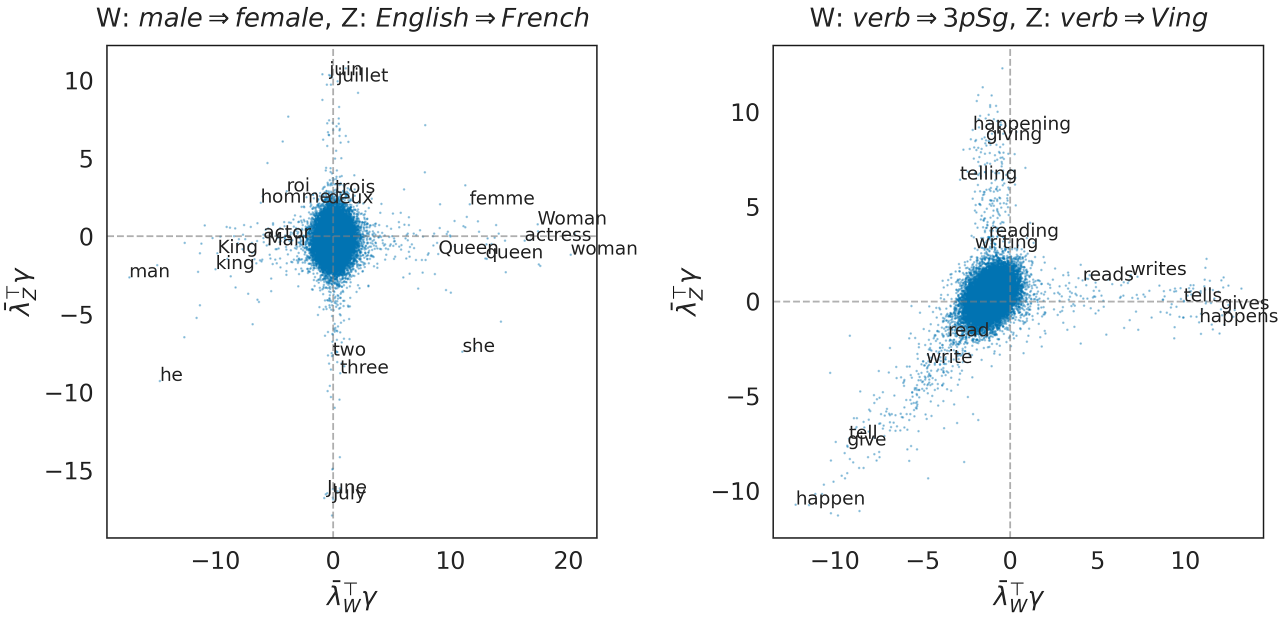

Figure 13. The left plot shows that $\bar{\lambda}_W^\top \gamma$ and $\bar{\lambda}_Z^\top \gamma$ are uncorrelated for the causally separable concepts $W = \texttt{male}{\Rightarrow}\texttt{female}$ and $Z = \texttt{English}{\Rightarrow}\texttt{French}$. On the other hand, the right plot shows that $\bar{\lambda}_W^\top \gamma$ and $\bar{\lambda}_Z^\top \gamma$ are correlated for the non-causally separable concepts $W = \texttt{verb}{\Rightarrow}\texttt{3pSg}$ and $Z = \texttt{verb}{\Rightarrow}\texttt{Ving}$. Each dot corresponds to the unembedding vector γ for each token in the vocabulary.