Unsupervised Turkish Morphological Segmentation for Statistical Machine Translation

Coskun Mermer and Murat Saraclar

Workshop on Machine Translation and
Morphologically-rich Languages
Haifa, 27 January 2011

Why Unsupervised?

- No human involvement
- Language independence
- Automatic optimization to task

Using a Morphological Analyzer

- Linguistic morphological analysis intuitive, but
  - language-dependent
  - ambiguous
  - not always optimal
    - manually engineered segmentation schemes can outperform a straightforward linguistic morphological segmentation
    - naive linguistic segmentation may result in even worse performance than a word-based system

Heuristic Segmentation/Merging Rules

- Widely varying heuristics:
  - Minimal segmentation
    - Only segment predominant & sure-to-help affixation
  - Start with linguistic segmentation and take back some segmentations
    - Requires careful study of both linguistics, experimental results
  - Trial-and-error
  - Not portable to other language pairs

Adopted Approach

- Unsupervised learning form a corpus
- Maximize an objective function (posterior probability)

Morfessor

- M. Creutz and K. Lagus, “Unsupervised models for morpheme segmentation and morphology learning,” ACM Transactions on Speech and Language Processing, 2007.

Probabilistic Segmentation Model

$$P(M_f, f) = P(M_f)P(f \mid M_f)$$

- f : Observed corpus
- M_f : Hidden segmentation model for the corpus (≈ “morph” vocabulary)

MAP Segmentation

$$\hat{M}_f = \arg\max_{M_f} P(M_f \mid f)$$

$$= \arg\max_{M_f} P(M_f, f)$$

$$= \arg\max_{M_f} P(M_f) P(f \mid M_f)$$

Probabilistic Model Components

$$P(M_f) = P(frequencies_{M_f})P(lengths_{M_f})$$

- $P(frequencies_{M_f})$: Uniform probability for all possible morph vocabularies of size M for a given morph token count of N (i.e., frequencies do not matter)
- $P(lengths_{M_f})$: For each morph, product of its character probabilities (including end-of-morph marker)
- $P(f \mid M_f)$: Product of probabilities for each morph token

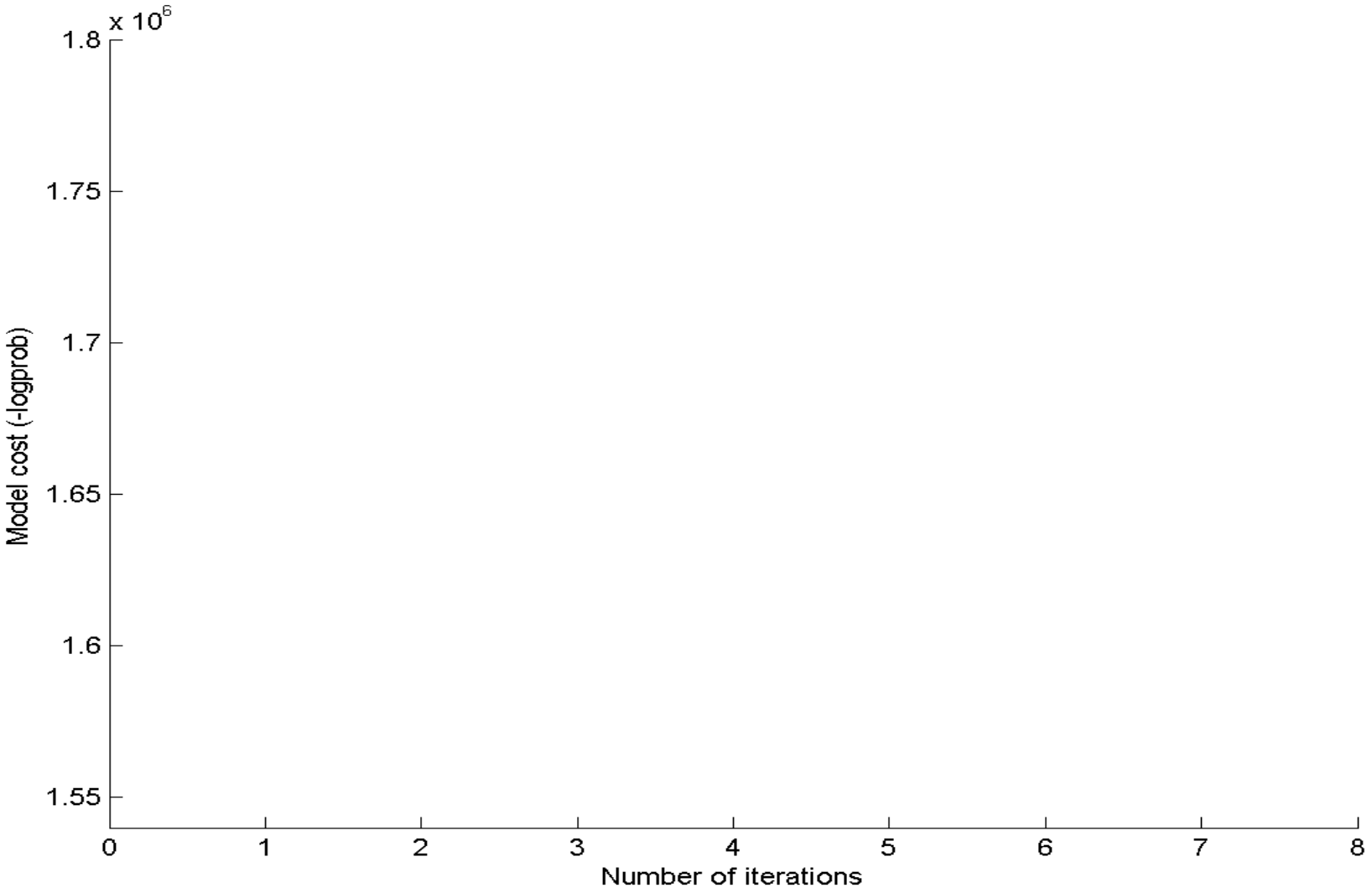
Original Search Algorithm

- Greedy
- Scan the current word/morph vocabulary
- Accept the best segmentation location (or non-segmentation) and update the model

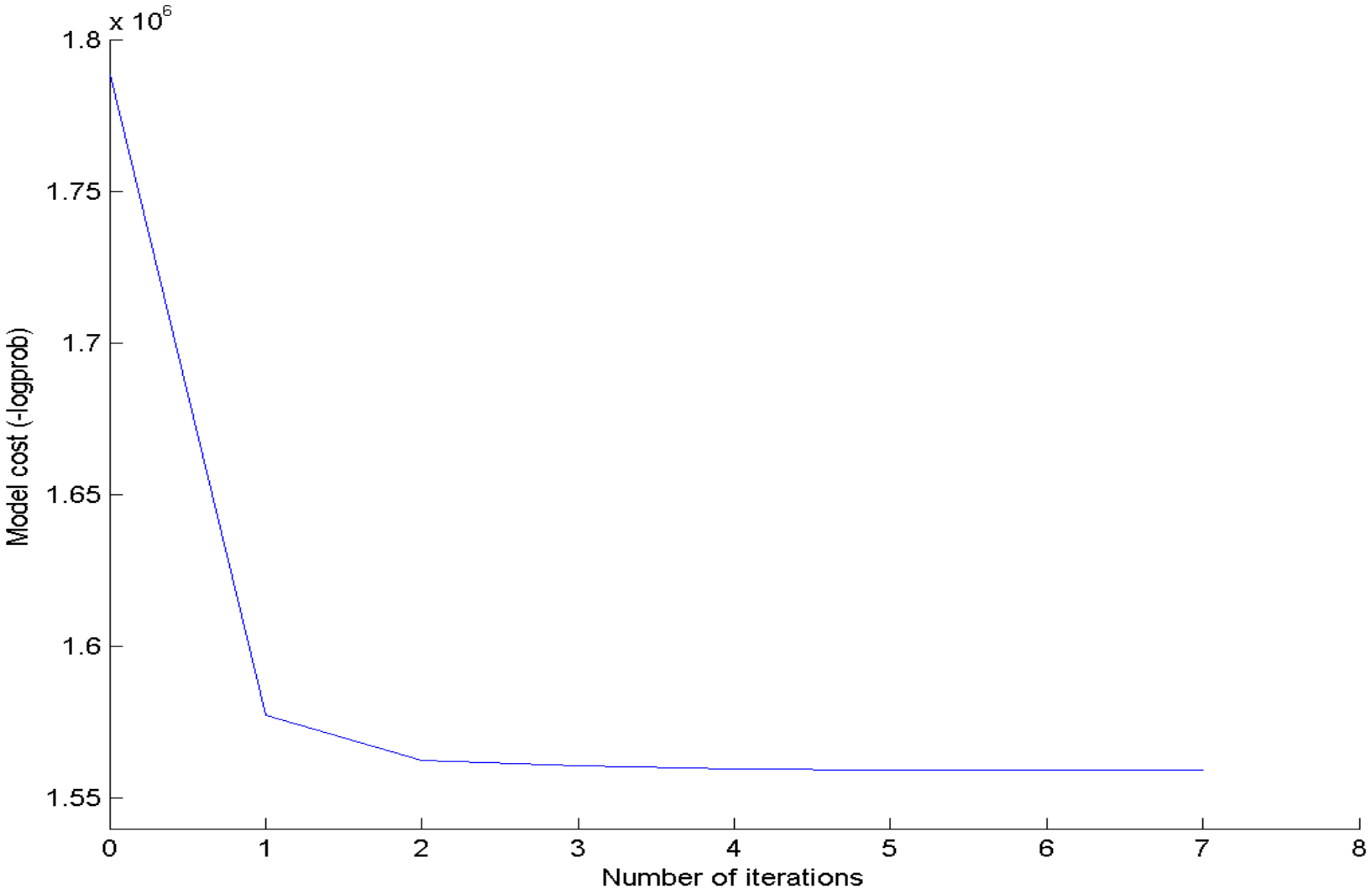
Parallel Search

- Less greedy
- Wait until all the vocabulary is scanned before applying the updates

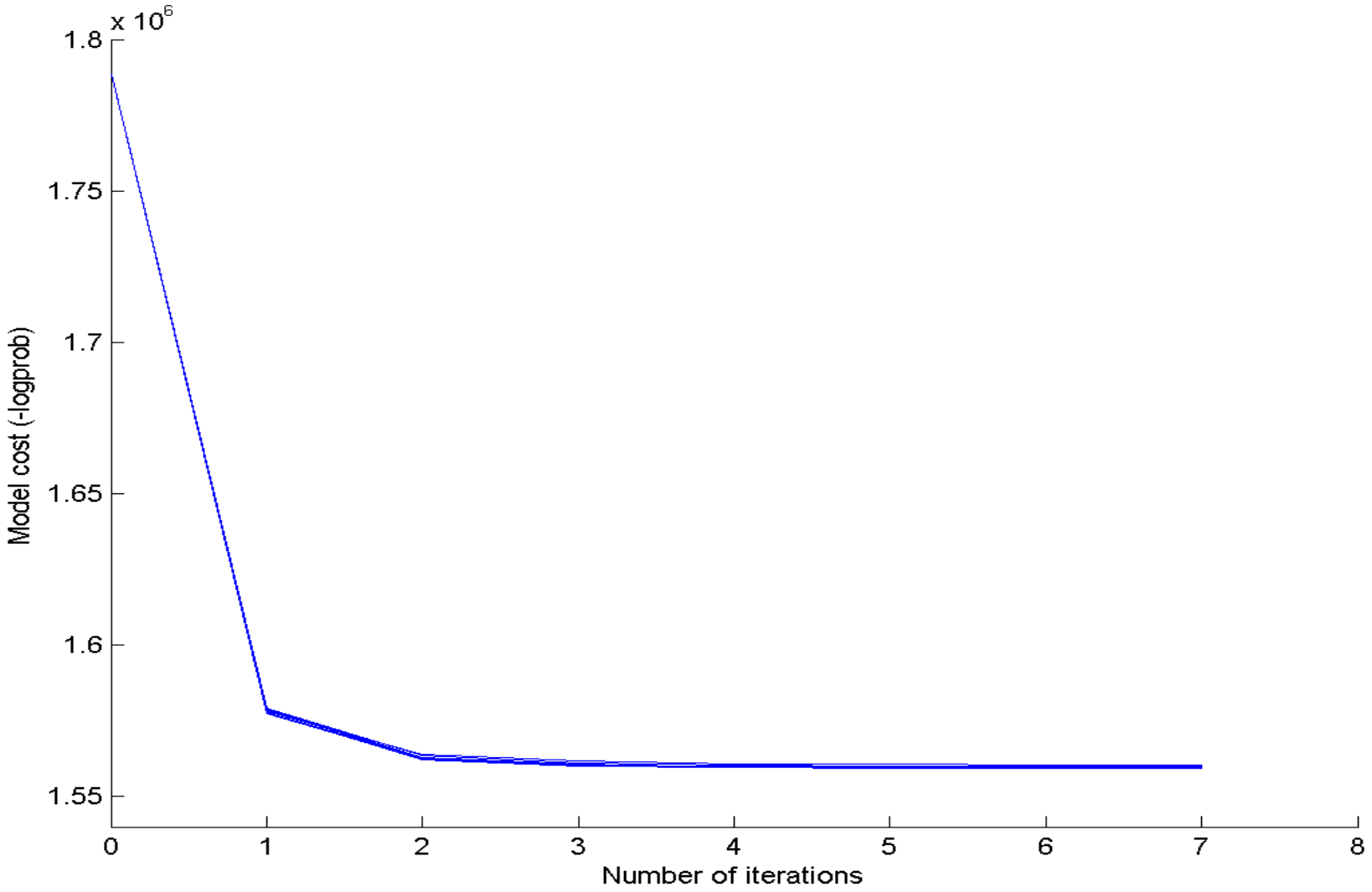
Sequential Search



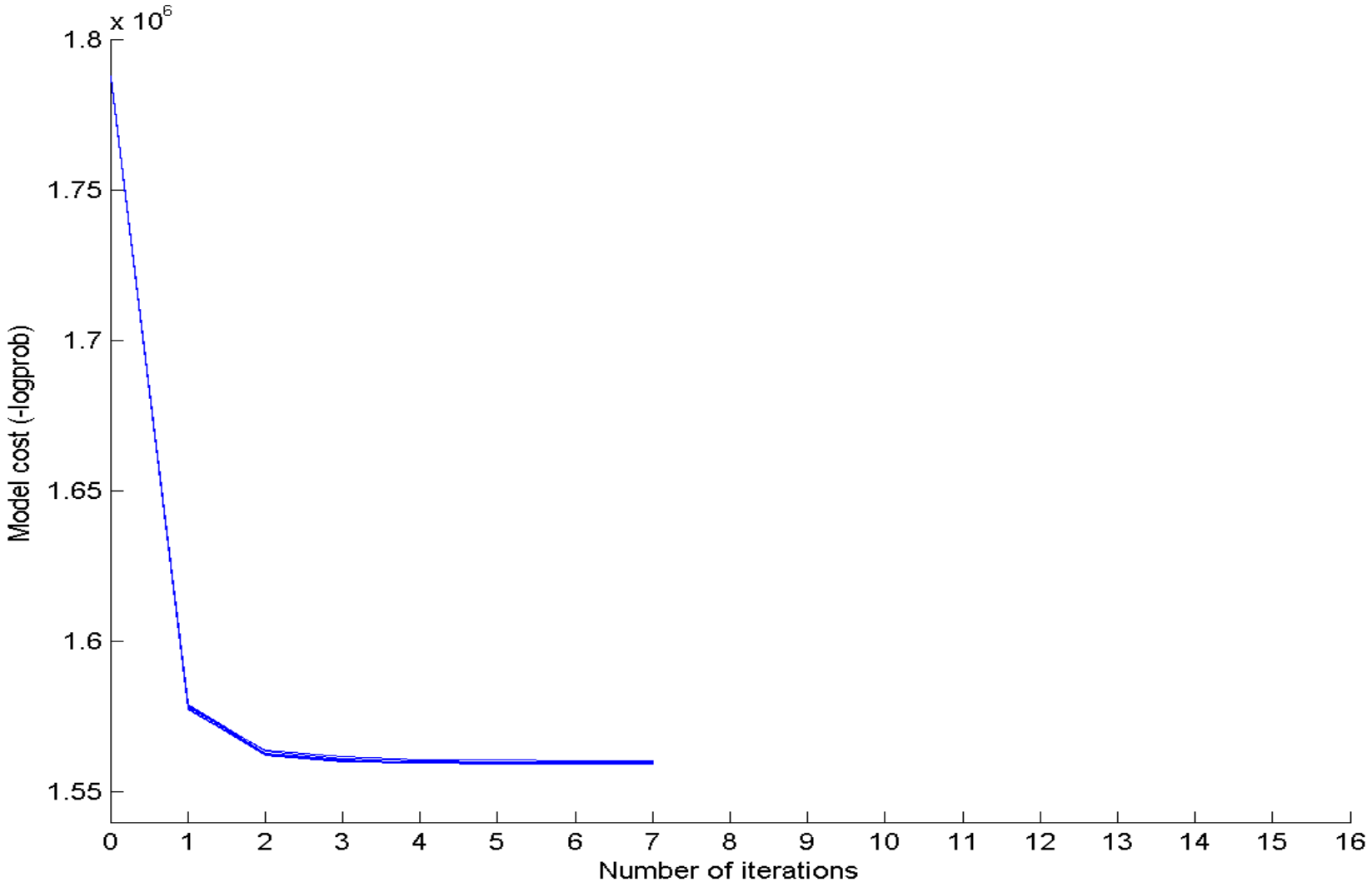
Sequential Search



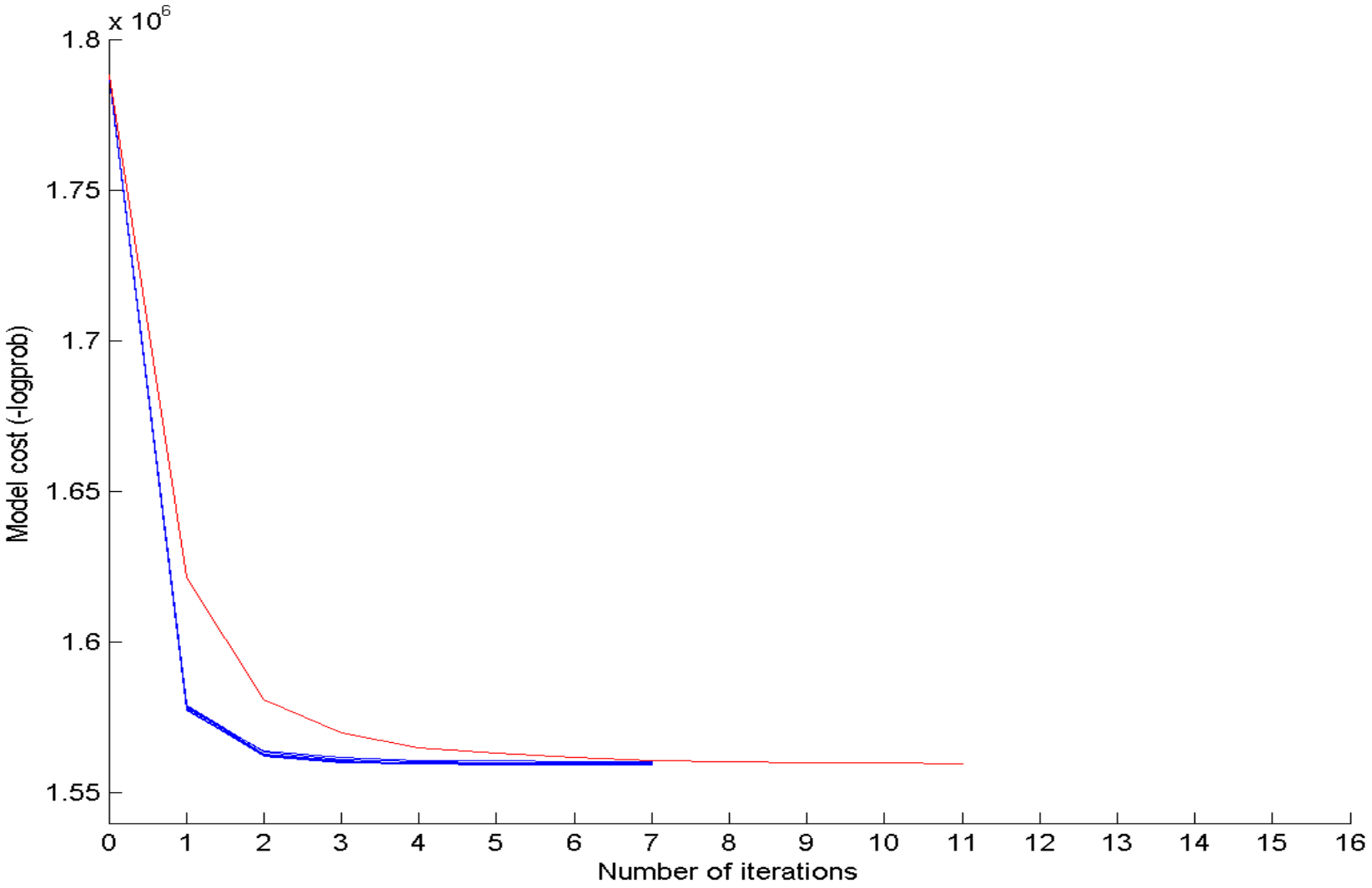
Sequential Search (different vocabulary scan orders)



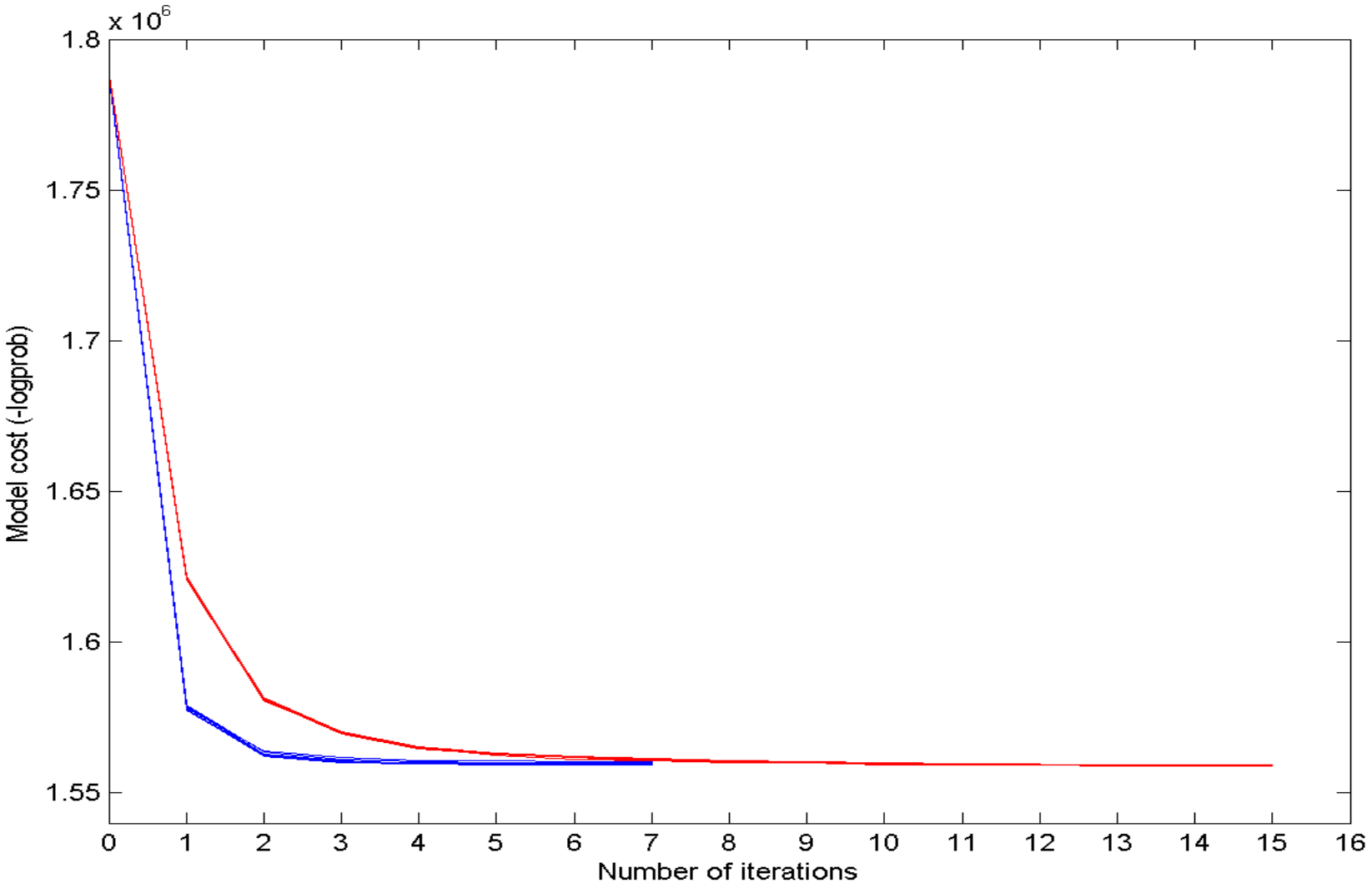
Sequential Search vs. Parallel Search



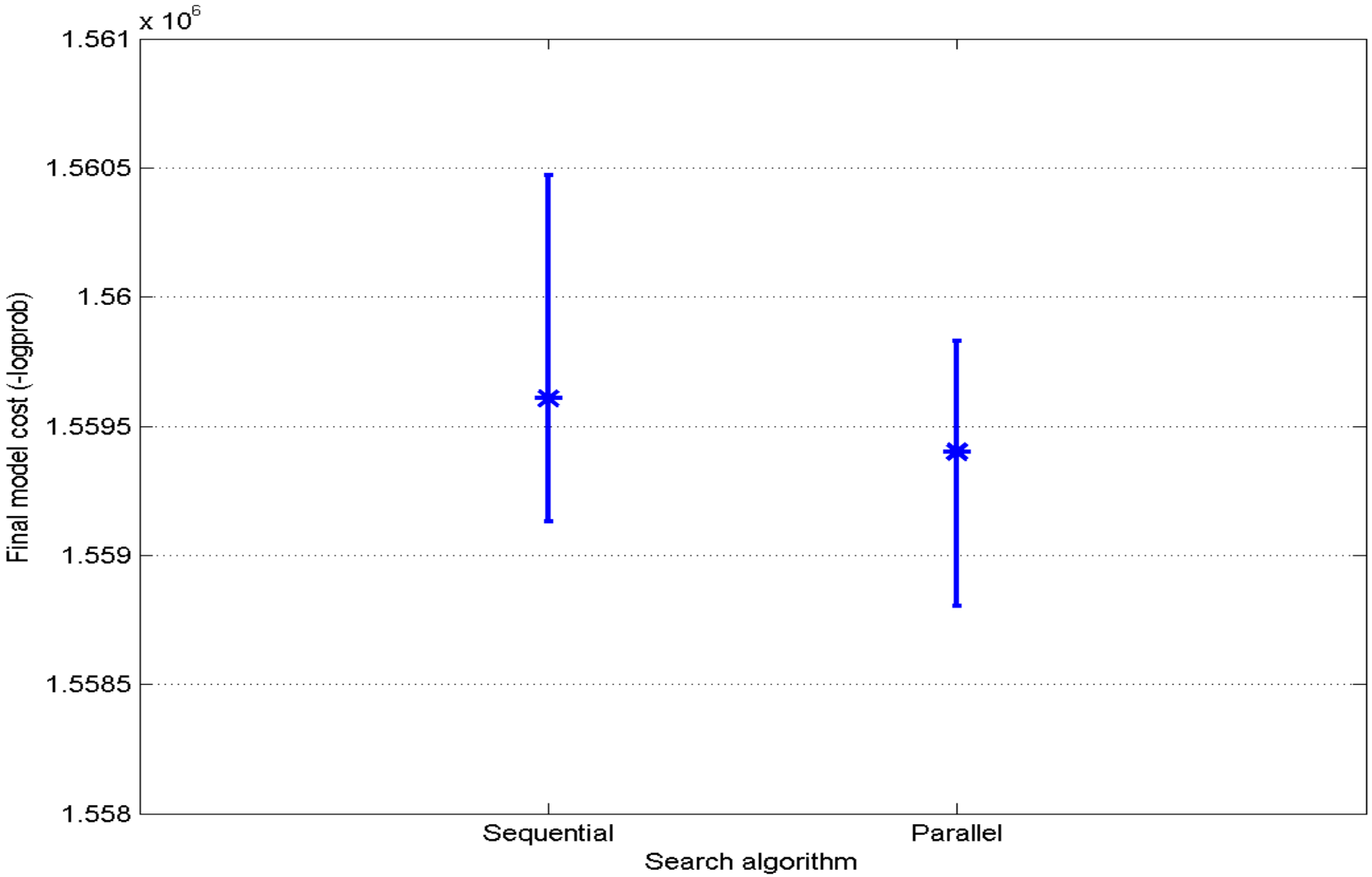
Sequential Search vs. Parallel Search



Sequential Search vs. Parallel Search



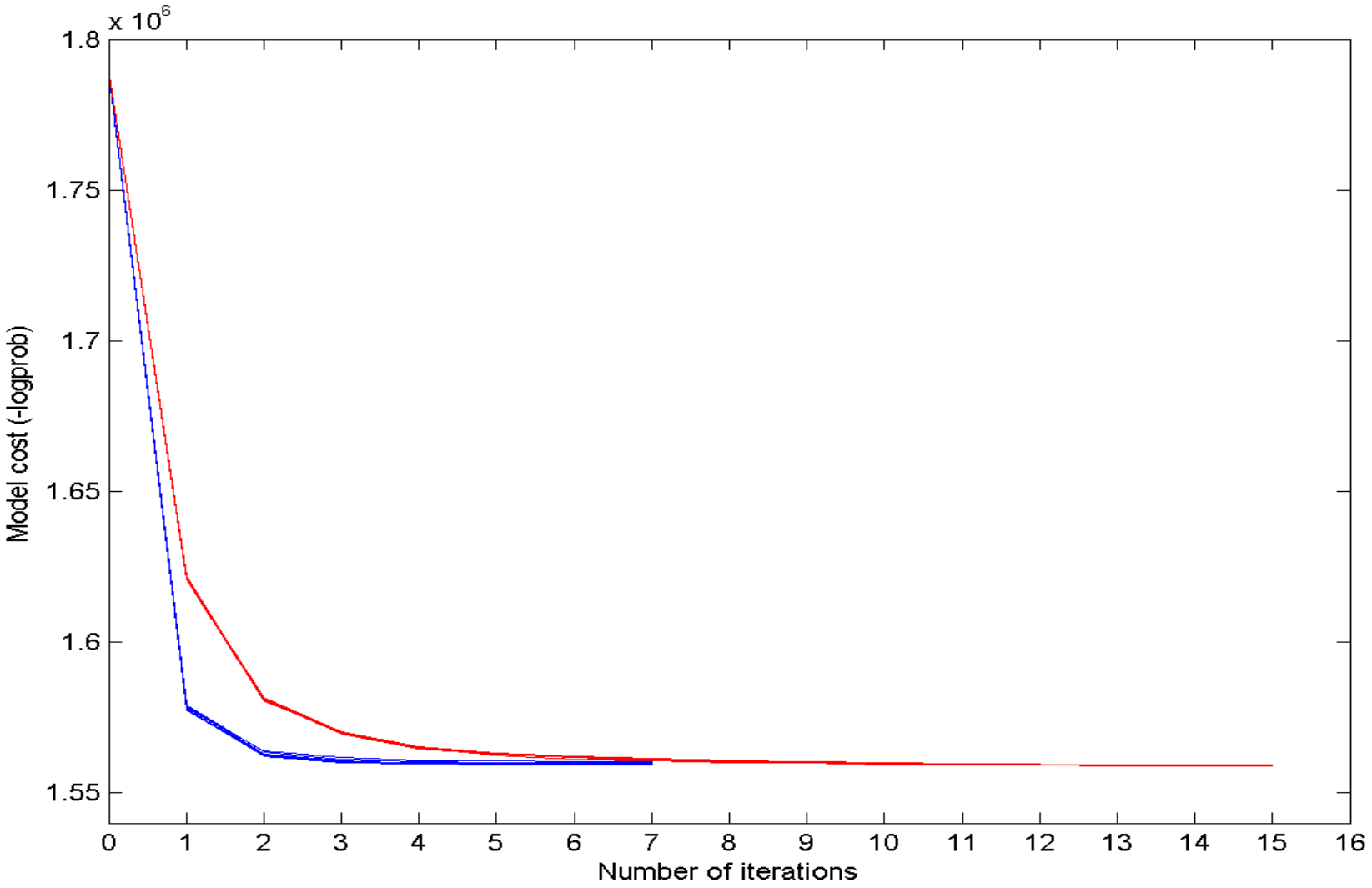
Sequential Search vs. Parallel Search



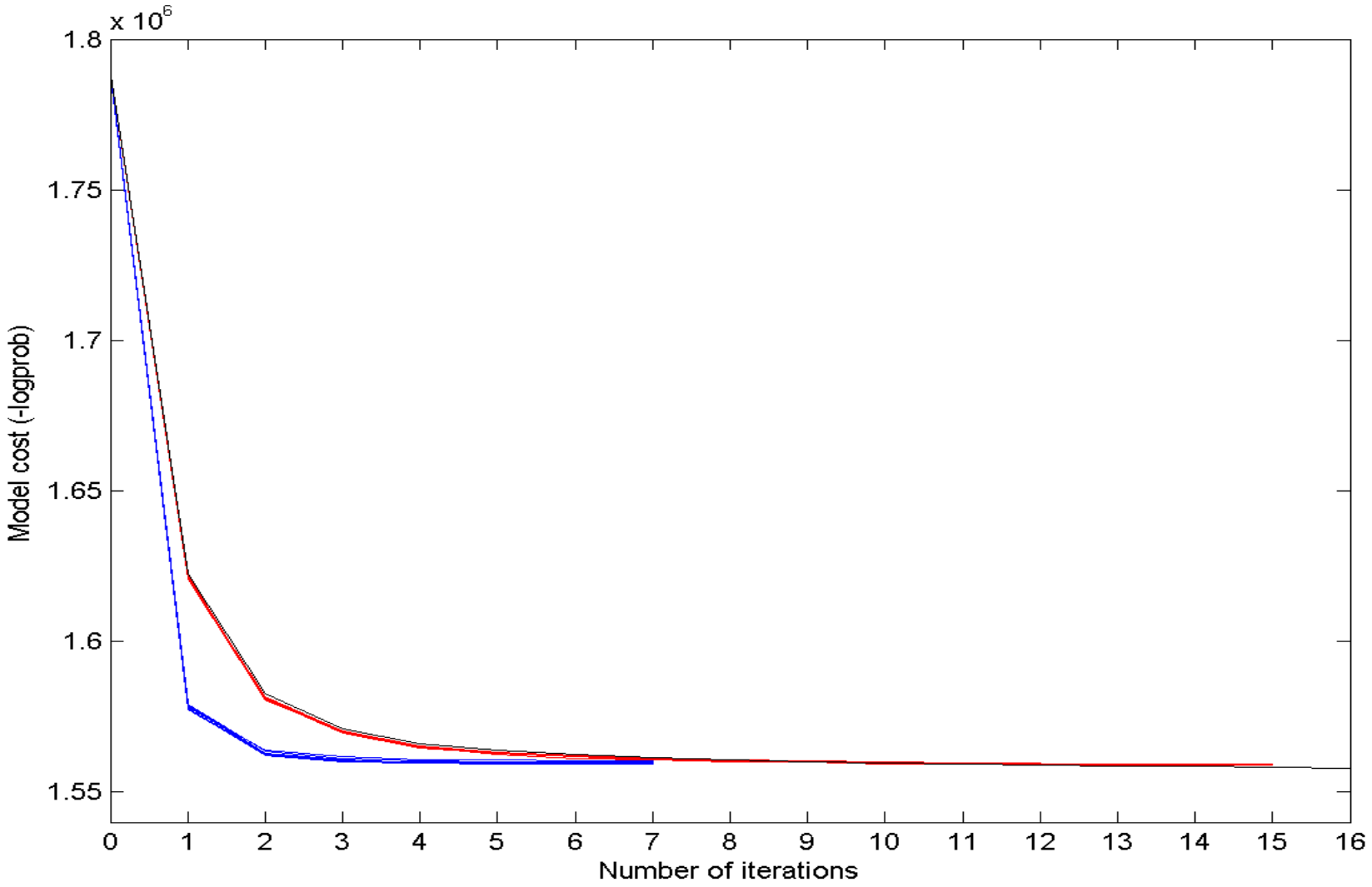
Random Search

- Even less greedy
- Do not automatically accept the maximum probability segmentation, instead make a random draw proportional to the posteriors
  - *cf.* Gibbs sampling

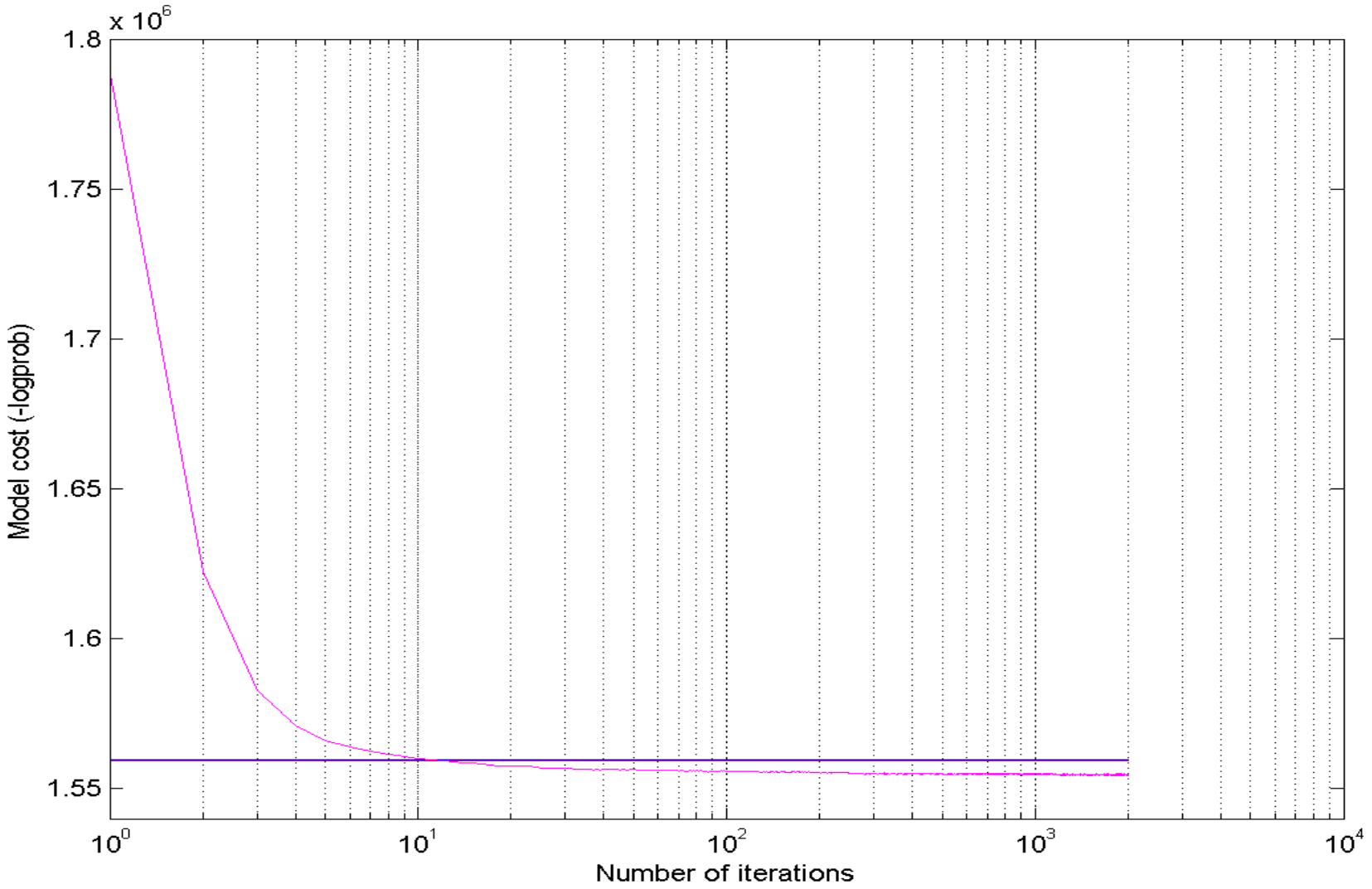
Deterministic vs. Random Search



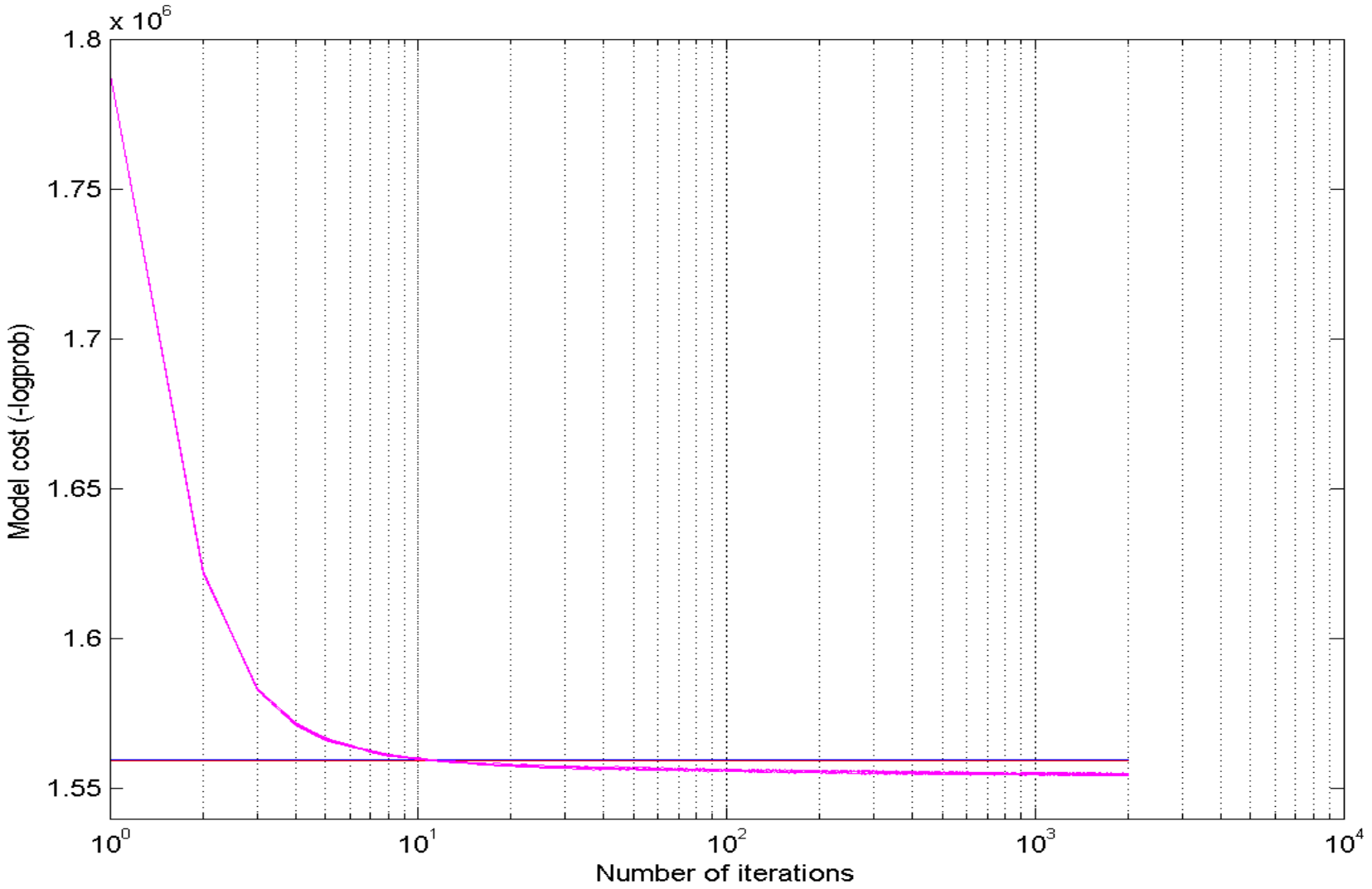
Deterministic vs. Random Search



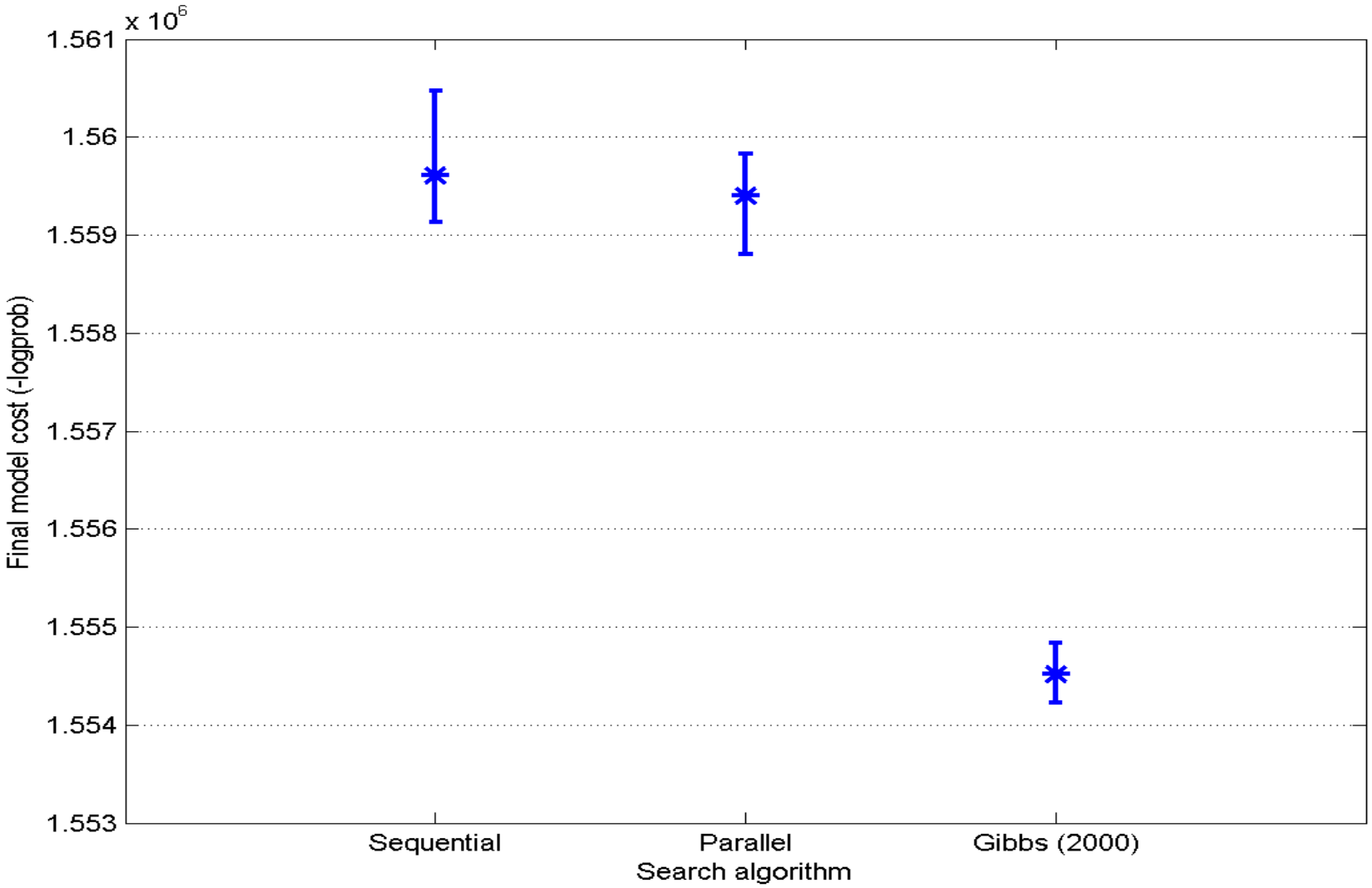
Deterministic vs. Random Search



Deterministic vs. Random Search



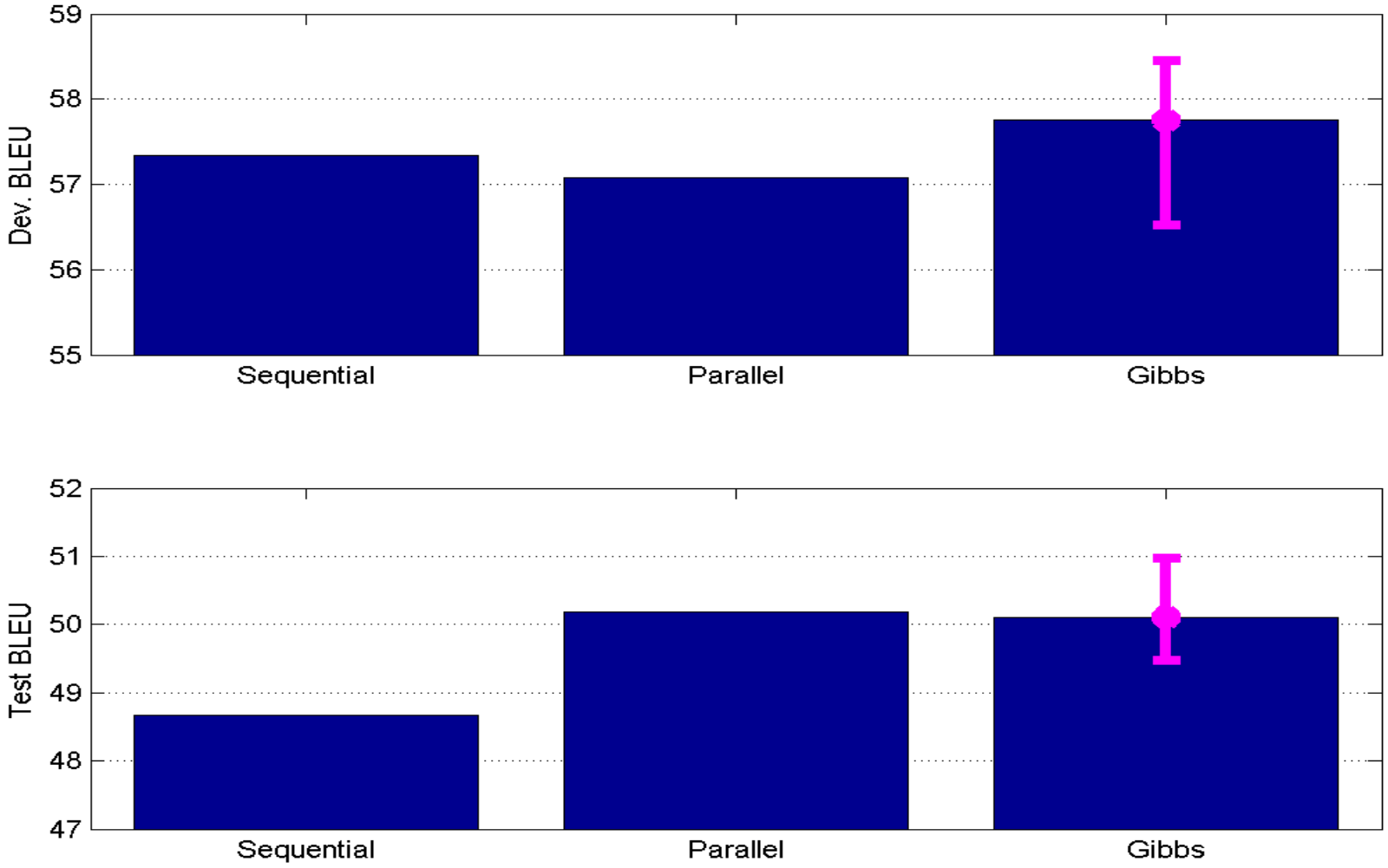
Deterministic vs. Random Search



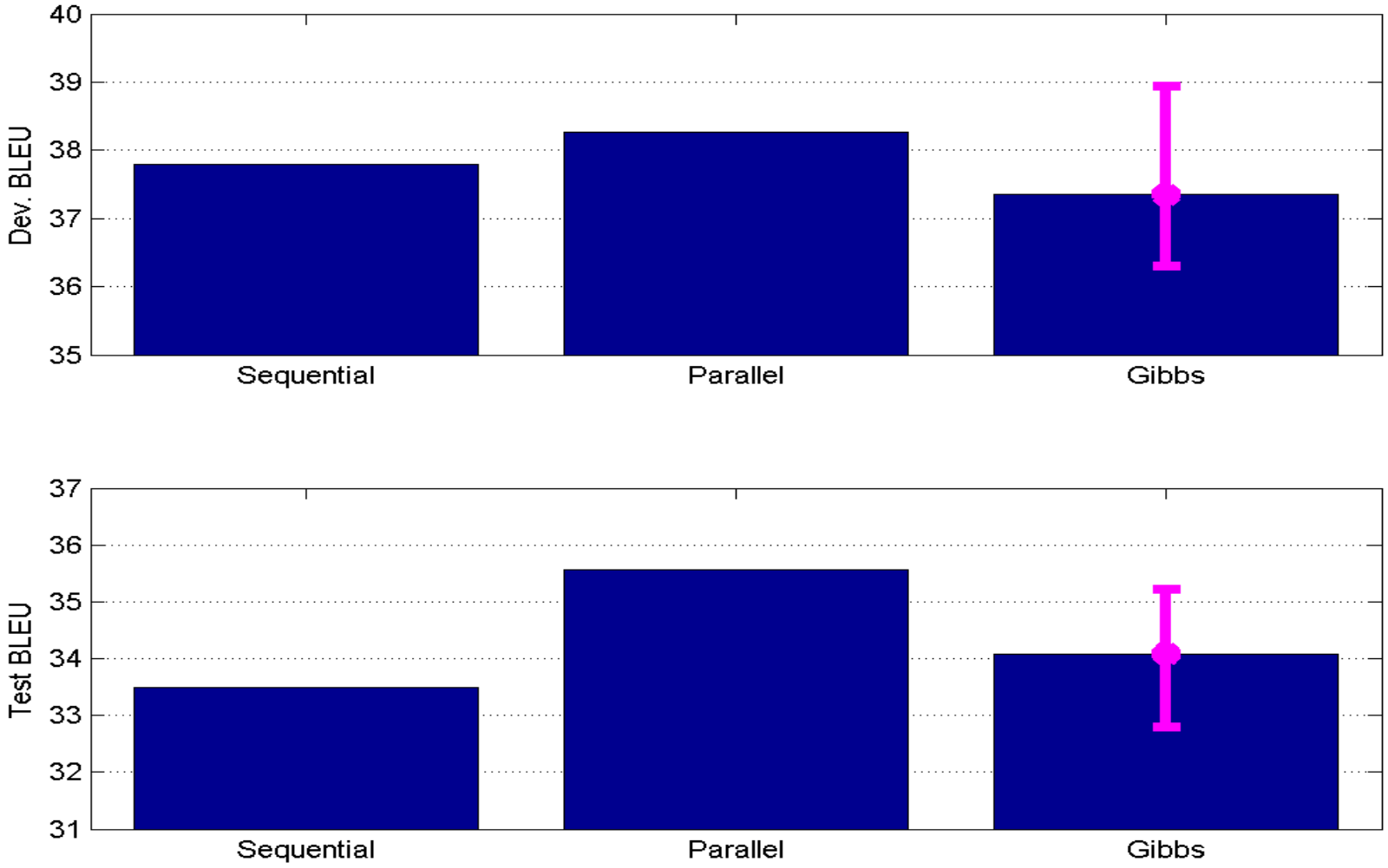
So far…

- We can obtain lower model costs by being less greedy in search
- Does it translate to BLEU scores?

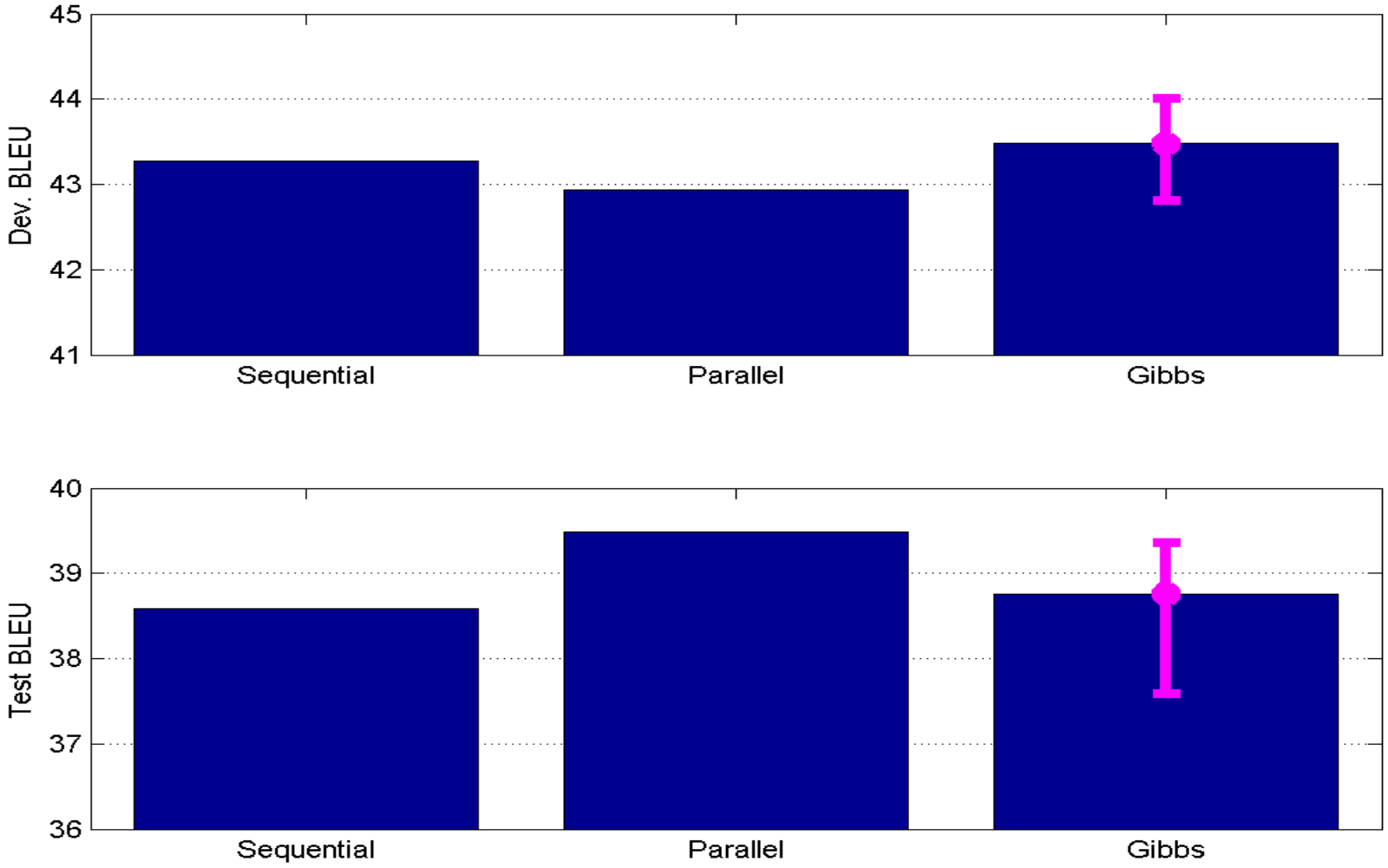
Turkish-to-English



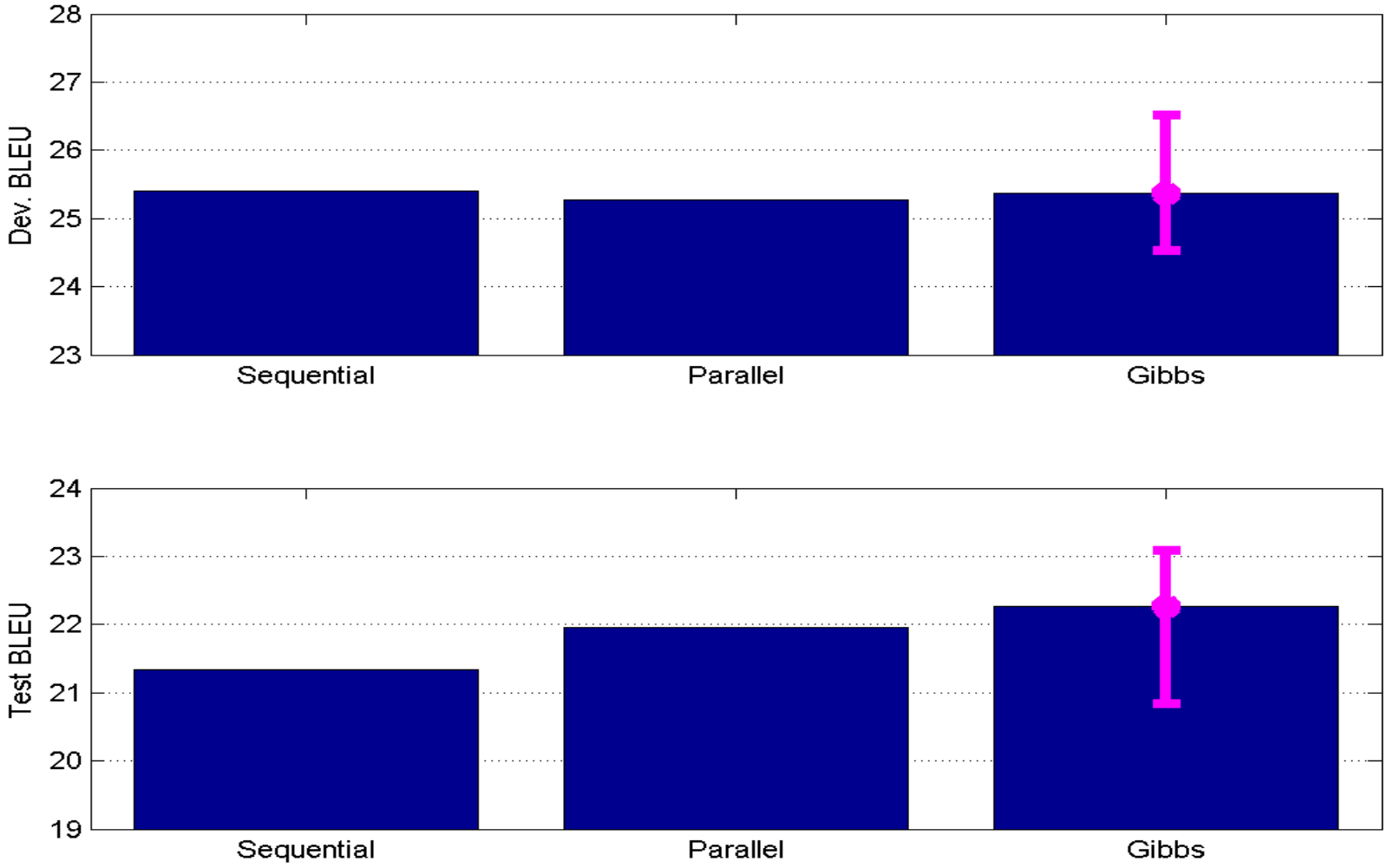
English-to-Turkish



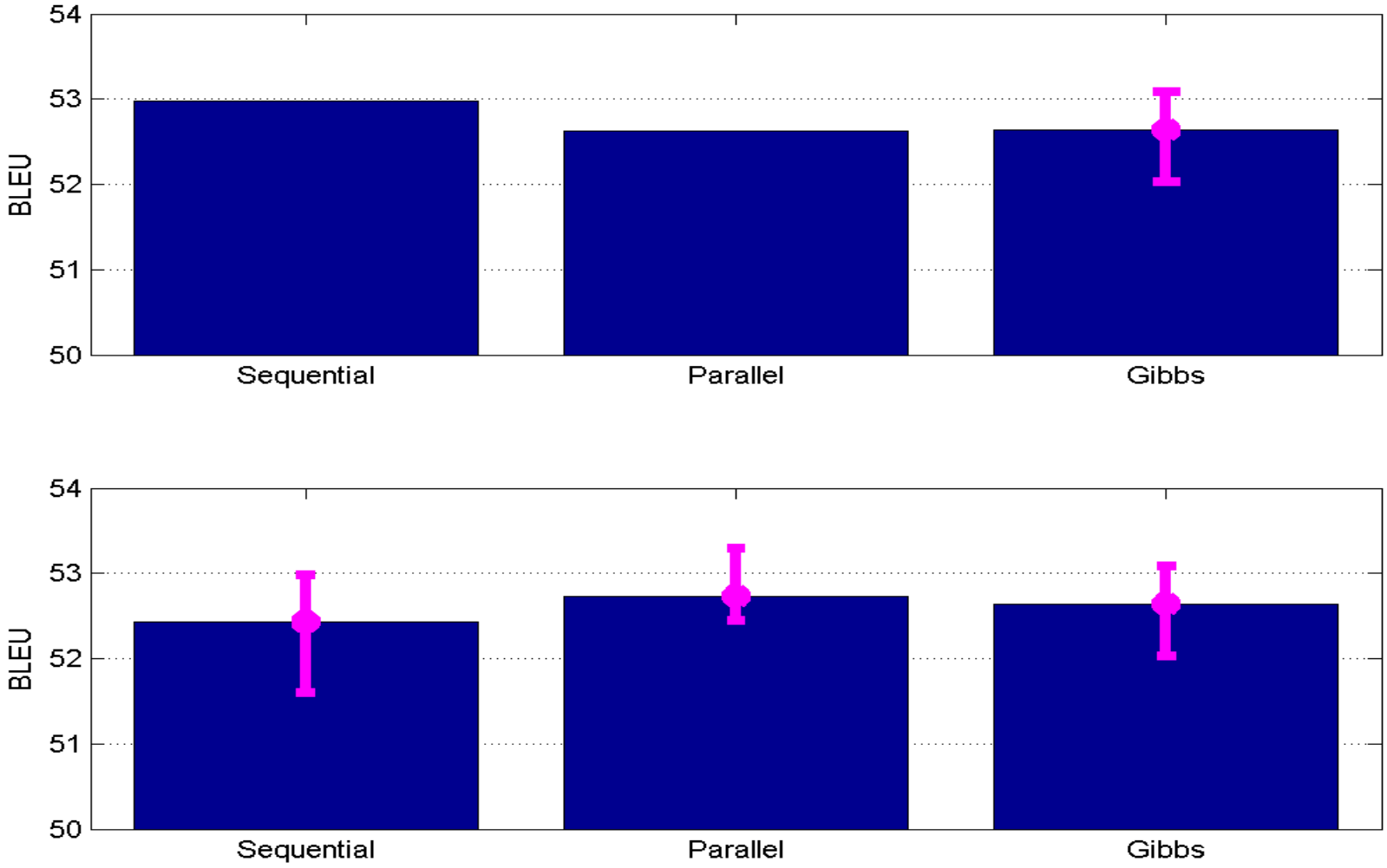
Turkish-to-English (1 reference)



English-to-Turkish (1 reference)



On a Large Test Set (1512 sentences) Turkish-to-English, No MERT



Optimizing Segmentation for Statistical Translation

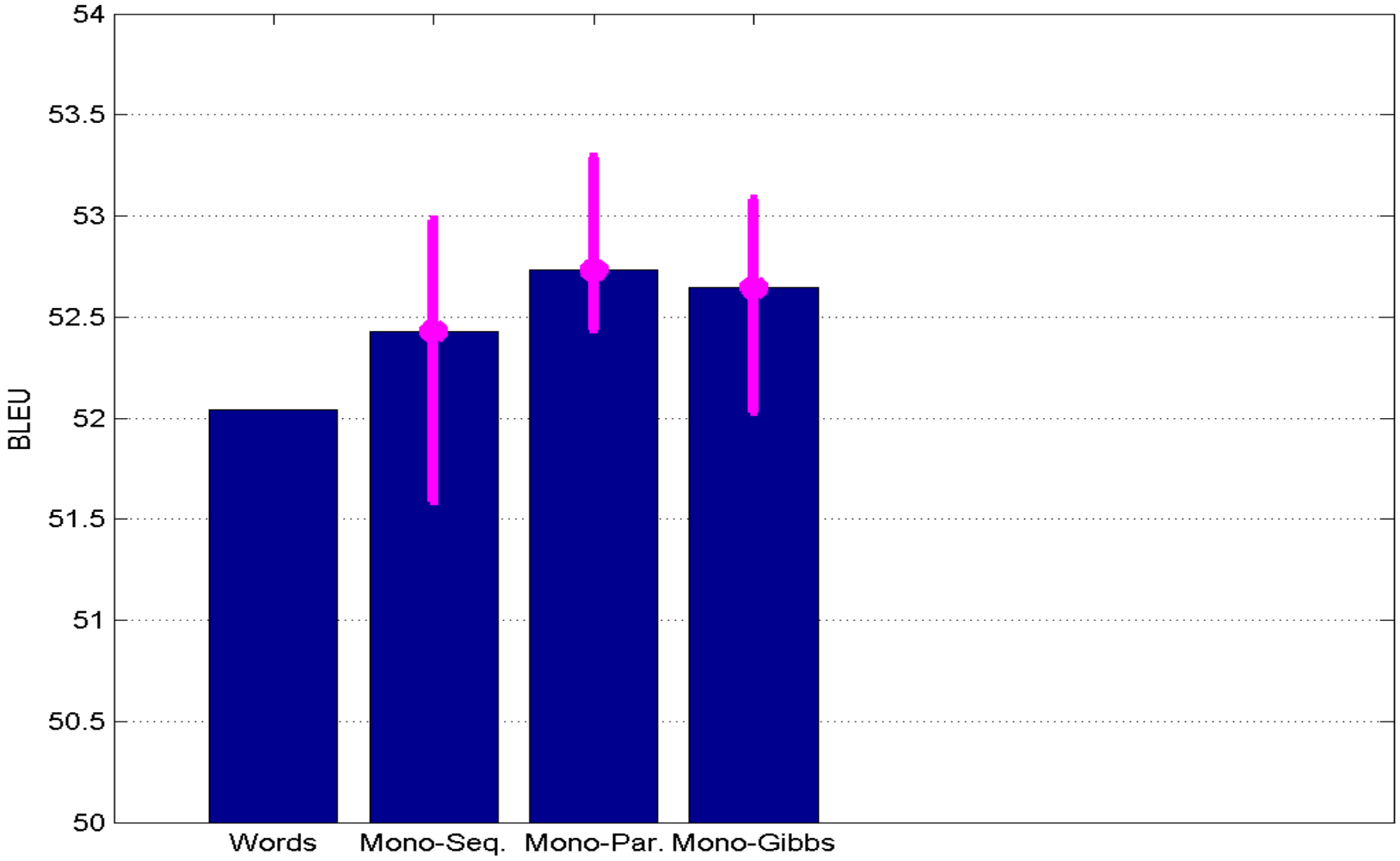
- The best-performing segmentation is highly task-dependent
  - Could change when paired with a different language
  - Depends on size of parallel corpora
- For a given parallel corpus, what is the optimal segmentation in terms of translation performance?

Adding Bilingual Information

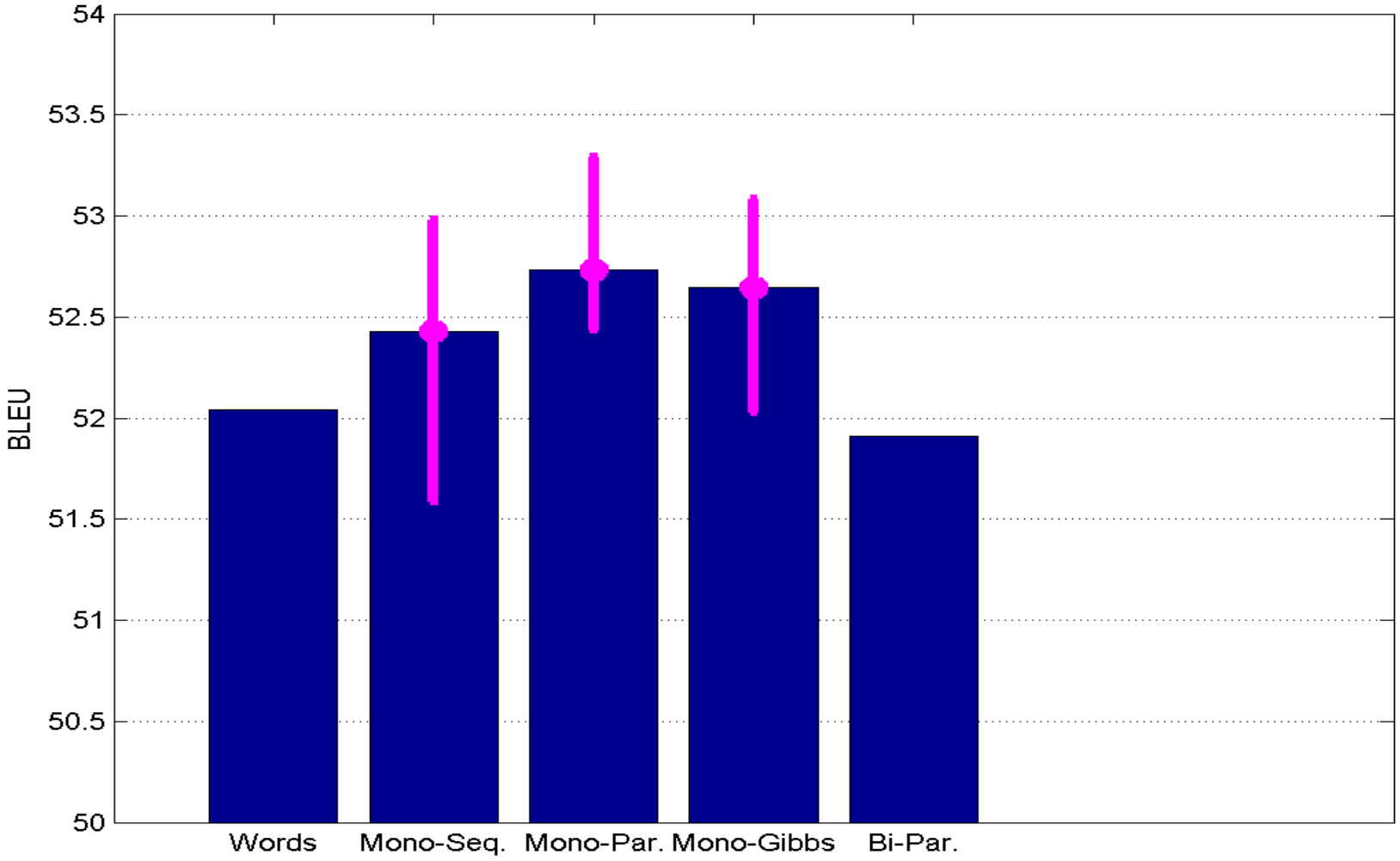
$$\hat{M}_f = \arg\max_{M_f} P(M_f)P(f \mid M_f)P(e \mid f)$$

- $P(e \mid f)$: Using IBM Model-1 probability
  - Estimated via EM

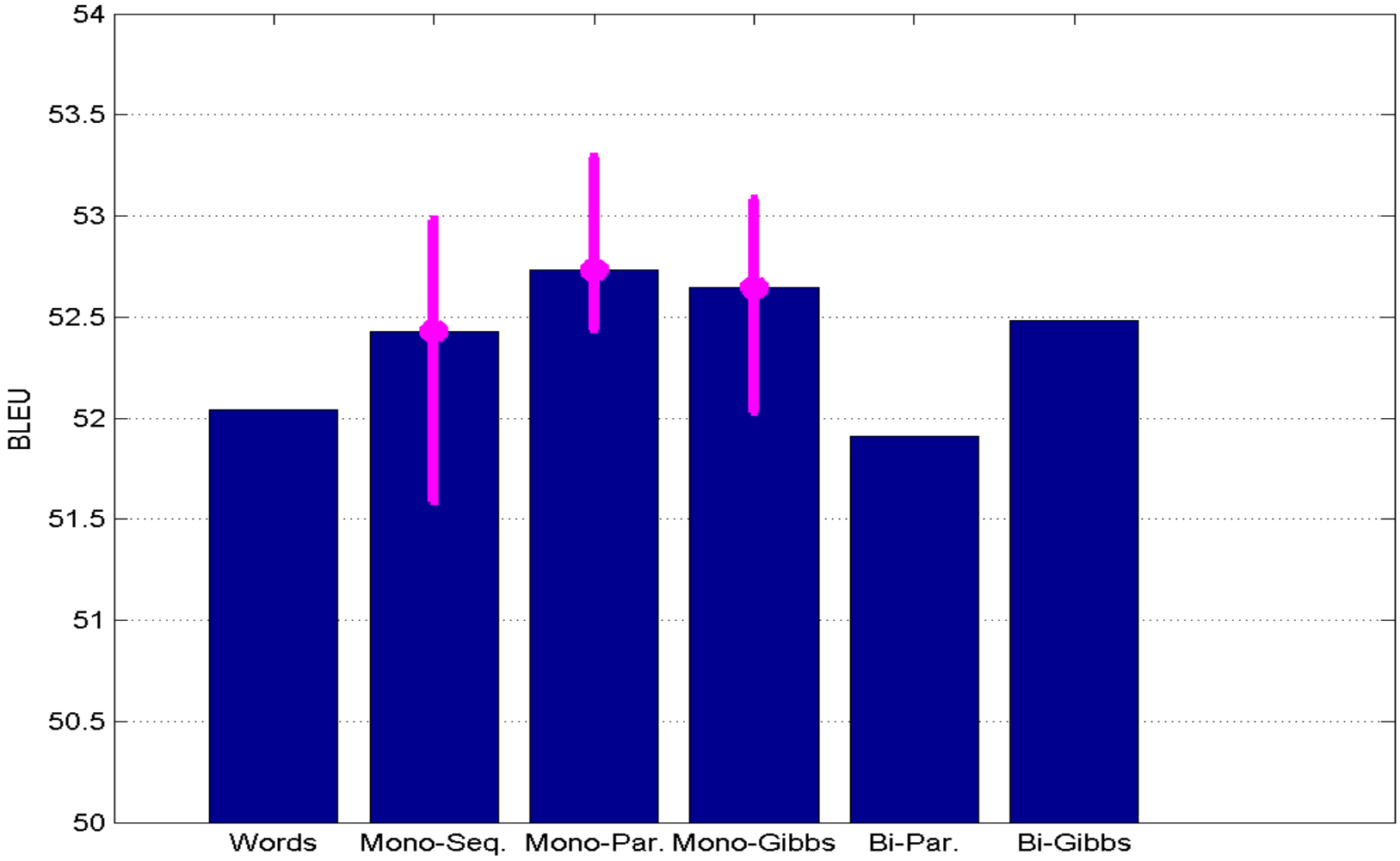
Results


54
53.5
53
52.5
52
51.5
51
50.5
50
BLEU
Words
Mono-Seq.
Mono-Par.
Mono-Gibbs

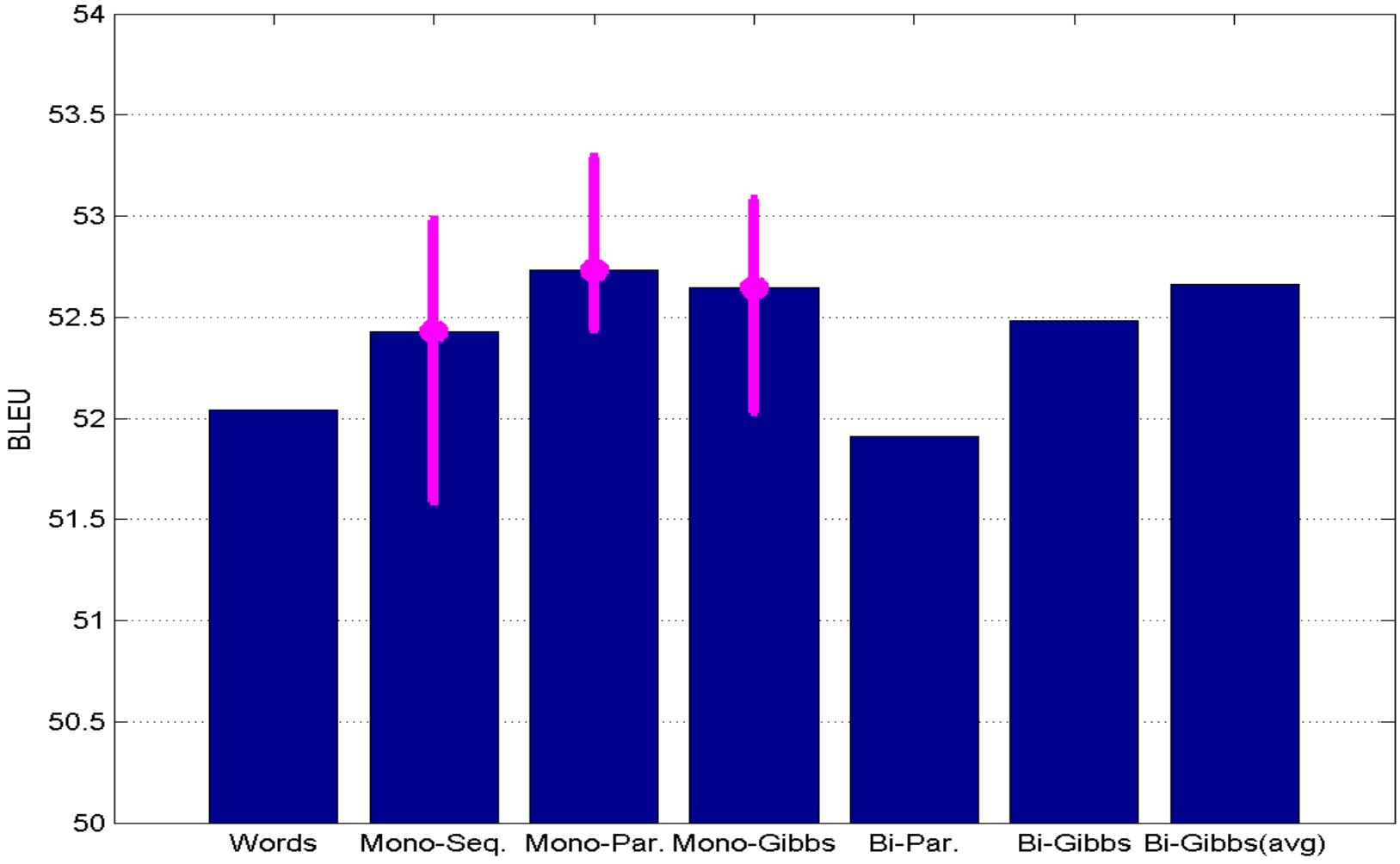
Results


54
53.5
53
52.5
52
51.5
51
50.5
50
BLEU
Words
Mono-Seq.
Mono-Par.
Mono-Gibbs
Bi-Par.

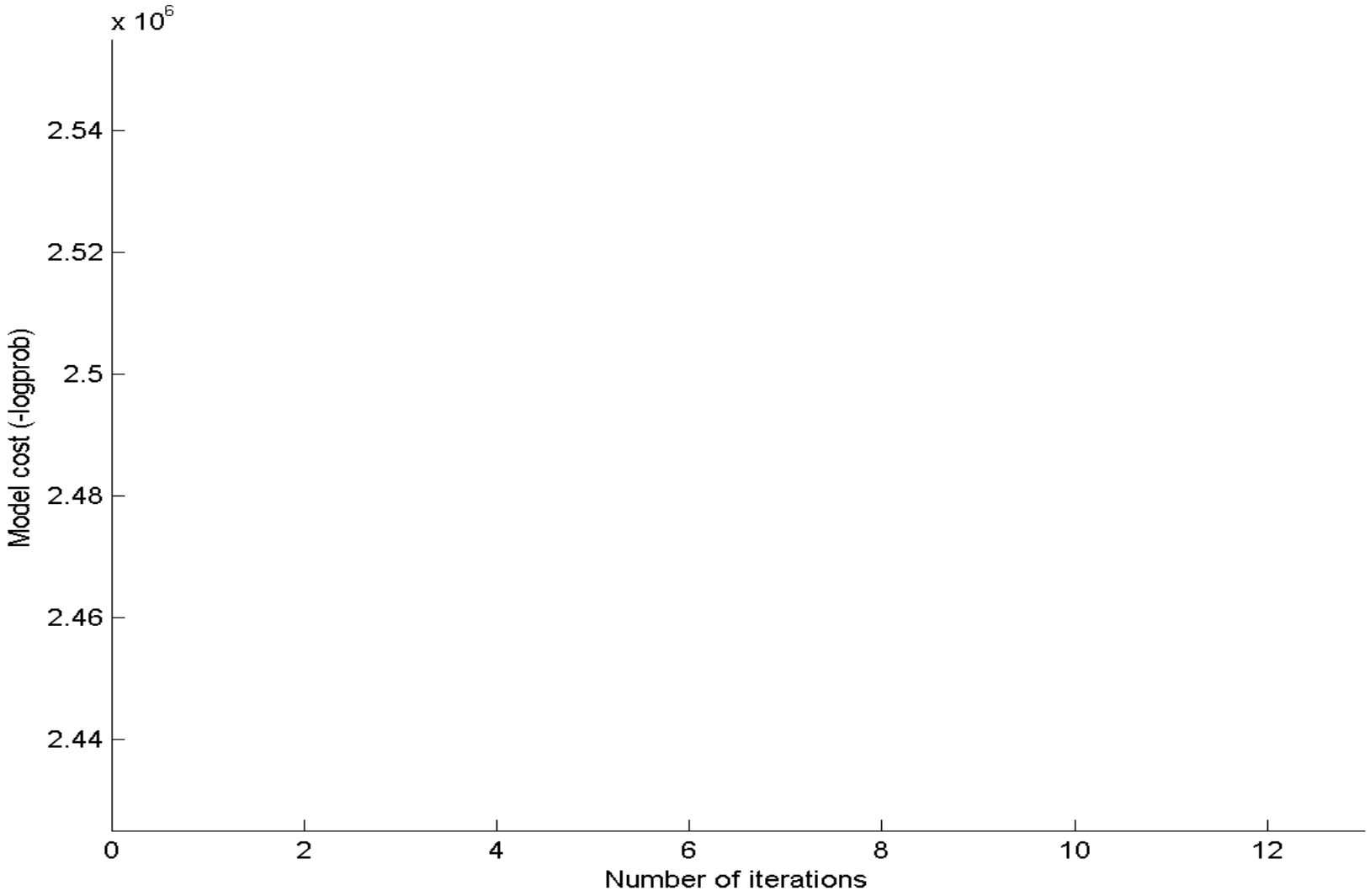
Results


54
53.5
53
52.5
52
51.5
51
50.5
50
BLEU
Words
Mono-Seq.
Mono-Par.
Mono-Gibbs
Bi-Par.
Bi-Gibbs

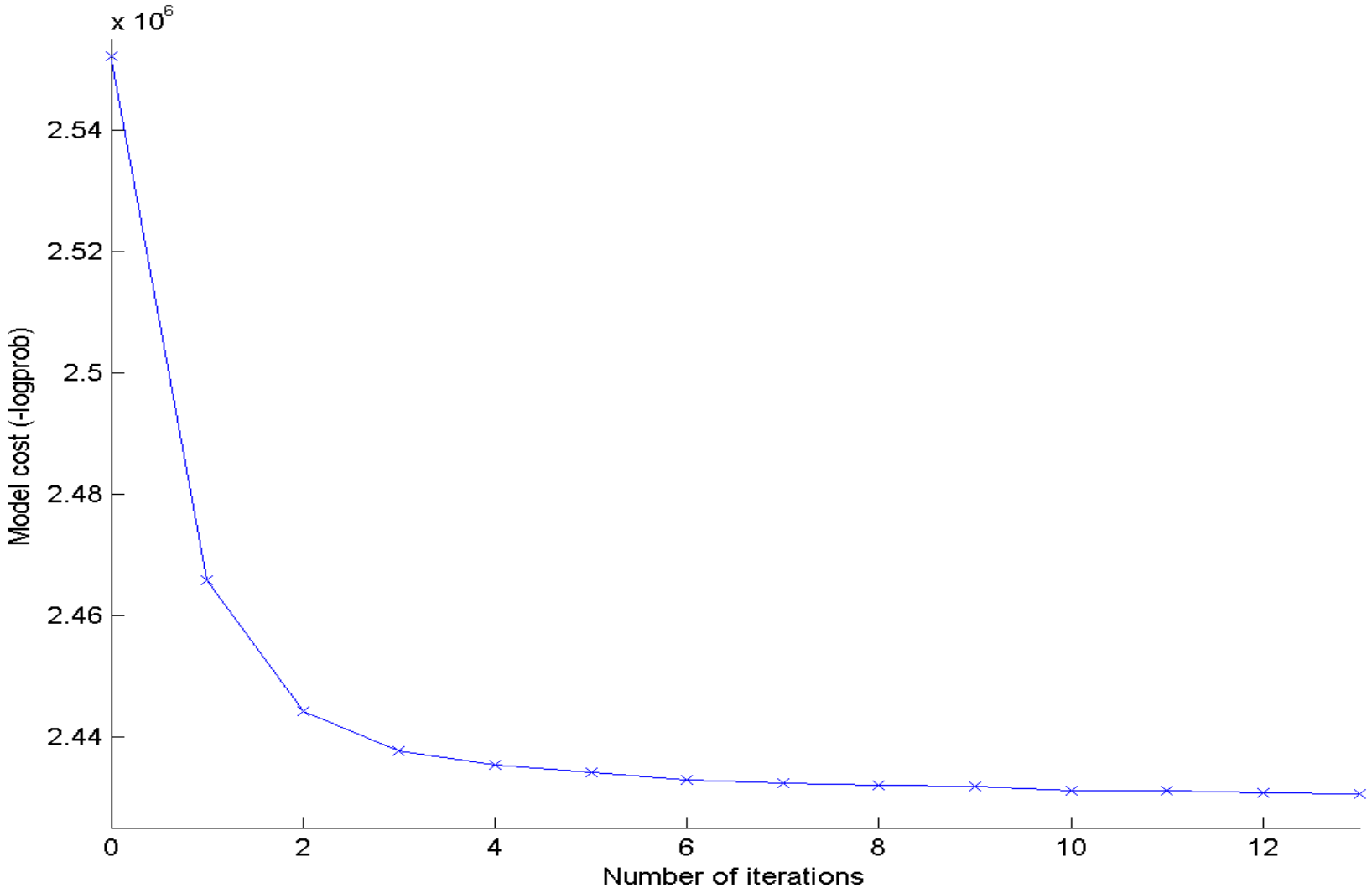
Results


54
53.5
53
52.5
52
51.5
51
50.5
50
BLEU
Words
Mono-Seq.
Mono-Par.
Mono-Gibbs
Bi-Par.
Bi-Gibbs
Bi-Gibbs(avg)

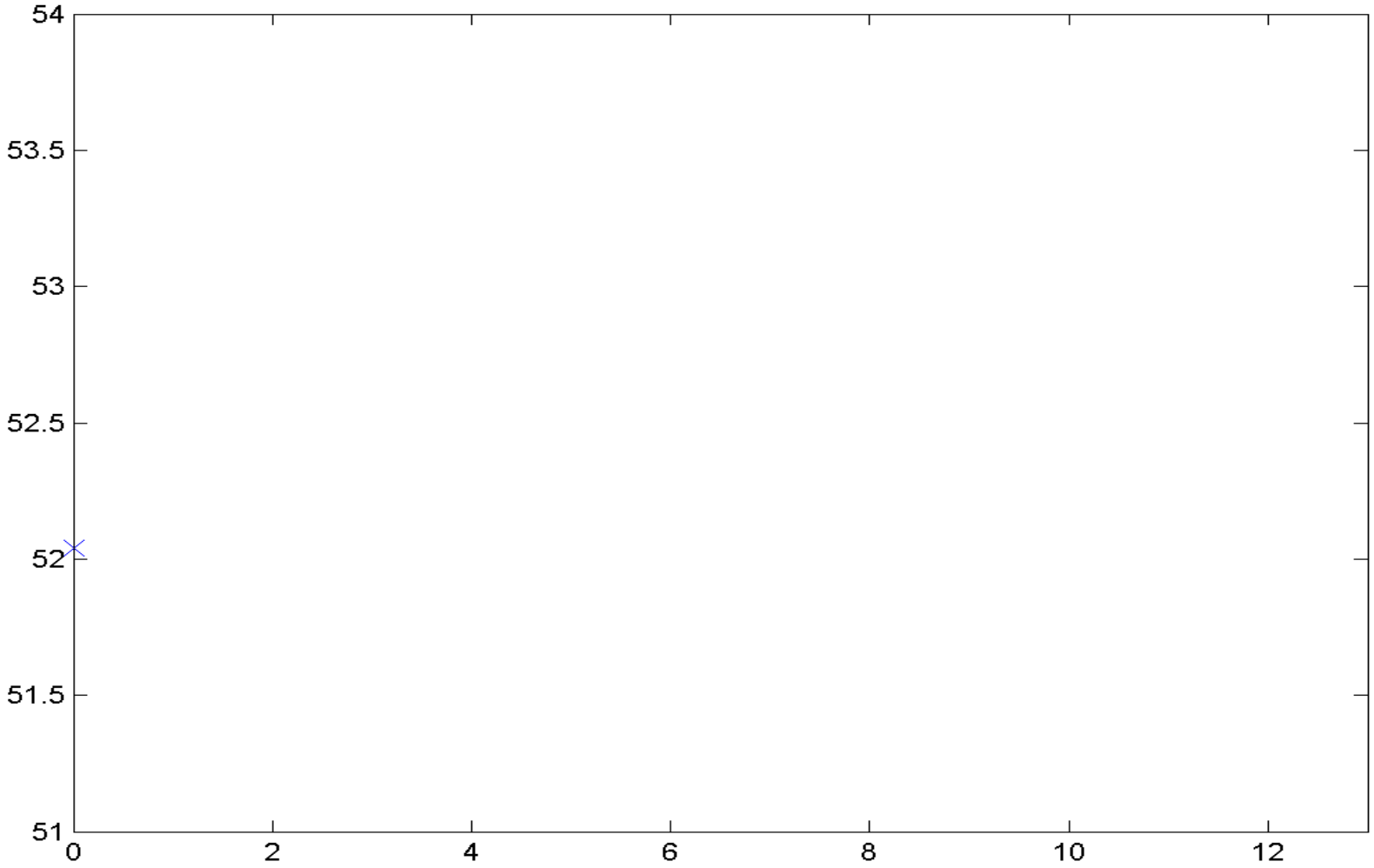
Evolution of the Gibbs Chain



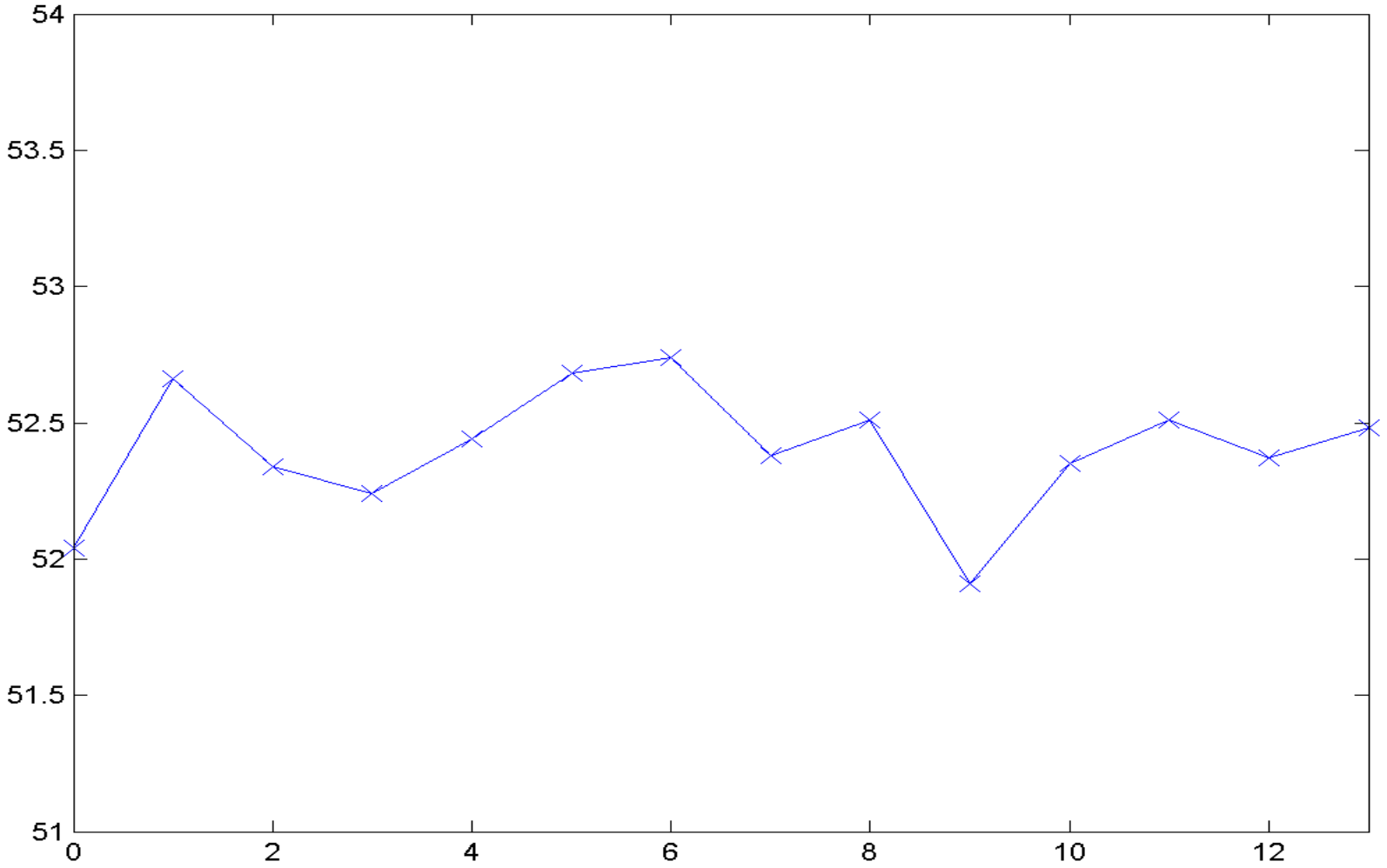
Evolution of the Gibbs Chain



Evolution of the Gibbs Chain (BLEU)



Evolution of the Gibbs Chain (BLEU)



Conclusions

- Probabilistic model for unsupervised learning of segmentation
- Improvements to the search algorithm
  - Parallel search
  - Random search via Gibbs sampling
- Incorporated (an approximate) translation probability to the model
- So far, model scores do not correlate well with BLEU scores