“A Quantitative Analysis of Reordering Phenomena”

Alexandra Birch, Phil Blunsom, and
Miles Osborne

presented at

WMT 2009

Greg Hanneman
11-734: Advanced MT Seminar
March 31, 2010

lti

Carnegie Mellon

Opening Questions

- Who's winning the fight between lexicalized reordering models and SCFGs?
  - Why is it that Hiero does better than Moses for some language pairs?
  - Why is it that Moses does better than Hiero for other language pairs?
- In comparing the amount of reordering between two language pairs, can we do better than comparing BLEU scores?

Outline

- Quantifying reordering in a language pair
- Reordering in manual data
- Reordering in MT systems
- Conclusions and discussion

Quantifying Reordering

- Reordering: Binary swap between two adjacent blocks or sibling nodes
- Extract reorderings from sentence pair
- Score according to RQuantity metric (range 0 to $\sum_{i=2}^{I} i$) [Birch et al., EMNLP 2008]

$$\frac{\sum_{r \in R} |r_A| + |r_B|}{|I|}$$

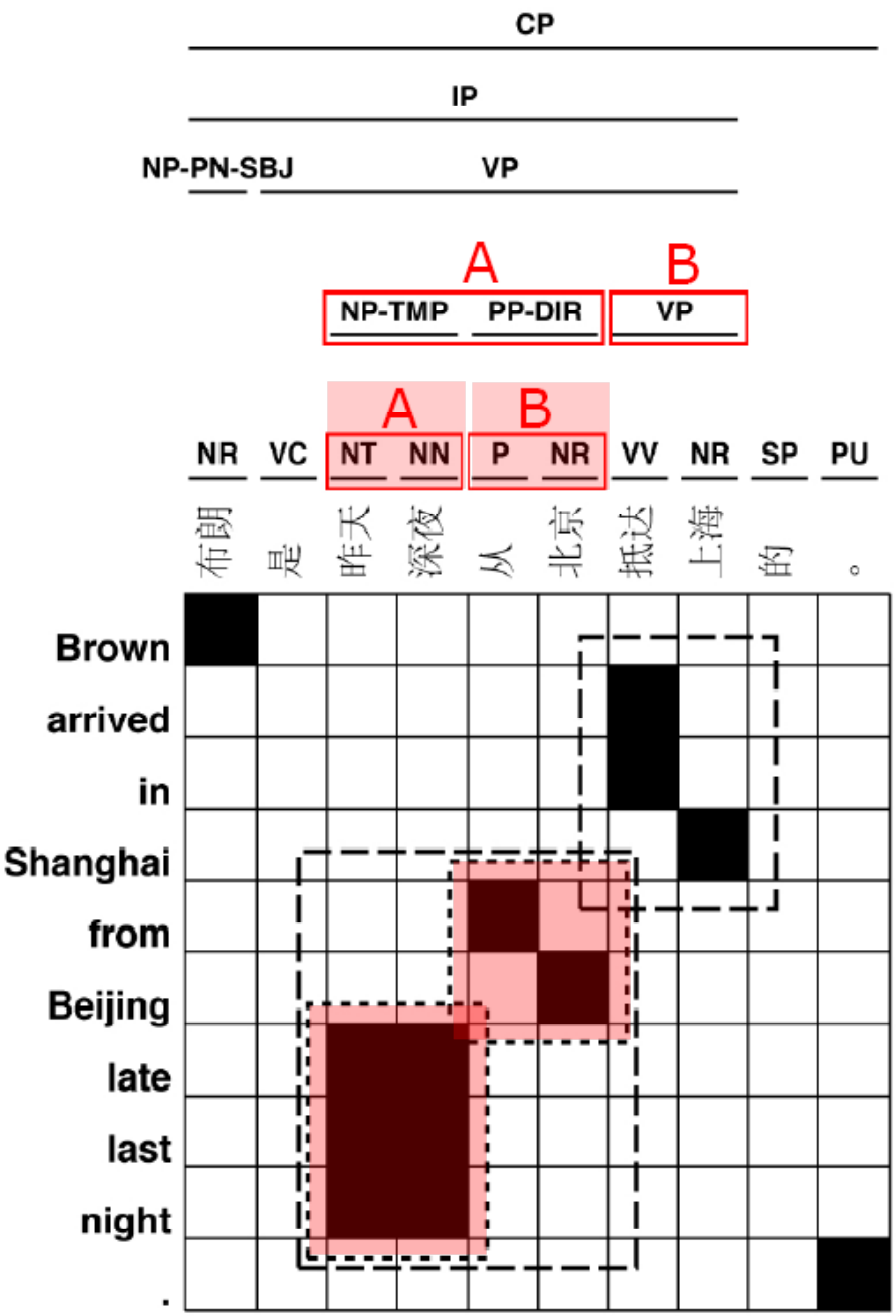
Reordering blocks r_A and r_B
Set of reorderings R in sentence pair
Target sentence of length $|I|$

$$\frac{\sum_{r \in R} |r_A| + |r_B|}{|I|}$$

Reordering r_1:

$|r_A| = 2$

$|r_B| = 2$



$$\frac{\sum_{r \in R} |r_A| + |r_B|}{|I|}$$

Reordering r_1:

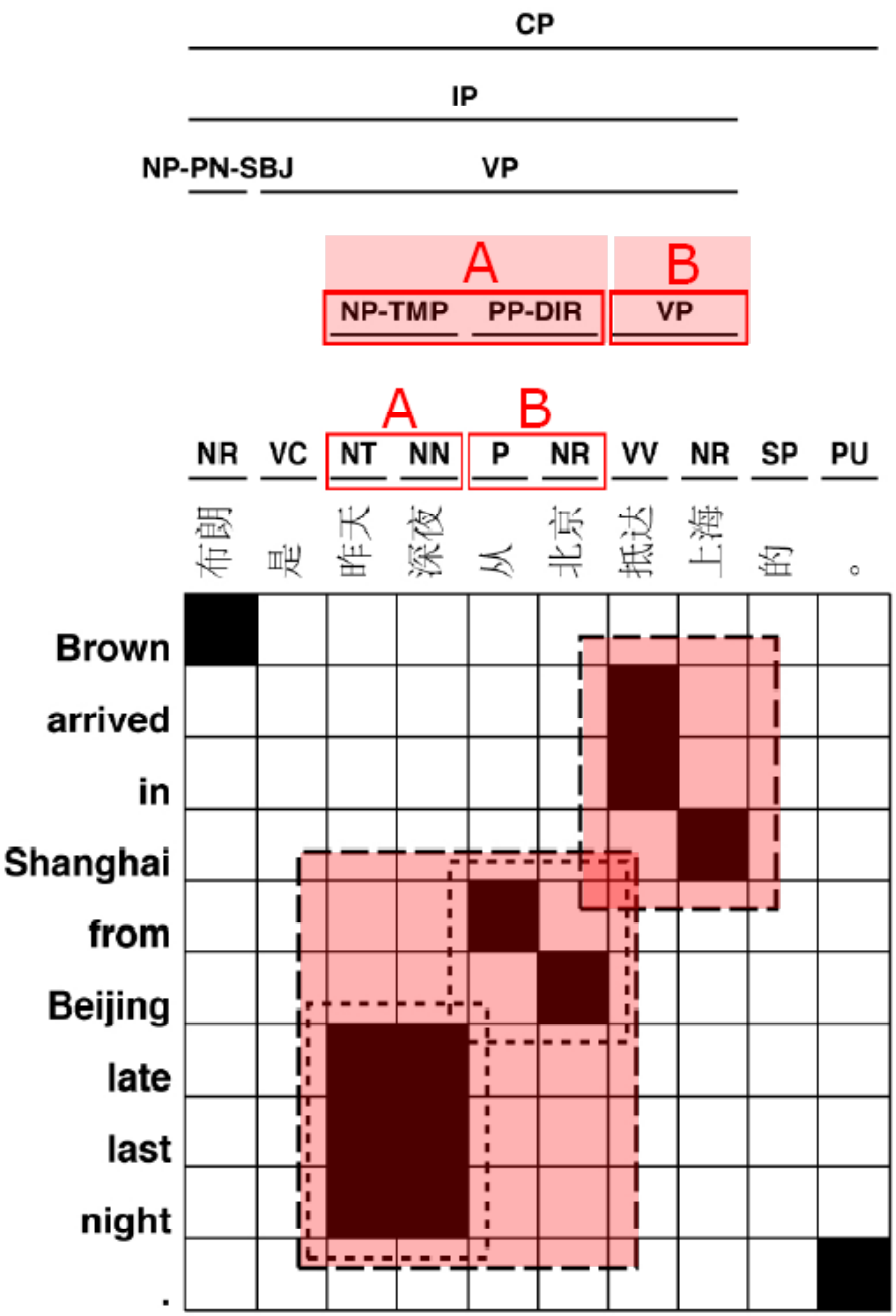
$|r_A| = 2$

$|r_B| = 2$

Reordering r_2:

$|r_A| = 4$

$|r_B| = 2$



$$\frac{\sum_{r \in R} |r_A| + |r_B|}{|I|}$$

Reordering r_1:

$|r_A| = 2$

$|r_B| = 2$

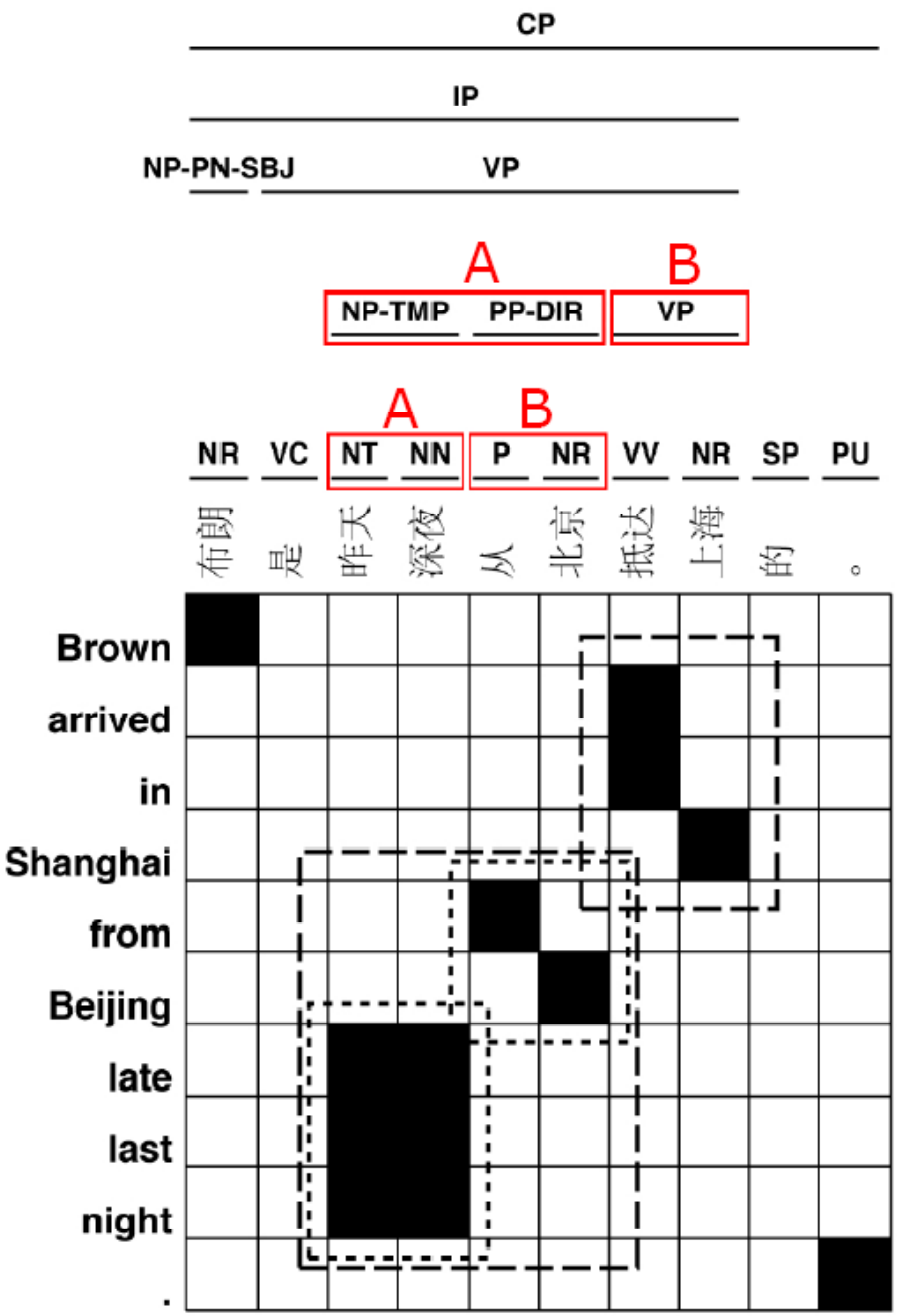
Reordering r_2:

$|r_A| = 4$

$|r_B| = 2$

RQuantity:

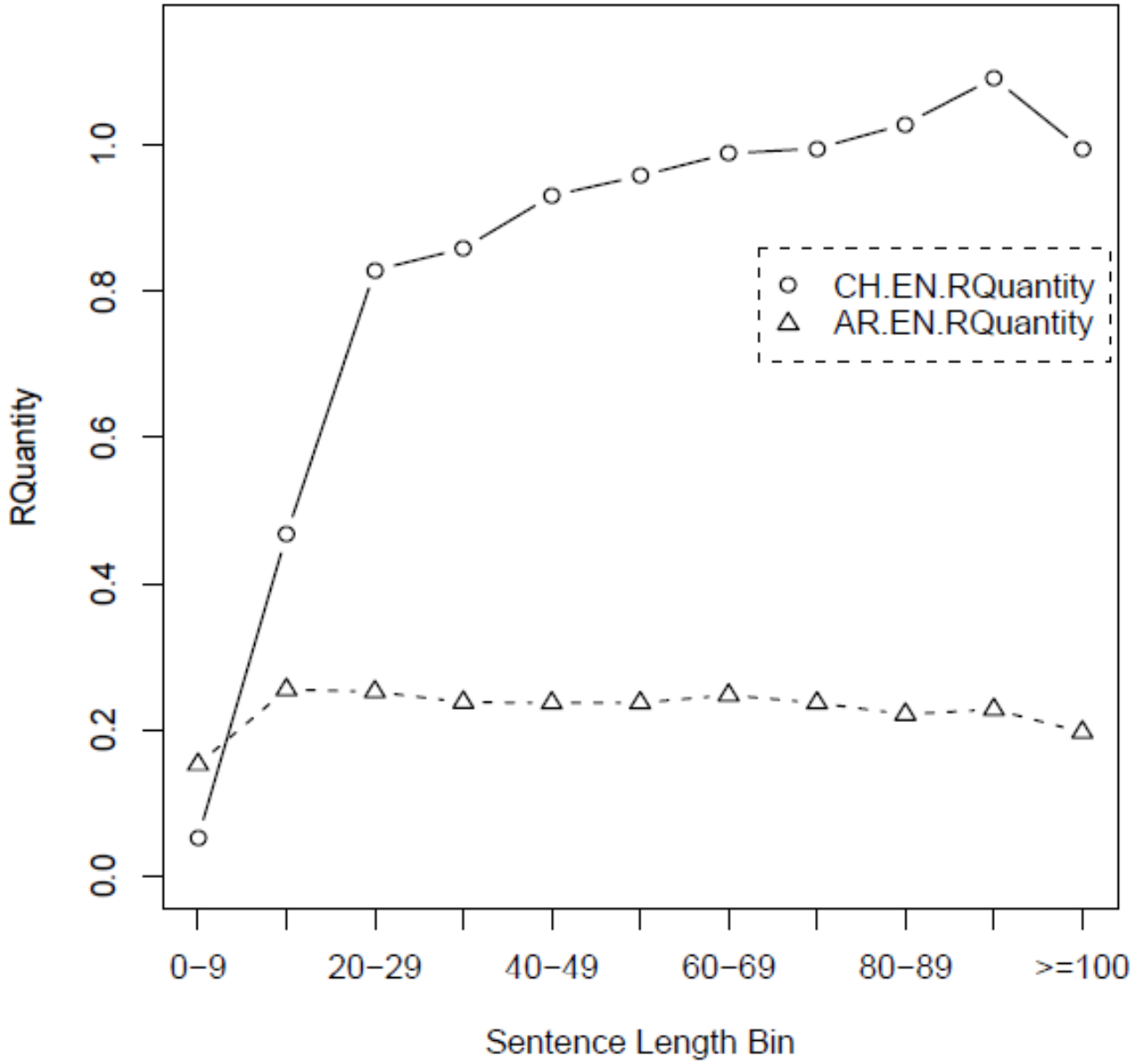
$$\frac{2+2+2+4}{10} = \frac{10}{10} = 1$$



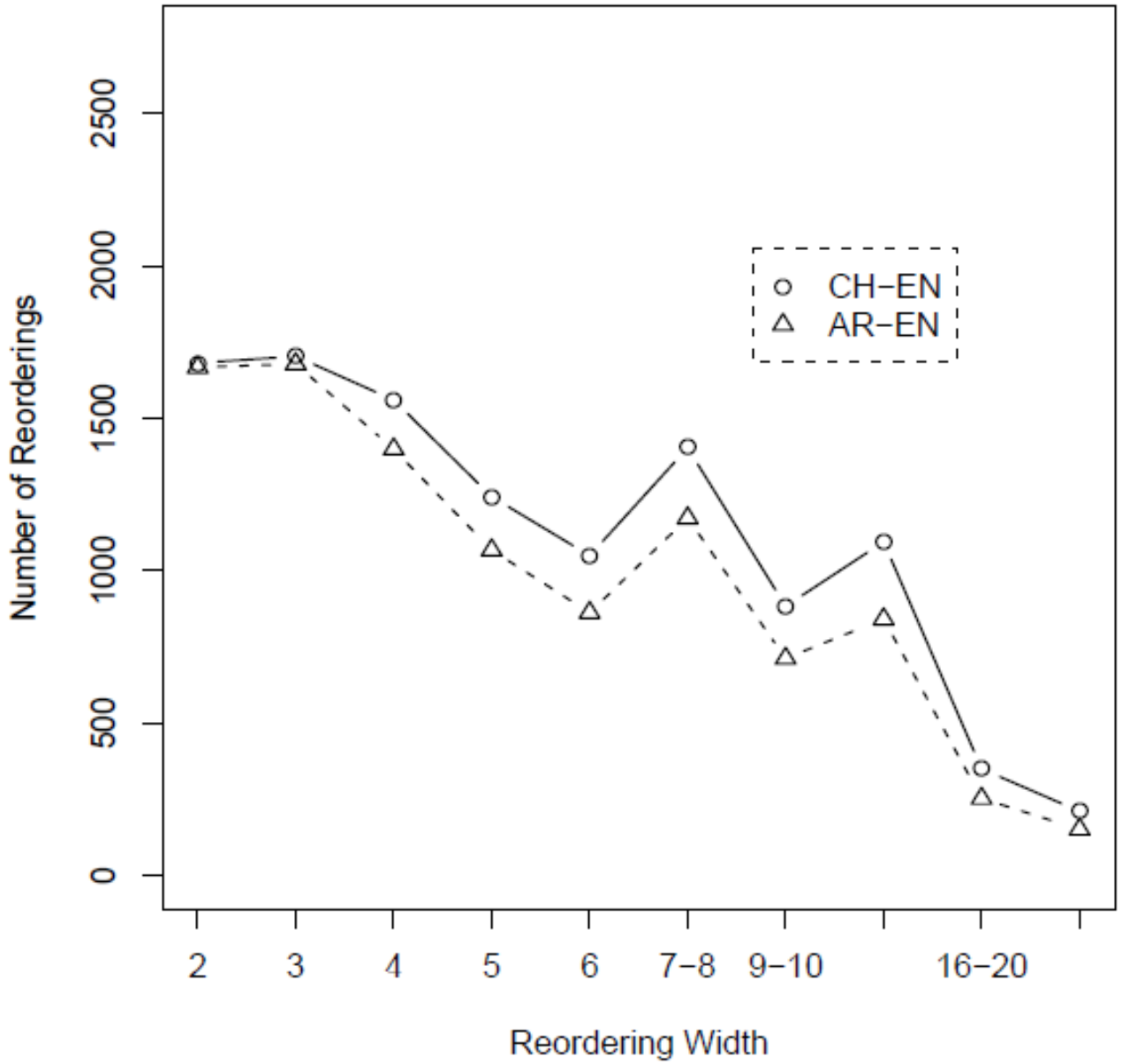
Reordering in Manual Data

- Gold-standard parses and alignments for
  - 3380 Chinese–English sentences
  - 4337 Arabic–English sentences
- Computed amount, width, and syntactic category of reordering
- Results mostly what you expect

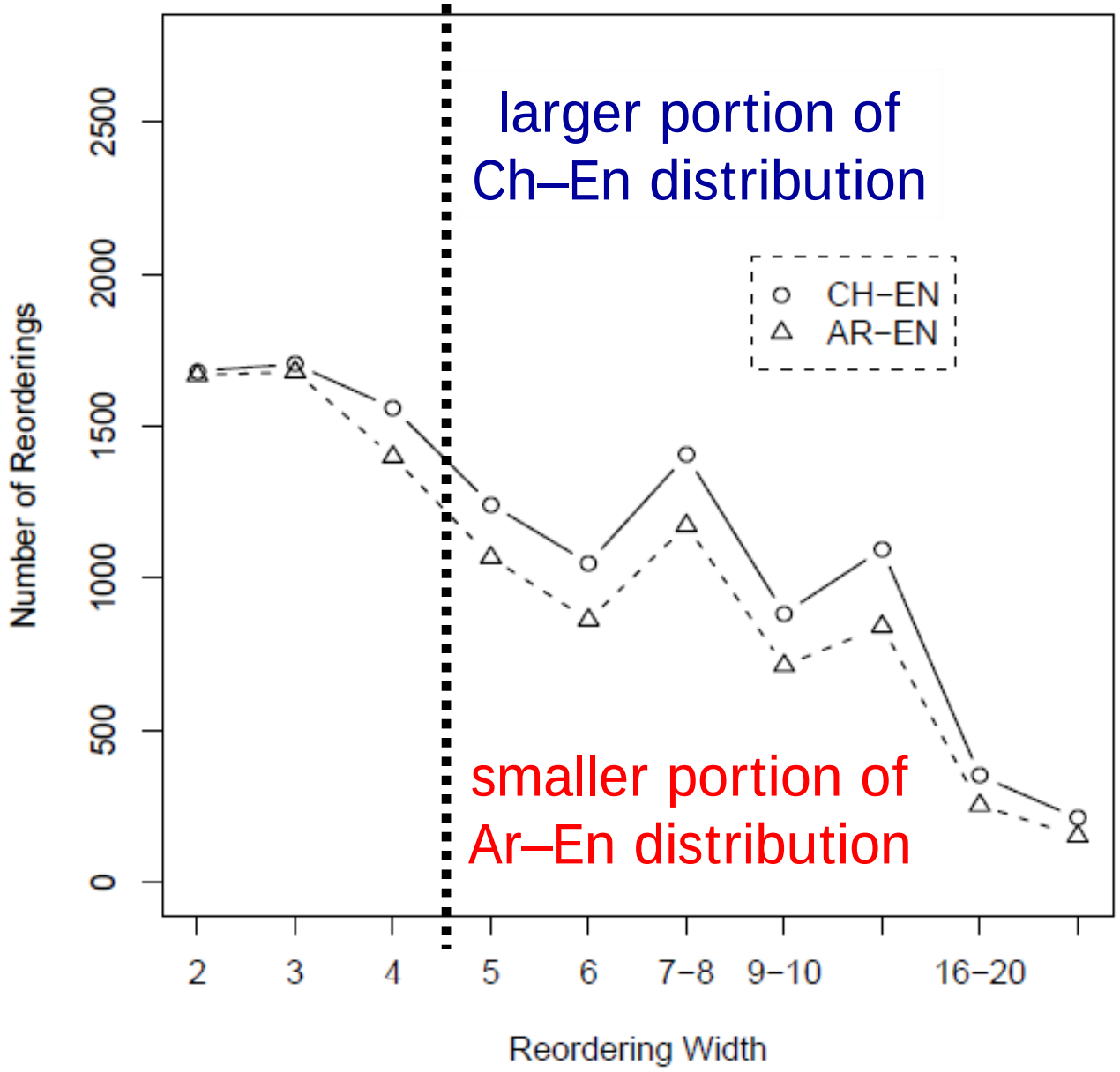
(1) Ch–En reorders more than Ar–En



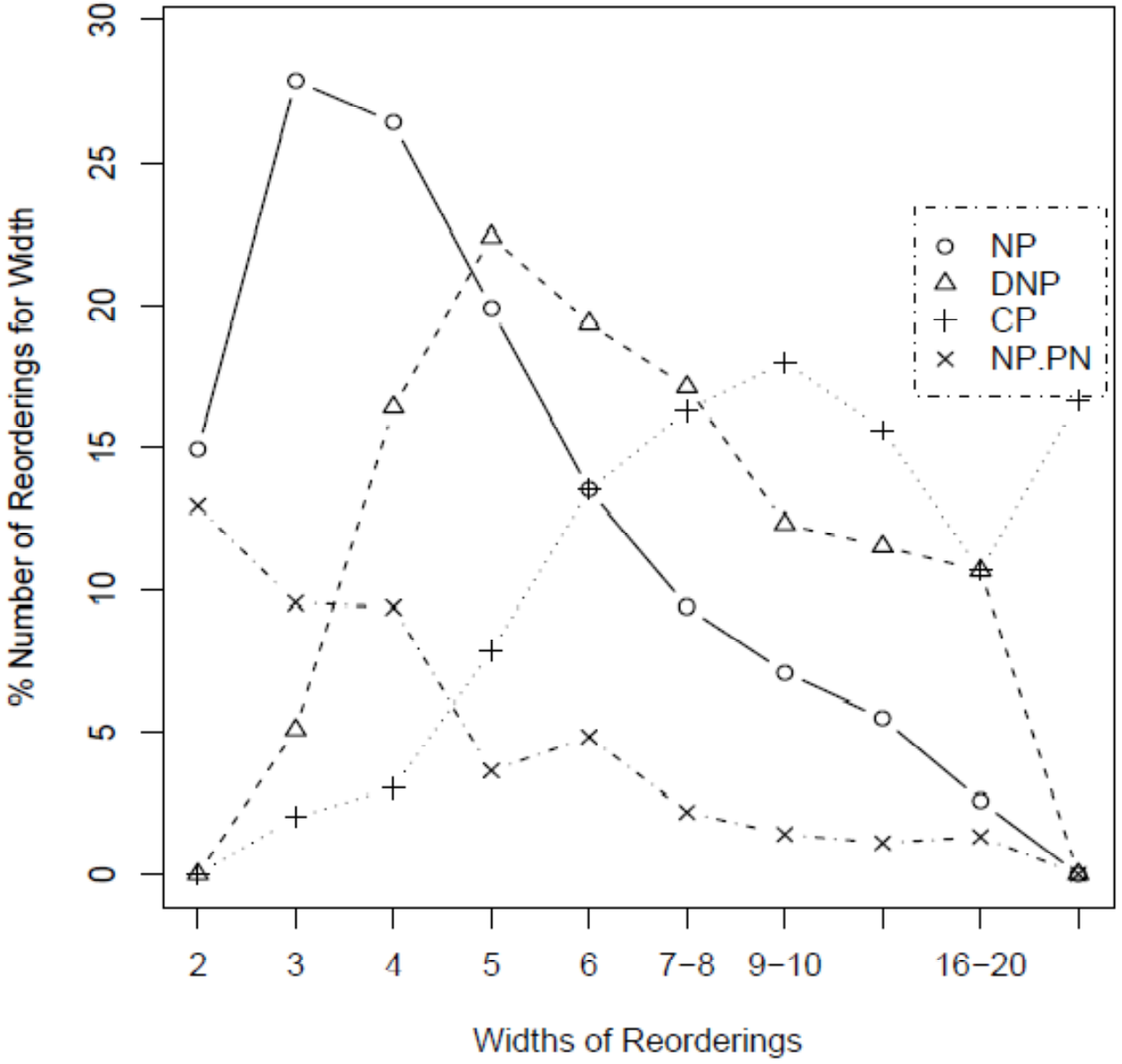
(2) Ch–En reorders longer than Ar–En



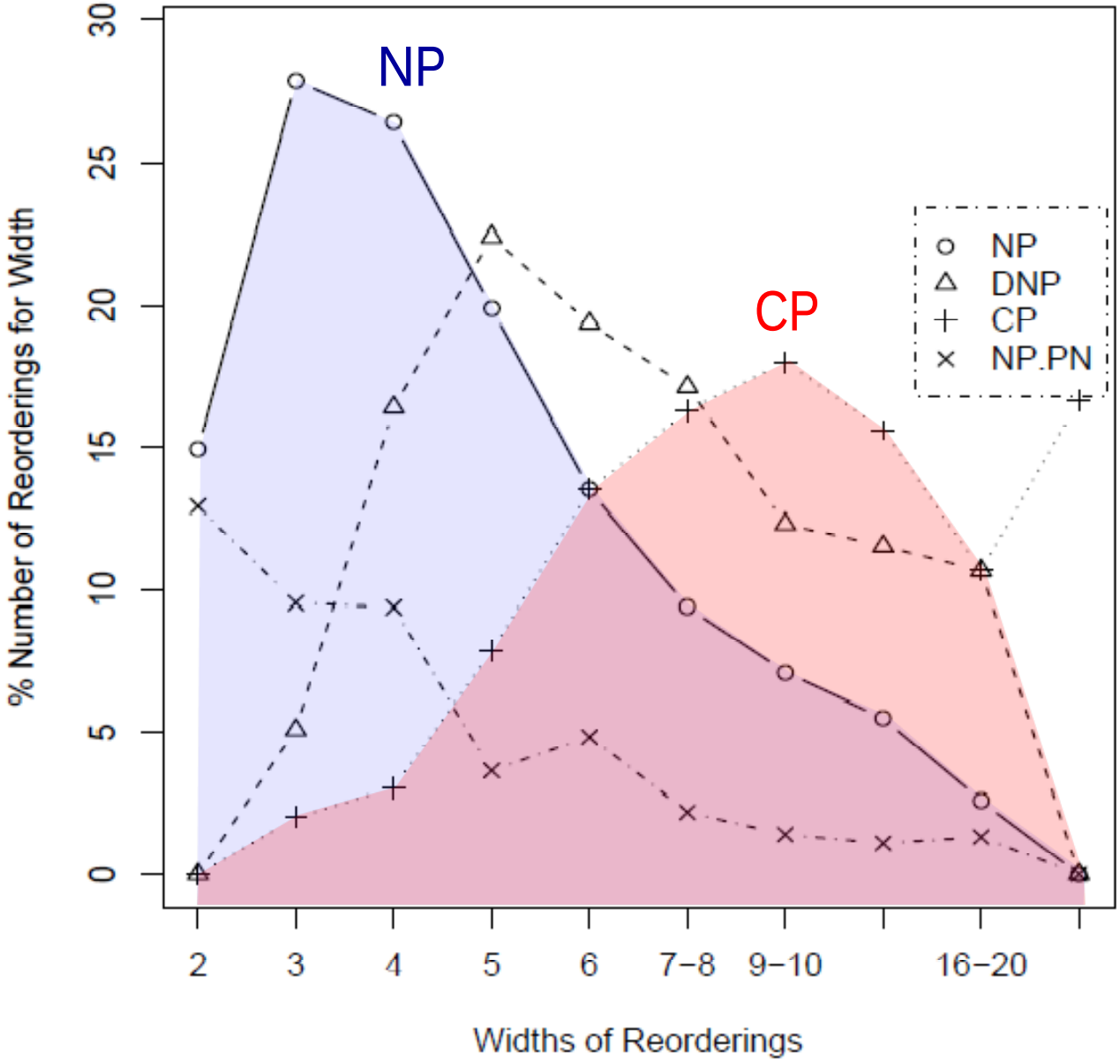
(2) Ch–En reorders longer than Ar–En



(3) Constituents reorder differently (Ch–En)



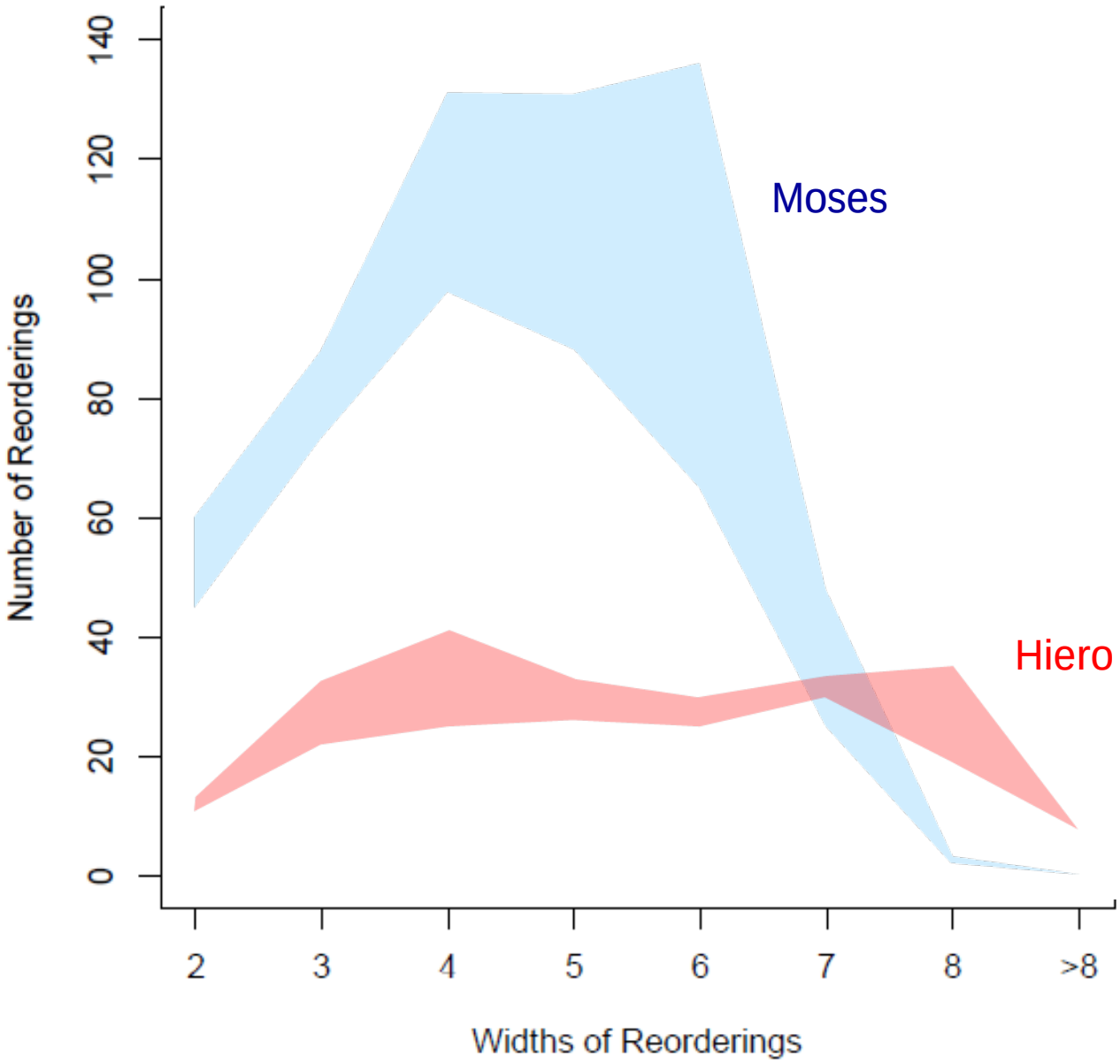
(3) Constituents reorder differently (Ch–En)



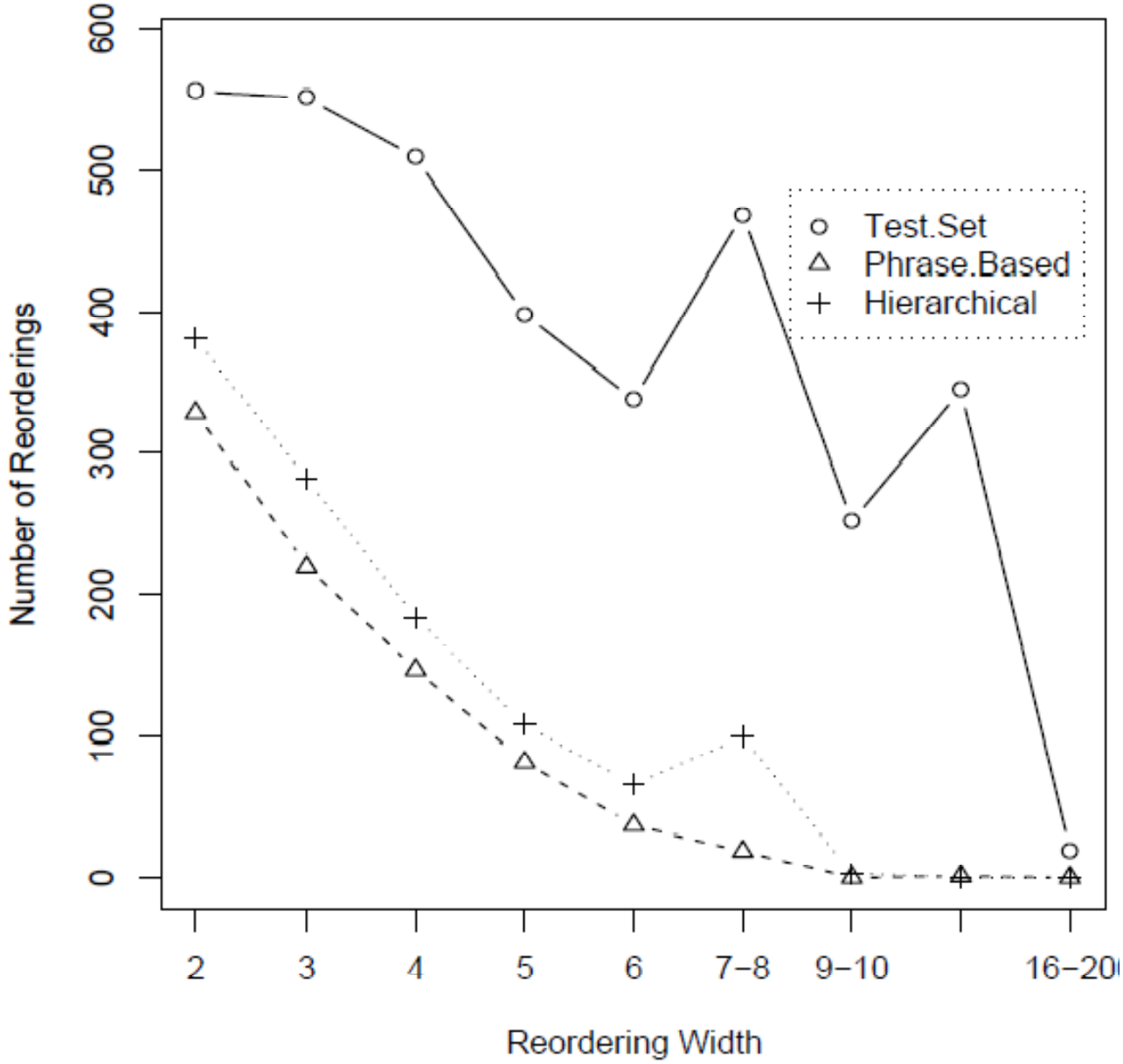
Reordering in MT Systems

- Partitioned 20- to 39-word sentences equally by RQuantity (none, low, mid, high)
- Translated with Moses and Hiero
- Computed characteristics of MT system reordering compared to reference (Fight!)

(1) Number of reorderings (Ch–En)



(2) Recall of reorderings (Ch–En)



Main Conclusions

- Chinese–English has more medium- and long-range reorderings
- Arabic–English has more short-range reorderings (as a proportion of total)
- Moses is better at the short range
- Hiero is better at the medium range
- Neither is good at the long range!

Other Points

- Constraints helpful when reordering beyond a small window, but locally they're worse than exhaustive search
- BLEU is not good at assessing reorderings because it only penalizes the boundary
- RQuantity useful for categorizing system and language pair behavior? [Koehn et al., MT Summit 2009]

Discussion Questions

- A lot of these graphs are “fun facts” – can they be put to any useful work?
  - Syntax-based reordering?
  - MT system construction/modeling decisions?
- Role of search space and constraints?
  - Brute force vs. constrained search vs. sparse data estimation