# The Effect of Unaligned Words in Phrase Based Statistical Machine Translation

DTC- Seminar

Francisco Guzman

October 2, 2009

Tec de Monterrey

CCIR
Centro de Computación Inteligente y Robótica

# Outline

1) Research stay @ Carnegie Mellon
2) Brief intro to PBSMT
   - Word Alignments
   - Phrase Extraction
3) Quality of word alignments
4) Analysis of phrase extraction
   - Manual evaluation of phrases
   - The effect of unaligned words
5) Lessons learned: Improving translation
6) Conclusions

# My Research Stay

- At Carnegie Mellon University (Pittsburgh)
- August 08 – August 09
- Statistical Machine Translation
- Collaborated mainly with Stephan Vogel
- Attended several courses and seminars
- Got involved in important projects (GALE, Avenue)
- Learned **a lot!!!**

# MT @ CMU

- Carnegie Mellon University (Pittsburgh)
  - School of Computer Science
    - Language Technologies Institute
      - Avenue Group (Xfer)
      - Interlingua
      - Example Based
    - Inter/ACT
      - CMU SMT

# CMU SMT

**CMU SMT**

Stephan Vogel

- Phrase Based/ Syntax Augmented
- Moses/STTK decoder/SAMT decoder

**CMU Avenue**

Alon Lavie

- Syntax Based Rules + Statistical Engine
- Xfer decoder

# Gale Project

- Funded by DARPA
- Three main consortia: Nightingale (SRI), Rosetta (IBM), Agile(BBN)
- Rosetta Team:
  - CMU
  - Apptek
  - IBM
  - JHU
  - Columbia
  - Stanford
  - RTWH
- GALE Conference in Tampa (May 09)

# Evaluation Campaigns

- Gale P3.5 (Chinese)
  - great
- NIST Eval (Arabic)
  - Not so
- Gale P4 (Arabic)
  - In progress.

# Publications

- **Gale-Book (to be published)**

  Word Alignment Revisited: we present the summary of our work in word alignment analysis. Shed some light in how Discriminative Models can be used to boost WA performance

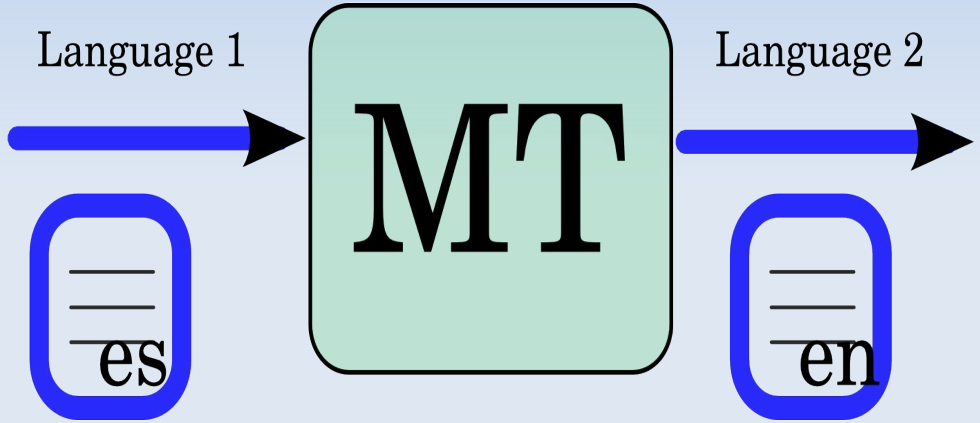
- **MTSummitXII**

  We follow an analysis from word alignment to phrase extraction in detail, and reveal how the former affects the latter. We also perform a manual evaluation that unveils the impact of unaligned words in exacted phrase pairs.

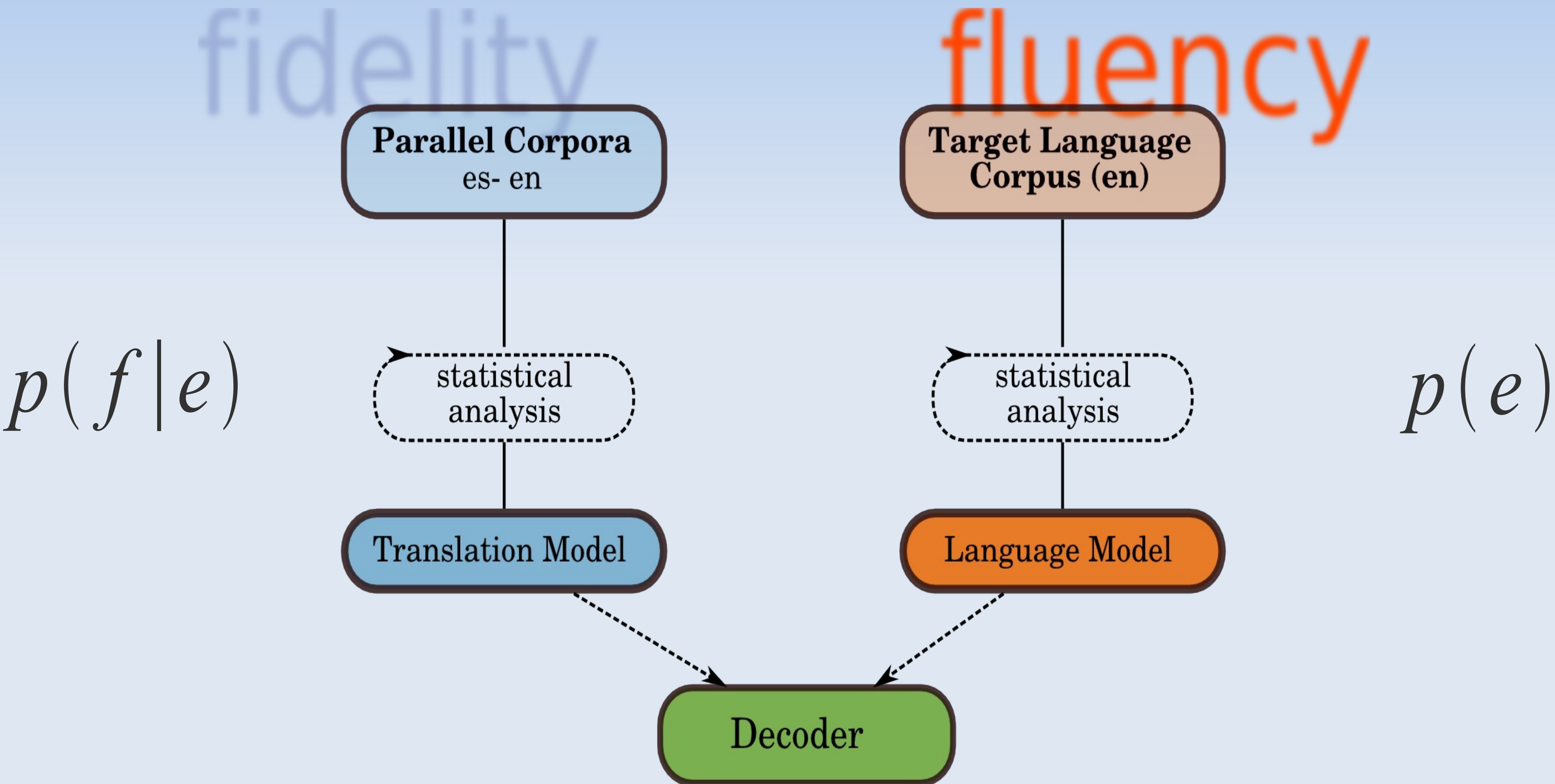
# Brief Introduction to SMT

# Statistical Machine Translation

- Started in 80's (IBM Candide).
- Phrase-based concept introduced (Och)
- Koehn et al. (2003) – Introduced the concept of Phrase based Statistical Machine Translation



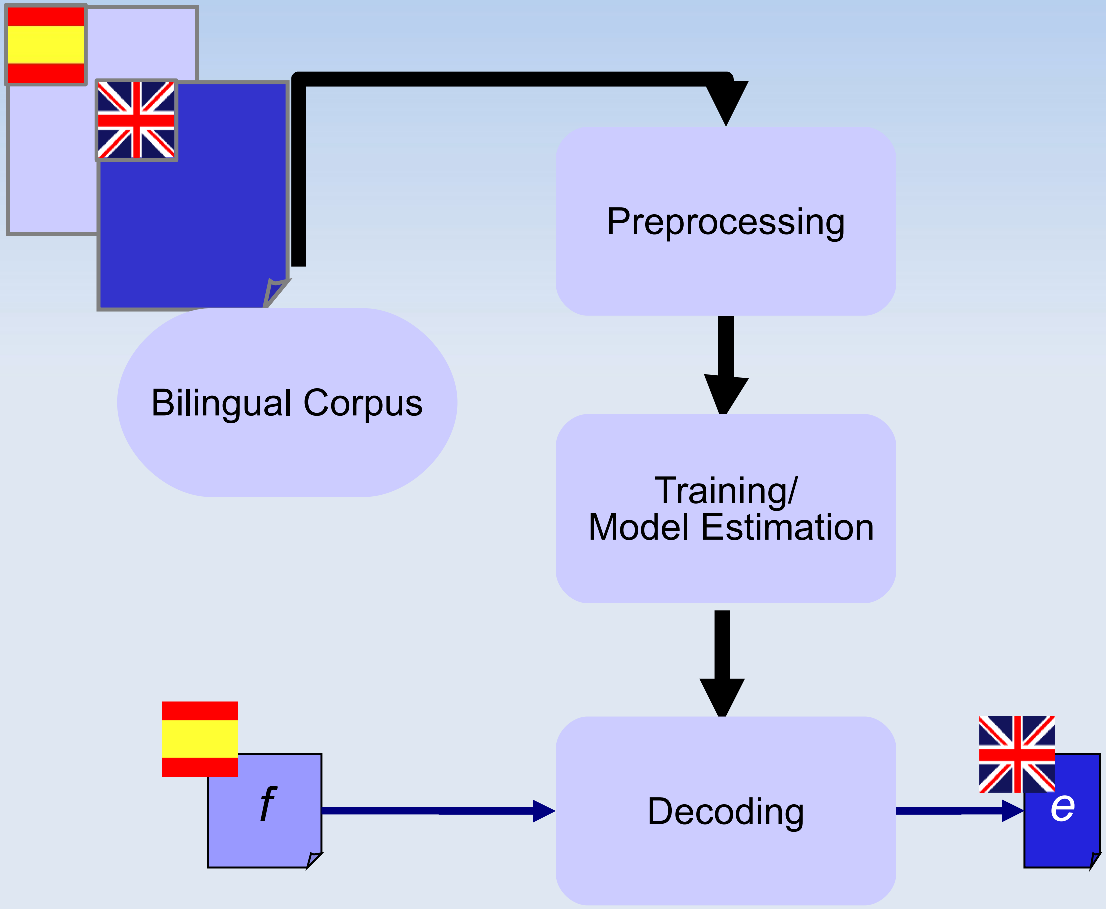
# How does it work?




$$\operatorname{argmax}_e p(e|f) = \operatorname{argmax}_e p(f|e)\, p(e)$$

# Translation model

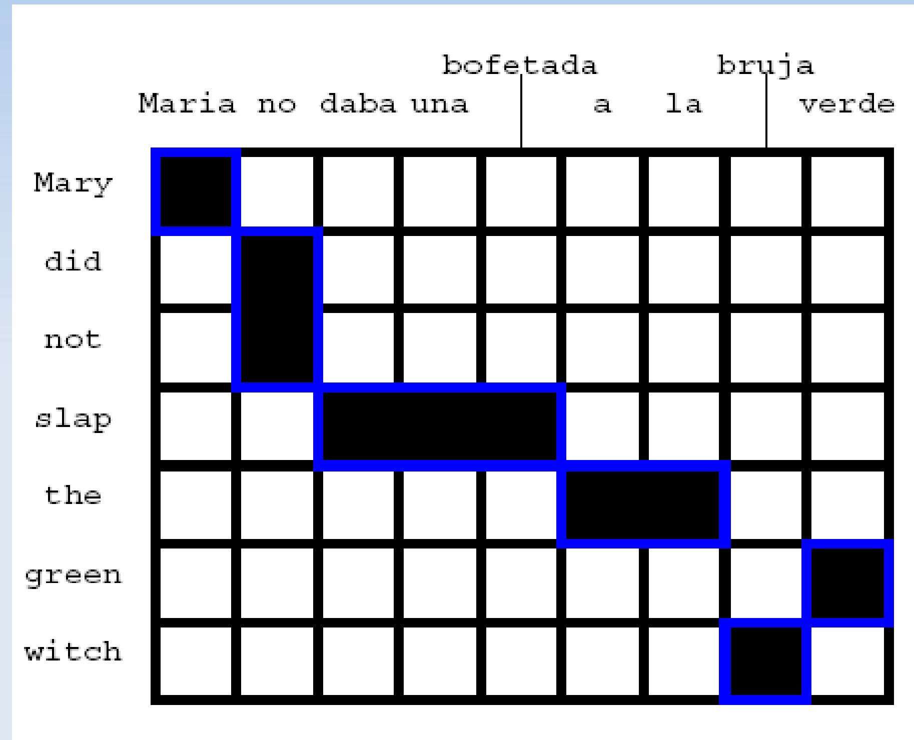
- Preprocessing
- Training
- Decoding

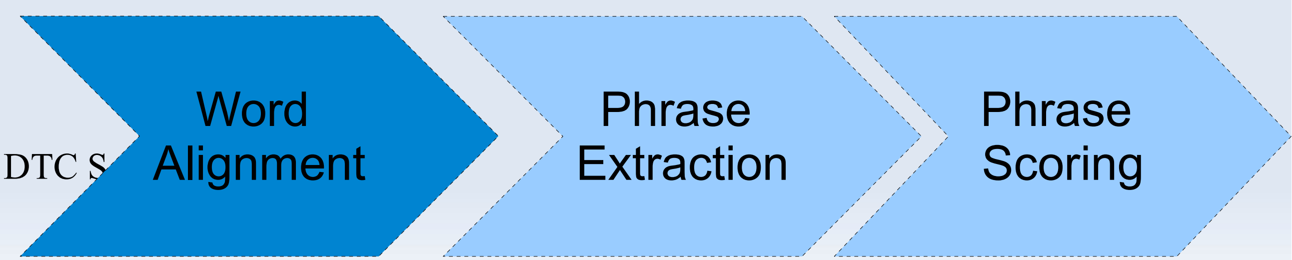


# Translation model estimation

# Word Alignment

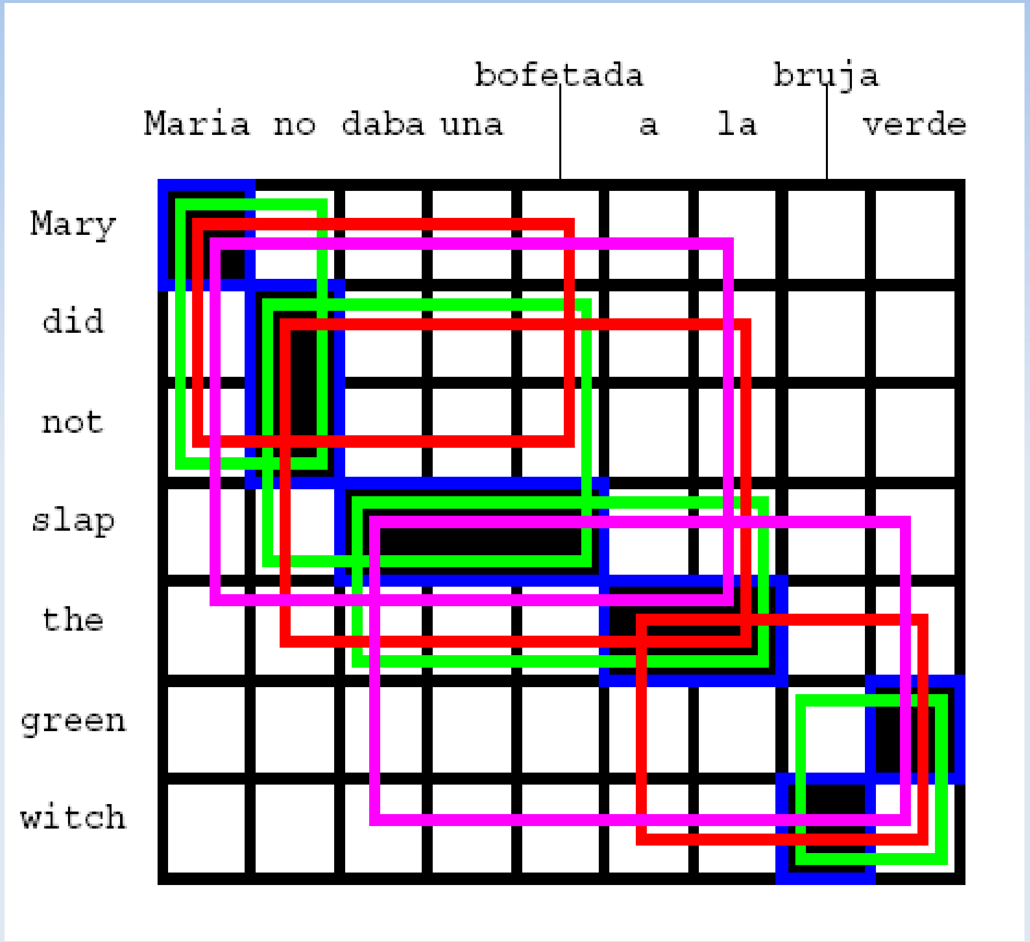
- Estimate the likelihood of individual words to be translated into each other
- Based on coocurrences
- IBM Models
- EM Algorithm



# Phase Extraction

- Use heuristics to extract phrases that are consistent with the word alignment

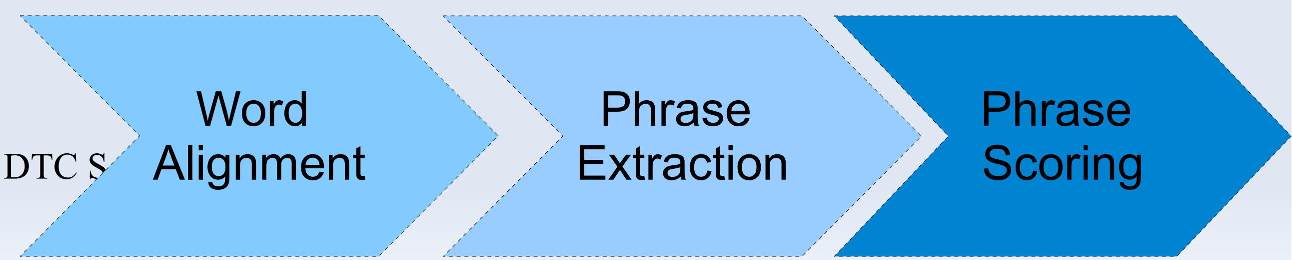



**Mary** ||| Maria
**Did not** ||| no
**Slap** ||| daba una bofetada
**The** ||| a la
**Green** ||| verde
**Witch** ||| bruja
**Mary did not** ||| Maria no
**Mary did not slap** ||| Maria daba una bofetada

# Phrase Scoring

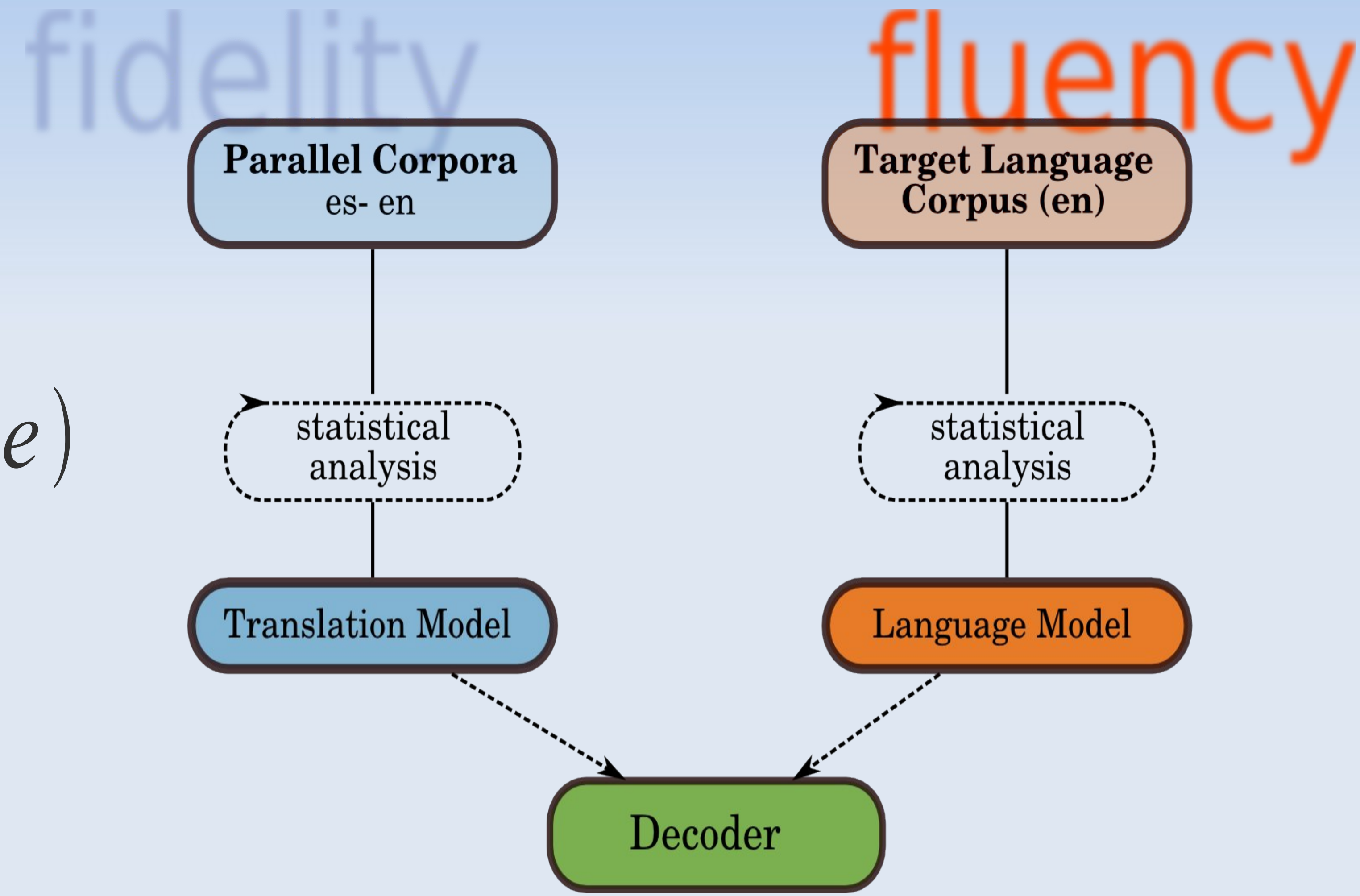
- Score each phrase-pair according to MLE

$$p(\text{la bruja verde} \mid \text{the green witch}) = \frac{count(\text{la bruja verde}, \text{the green witch})}{count(\text{the green witch})}$$



# How does it work?

$p(f|e)$




$p(e)$

# Jargon

- Source Language
  - Language from which we want to translate ( es)
- Target Language
  - Language to which we want to translate (en)
- Phrase Pair
  - Source phrase || Target phrase
- Phrase Table
  - Database of phrases (lexicon), with scores (probs)
- SMT
  - Statistical Machine Translation
- BLEU
  - De facto translation quality metric
- AER
  - De facto alignment quality metric
- Gaps
  - Unaligned words at a phrase level

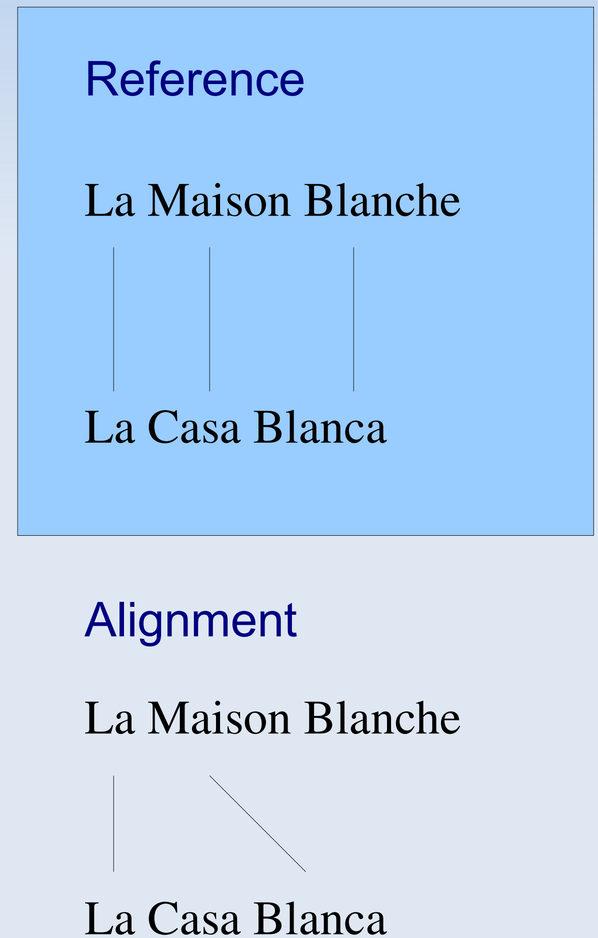
# Alignment Quality

# Word Alignment

- Beginning of SMT pipeline.
- Most subsequent steps based on WA.
- A lot of work to improve WA quality.
- Widely available Hand Alignments enabled discriminative approaches.
- New discriminative models based on metrics such AER.

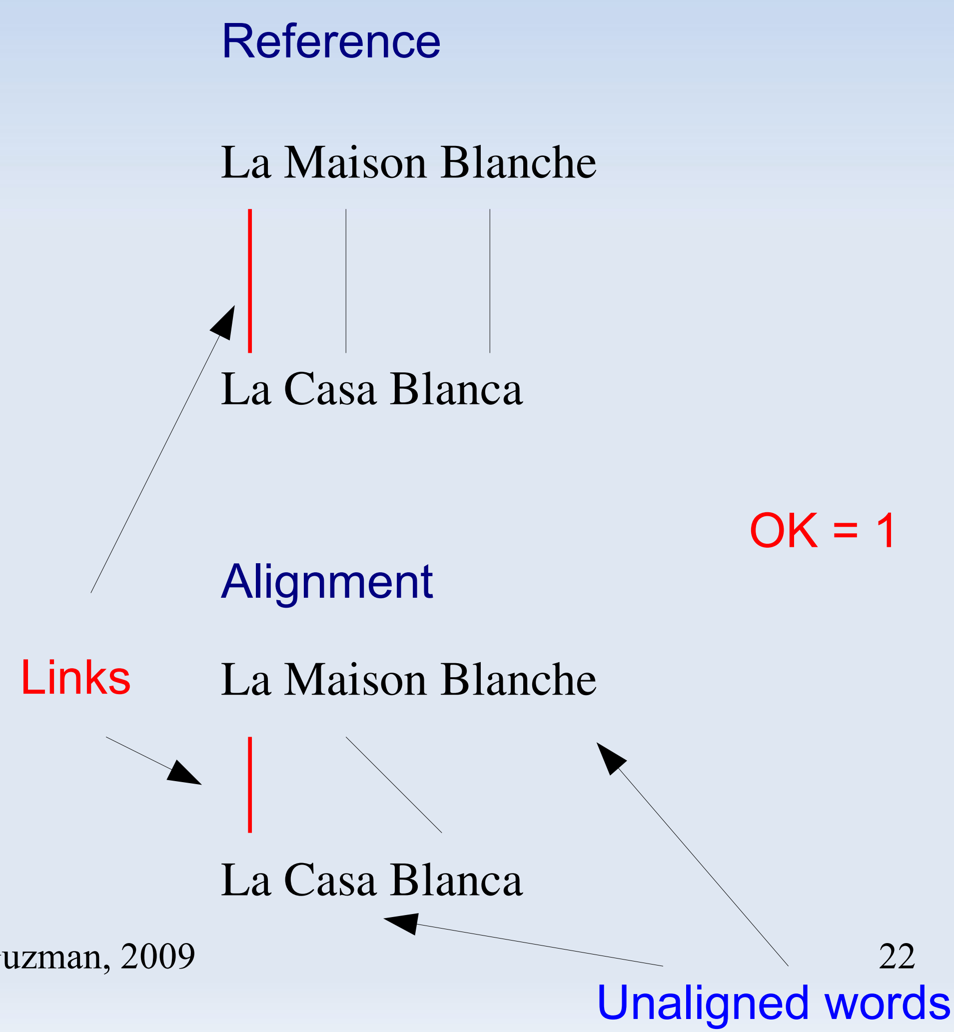
# Good vs. Bad alignments

- To get a quality score, a reference is needed



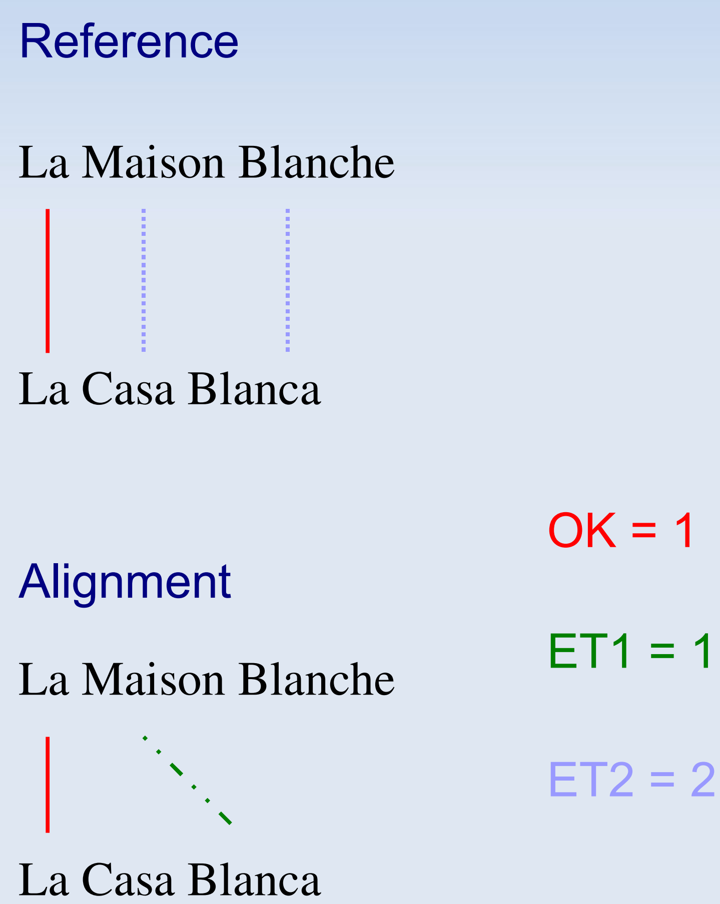
# Good vs. Bad alignments

- To get a quality score, a reference is needed
- We get the agreement in number of links



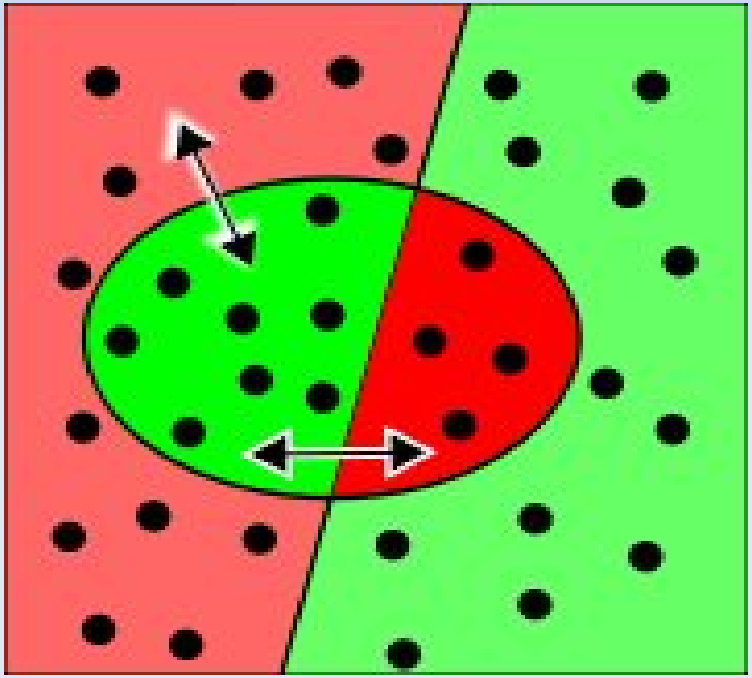
# Good vs. Bad alignments

- To get a quality score, a reference is needed
- We get the agreement in number of links
- We extract errors type I and II
- Keep the count!



# Different types of Metrics

- AER
- Precision
- Recall
- F-score
  - Alpha F-Score.
- etc



# AER vs BLEU

- Fraser and Marcu, 2004

$$AER \neq BLEU$$

  - Evaluated **correlation** between BLEU and AER.
  - Possible links => flaws.
  - Variation of **F-measure**, uses a coefficient to **modify balance** between **precision** and **recall**.
  - The optimal **coefficient** depends on the **corpus**.
- Vilar et al., 2004
  - **Better BLEU** scores can be obtained with **"degraded" alignments.**
  - **Mismatch** between **alignment** and **translation** models.
  - Support the **use of AER**.

$$\downarrow AER \Rightarrow \uparrow BLEU$$

# Going beyond

- Ayan and Dorr, 2004
  - Analyze the quality of the alignments and resulting phrase tables.
  - Several types of alignments.
  - Several lexical weightings
  - CPER (Consistent Phrase Error Rate)
  - Do not analyze other characteristics of the alignment/ phrase table

# Word Alignment & Phrase Extraction Analysis

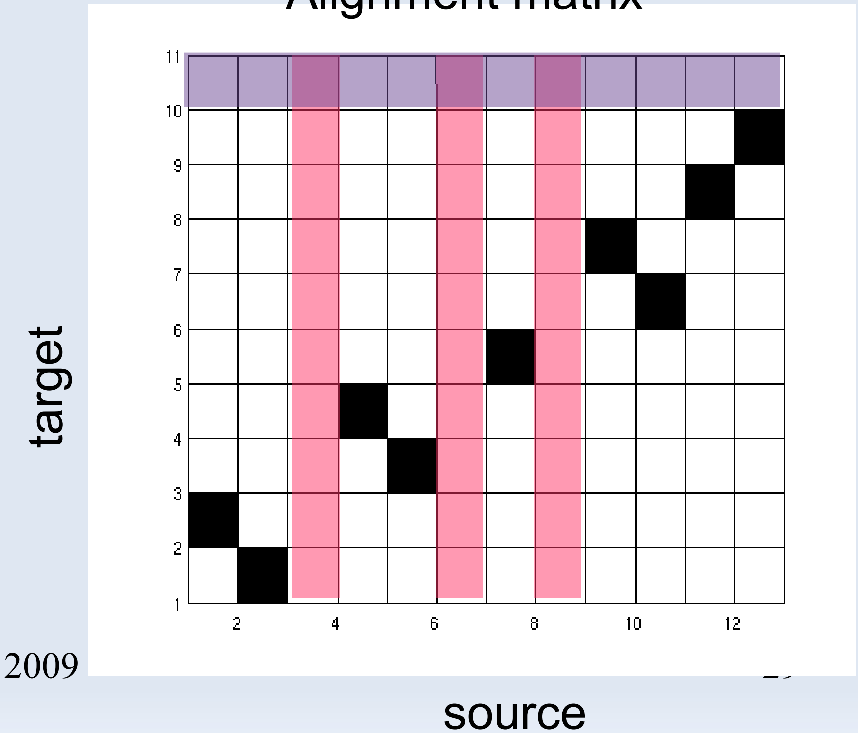
# Setup

- We analyzed different types of alignments
- Chinese - English
- The objective was to determine which characteristic was more relevant
  - Quality?
  - Structure?
- Analysis beyond alignment quality.

# Word Alignment Metrics

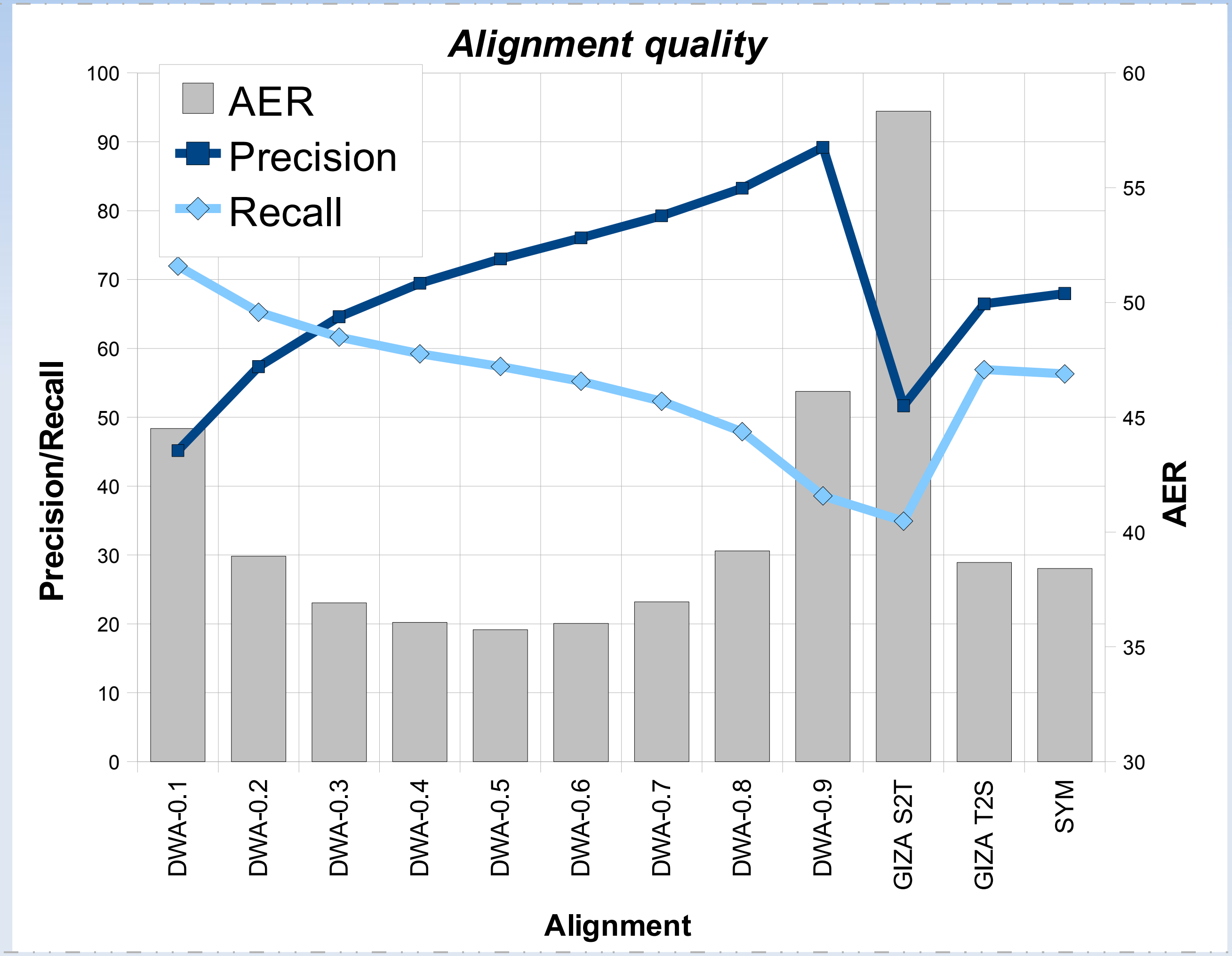
- Qualitative:
  - AER (F-measure)
  - Precision
  - Recall
- Quantitative:
  - Number of links
  - Number of unaligned words

Alignment matrix



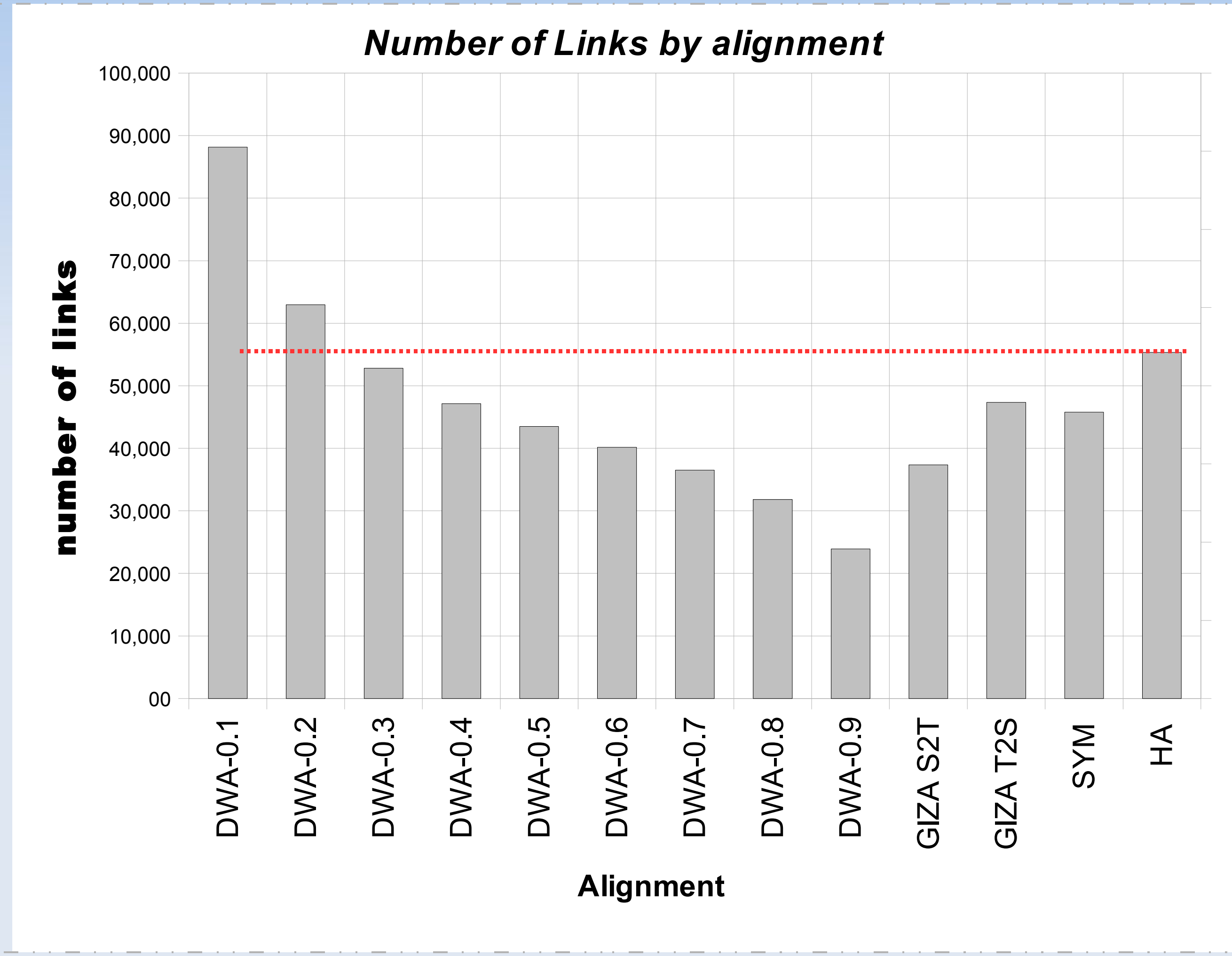
# Alignment quality results

- DWA alignments: higher threshold=>more precision
- Best AER from slightly more precise alignment (DWA-0.5)
- GIZA=> more precision than recall.
- SYM lower AER than GIZA.



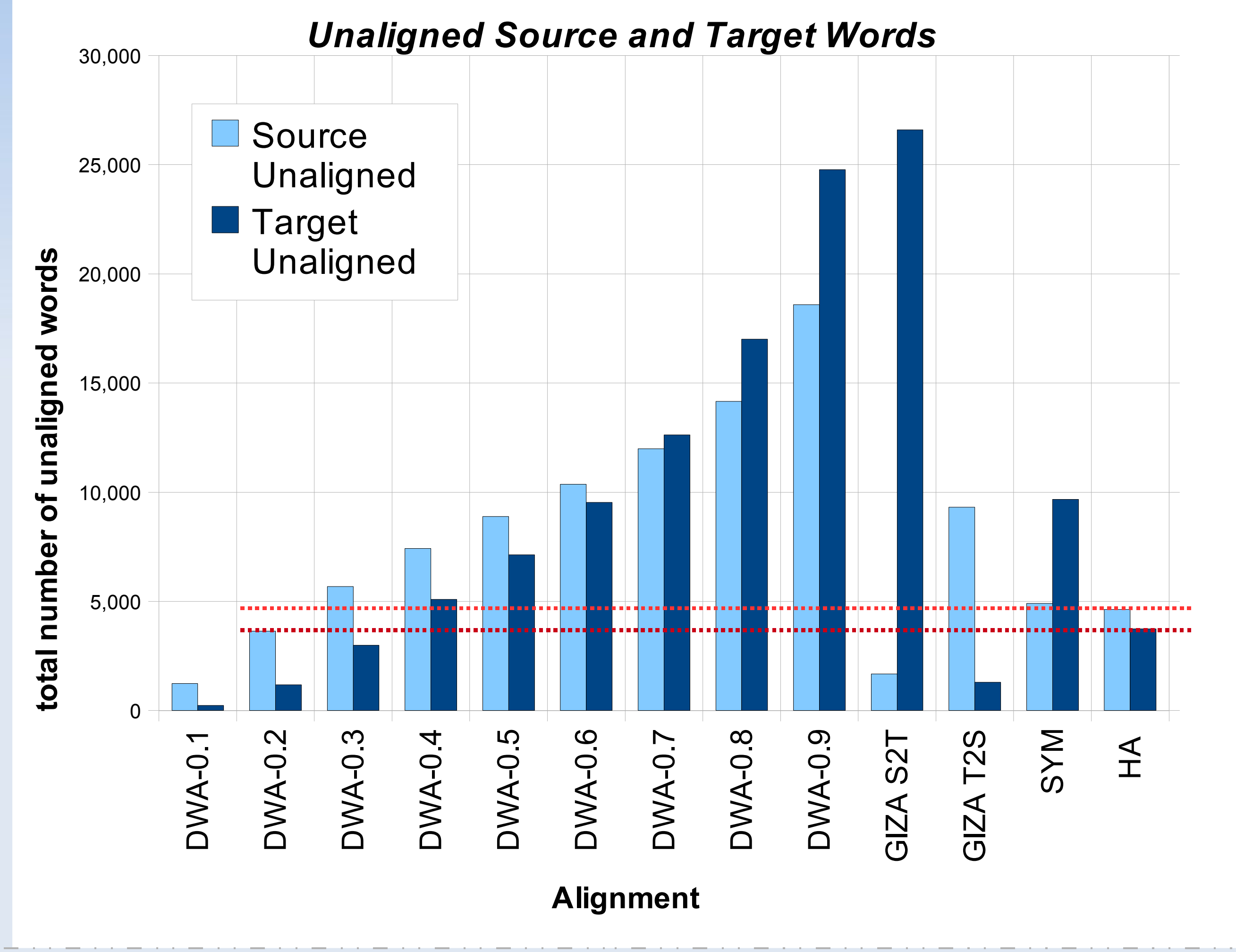
# Links

- Hand Align closer to (DWA-0.3)
- DWA aligns: high threshold=>fewer links
- Best AER (DWA-0.5) fewer links than HA



# Unaligned Words

- HA Source: closer to SYM, DWA-0.3
- HA Target: closer to DWA-0.4, DWA-0.3
- GIZA asymmetry
- DWA: higher threshold, more unalignments.
- DWA: lower threshold=> more proportion Chinese words unaligned.



# Word Alignment: Summary

- Diversity of balance precision/recall between alignments.
- Usually precision prevails over recall.
- Two ways of describing an alignment: links and unaligned words.
- In next section, we'll observe the importance of such factors in the generation of phrase pairs.

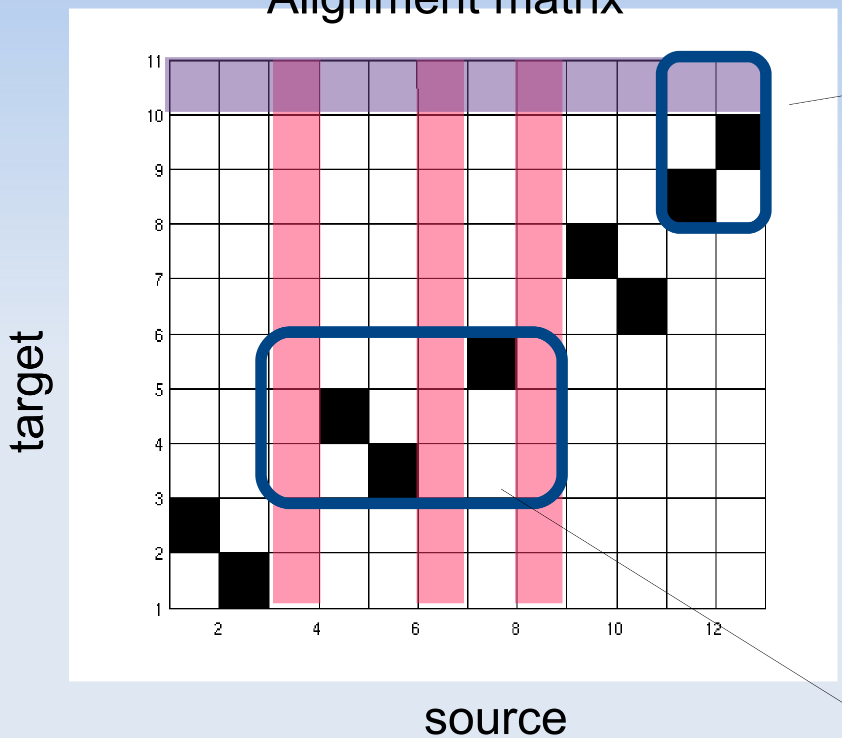
# Analysis II:
# Phrase Extraction

# Metrics

- Quantitative:
  - Number of phrases
  - Singletons (unique entries)
  - Phrase lengths
  - Gaps (unaligned words inside phrase pair)
- Qualitative:
  - Manual Evaluation

# What do we mean by gaps?

Alignment matrix

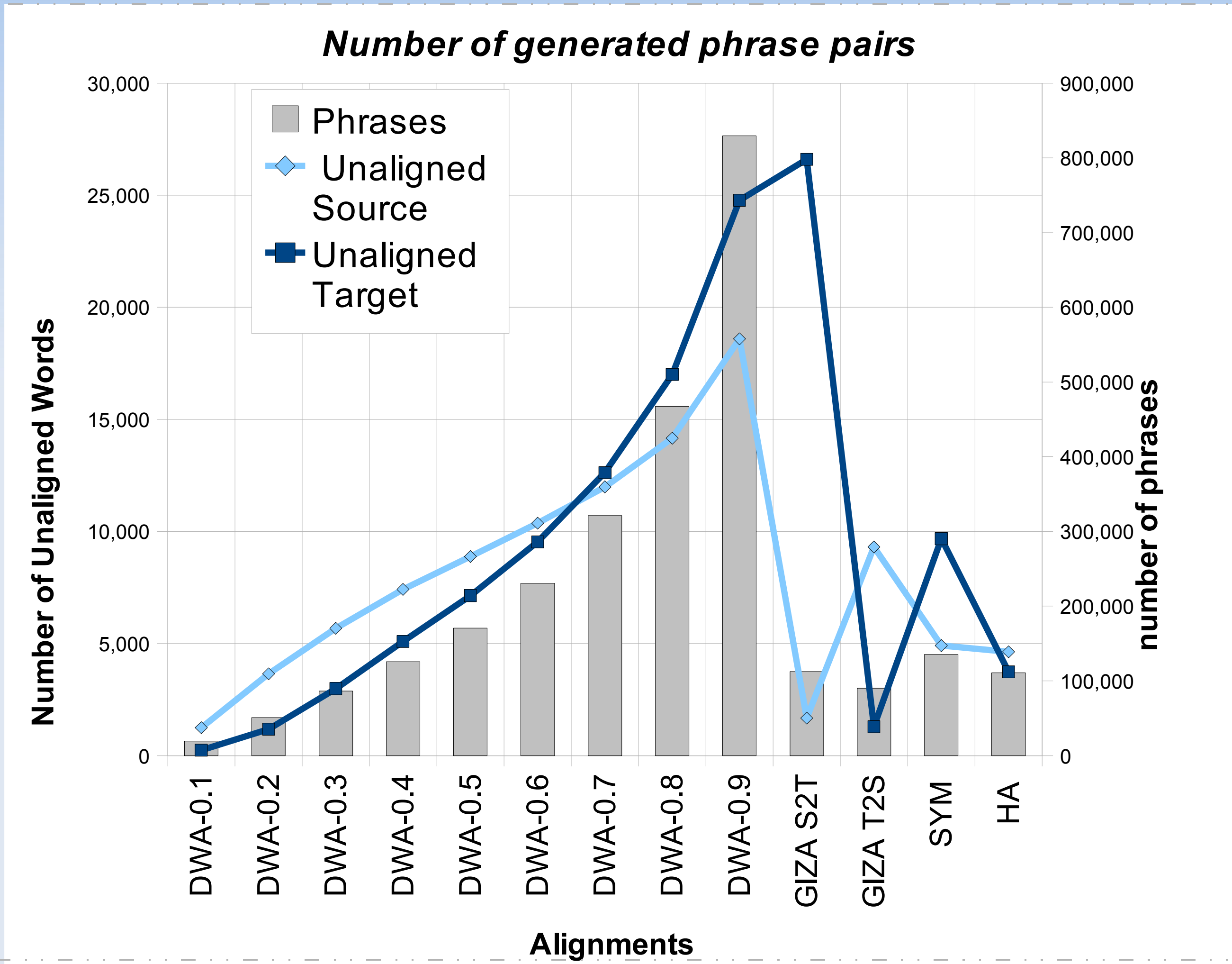



In this phrase pair:
1 gap on the target phrase

In this phrase pair:
3 gaps on the source phrase

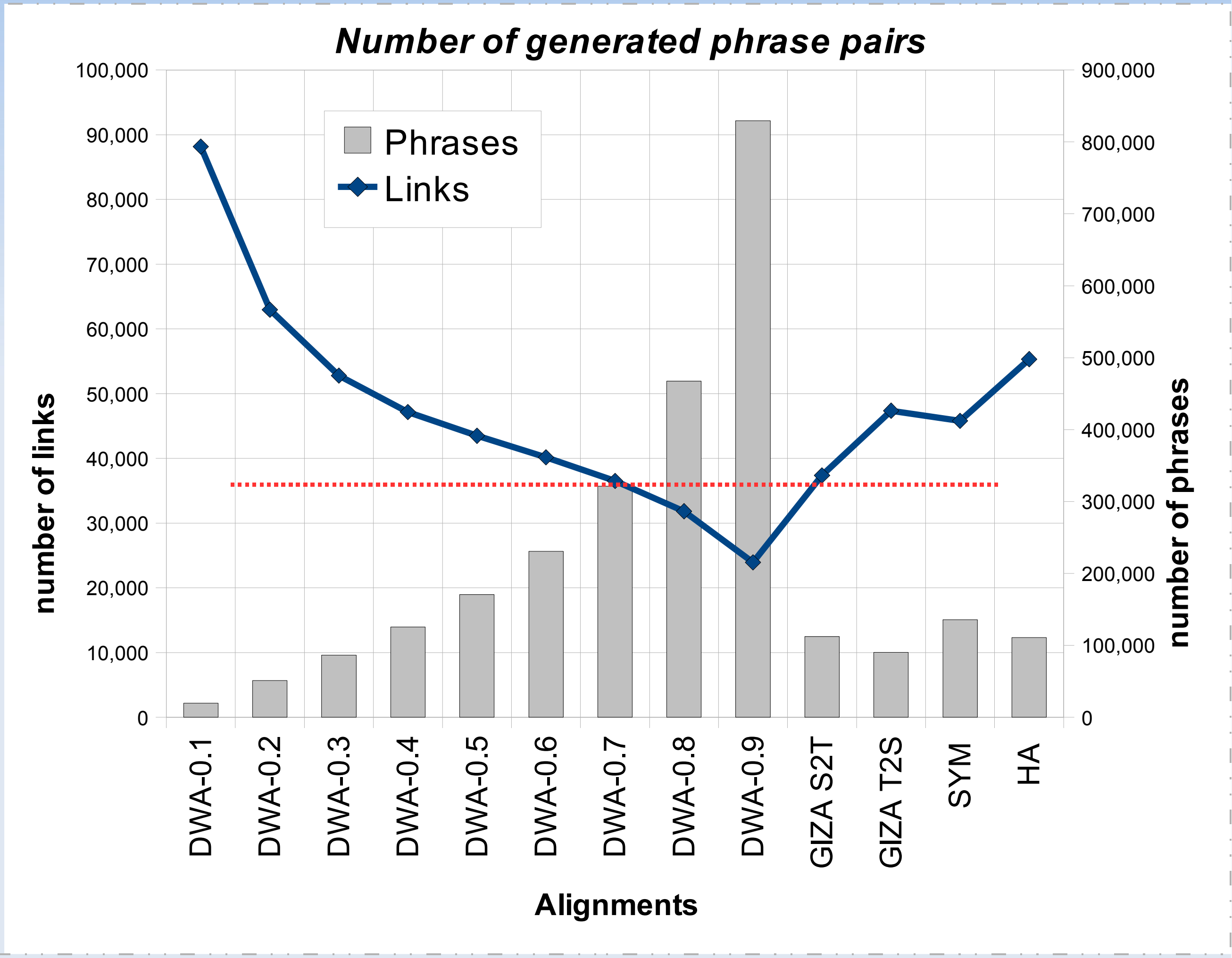
# Number of Phrases

- PT **grows** as our alignment gets sparser
- Related to unaligned words rather than number of links



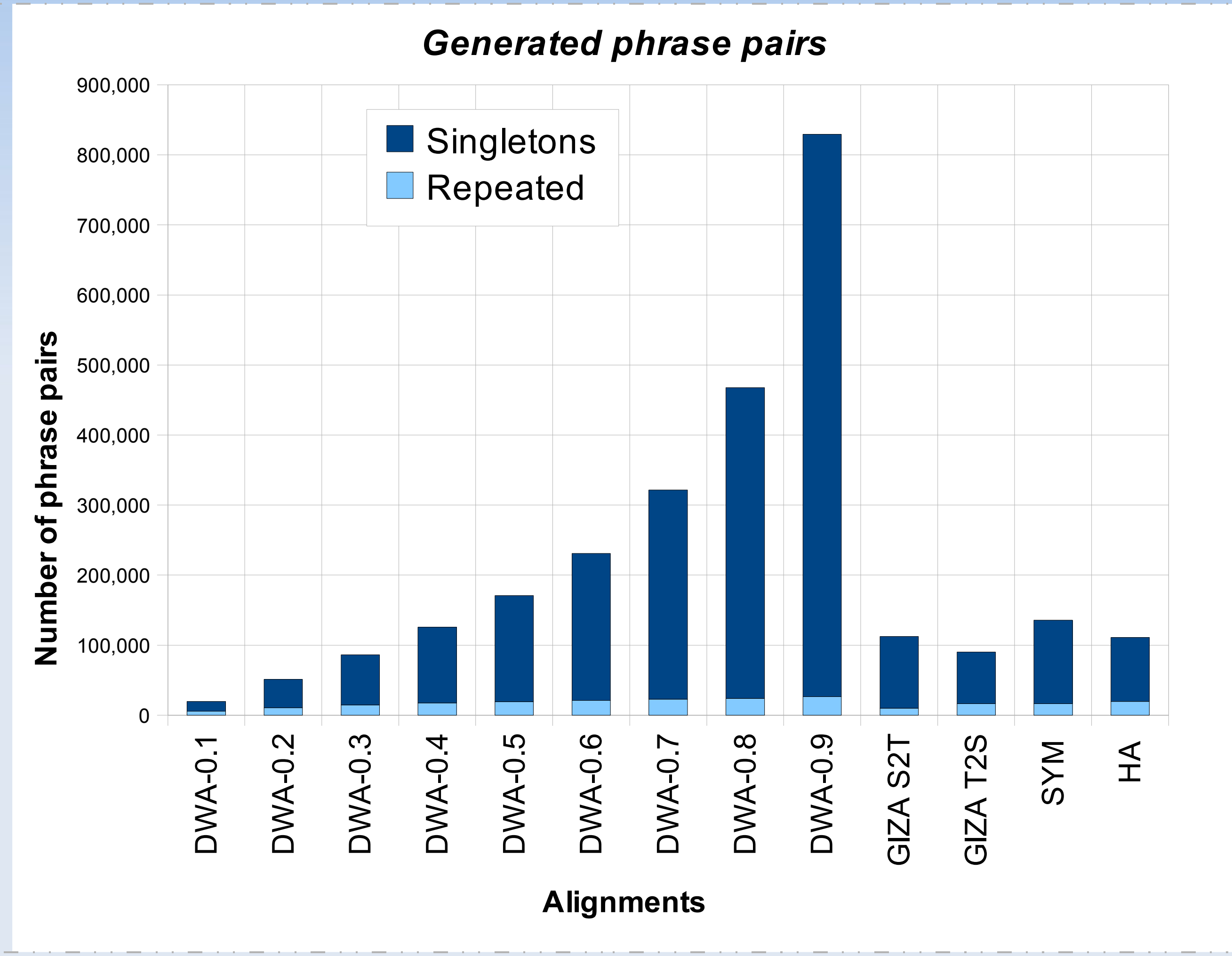
# Number of Phrases

- PT **grows** as our alignment gets sparser
- Related to unaligned words rather than number of links



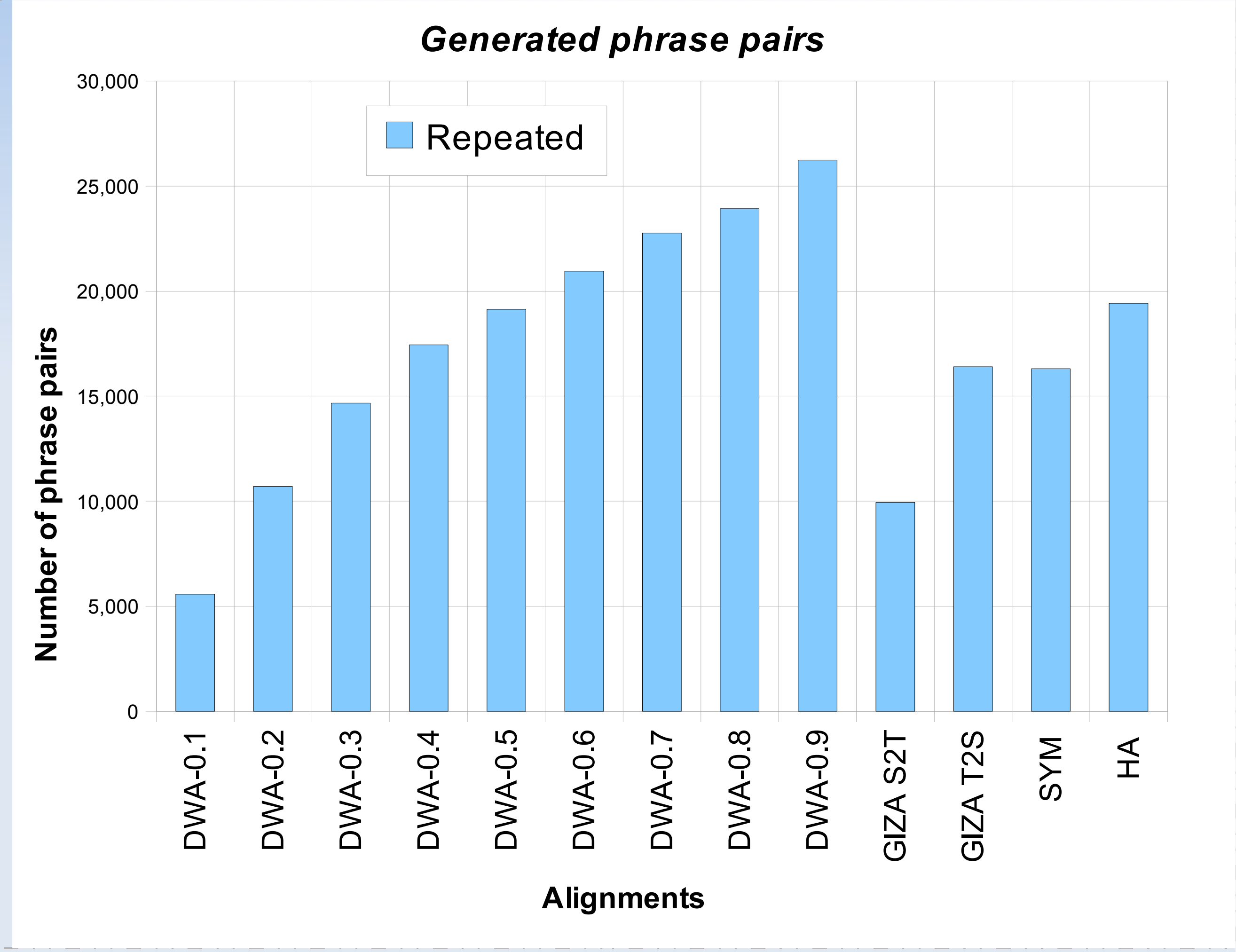
# Singletons

- Most of the phrase-pairs are **singletons**.
- Repeated phrase-pairs grow at a slower rate.



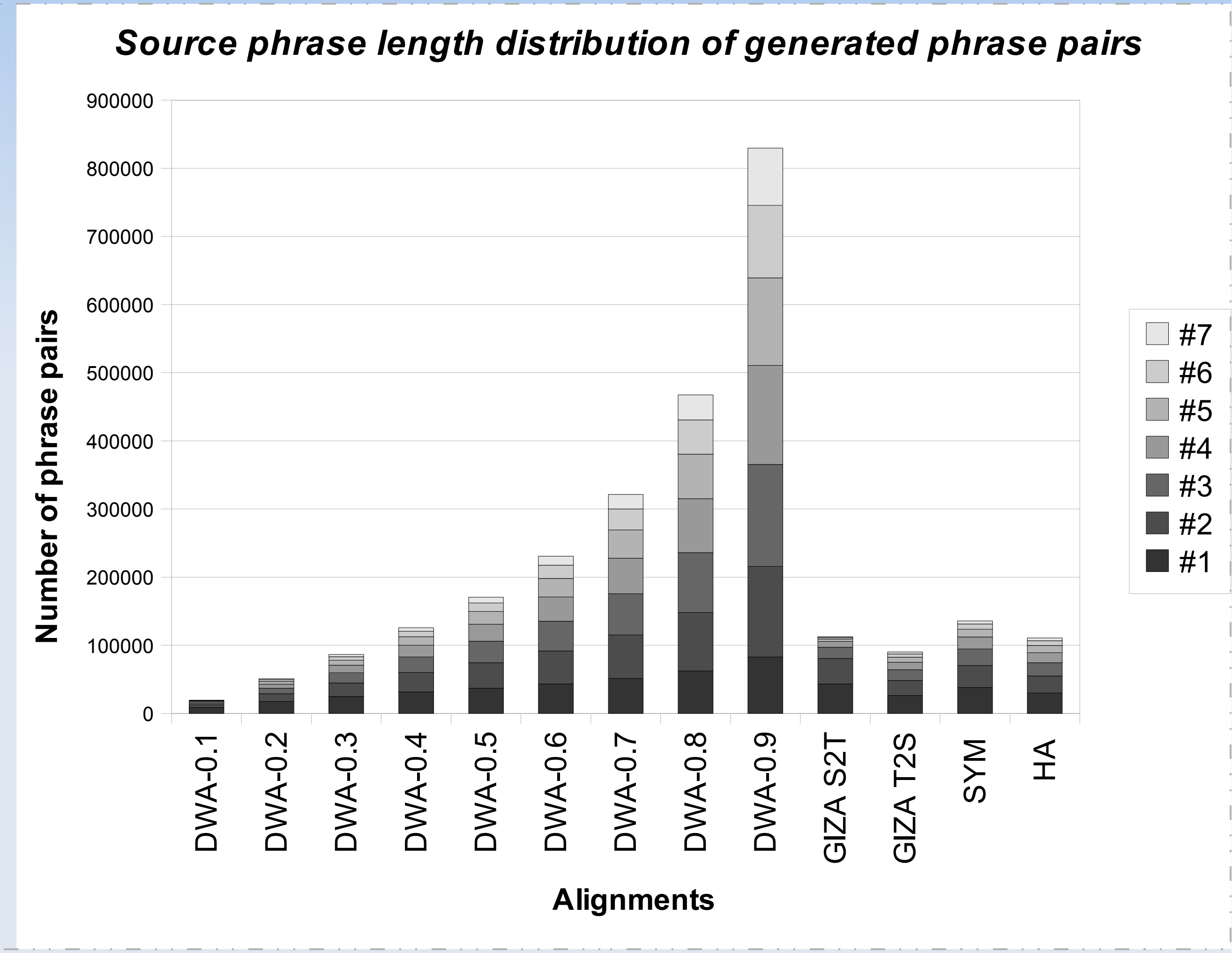
# Singletons

- Most of the phrase-pairs are **singletons**.
- Repeated phrase-pairs grow at a slower rate.



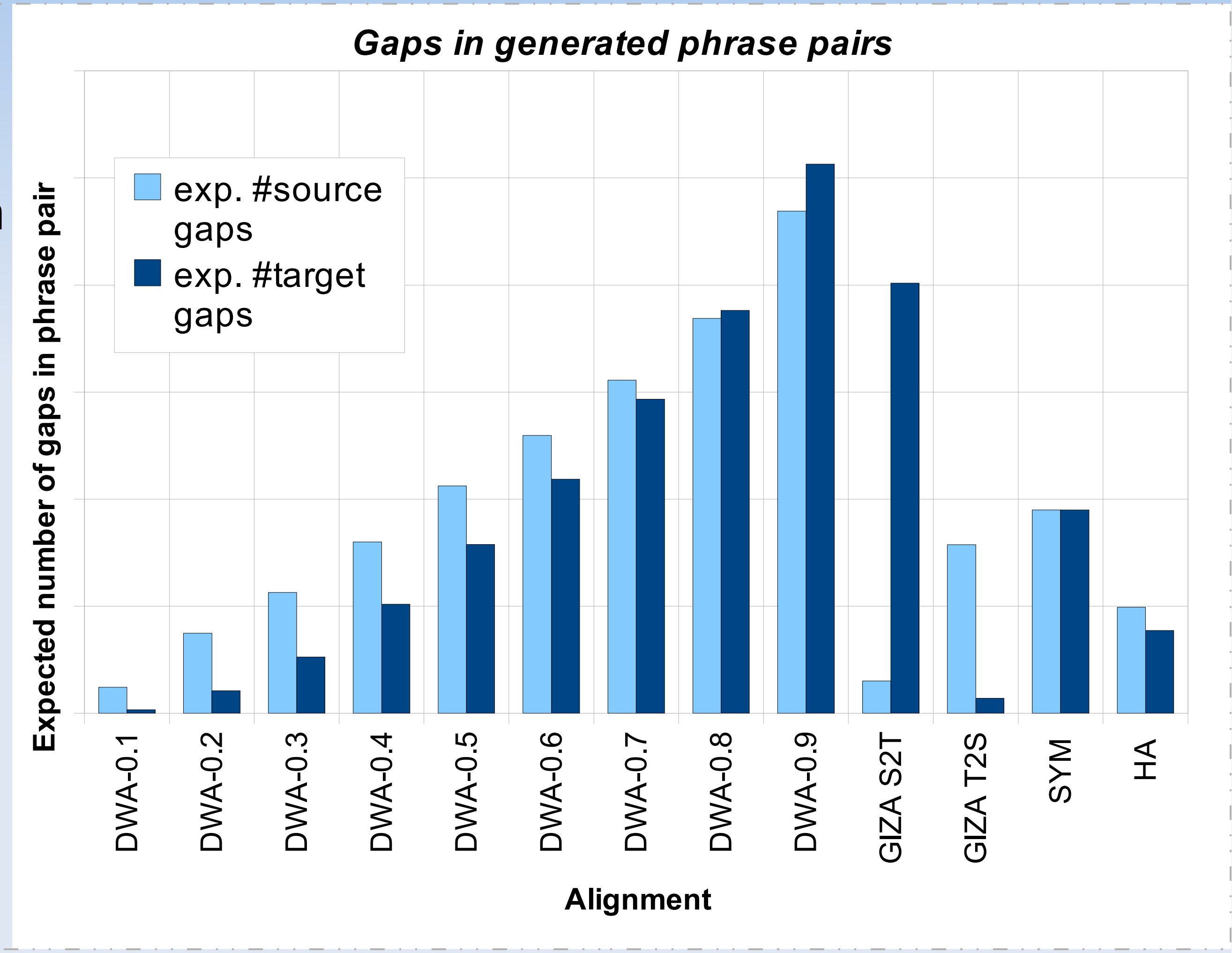
# Phase Length

- As our PT grows, entries become longer and longer.



# Gaps

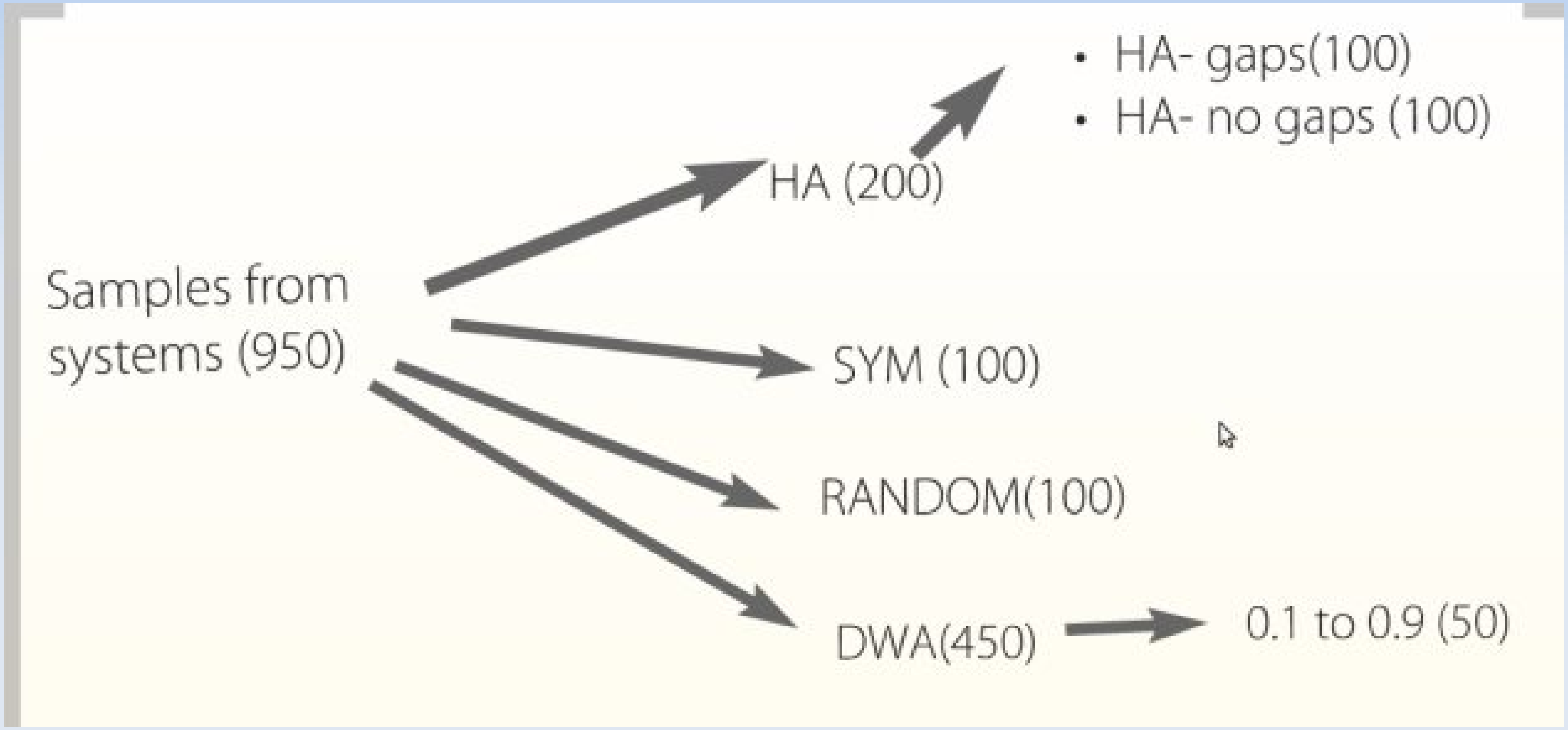
- The gaps inside a phrase pair increases too.
- The distribution of gaps in the generated phrases follows the distribution of unaligned words in the alignment



# Human Evaluation of Phrase Pairs

- Setup:
  - Native Chinese Speakers
  - Each subject was asked whether a phrase pair was adequate
  - No contextual information
  - Included a noisy input
  - Included phrases extracted from Hand Aligned data.

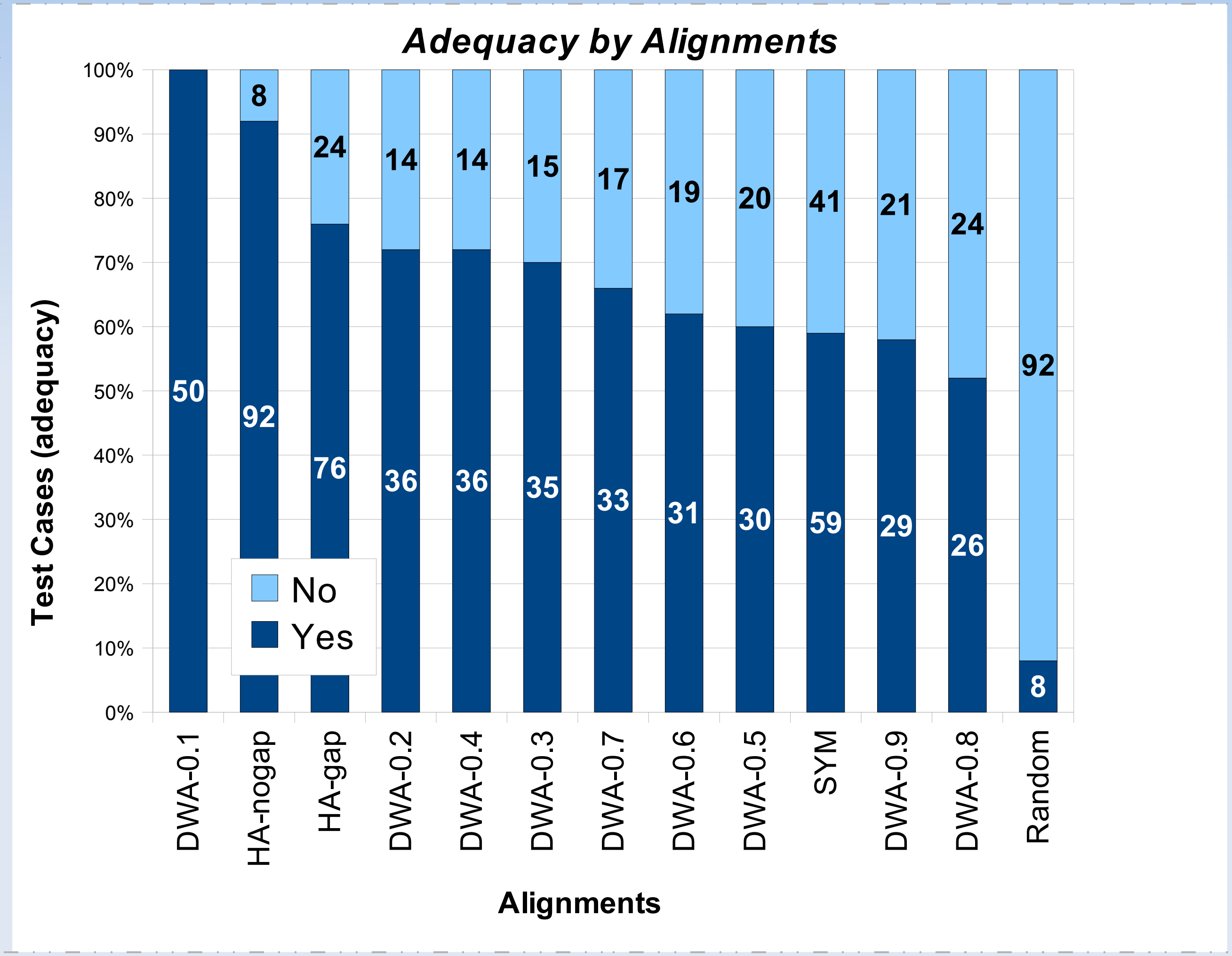
# Sampling



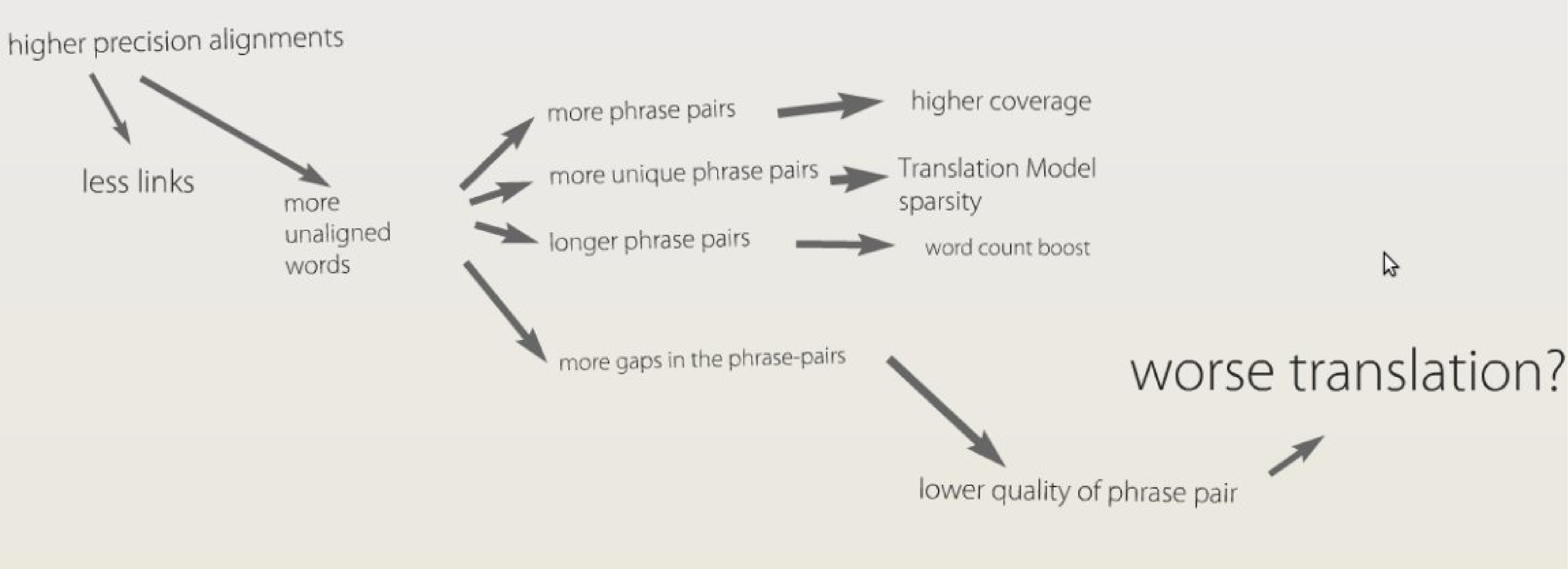
HA- gaps(100)
HA- no gaps (100)
HA (200)
Samples from systems (950)
SYM (100)
RANDOM(100)
DWA(450)
0.1 to 0.9 (50)

# Results

- HA w/o gaps yields better results than HA w/ gaps.
- DWA-0.1 very good (short phrase pairs)
- DWA-0.5 not so great
- Random pairings are usually bad



# Phrase Extraction: Summary



# Lessons Learned: Mind your gaps

# Taking into account GAPS

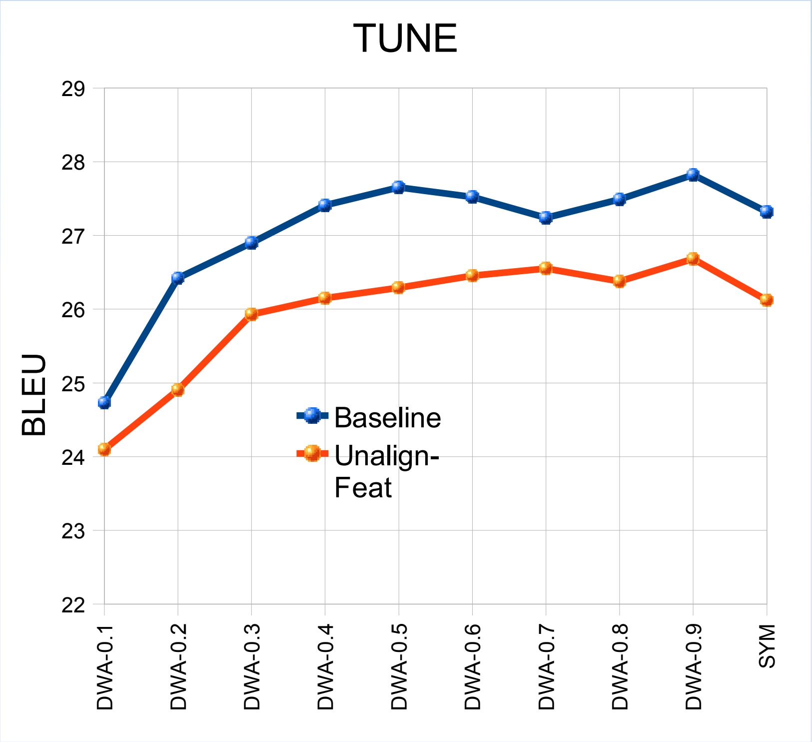
- Gaps inside phrase pairs have considerable impact on human perceived quality of phrase pair.
- Do they affect translation?
- Translation Experiment:
  - Include gap count as a feature (similar to WC)
  - Compare the performance of the different systems w/o the features

# Setup

- Training
  - GALE P3 Data
  - Maximum sentence length 30
  - 1 Million sentences (random)
- Tuning
  - MT05
- Test
  - GALE DEV07-Blind

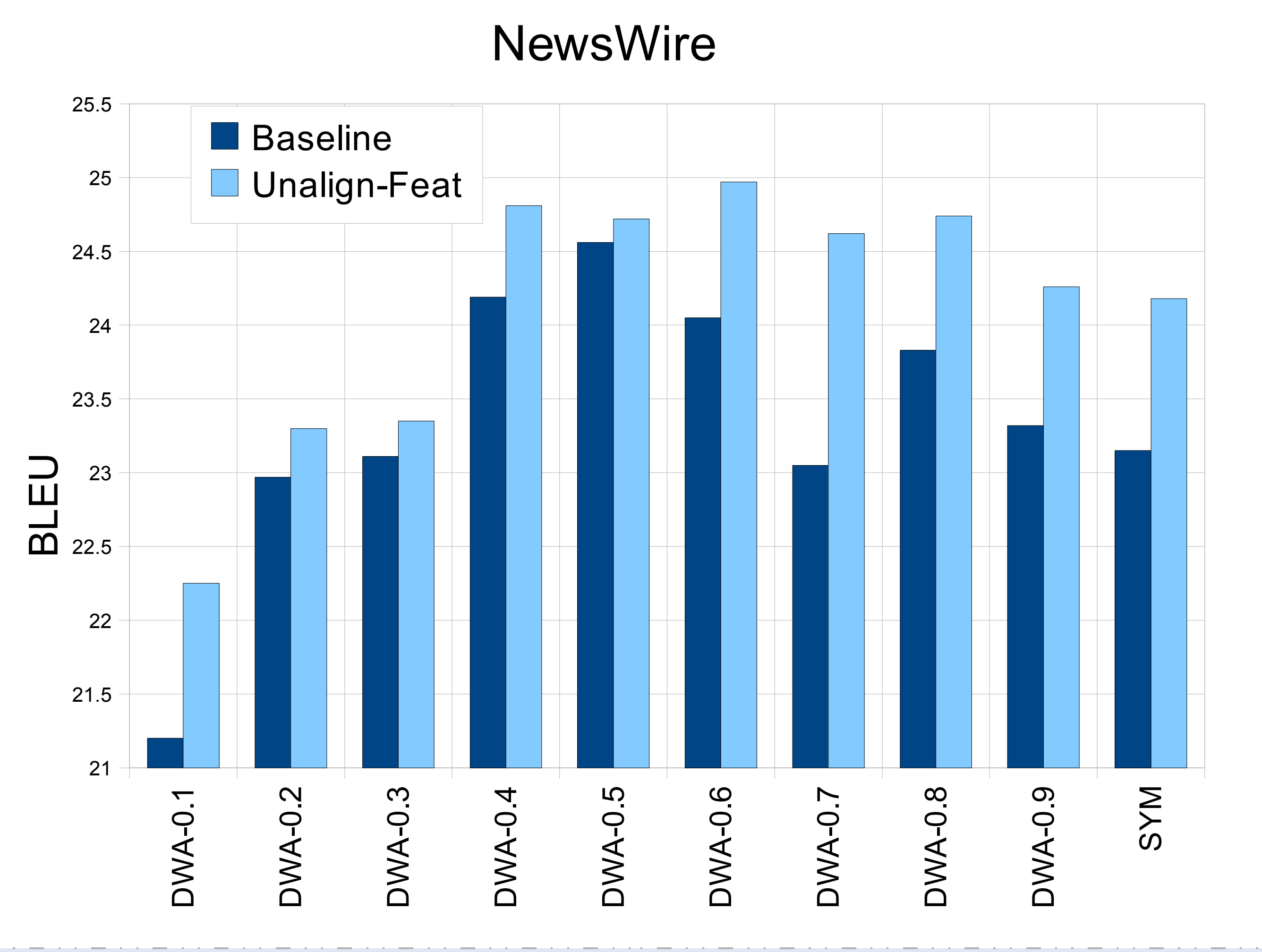
# Tuning

- Baseline gets get better results
- Over-fitting?



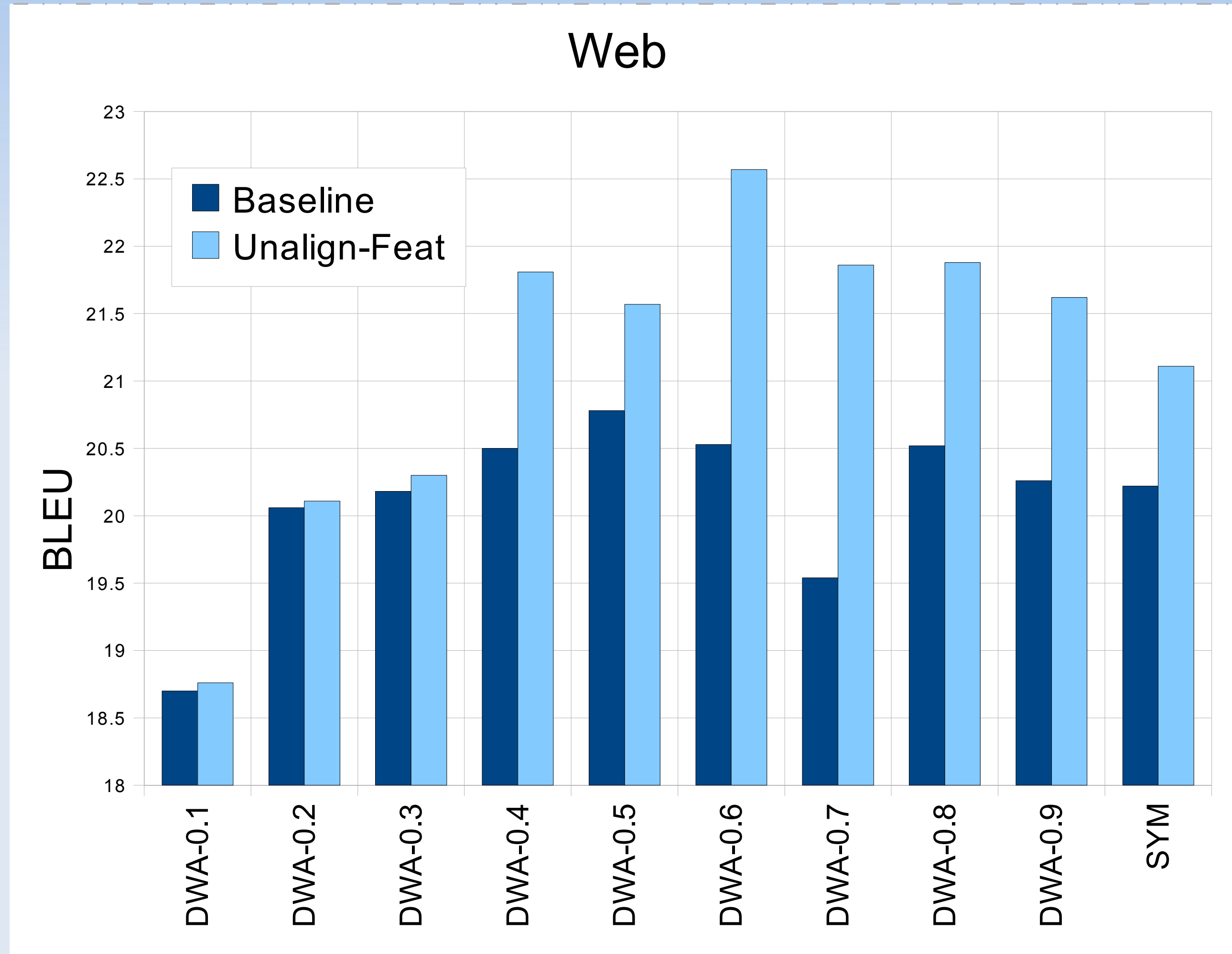
# Experimental Results

- Overall gains
- Web performs better (~2 BP)
- Best system shifts to a higher precision alignment (DWA-0.5 => DWA-0.6)
- Higher recall alignments without much change



# Experimental Results

- Overall gains
- Web performs better (~2 BP)
- Best system shifts to a higher precision alignment (DWA-0.5 => DWA-0.6)
- Higher recall alignments without much change

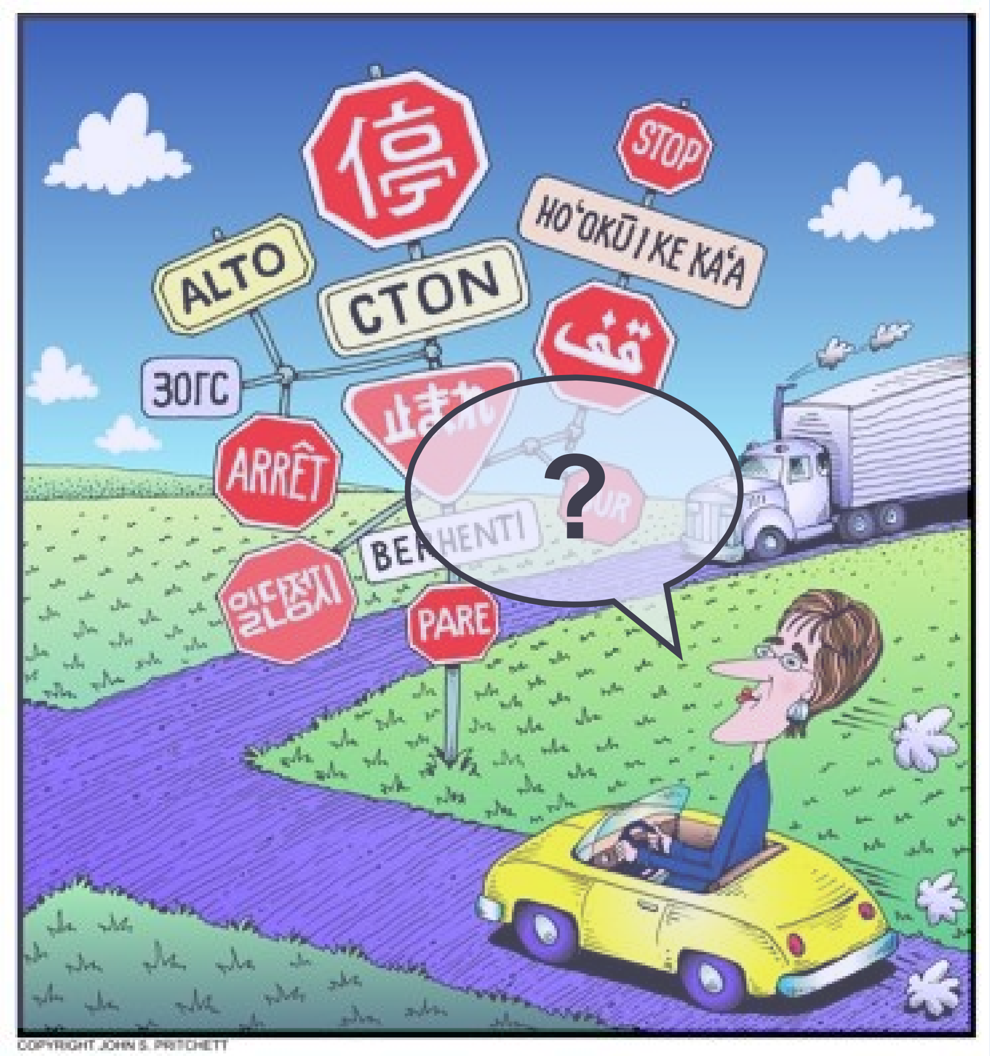


# Conclusions

- We can describe an **alignment** by its **quality** and its **structure** (links, unaligned words).
- **Unaligned words** have an important role in **phrase extraction** (more than number links).
- The distribution of the **gaps inside a phrase pair** is related to the distribution of **unaligned words** in the alignment.
- Extracted phrase pairs with **more gaps** have **lower** human perceived **quality**.
- Taking into account the **number of gaps** in an extracted phrase pair as features achieved **overall improvements**.

# What's next?

- Determine which phrases are now chosen by the decoder and why.
- Determine if improvement holds for other language pairs.
- Incorporate unalignment information in other stages of SMT (phrase extraction, scoring).



# About a research stay

If you can, take the chance.