

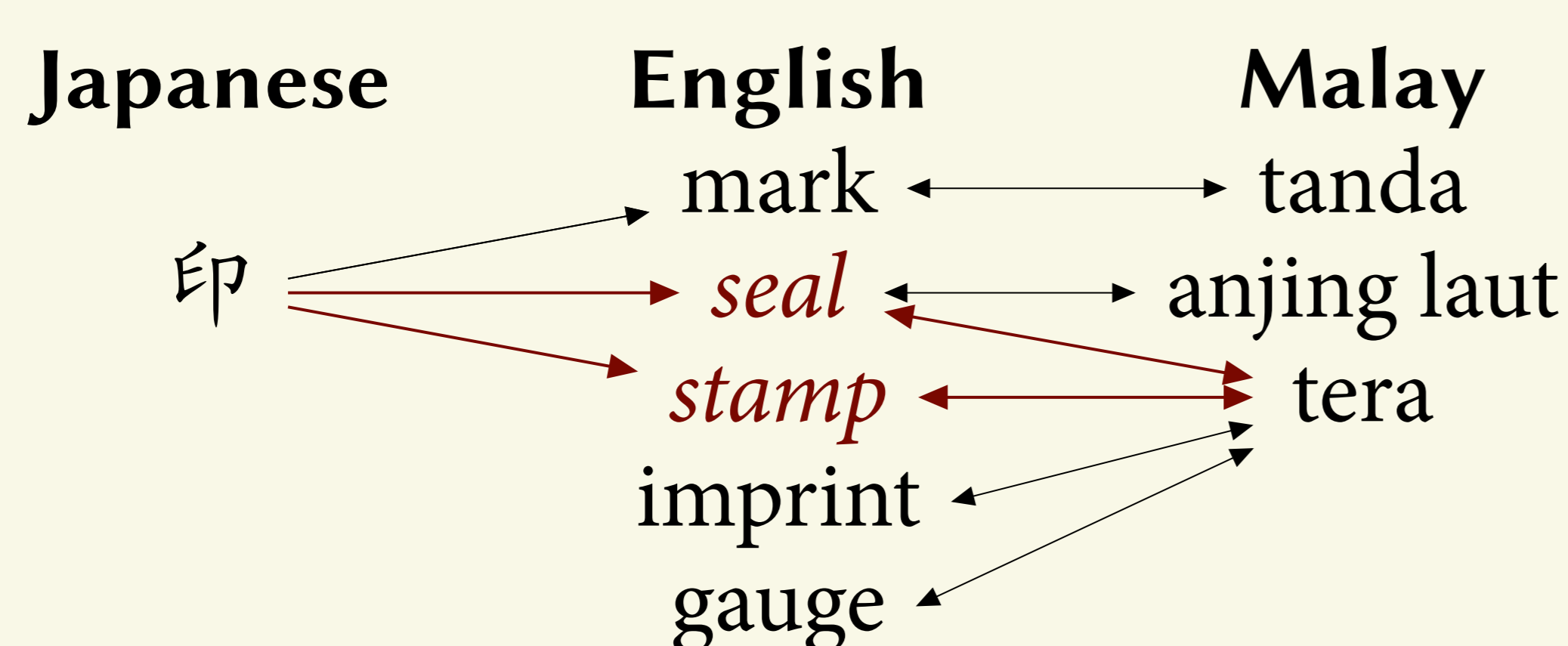
Low-Cost Construction of a Multilingual Lexicon from Bilingual Lists

Introduction

- ▶ Bilingual MRDs are good resources for building multilingual lexicons, but heterogeneous structures
- ▶ Lowest common denominator: list of **source language item** → **target language item(s)**
- ▶ **Proposal:** Multilingual lexicon construction using only simple bilingual lists

One-time Inverse Consultation [1]

- ▶ Generates a bilingual lexicon for new language pair from existing bilingual lists
- ▶ JP-EN, EN-MS, MS-EN lexicons ⇒ JP-MS

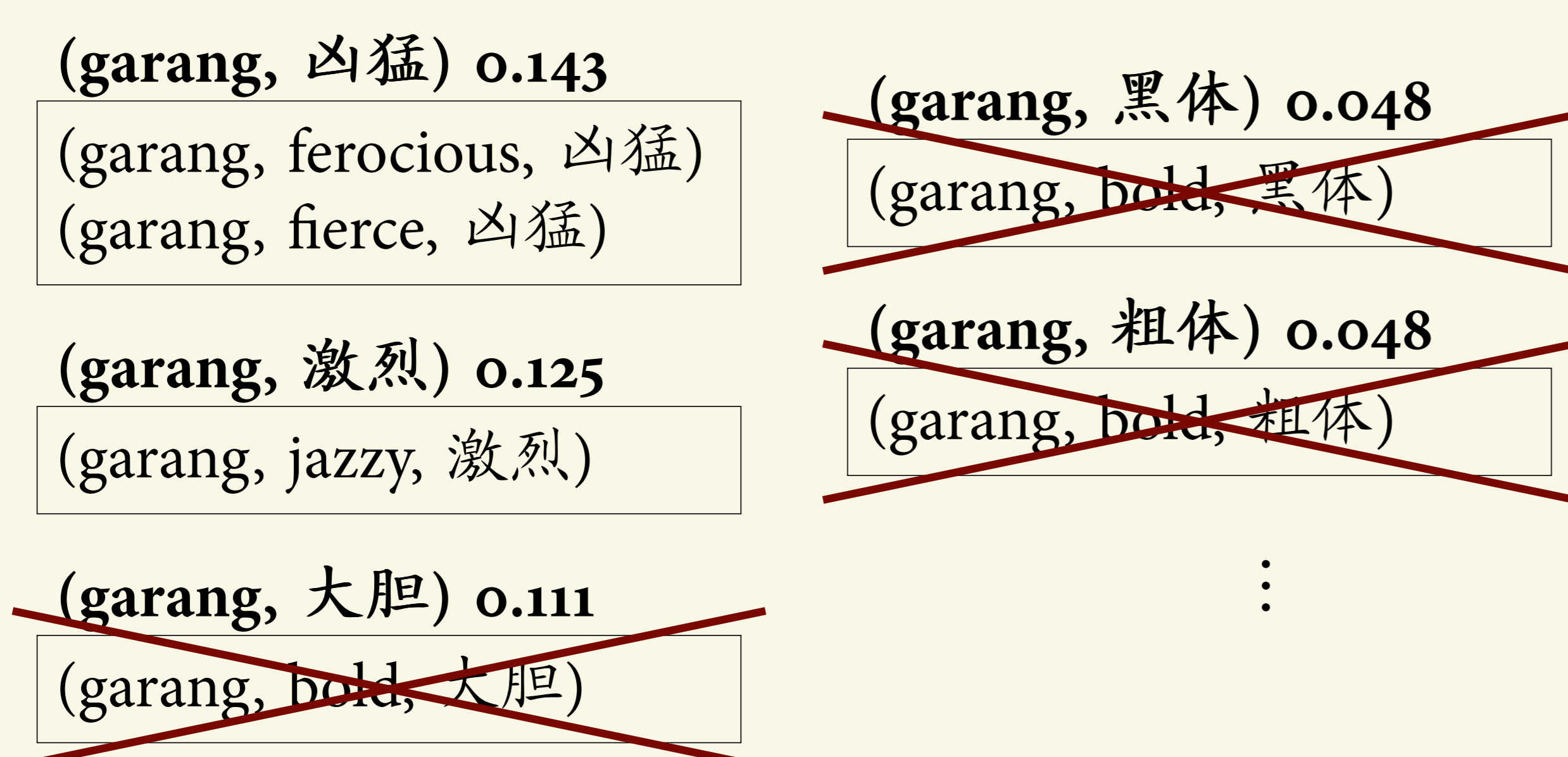


$$\text{score}(\text{'tera'}) = 2 \times \frac{|\mathbb{E}_1 \cap \mathbb{E}_2|}{|\mathbb{E}_1| + |\mathbb{E}_2|} = 2 \times \frac{2}{3 + 4} = 0.57$$

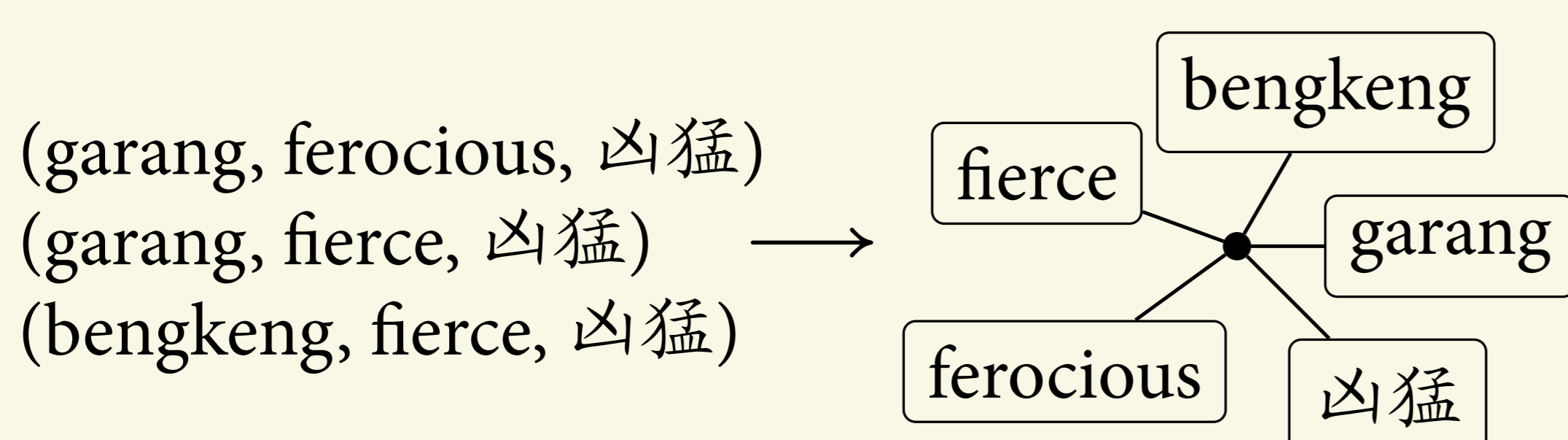
∴ '印' ↔ 'tera' is most likely valid

Merging Translation Triples into Sets

- ▶ (Example: Malay-English-Chinese)
- ▶ Retain OTIC 'middle' language links
- ▶ For each 'head' language LI, discard triples with score $< \alpha X$ or $\text{score}^2 < \beta X$, where $X = \text{max score of all triples containing that LI}$

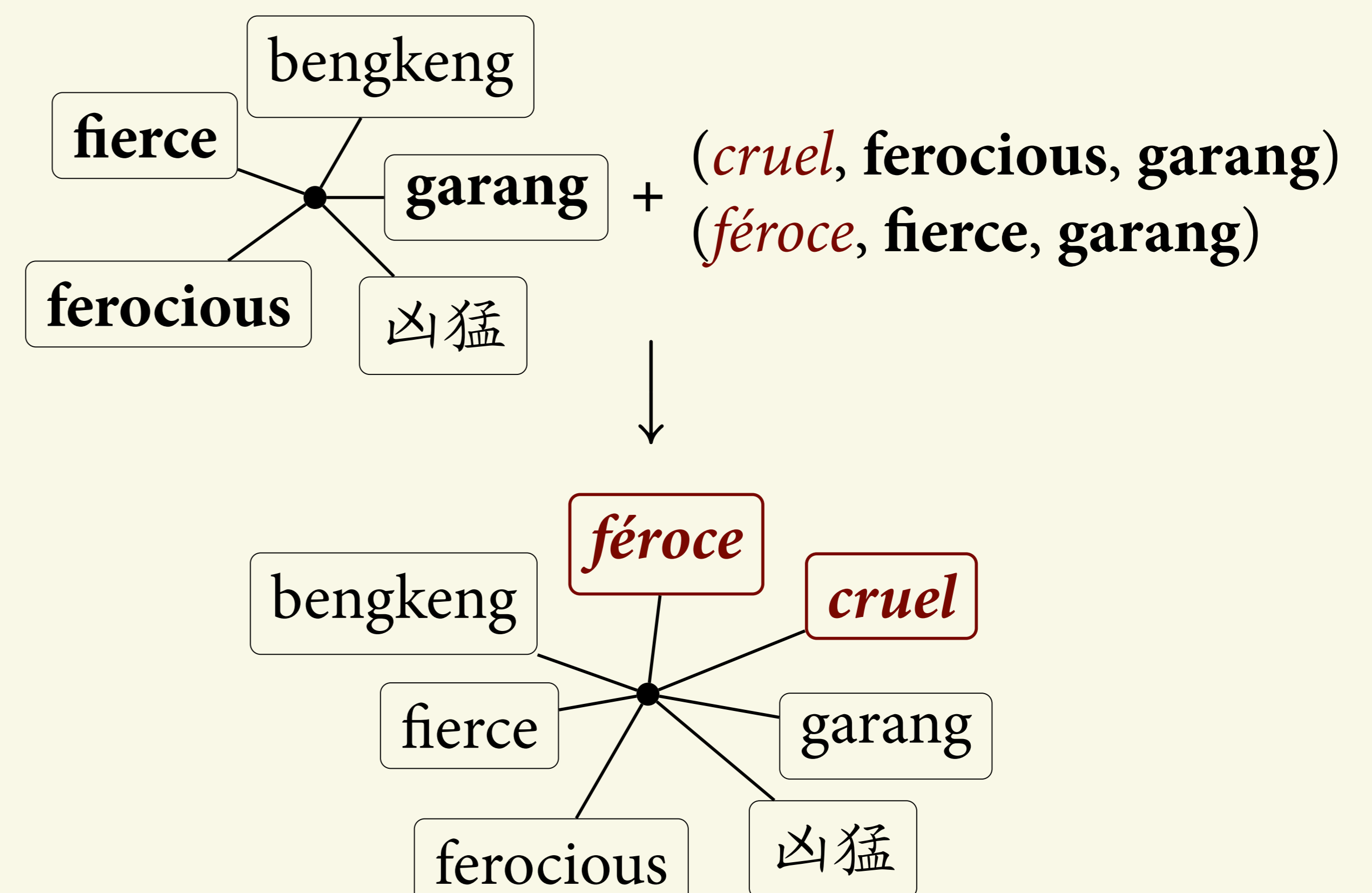


- ▶ Merge all triples with common bilingual pairs

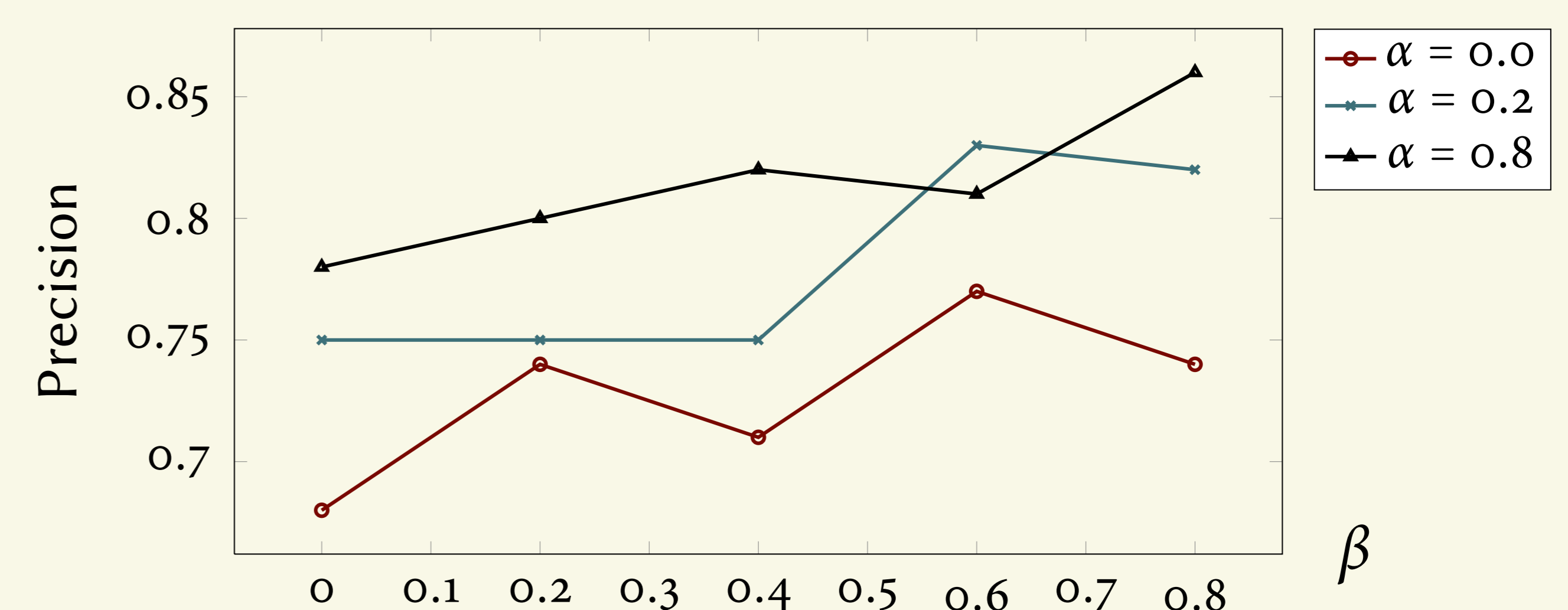


Adding a New Language

- ▶ (Example: Malay-English-Chinese + French)
- ▶ Construct also French-English-Malay triples
- ▶ Add French members to existing M-E-C clusters with common English & Malay members



Precision of 100 Random Translation Sets



- ▶ Precision generally around 0.70–0.82; max 0.86

F₁ and Rand Index of Selected Translation Sets

- ▶ Evaluating accuracy of sets with polysemous 'middle' language members, e.g. 'plant', 'target'

Test word	Rand Index		F ₁		Best accuracy when	
	min	max	min	max	α	β
'bank'	0.417	0.611	0.588	0.632	0.6	0.4
'plant'	0.818	0.927	0.809	0.913	0.6	0.2
'target'	0.821	1.000	0.902	1.000	0.4	0.2
'letter'	0.709	0.818	0.724	0.792	0.8	0.2

Discussion and Conclusion

- ▶ Low thresholds (α, β): more coverage; low precision
- ▶ High thresholds: good precision; low coverage
- ▶ $\alpha = 0.6, \beta = 0.2$ gives good trade-off between coverage, precision and recall
- ▶ Results are encouraging for such simple input data!
- ▶ **Future plan:** Integrate lexicon into an MT system with WSD

References

- [1] F. Bond and K. Ogura. "Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary". In: *Language Resources and Evaluation* 42 (2008), pp. 127–136.

