“Drinking Maicaraosaft…”

Talk on Language Neutral Statistical Machine Transliteration

by Tobias Kellner

ประเทศอินเดีย
ལོན་ཁོན་
لندن
ლონდონი
லண்
இலண்
लंदन
ಇന್ಡ

On work carried out in summer 2006 with the
Multilingual Systems Group
at **Microsoft Research India, Bangalore**

Supervised by A. Kumaran <a.kumaran@microsoft.com>

ประเทศอินเดีย
ལོན་ཁོན་
لندن
لوندون
ლონდონი
இலண்
लंदन
ಇന್ಡ್

मेरा नाम तोबि है

my name [illegible] is

ประเทศอินเดีย
ལོན་ཁོན་
ლონდონი
لندن
לונדון
இலண்
लंदन
ಇಂಡ್ಯ

Outline

- Introduction to Transliteration
  - What are we trying to do?
- Machine Transliterations
  - How can computers transliterate?
  - What is our approach?
- Results obtained
  - Does it work ?!?

ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇಂಡಿಯಾ

What is Transliteration?

- Definition (short):
  "*Converting from one orthography to another*"

- Definition (longer)
  "*Replacing a word in the source language/script with a phonetically identical word in the target language/script*"

ประเทศอินเดีย לונדון ལོན་ཁོན་ لندن ლონდონი இலண்

Transliteration: Example

लंदन

ഇന്ത്യ

नमस्ते ➔ Namaste

is a transliteration, whereas

नमस्ते ➔ Hello

is a translation


PUREVEG
Hotel Swathi
ಹೋಟೆಲ್ ಸ್ವಾತಿ
ಹೊಟೆಲ್ ಸ್ವಾತಿ

Transliteration in MT

- Out-Of-Vocabulary Words
  - many (most?) OOV words are named entities
  - Named Entities should be transliterated

- Text alignment
  - identifying cognates ("obvious translations") for alignment of words and sentences

ประเทศอินเดีย
לונדון
ལོན་ཁོན་
لندن
ლონდონი
இலண்
लंदन
ഇന്ത്യ

“Iraq's Foreign Minister Hoshyar Zebari said...”

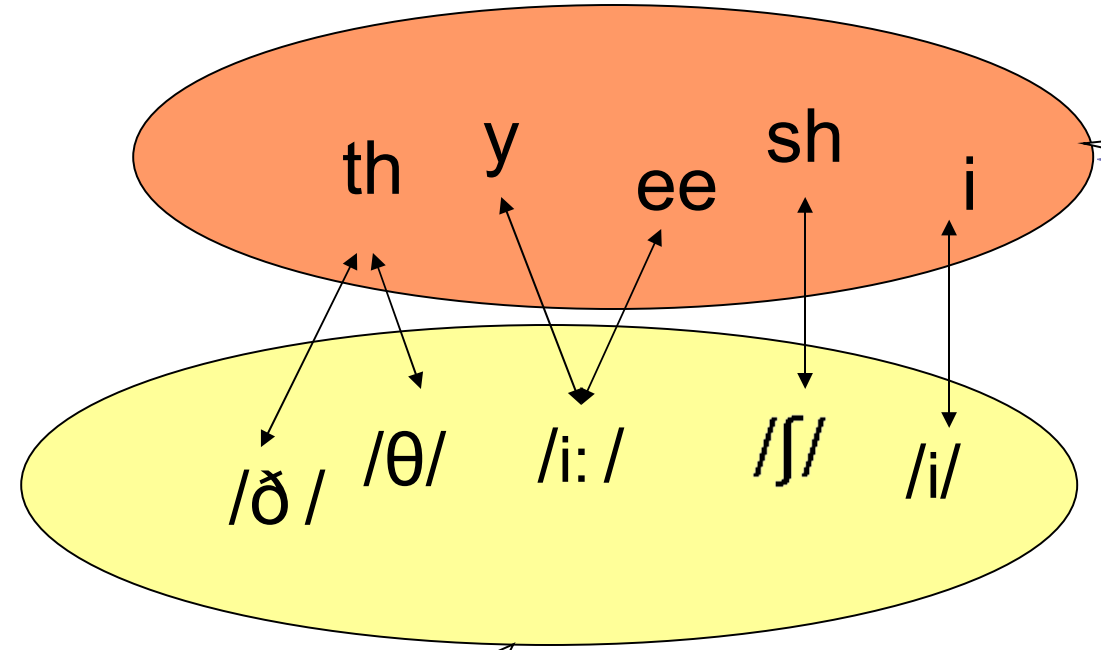
"उधर इराक के विदेश मंत्री होशियार ज़ेबारी ..”

“udhar irak ke videsh mentri hoshiyar zebari...”

ประเทศอินเดีย
ལོན་ཁོན་
لندن
لوندان
ლონდონი
இலண்
लंदन
ഇന്ത്യ

A closer look at the problem

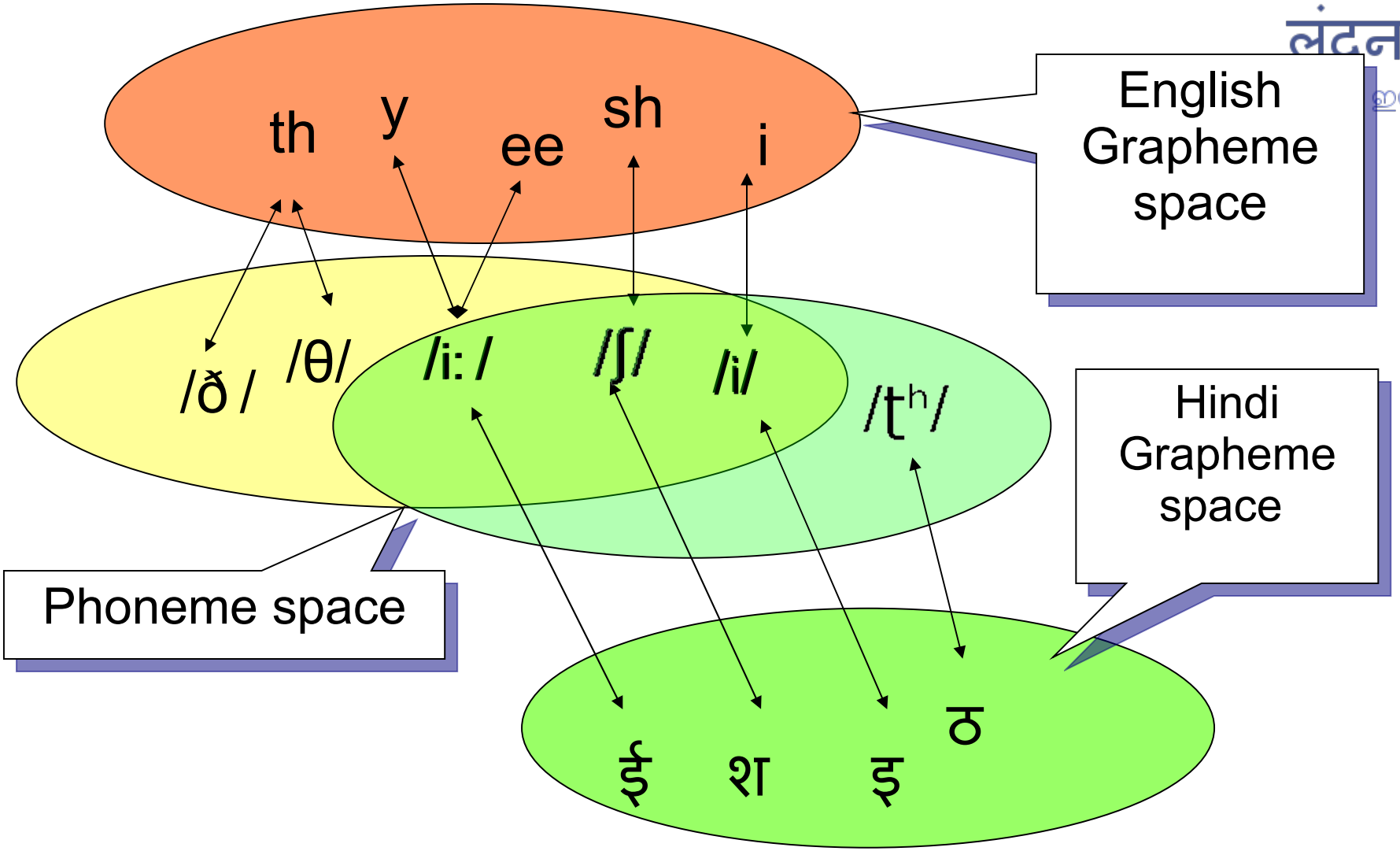
A closer look at language



English Grapheme space

Phoneme space

A closer look at language



ประเทศอินเดีย ལོན་ཁོན་ לונדון لندن ლონდონი இலண்

Direction of transliteration

लंदन

ഇന്ത്യ

- Forward Transliteration
  - Word does not exist in Target Language
  - results (usually) not in any dictionary/list
  - (usually) all phonetically close results ok
  - मीना → Mina
    → Minaa
    → Meena
    → Meenaa...

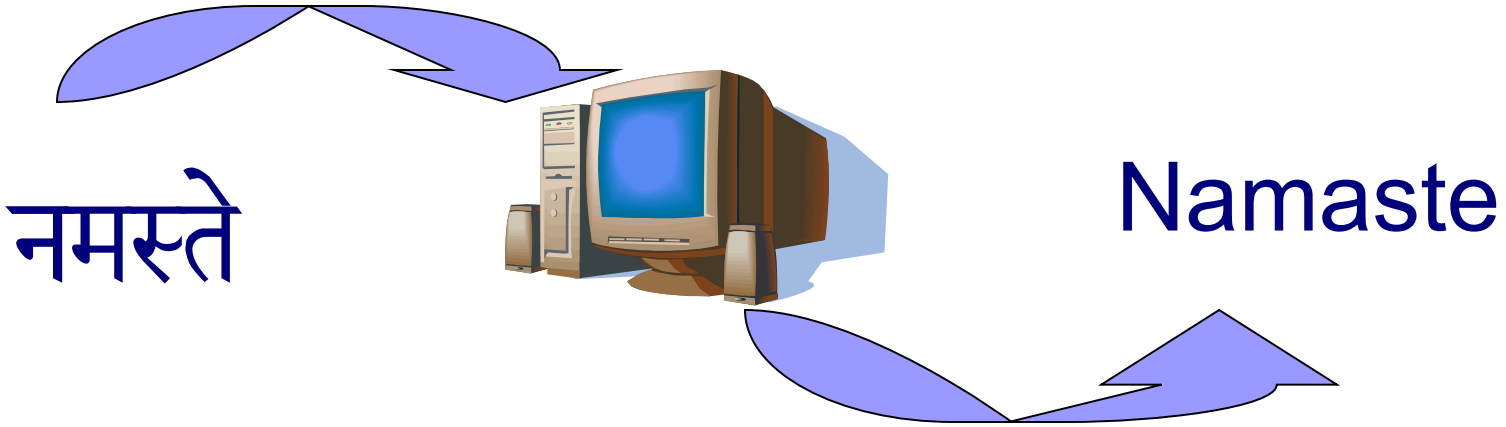
ประเทศอินเดีย ལོན་ཁོན་ لندن לונדון
ლონდონი இலண்
लंदन
இந்

Direction of transliteration

- Backward Transliteration
  - Word exists in Target Language
  - Correct form can be found in word lists
  - Only one form is acceptable
  - लंदन → London
    → *Lundan
    → *Lundon…

ประเทศอินเดีย
ལོན་ཁོན་
لوندان
لندن
ლონდონი
இலண்
लंदन
ഇന്ത്യ

PART II:
MACHINE TRANSLITERATION



ประเทศอินเดีย ལོན་ཁོན་ לונדון لندن ლონდონი இலண்

लंदन ಇಂಡಿಯಾ

Machine Transliteration

- most work on Chinese, Arabic, Korean, Katakana, Thai and Indic languages
- Three basic strategies:
  1) Manually compiled rules/tables (not discussed here)
  2) **Phonetic-based model** ("6-fold path") (Knight & Graehl, 1998)
  3) **Direct Orthographical Mapping** (Al-Onaizan & Knight, 2002)

ประเทศอินเดีย
ལོན་ཁོན་
لندن
لوندون
ლონდონი
இலண்
लंदन
ಇಂ

Source Word – “Athens”

Source Segments – “a” “th “e” “n” “s”

Source Phonemes – /æ/ /θ/ /ə/ /n/ /s/

Target Phonemes - /e/ /tʰ/ /e/ /n/ /s/

Target Segments – “ए” “थ” “ए” “ं” “स”

Target Word – "एथेंस"

“The six-fold path”

ประเทศอินเดีย ལོན་ཁོན་ לונדון لندن ლონდონი இலண்

லंदन

ಇಂಡಿಯಾ

Phonetic-based models

- initially: K. Knight, J. Graehl (1998): "*Machine Transliteration*"
- several other papers for different languages, using
  - Hand-crafted Rules
  - Machine Learning from annotated data
- requires language-specific, linguistic knowledge, or data about segmentation & pronunciation

ประเทศอินเดีย לונדון ལོན་ཁོན་ لندن ლონდონი இலண்

லंदन

ಇന್ಡ್

Direct Orthographic Modelling

- Y. Al-Onaizan, K. Knight (2002) "*Machine Translation of Names in Arabic Text*"
- direct mapping between orthographic souce- and target-segments
- Al-Onaizan & Knight: "[...] The spelling based model was by far more accurate than the phonetic-based model [...]"

ประเทศอินเดีย
ལོན་ཁོན་
ლონდონი
لندن
לונדון
இலண்
लंदन
ಇಂಡಿಯಾ

Our strategy:

- Goals at MSRI: Transliteration which
  - is Language-neutral → no language-specific knowledge or rules required, "One code for all languages"
  - requires no annotated training data for segmentation or pronunciation
- strategy: Direct Orthographic Modelling with automated segmentation

Microsoft Excel - WebDunia_ParalleNamesCorpus_5000

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial Unicode MS 18

E4686 ಖಿರ್ವಾಡ್ಕರ್

	B	C	D	E
4686	Khirwadkar	खिरवडकर	கிர்வாத்கர்	ಖಿರ್ವಾಡ್ಕರ್
4687	Ghule	घुले	குலே	ಘುಲೆ
4688	Thipse	टिप्से	திப்ஸே	ತಿಪ್ಸೆ
4689	Jambhale	जांभले	ஜம்பலே	ಜಾಂಭಾಲೆ
4690	Mutatkar	मुटाटकर	முதாத்கர்	ಮುತಾತ್ಕರ್
4691	Tungaare	तुंगारे	துன்காரே	ತುಂಗಾರೆ
4692	Urdhvareshe	ऊर्ध्वरेशे	உர்த்வரேஷே	ಊರ್ಧ್ವರೇಶೆ
4693	Telang	तेलंग	தெலன்க்	ತೆಲಾಂಗ್
4694	Gawarikar	गवारीकर	கவாரிகர்	ಗವಾರಿಕರ್
4695	Raverkar	रावेरकर	ரவேர்கர்	ರಾವೇರ್ಕರ್
4696	Yavalkar	यावलकर	யவல்கர்	ಯವಾಲ್ಕರ್
4697	Kendurkar	केंदूरकर	கெந்துர்கர்	ಕೆಂಡೂರ್ಕರ್
4698	Hedgewar	हेडगेवार	ஹெட்கேவர்	ಹೆಡ್ಗೆವಾರ್
4699	Wavde	वावदे	வாவ்தே	ವಾವ್ದೆ
4700	Mande	मांडे	மண்டே	ಮಂಡೆ
4701	Mendhe	मेंढे	மேந்தே	ಮೆಂಢೆ
4702	Indapurkar	इंदापूरकर	இந்தபுர்கார்	ಇಂದಾಪುರ್ಕರ್

Sheet1 Sheet2 Sheet3

Ready

ประเทศอินเดีย ལོན་ཁོན་ לונדון لندن ლონდონი இலண்

लंदन ಇಂಡಿಯಾ

Our approach

- Uses "segments": chunks of graphemes (usually representing one or more speech sounds)

  e.g. "na", "न"

- Expresses transliterations between a source- and a target-word as sequences of segment pairs:

  Namaste ➔ नमस्ते

  as < न ,na> < म , ma> < स्त, st> <े,e>

- Calculates probabilities for transliterations between source- and target words based on transliteration probabilities for segment pairs

  e.g. P(न| na)

ประเทศอินเดีย
ལོན་ཁོན་
لندن
لوندون
ლონდონი
இலண்
लंदन
ಇಂಡ್ಯ

Our approach

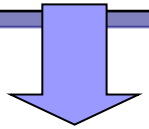
- BUT:
  - What are the segments?
  - What are the probabilities?

- Solution: Training through iterative alignment (Viterbi Training)

ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇಂಡಿಯಾ

Iterative Training

Generate Initial Segment-Pairs



ประเทศอินเดีย
ལོན་ཁོན་
لندن
ლონდონი
லண்டன்
இலண்
लंदन
ಇಂಡಿಯಾ

Training: Initialization

नमस्ते
namaste

- We only know two points: beginning of first segment, end of last segment
- Assumption: source- and target segments are at most *n* and *m* characters long
- We can create *n* x *m* segment pairs as hypotheses for first and last segment pair

ประเทศอินเดีย
ལོན་ཁོན་
ლონდონი
لندن
லண்
இலண்
लंदन
ಇಂಡಿಯಾ

Training: Initialization

नमस्ते

namaste

<n, न>, <n, नम>, <n, नमस>,

<na, न>, <na, नम >, <na, नमस >

<nam, न>, <nam, नम >, <nam, नमस >

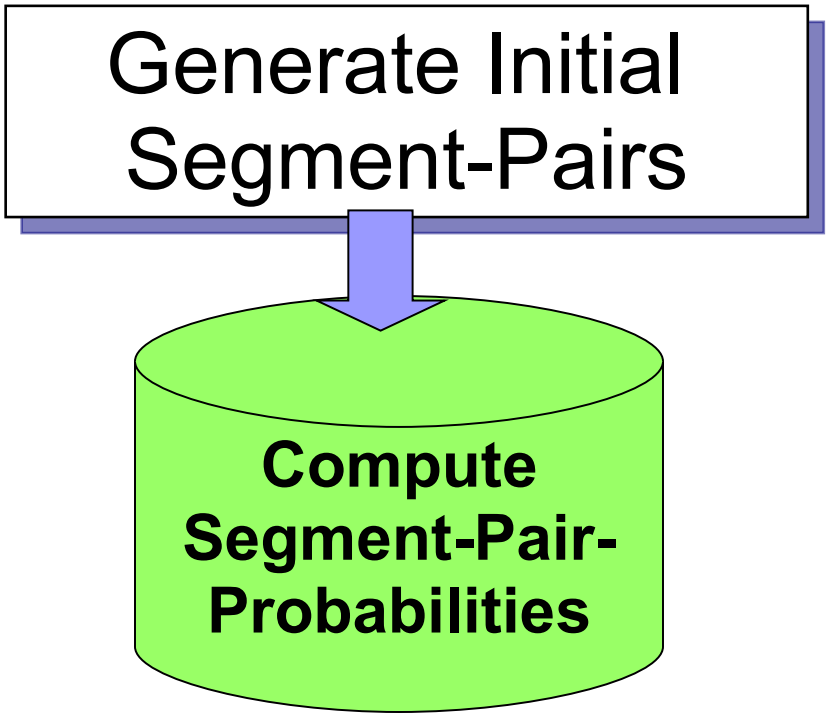
- do for all word pairs, count segment pairs
- Idea: Good pairs like <na,न> will have higher counts than artefacts like <n,नम>

ประเทศอินเดีย לונדון ལོན་ཁོན་ لندن ლონდონი இலண்

லंदन

ಇന್ಡ್

Iterative Training



Training: Probabilities

- A mapping between a source- and a target word can be represented through a sequence of n segment-pairs

  $< t_1,s_1>, <t_2,s_2> \ldots < t_n,s_n>$

  e.g. **< न ,na> < म , ma> <स्त, st> <े,e>**

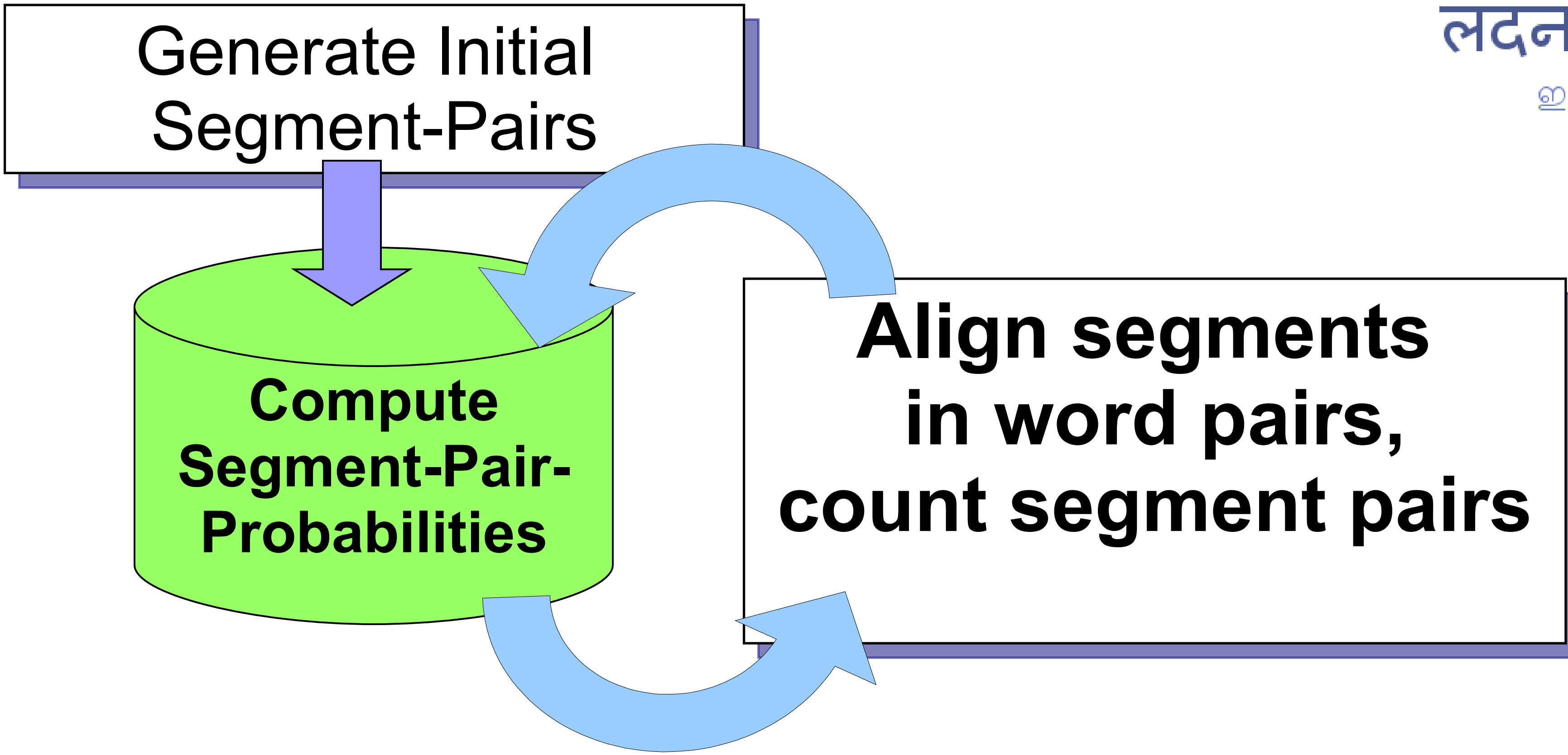
- The probability of this sequence can be calculated as

$$\prod_{n} P(t_n \mid s_n) P(t_n \mid t_{n-1})$$

- **$P(t_n|s_n)$ = count($<t_n ,s_n>$) / count(s_n)**

- **$P(t_n|t_{n-1})$** is a (character-) bi-gram language model

Iterative Training



Training: Alignment

- Input is Word-pairs:
  Source-words and their transliterations
- A Viterbi- type algorithm is used to find the most probable segment-sequence for the source-target-word pair

Align: नमस्ते <> namaste

◌े							
त							
◌							
स							
म							
न							
	n	a	m	a	s	t	e

Align: नमस्ते <> namaste

े							P(े|e)
त						P(त|t)	
◌					P(◌|s)		
स		P(स|ama)					
म	P(नम|n)						
न							
	n	a	m	a	s	t	e

Align: नमस्ते <> namaste

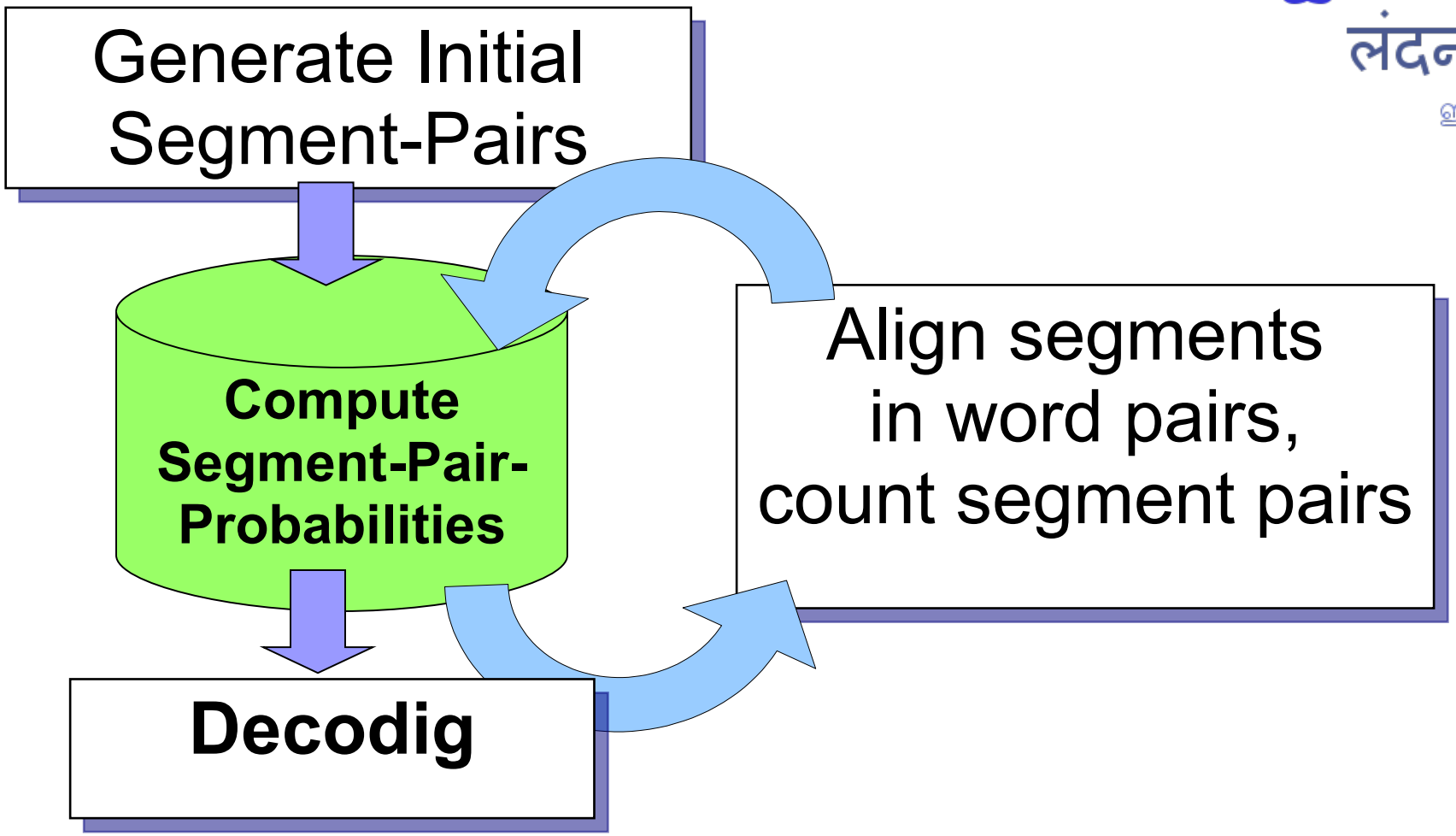
े							P(े|e)
त					P(स्त|st)		
्							
स							
म			P(म|ma)				
न	P(न|na)						
	n	a	m	a	s	t	e

Training: Alignment

- This is done for all examples in the training data
- Smoothing guarantees that previously unseen segment pairs are also considered
- Pairs used in the alignments are then counted
- These counts are used to update probabilities

ประเทศอินเดีย
ལོན་ཁོན་
لندن
லண்டன்
ლონდონი
இலண்
लंदन
ಇന್ಡ್

Iterative Training



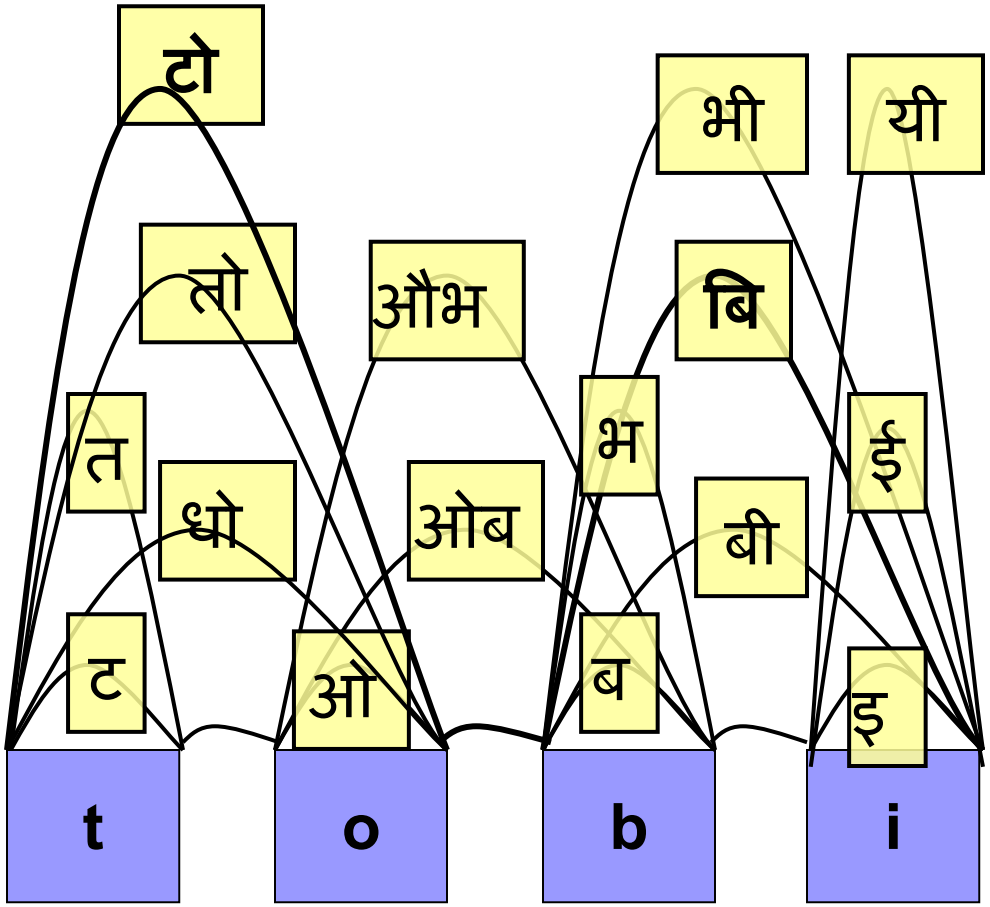
ประเทศอินเดีย
ལོན་ཁོན་
لندن
لوندون
ლონდონი
இலண்
लंदन
ಇന್ಡ್

Decoding

- Creates the best target-language transliteration for a given source-language word
- Based on the pair-probabilities estimated in training
- Uses a Viterbi-type algorithm

ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇന್ಡ್

Decoding



Ranked list of results:

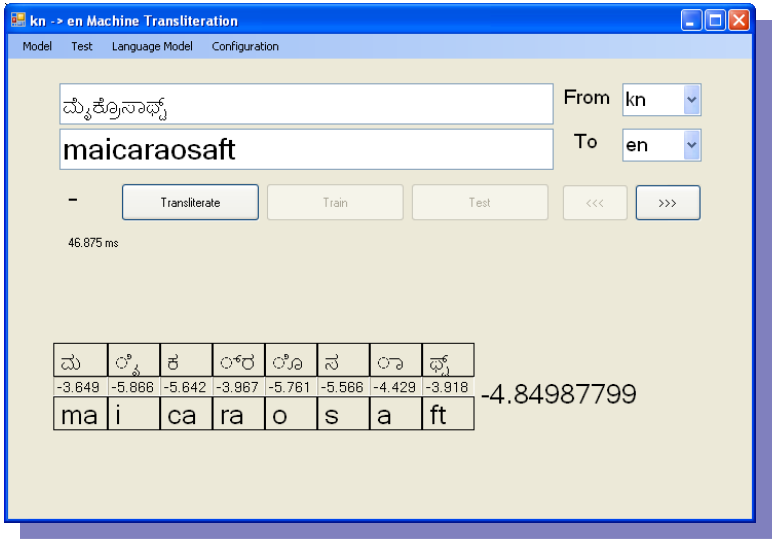
- टोबी
- तोबी
- टोबि
- …
- …
- टऔभयी

ประเทศอินเดีย
ლონდონი
இலண்

PART III: Results

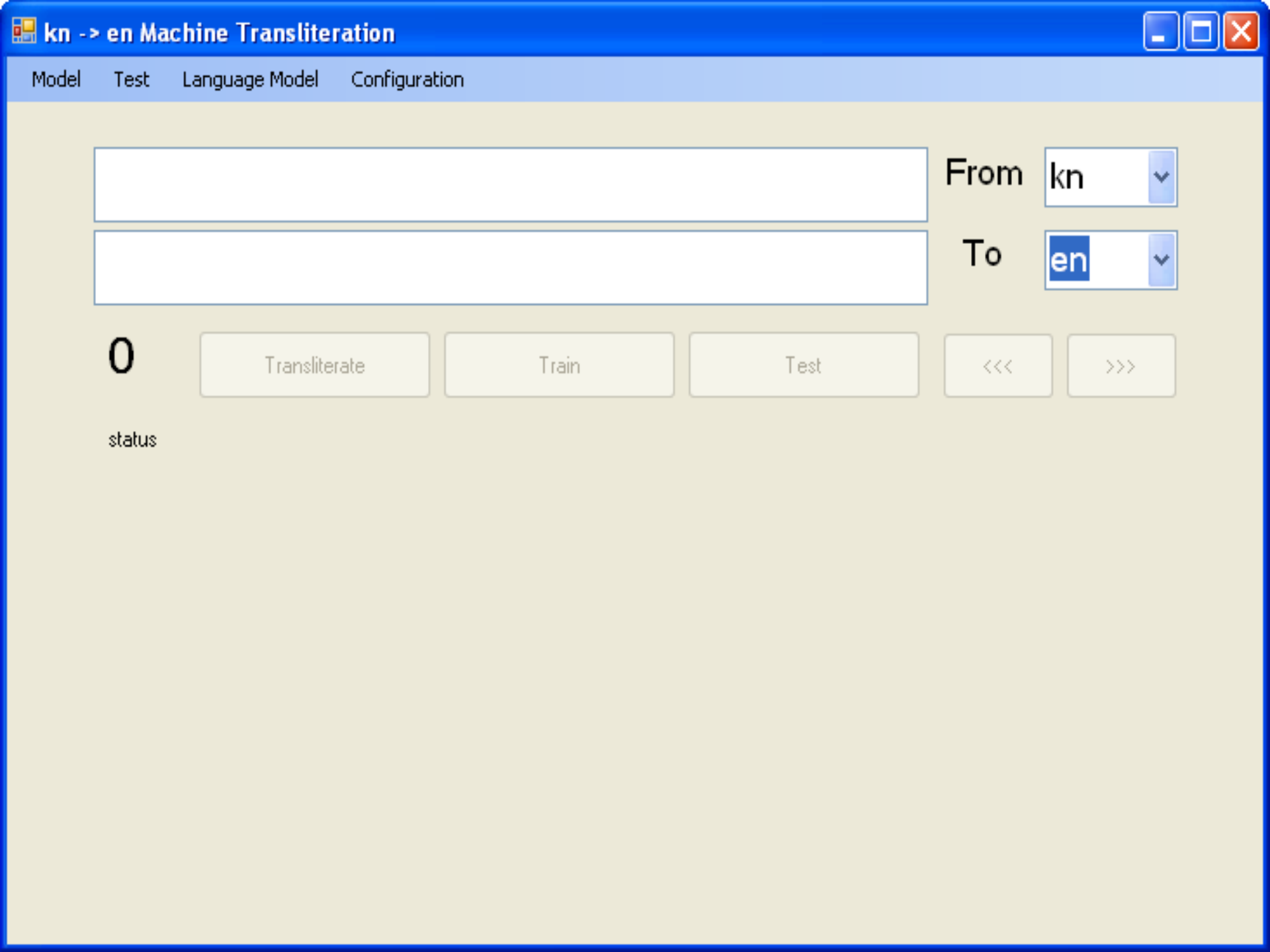
Part III:

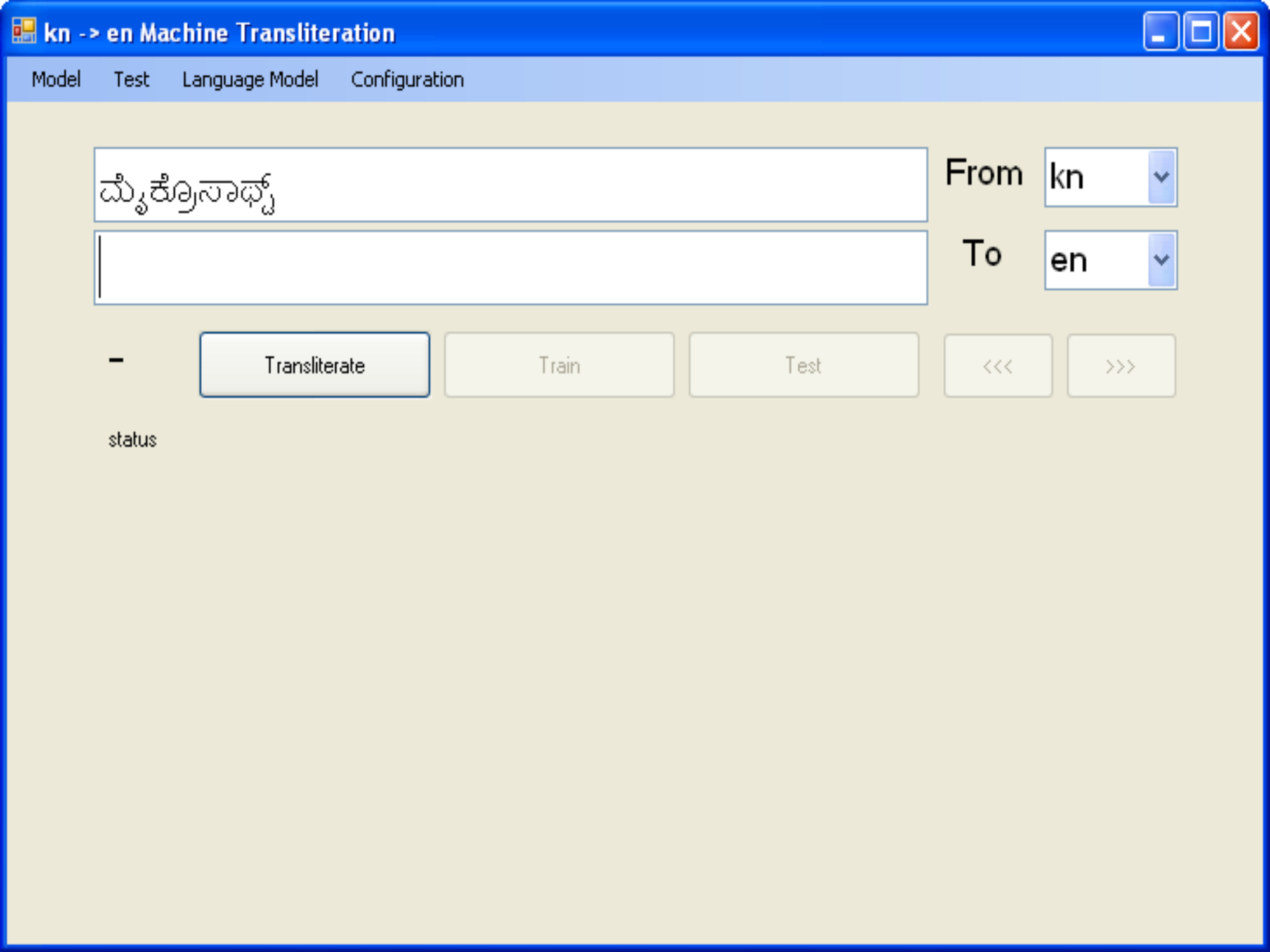
Results & Prototype-Demo

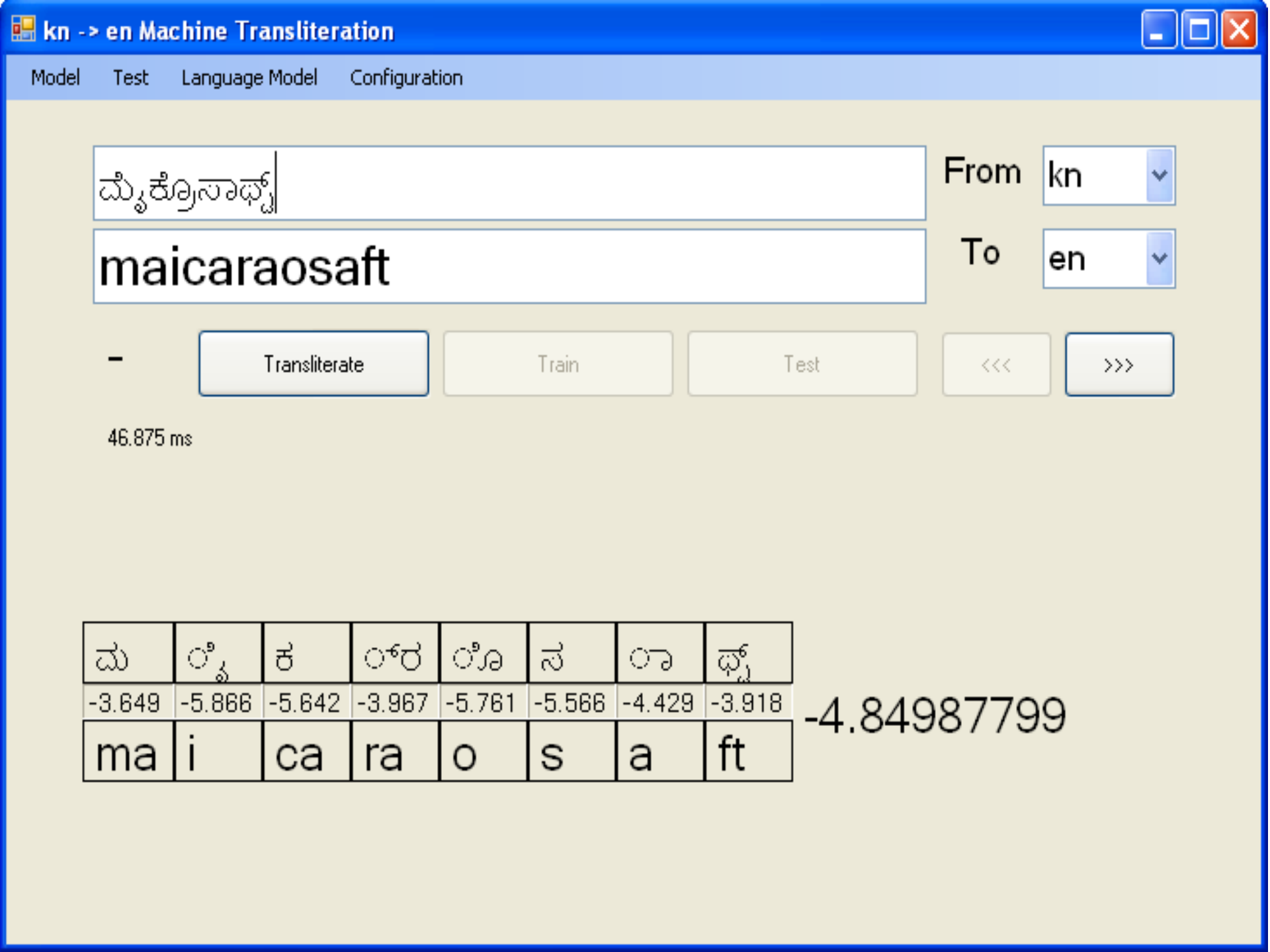


ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ഇന്ത്യ

Generation

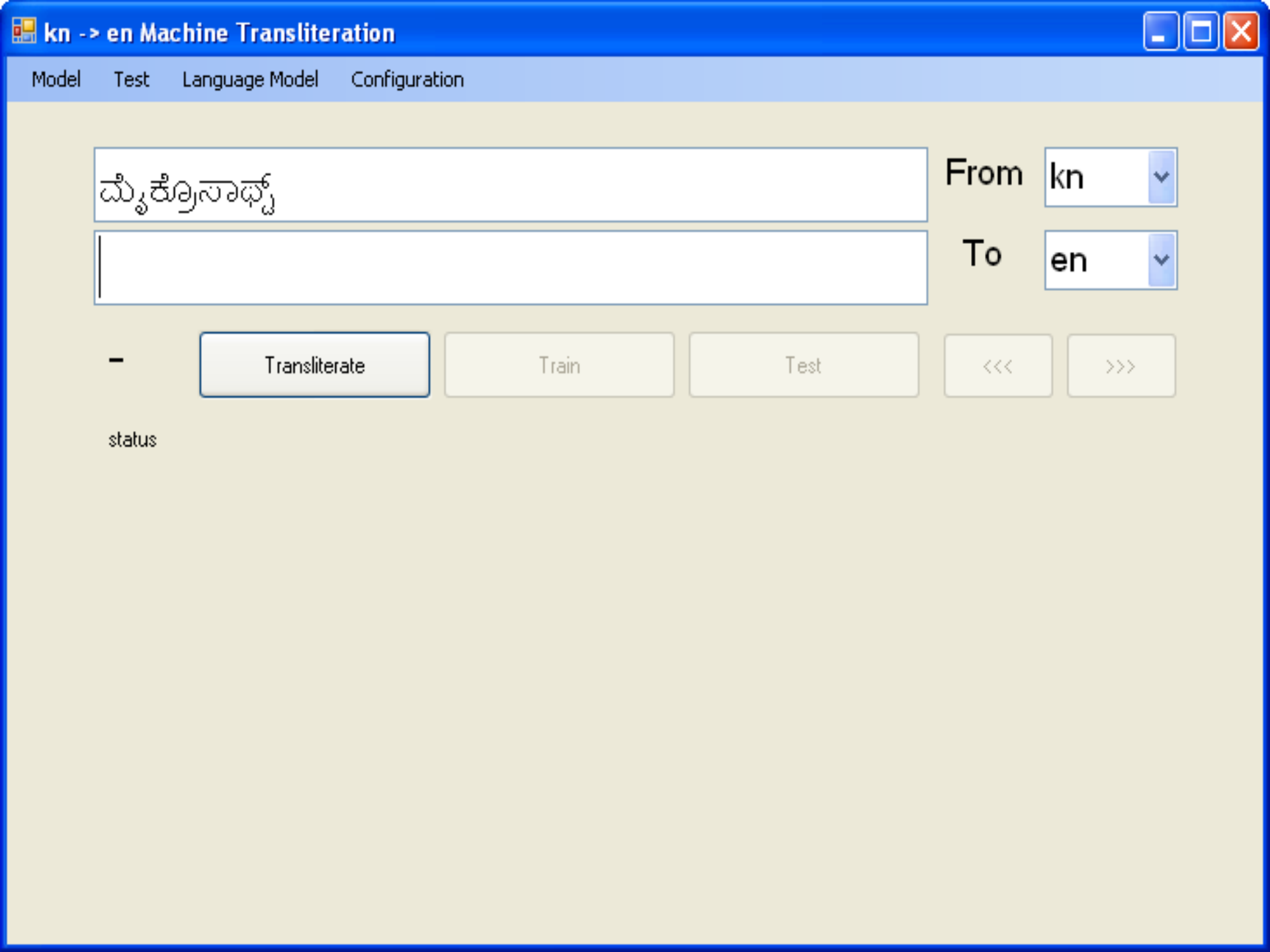

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
From
kn
To
en
0
Transliterate
Train
Test
<<<
>>>
status

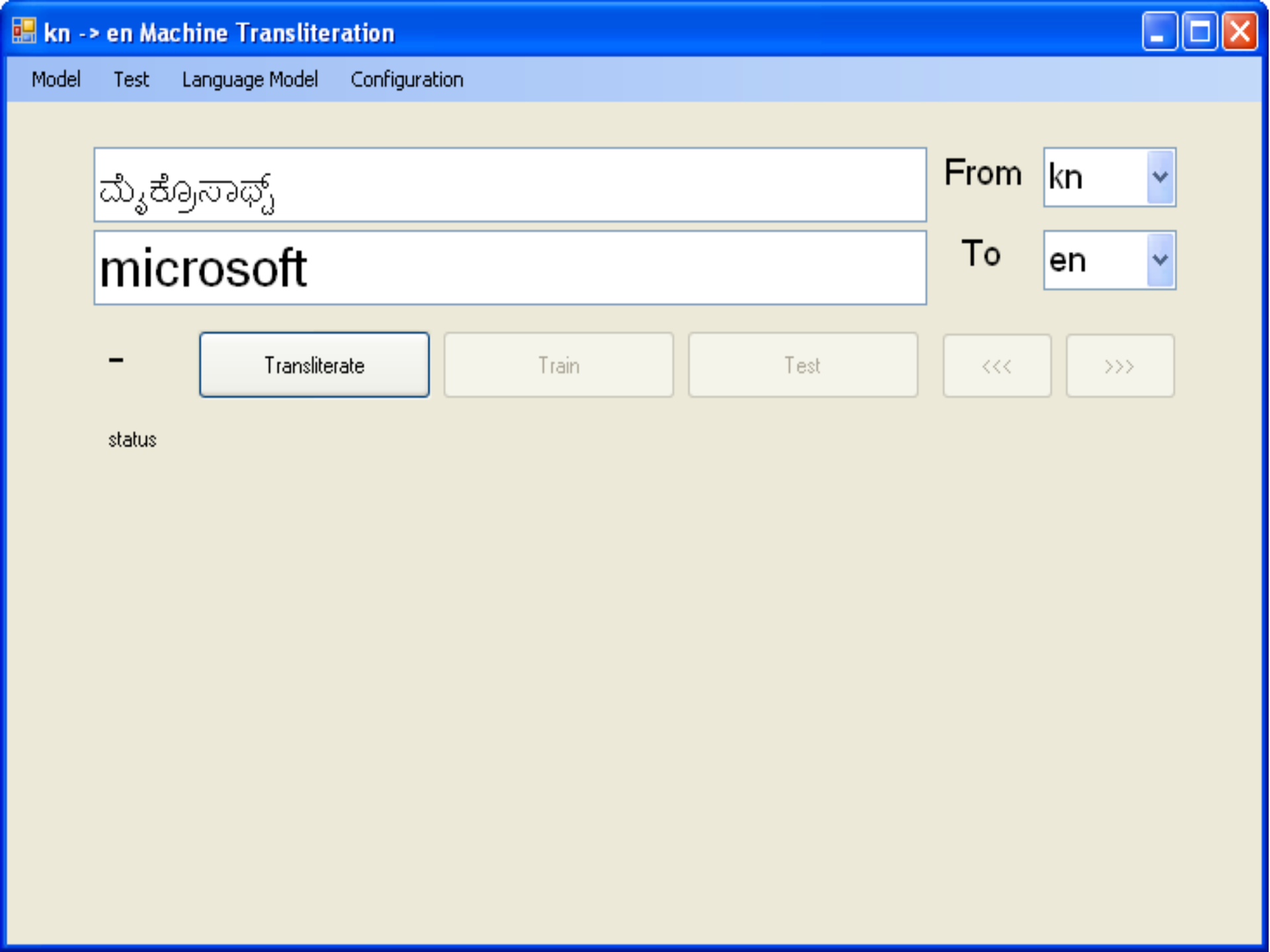

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
From
kn
To
en
Transliterate
Train
Test
<<<
>>>
status

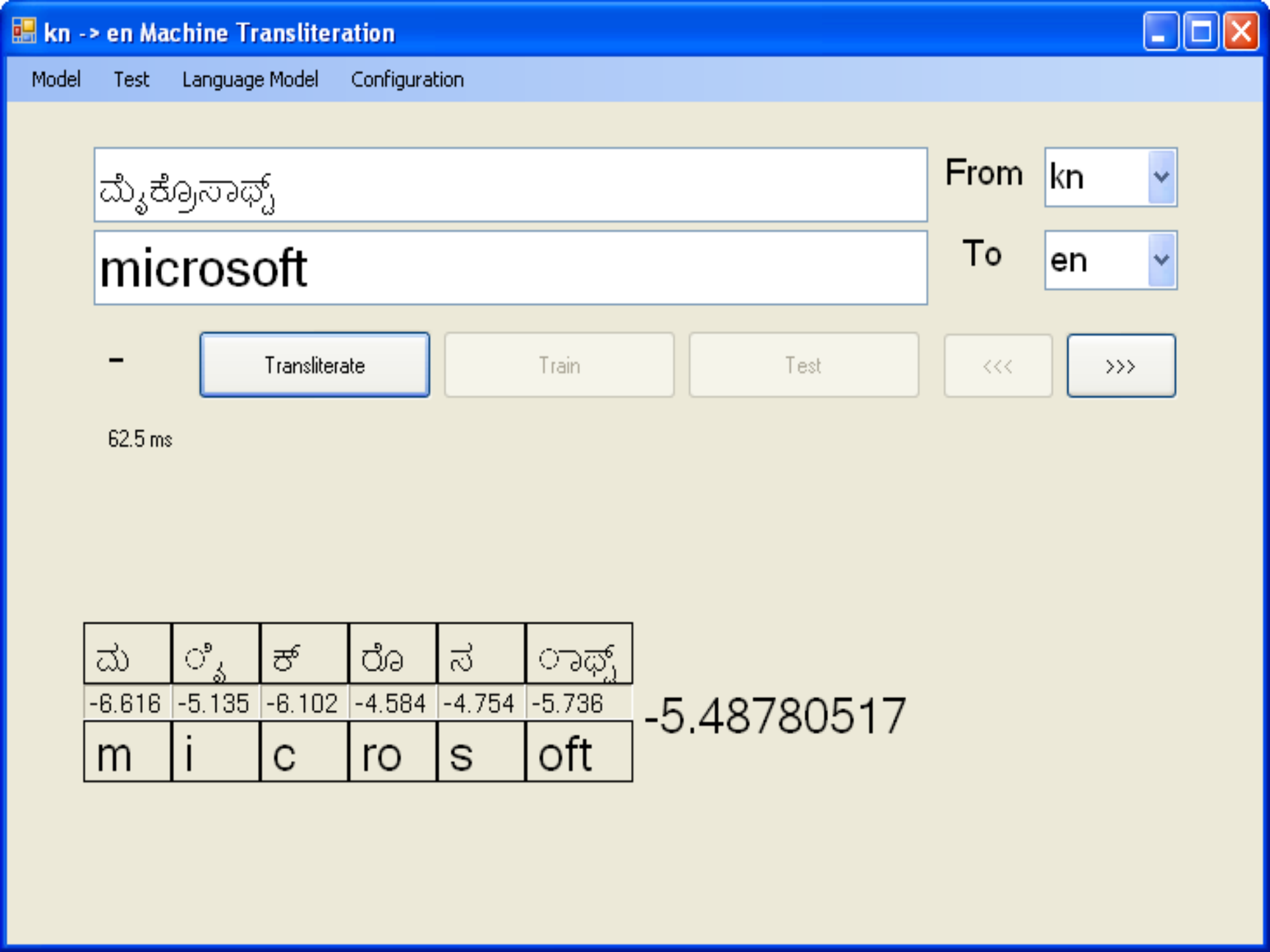

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕೊರಾಫ್ಟ್
From
kn
maicaraosaft
To
en
-
Transliterate
Train
Test
<<<
>>>
46.875 ms
ಮ
್ಯ
ಕ
ರ್ರ
ೊ
ಸ
ಾ
ಫ್ಟ್
-3.649
-5.866
-5.642
-3.967
-5.761
-5.566
-4.429
-3.918
-4.84987799
ma
i
ca
ra
o
s
a
ft

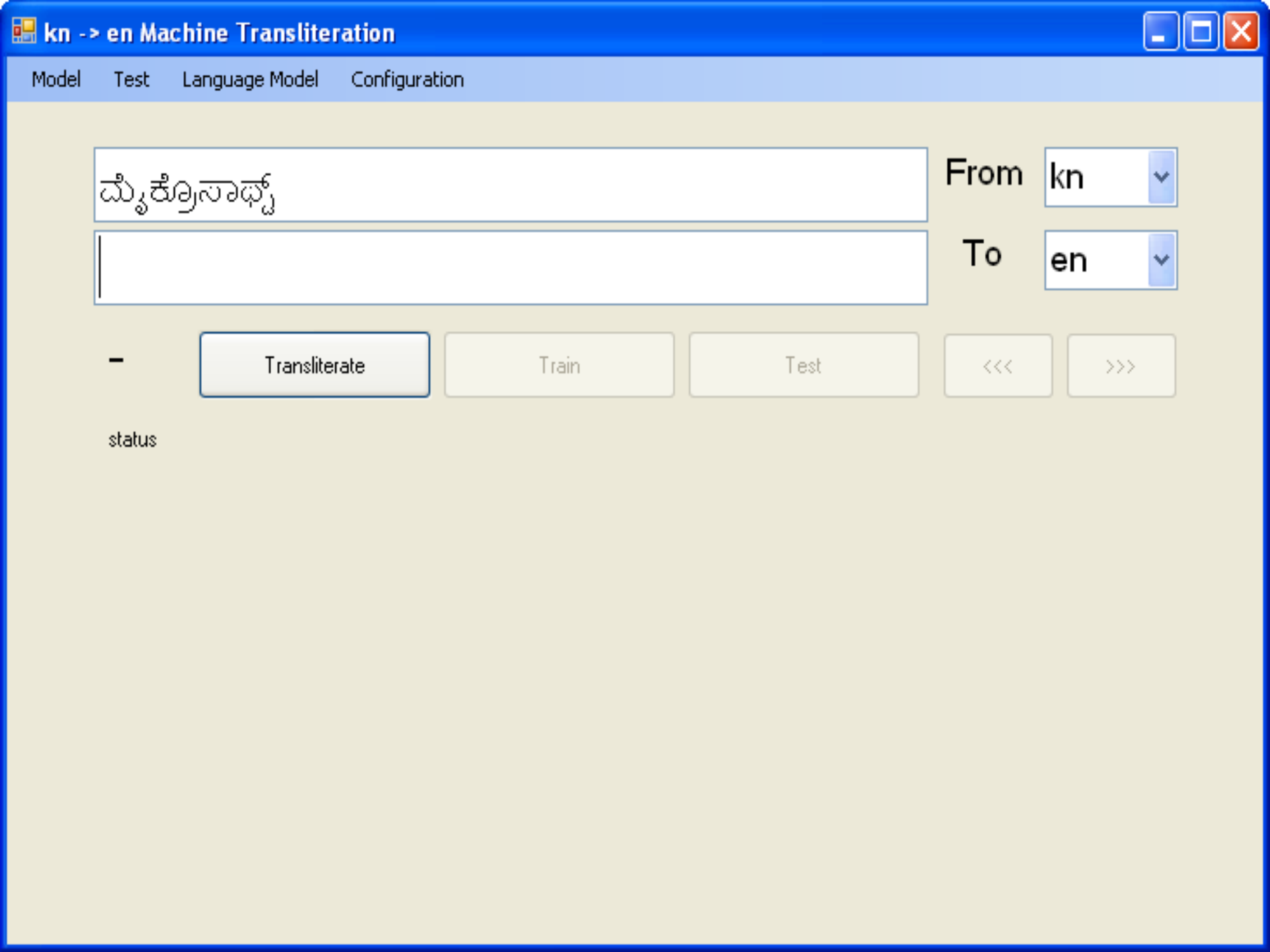
ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇಂಡಿಯಾ

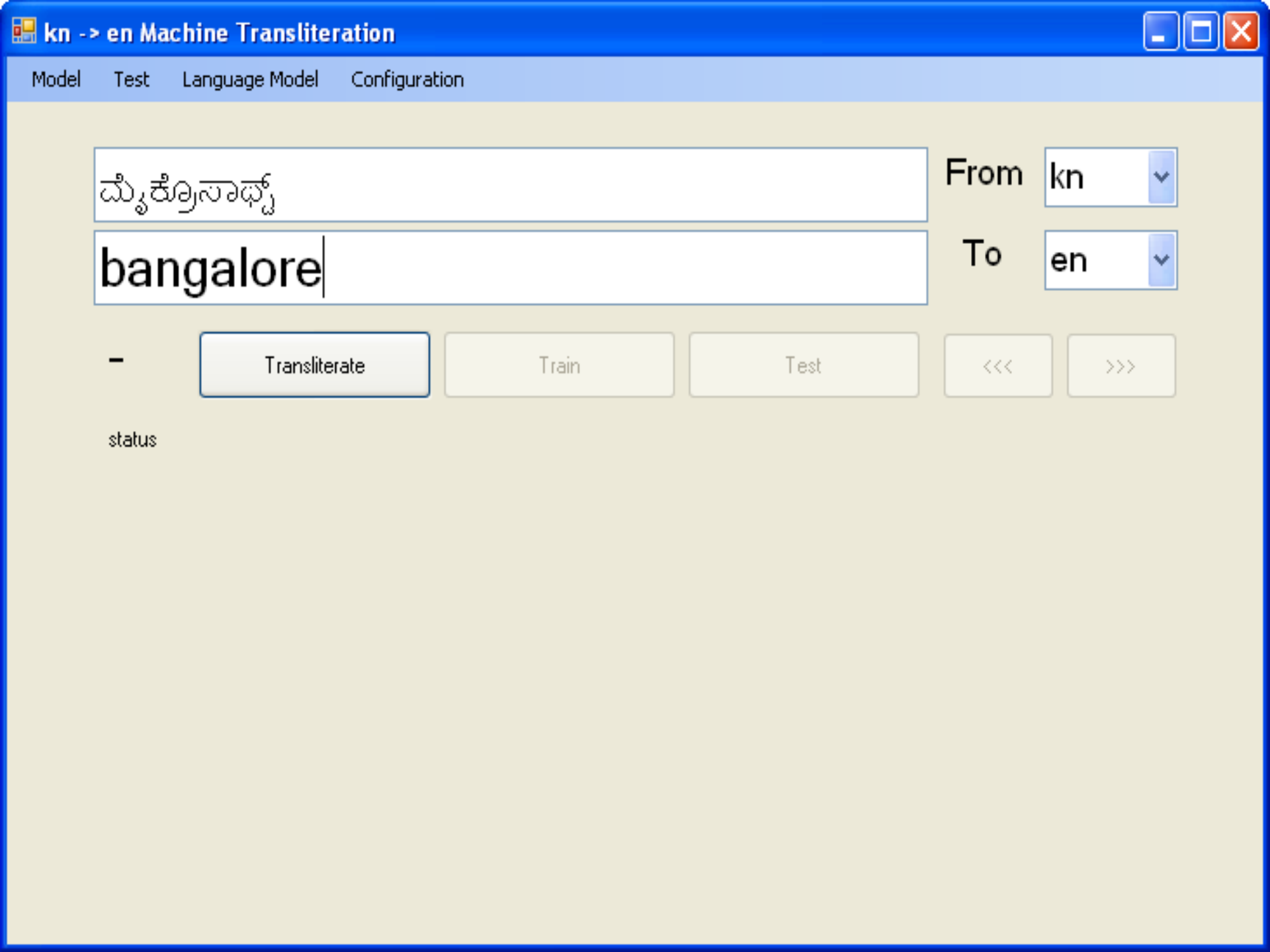
Comparison of candidates

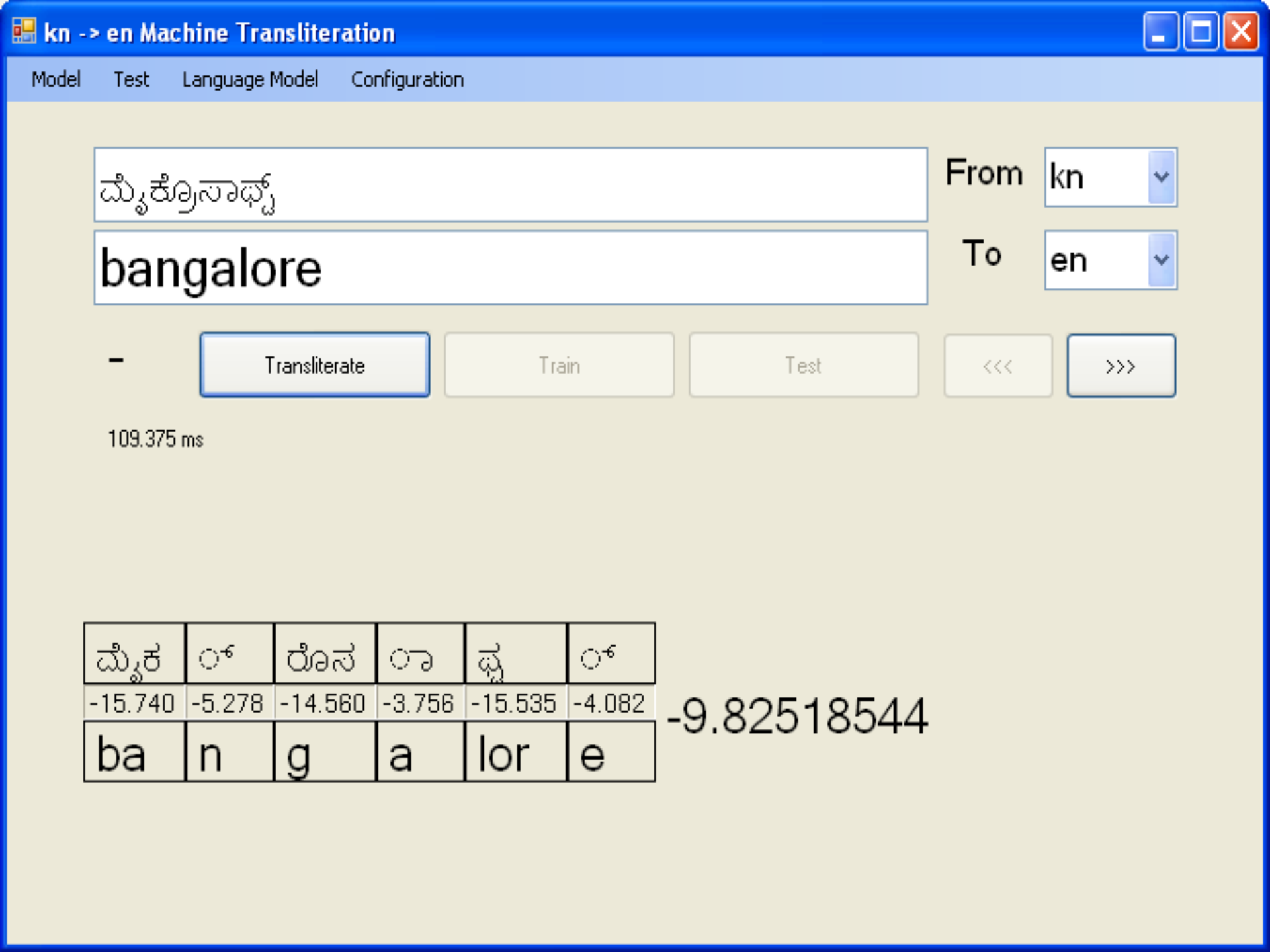

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
From
kn
To
en
Transliterate
Train
Test
<<<
>>>
status


kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
microsoft
From
kn
To
en
-
Transliterate
Train
Test
<<<
>>>
status


kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
microsoft
From
kn
To
en
-
Transliterate
Train
Test
<<<
>>>
62.5 ms
ಮ
ೈ
ಕ್
ರೊ
ಸ
ಾಫ್ಟ್
-6.616
-5.135
-6.102
-4.584
-4.754
-5.736
m
i
c
ro
s
oft
-5.48780517

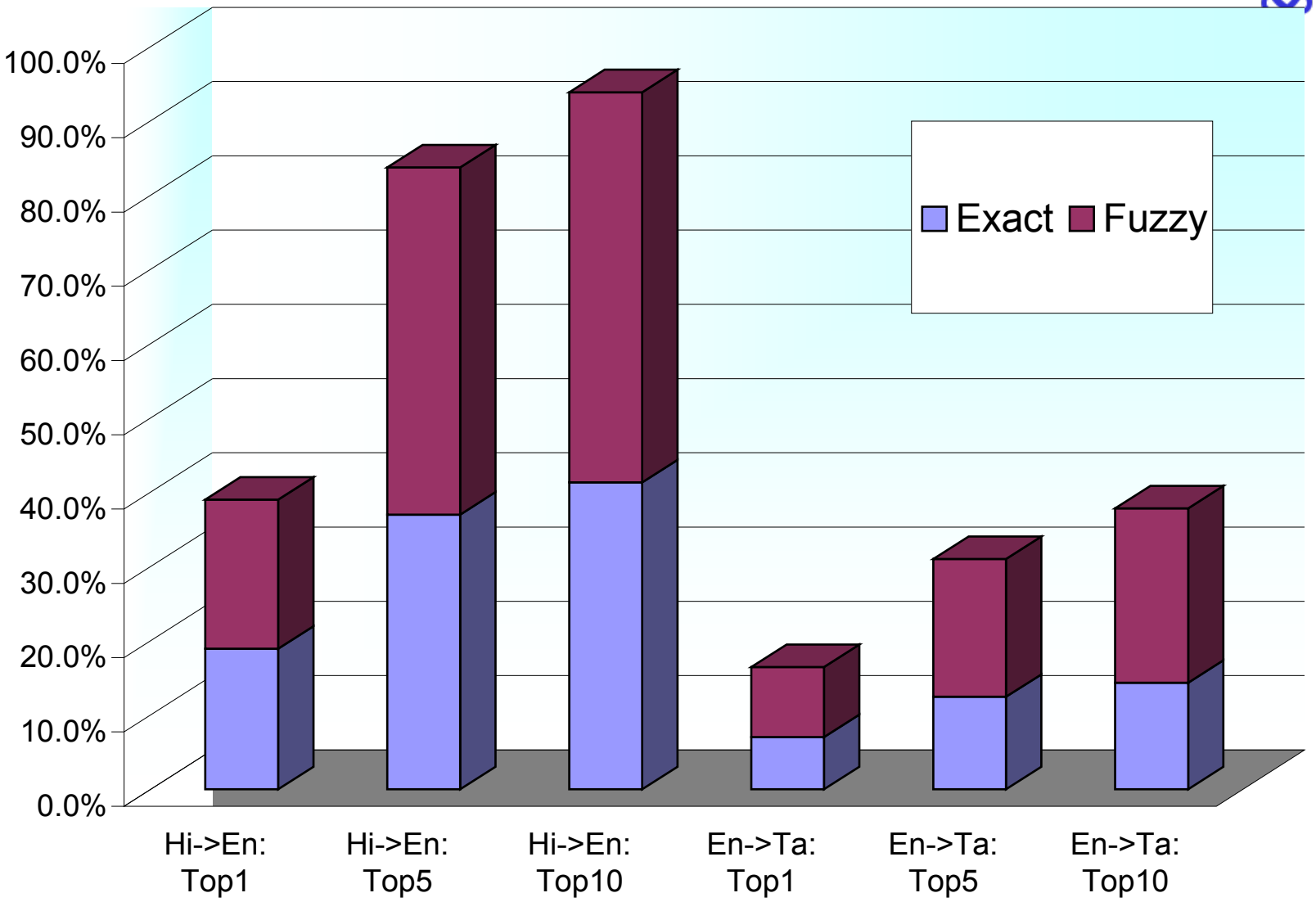

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
From
kn
To
en
Transliterate
Train
Test
<<<
>>>
status


kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕೊ್ರಸಾಫ್ಟ್
From
kn
bangalore
To
en
-
Transliterate
Train
Test
<<<
>>>
status

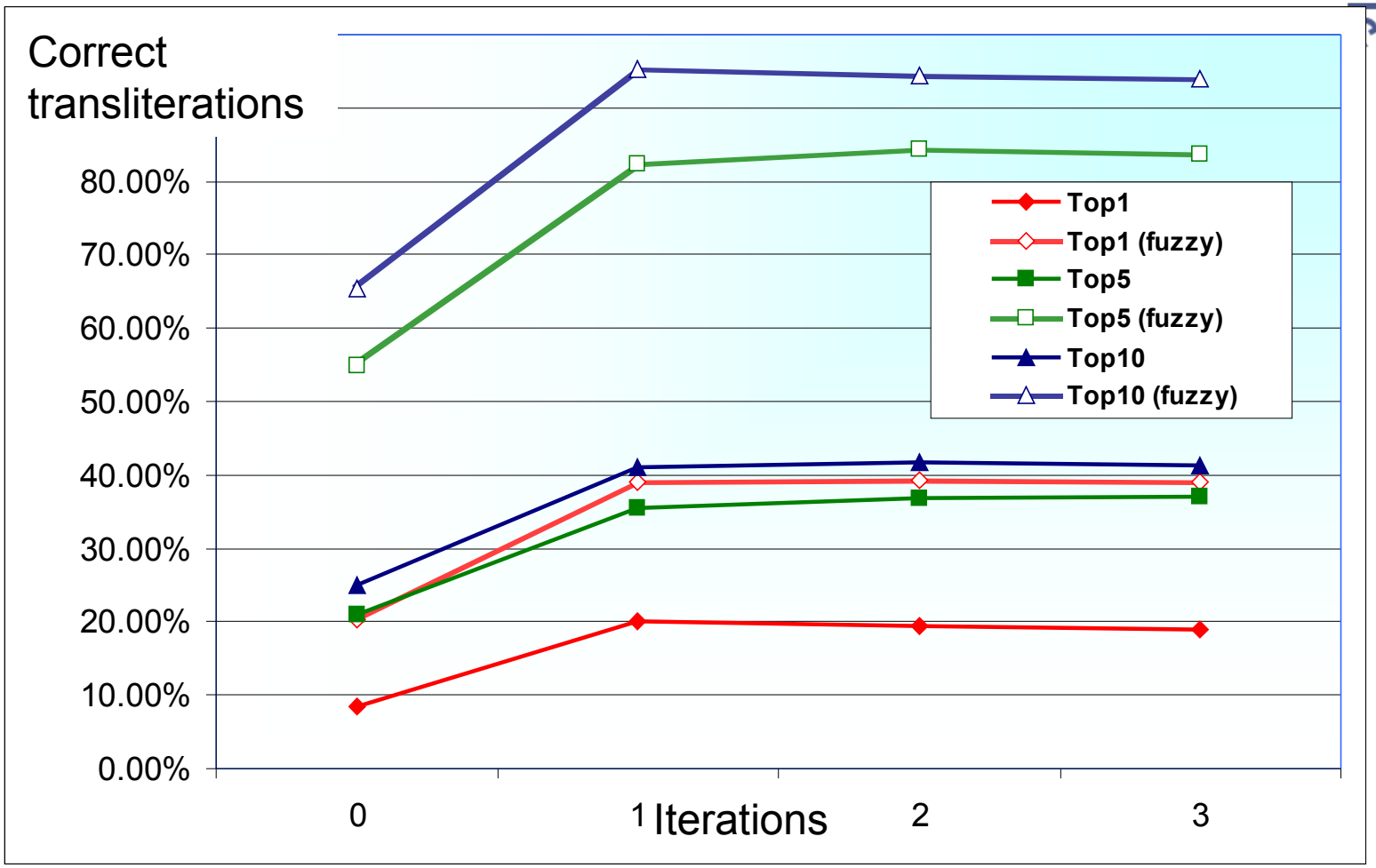

kn -> en Machine Transliteration
Model
Test
Language Model
Configuration
ಮೈಕ್ರೊಸಾಫ್ಟ್
From
kn
bangalore
To
en
-
Transliterate
Train
Test
<<<
>>>
109.375 ms
ಮೈಕ
್
ರೊಸ
ಾ
ಫ್ಟ
್
-15.740
-5.278
-14.560
-3.756
-15.535
-4.082
ba
n
g
a
lor
e
-9.82518544

ประเทศอินเดีย
ლონდონი
لندن
இலண்
लंदन
ಇಂಡಿಯಾ

Results (1)

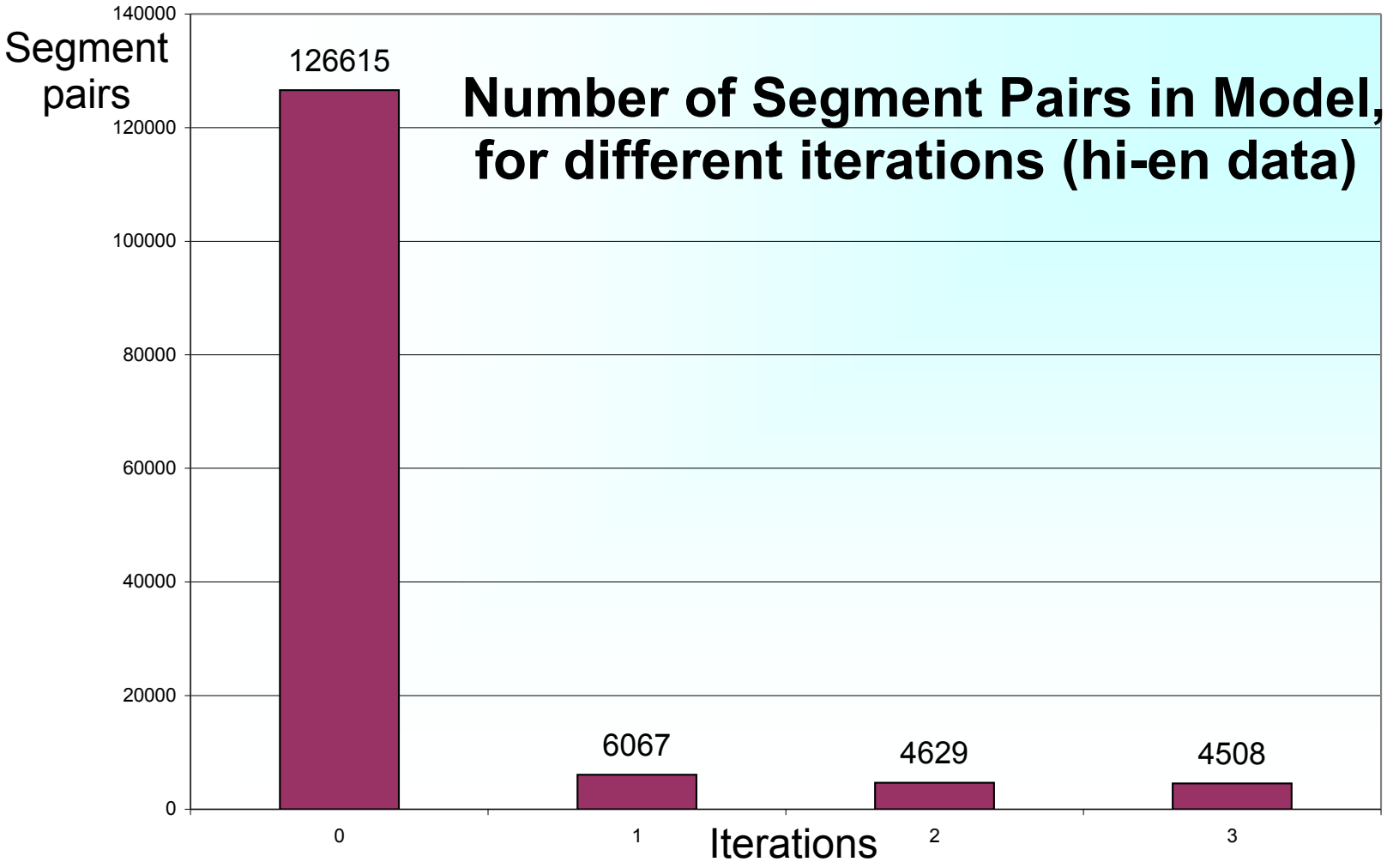


Results (2)


Correct transliterations
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
1
2
3
Iterations
Top1
Top1 (fuzzy)
Top5
Top5 (fuzzy)
Top10
Top10 (fuzzy)

ประเทศอินเดีย
ལོན་ཁོན་
ლონდონი
لندن
லண்டன்
இலண்
लंदन
ಇಂಡಿಯಾ

Results (3)

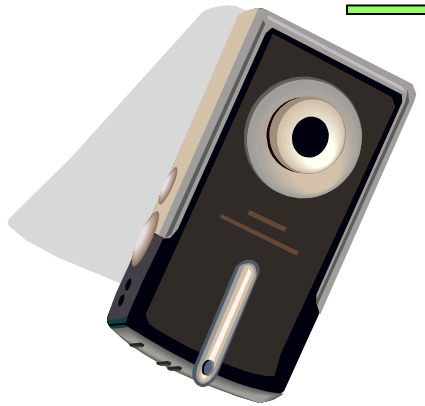


ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇಂಡಿಯಾ

Current trends:

- Transliteration and Named Entity Recognition (NRE) in translated text:
  - complementing NER and transliteration to identify cognates in bi-text

The future



ประเทศอินเดีย
לונדון
لندن
ლონდონი
இலண்
लंदन
ಇಂಡಿಯಾ

THE END

Thank you!

थानक यो

ತ್ಯಾಂಕ್ ಯೂ

தான்க் யூ