



Figure 5: These two plots show the same candidates (also randomly sampled 30% of all candidates) used in Figure 3 except now the y axis is phonological similarity. The computation of phonological similarity for the plot ① deploys only the pronunciation in Mandarin Chinese; whereas plot ② deploys pronunciations in all CJK languages as in FASpell.

## A Preliminary experiments

This appendix is about the preliminary experiments used to justify some settings of FASpell in the main article.

### A.1 Using only one CJK language

The preliminary experiment using only one CJK language (e.g., Mandarin Chinese) in the computation of phonological similarity results in a confidence-similarity graph like the plot ① in Figure 5. As pointed out in the main article, one can observe that the values of such similarity are limited to a few discrete values, which makes it harder to find a smooth filtering curve. Therefore, in FASpell, as shown in the plot ② in Figure 5, all CJK languages are used, which makes the graph resemble those in Figure 3 more.

### A.2 Using tree edit distance

The preliminary experiment using tree edit distance is evaluated on  $Tst_{15}$  following the same

Table 7: The performance of FASpell on  $Tst_{15}$  using tree edit distance.

Level	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Detection	74.4	67.6	60.4	63.8
Correction	73.7	66.2	59.1	62.4

Table 8: The performance of FASpell on  $Tst_{ocr}$  when the masked language model is fine-tuned.

Level	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Detection	14.4	73.1	14.4	24.1
Correction	13.8	70.1	13.8	23.1

configurations in Table 4. As shown in Table 7, the performance is just slightly different from its counterpart in Table 6. Yet, the filtering speed is 32,143 ms/sent in each round, which is much slower than that in Table 5. Therefore, FASpell is configured to use edit distance of strings instead of tree edit distance.

### A.3 Fine-tuning for OCR data

As shown in Table 8, the performance is much worse (in comparison with Table 6) if the masked language model is fine-tuned on  $Trn_{ocr}$  plus 65,900 more sentences harvested from OCR results. Therefore, for formal experiments in the main article, FASpell is configured to use the pre-trained masked language model without any fine-tuning.