An Introduction to Machine Translation & Transliteration

Anoop Kunchukuttan

Research Scholar
Center for Indian Language Technology
www.cfilt.iitb.ac.in

Department of Computer Science & Engineering
IIT Bombay

anoopk@cse.iitb.ac.in

www.cse.iitb.ac.in/~anoopk

Machine Translation

What is Machine Translation?

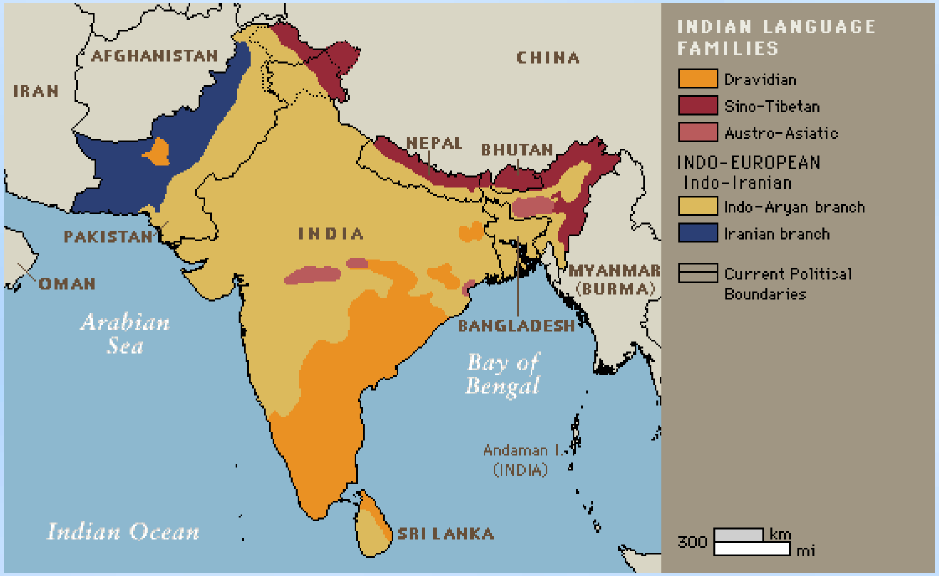
Automatic conversion of text/speech from one natural language to another

e.g.

- `Be the change you want to see in the world`
- वह परिवर्तन बनो जो संसार में देखना चाहते हो

Why do we need machine translation?

- 4 language families
- 22 scheduled languages
- 11 languages with more than 25 million speakers
- 30 languages with more than 1 million speakers
- Only India has 2 languages in the world’s 10 most spoken languages
- 7-8 Indian languages in the top 20 most spoken languages



Translation Usecases

- Government
  - Administrative requirements
  - Education
  - Security
- Enterprise
  - Product manuals
  - Customer support
- Social
  - Travel (signboards, food)
  - Entertainment (books, movies, videos)

Translation under the hood

- Cross-lingual Search
- Cross-lingual Summarization
- Building multilingual dictionaries

Any multilingual NLP system will involve some kind of machine translation at some level

Why study machine translation?

- One of the most challenging problems in Natural Language Processing
- Pushes the boundaries of NLP
- Involves analysis as well as synthesis
- Involves all layers of NLP: morphology, syntax, semantics, pragmatics, discourse
- Theory and techniques in MT are applicable to a wide range of other problems like transliteration, speech recognition and synthesis

Why is machine translation difficult?

Language Divergence: the great diversity among languages of the world

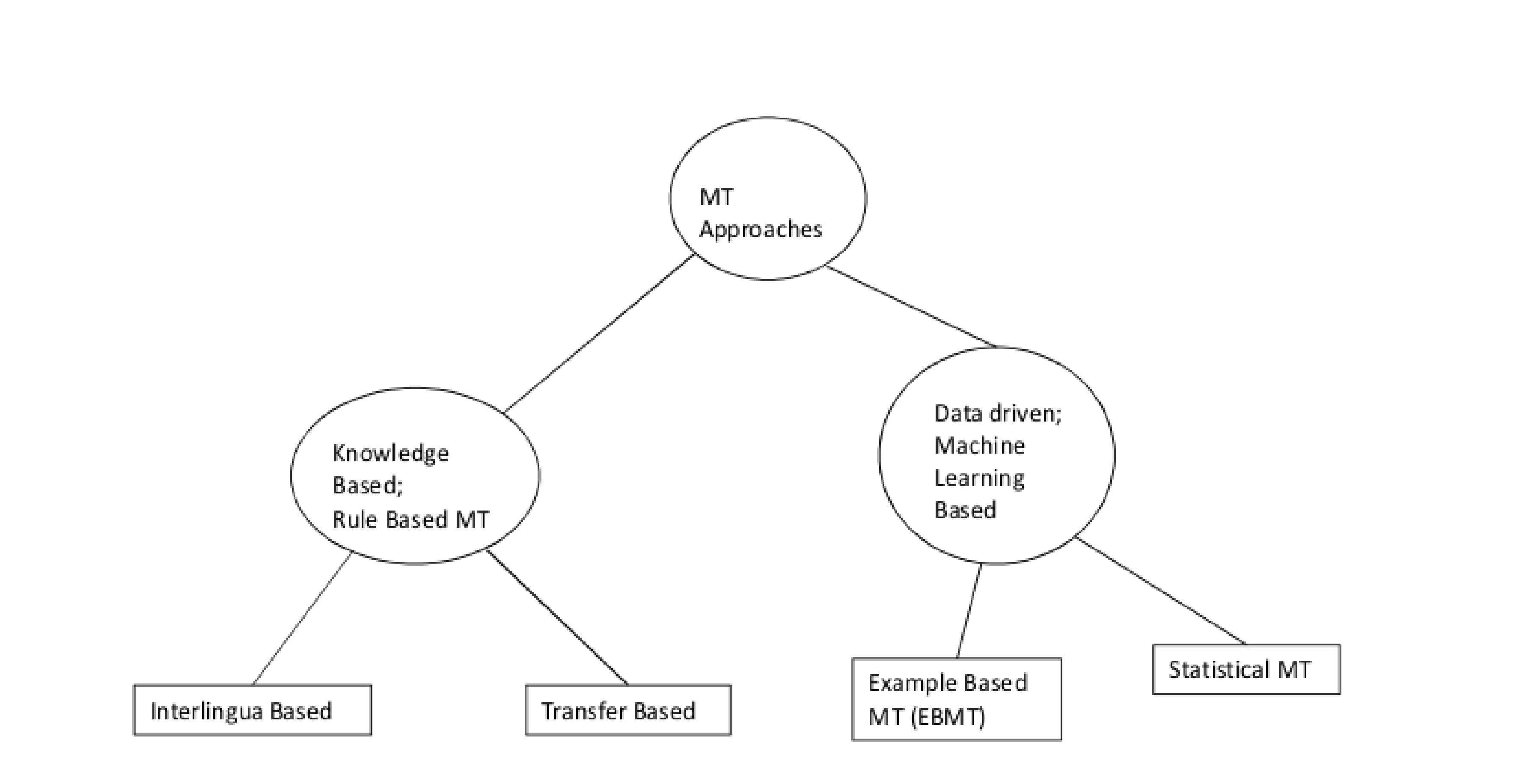
- Word order: SOV (Hindi), SVO (English), VSO, OSV,
- Free (Sanskrit) vs rigid (English) word order
- Analytic (Chinese) vs Polysynthetic (Finnish) languages
- Different ways of expressing same concept
- Case marking systems
- Language registers
- Inflectional systems [infixing (Arabic), fusional (Sanskrit), agglutinative (Marathi)]

… and much more

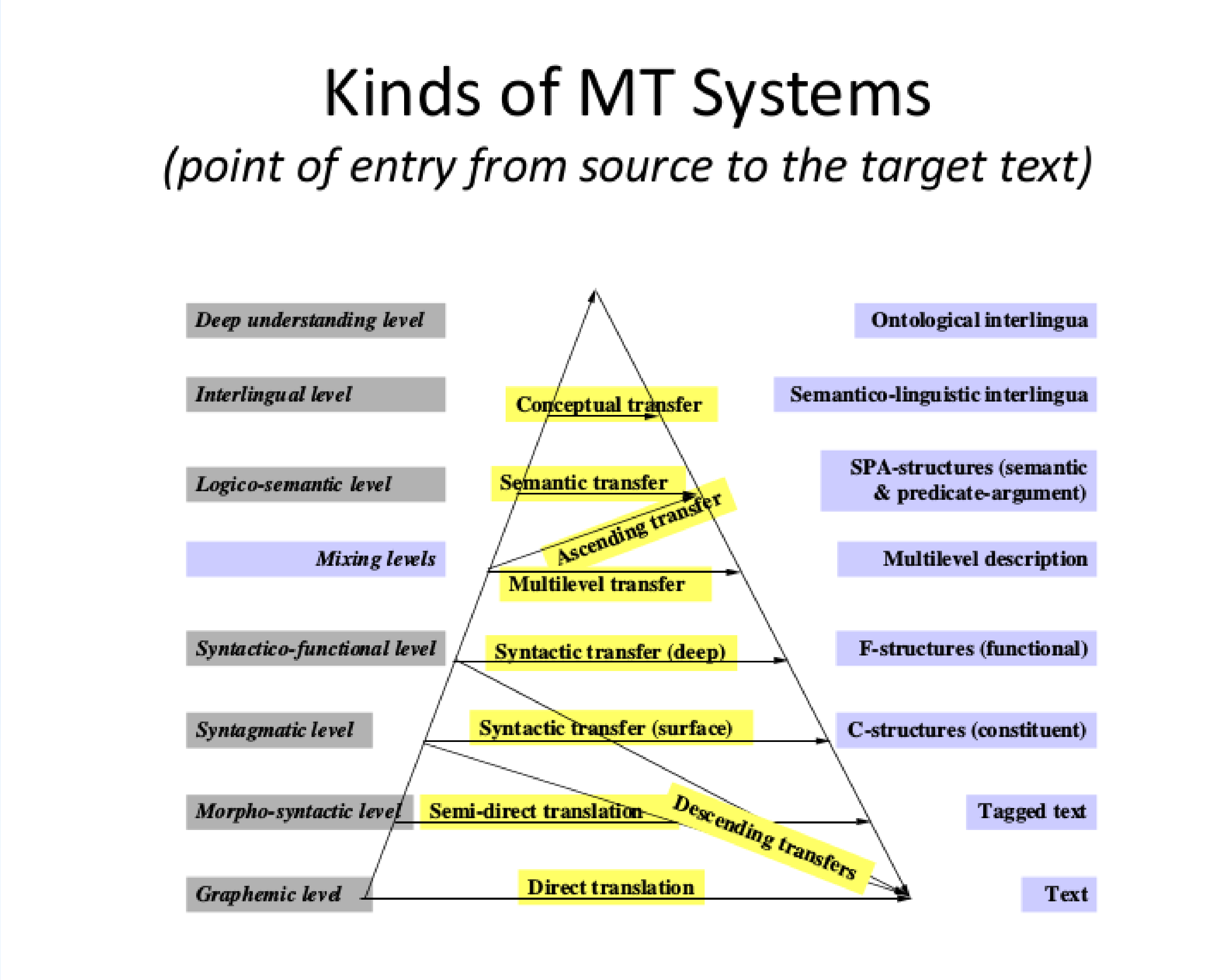
Why is machine translation difficult?

- Ambiguity
  - Same word, multiple meanings:
  - Same meaning, multiple words: जल, पानी,नीर (water)
- Word Order
  - Underlying deeper syntactic structure
  - Phrase structure grammar?
  - Computationally intensive
- Morphological Richness
  - Identifying basic units of words

Taxonomy of MT systems



Vauquois Triangle



Kinds of MT Systems

(point of entry from source to the target text)

Deep understanding level

Ontological interlingua

Interlingual level

Conceptual transfer

Semantico-linguistic interlingua

Logico-semantic level

Semantic transfer

SPA-structures (semantic & predicate-argument)

Ascending transfer

Mixing levels

Multilevel transfer

Multilevel description

Syntactico-functional level

Syntactic transfer (deep)

F-structures (functional)

Syntagmatic level

Syntactic transfer (surface)

C-structures (constituent)

Morpho-syntactic level

Semi-direct translation

Descending transfers

Tagged text

Graphemic level

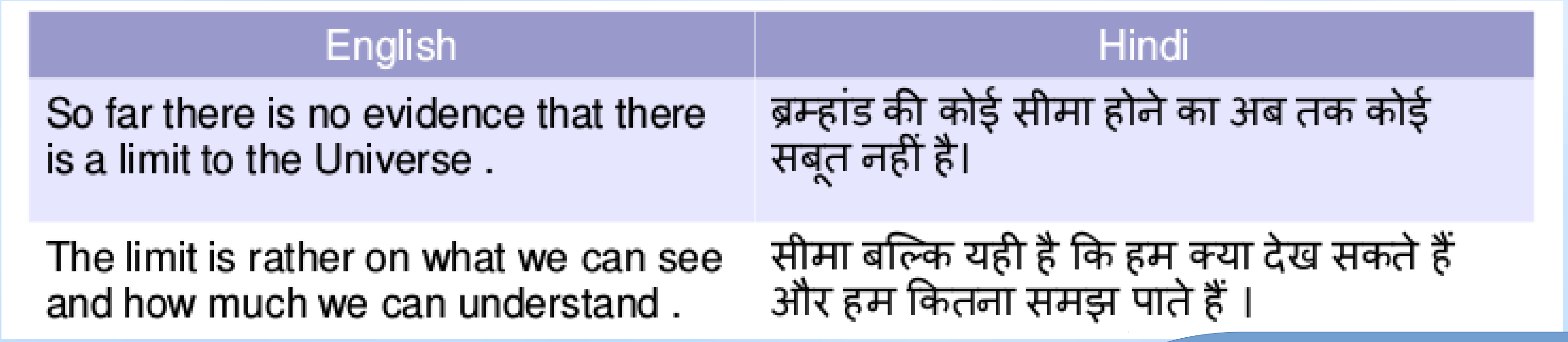
Direct translation

Text

Statistical Machine Translation

Data driven translation

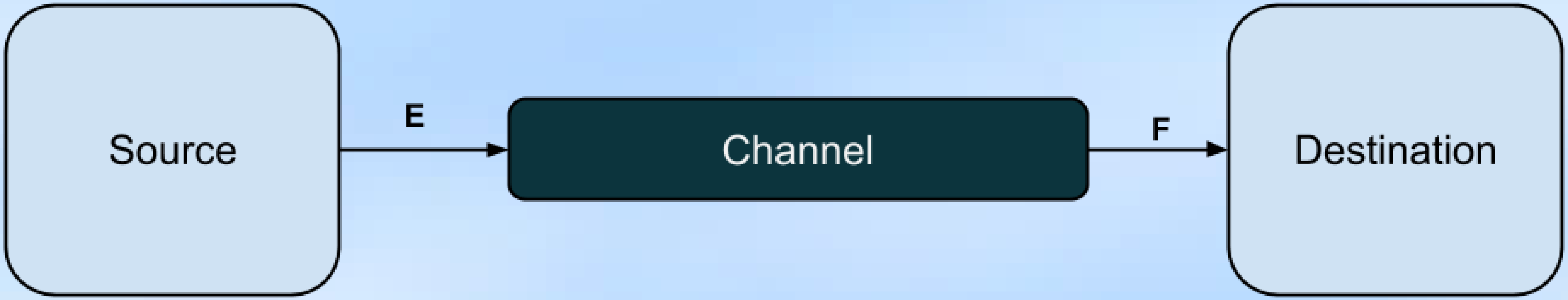
Parallel Corpus



English	Hindi
So far there is no evidence that there is a limit to the Universe .	ब्रम्हांड की कोई सीमा होने का अब तक कोई सबूत नहीं है|
The limit is rather on what we can see and how much we can understand .	सीमा बल्कि यही है कि हम क्या देख सकते हैं और हम कितना समझ पाते हैं ।

The Noisy Channel Model

A very general framework for many NLP problems



Language Model (LM)
P(e)

Captures fluency

Translation Model (TM)
P(f|e)

Captures fidelity

The SMT Process

Training

- Given: Parallel Corpus
- Output: P(e), P(f|e)
  - This is model learning
- Learning Objective: Maximize Likelihood
- Offline, one-time process
- Different translation models from different choice of P(f|e)

$$P^*(f|e) = \arg\max \mathbf{Likelihood}(data; P(f|e))$$

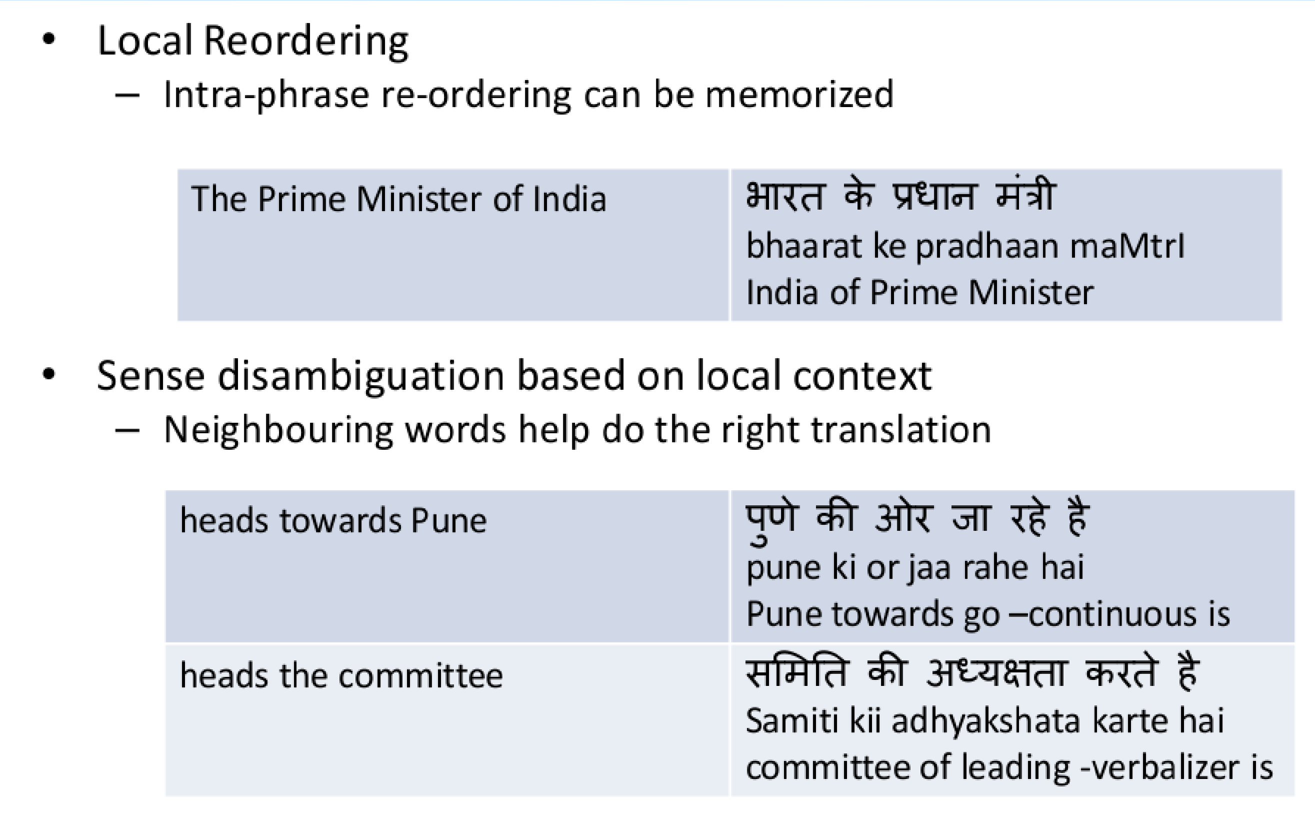
Decoding

- Given:
  - Sentence **f** in language **F**
  - P(e) and P(f|e)
- Output: Translation **e** for **f**
- Online process, should be fast
- TM & LM are used for scoring translation candidates

$$e^* = \arg\max_e P(f|e)P(e)$$

Phrase-based Translation Model

- One of the most successful models
- Widely used in commercial systems like Google Translate
- Basic unit of translation is a phrase
- A *phrase* is just a sequence of words



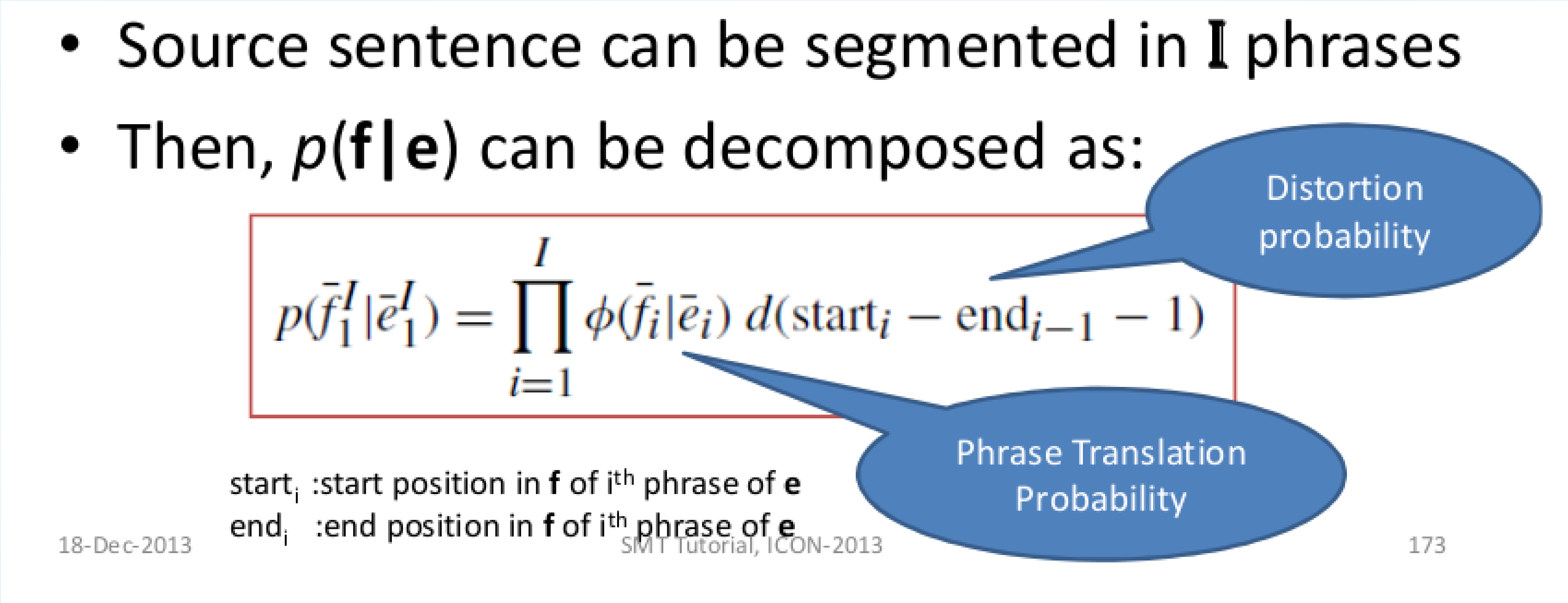
- Local Reordering
  – Intra-phrase re-ordering can be memorized

The Prime Minister of India	भारत के प्रधान मंत्री bhaarat ke pradhaan maMtrI India of Prime Minister

- Sense disambiguation based on local context
  – Neighbouring words help do the right translation

heads towards Pune	पुणे की ओर जा रहे है pune ki or jaa rahe hai Pune towards go –continuous is
heads the committee	समिति की अध्यक्षता करते है Samiti kii adhyakshata karte hai committee of leading -verbalizer is

So how the model look now?



- Source sentence can be segmented in I phrases
- Then, *p*(**f**|**e**) can be decomposed as:

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(\text{start}_i - \text{end}_{i-1} - 1)$$

Distortion probability

Phrase Translation Probability

start_i :start position in **f** of ith phrase of **e**

end_i :end position in **f** of ith phrase of **e**

Training a Phrase-based SMT system

- Building the Language Model
- Building the Translation Model
  - Word Alignment (find word-level correspondences)
  - Phrase Extraction (extract phrase pairs)
- Tuning

Building the Language Model

- Probability of a sentence **e**
  - P(**e**)=P$(e_1,e_2,...,e_k)$

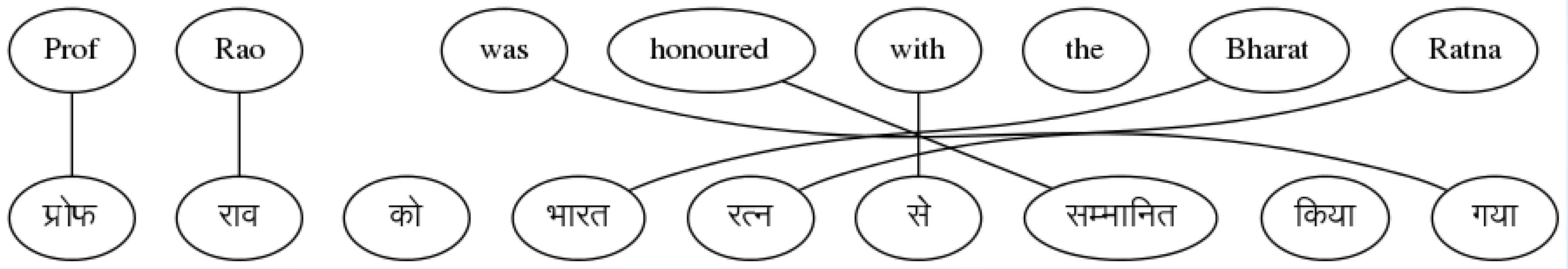
    $=\Pi_{i=1..k}P(e_i|e_{i-1..i-n+1})$
  - Apply Chain Rule of probability
  - Markov Assumption: i^{th} words depends only previous *n-1* words (n^{th} order Markov model)

- Estimate P$(e_i|e_{i-1..i-n+1})$ from a monolingual corpus

  e.g. of a bigram (2-gram) language model
  - P(book|the)=c(the,book)/c(the)
  - A little complication: what happens if *book* never comes in the training corpus
  - That's the complicated part of language modelling, let's skip it for now!

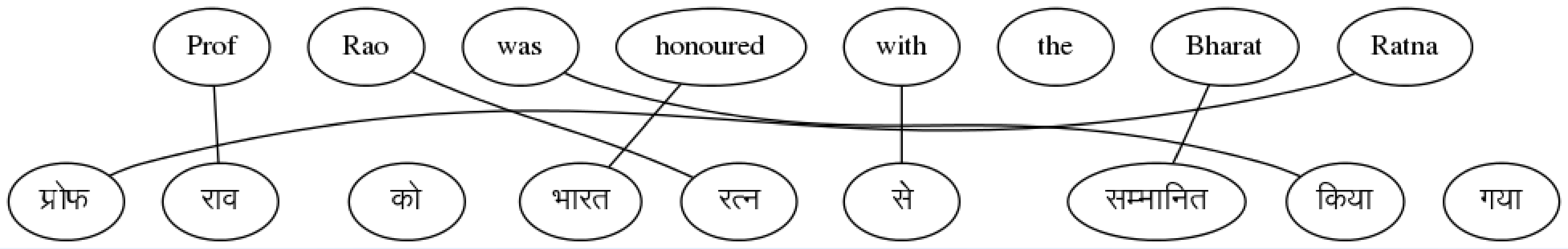
Word Alignment

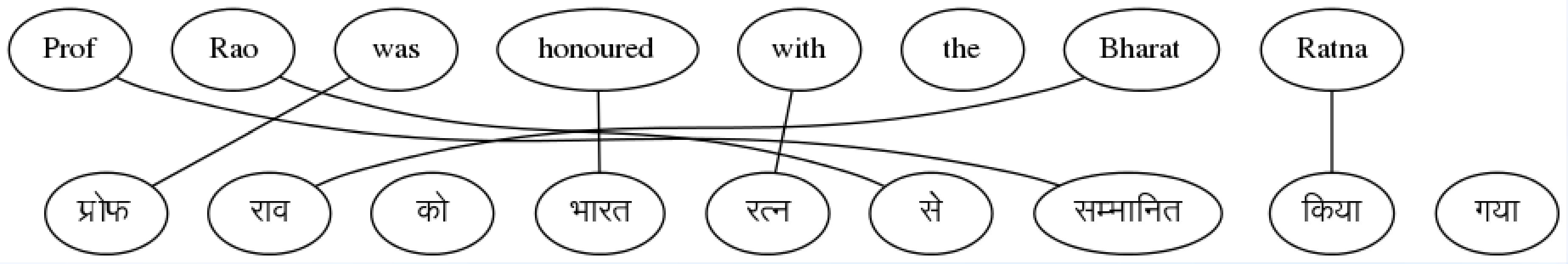
- Central Task in Statistical Machine Translation
- Given a parallel sentence pair, find word level correspondences *(alignment, let's say a)*



But there are multiple possible alignments

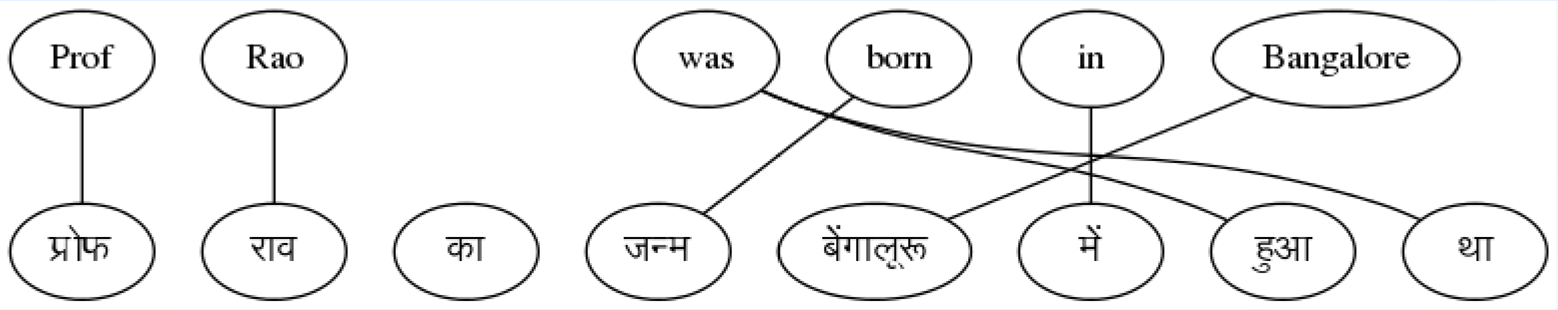
Sentence 1

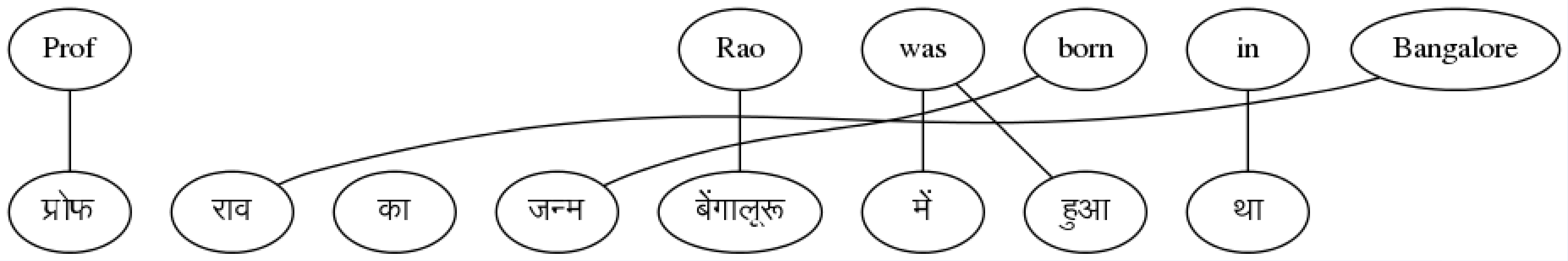




But there are multiple possible alignments

Sentence 2





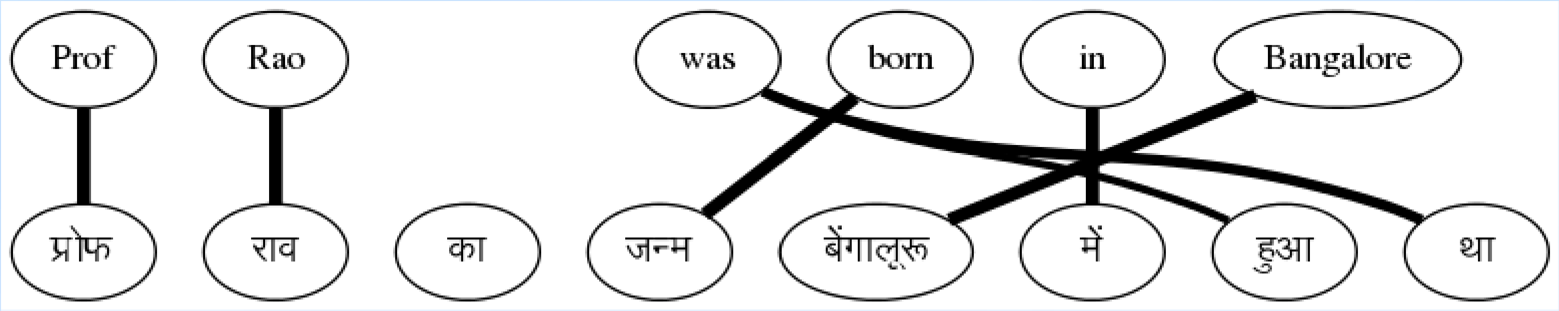
How do we find the correct alignment?

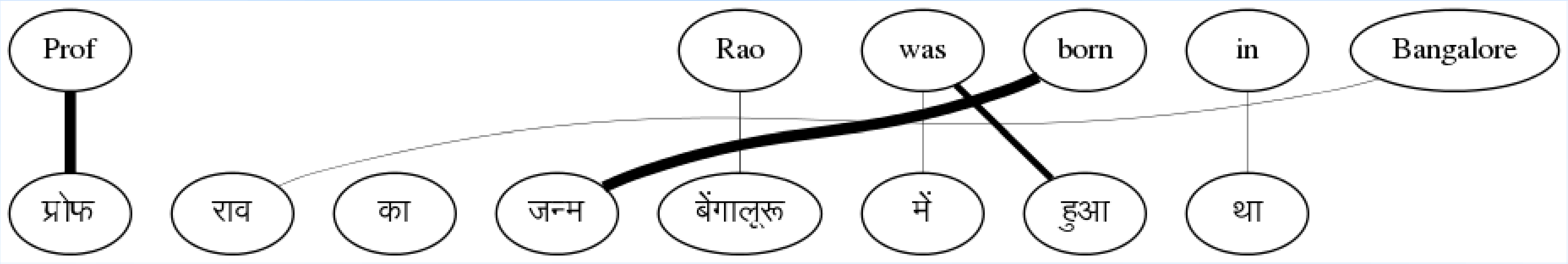
Key ideas

- **Co-occurrence of words**
  - Words which occur together in the parallel sentence are likely to be translations (*higher P(f|e)*)
  - Alignments which have more likely word-translation pairs are more likely (*higher P(a)*)
  - Its a chicken-and-egg problem!
  - How to actually find the best alignment?
- **Expectation-Maximization Algorithm**
  - Find the best ***hidden*** alignment
  - A key algorithm for various machine learning problems
    - Start with a random alignment
    - Find P(f|e) given the alignments
    - Now compute alignment probabilities P(a) with these new translation probabilities
    - Do this repeatedly till P(f|e) does not change

At the end of the process

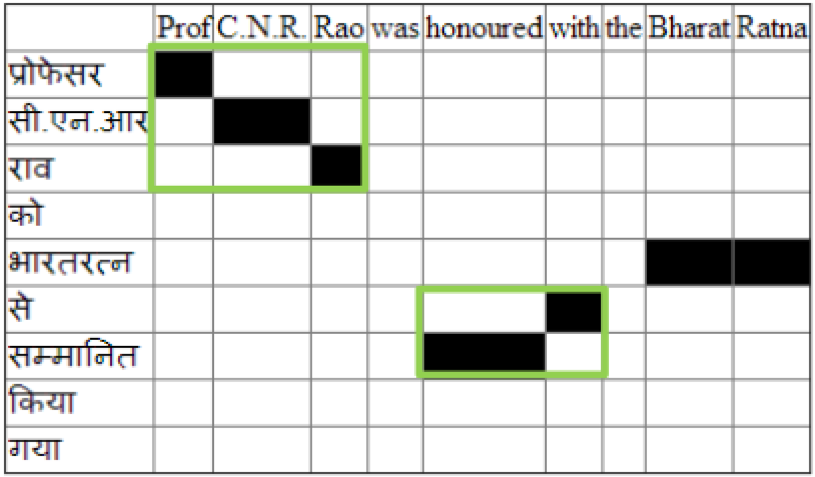
Sentence 2





Learning Phrase Tables from Word Alignments

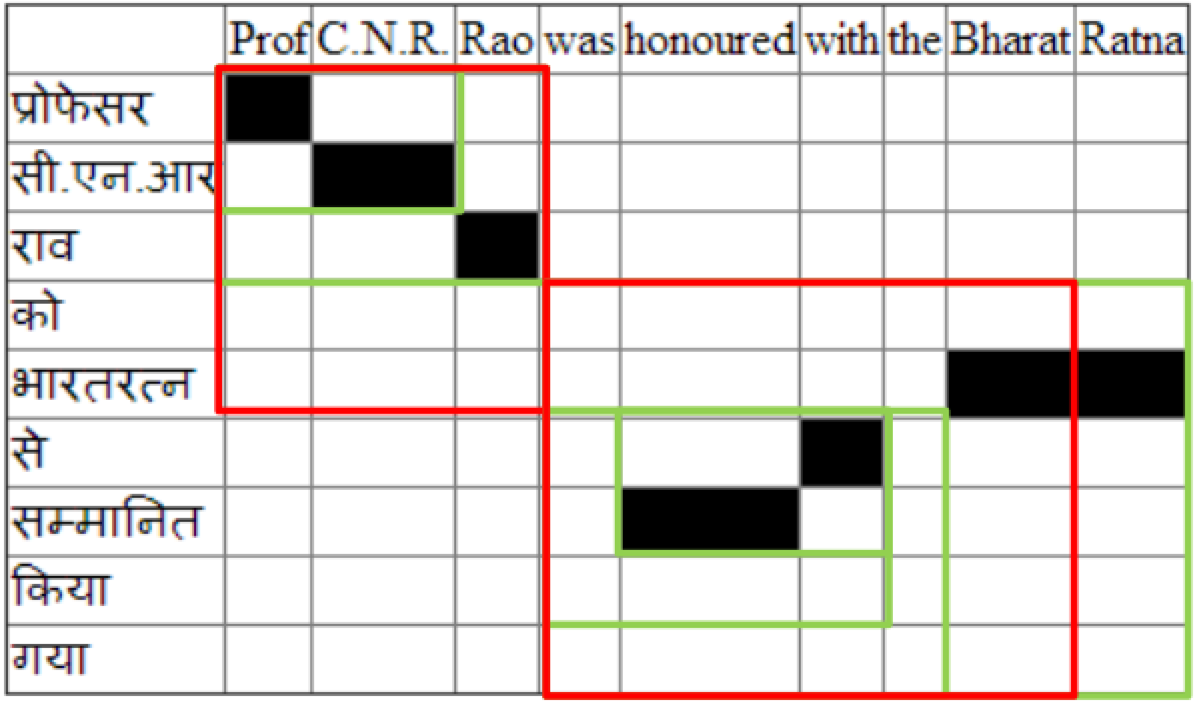
- Leverages word alignments learnt from IBM models
- Word Alignment : reliable input for phrase table learning
  - high accuracy reported for many language pairs
- Central Idea: A consecutive sequence of aligned words constitutes a “phrase pair”



	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न								■	■
से						■			
सम्मानित					■				
किया									
गया									

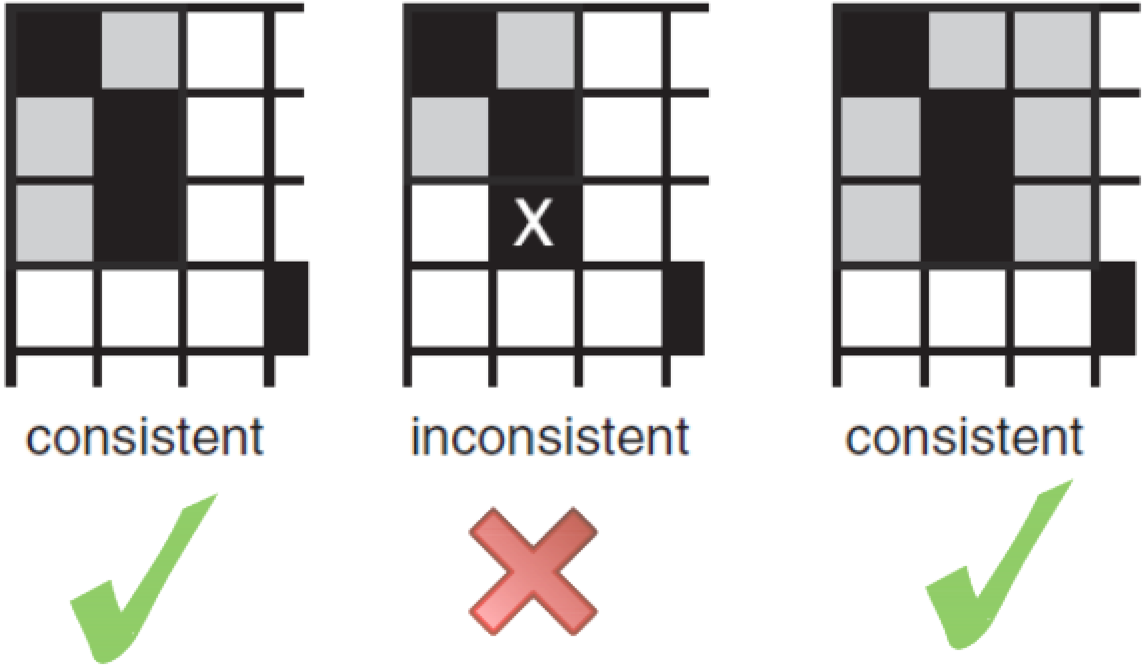
Which phrase pairs to include in the phrase table?

Extracting Phrase Pairs



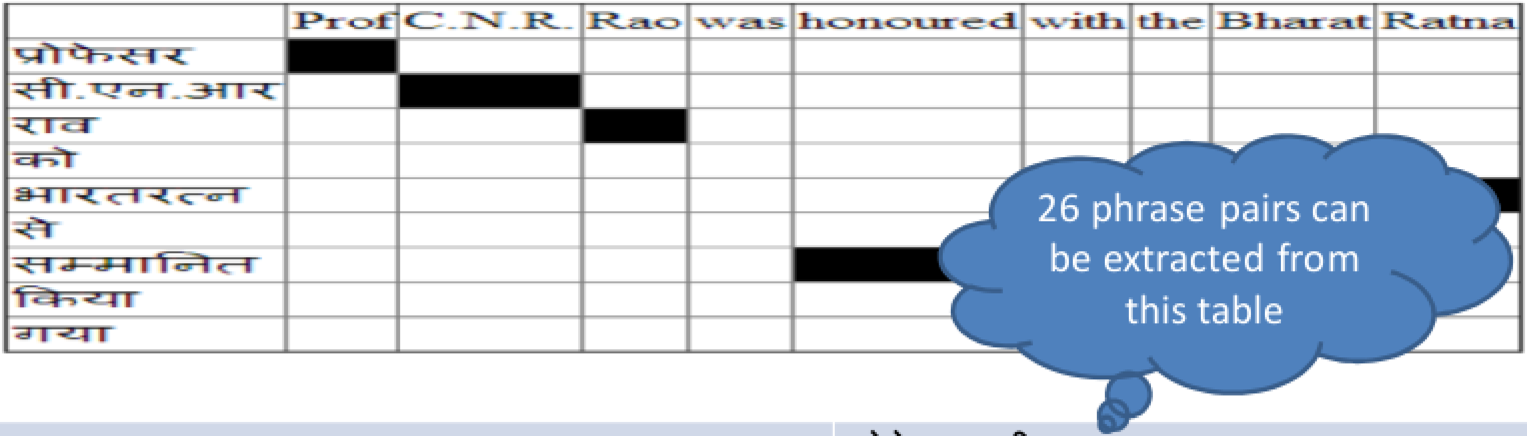
	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न								■	■
से						■			
सम्मानित					■				
किया									
गया									

Phrase Pairs “consistent” with word alignment



Source: SMT, Phillip Koehn

Examples



	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न									■
से									
सम्मानित					■				
किया									
गया									

26 phrase pairs can be extracted from this table

Professor CNR	प्रोफेसर सी.एन.आर
Professor CNR Rao	प्रोफेसर सी.एन.आर राव
Professor CNR Rao was	प्रोफेसर सी.एन.आर राव
Professor CNR Rao was	प्रोफेसर सी.एन.आर राव को
honoured with the Bharat Ratna	भारतरत्न से सम्मानित
honoured with the Bharat Ratna	भारतरत्न से सम्मानित किया
honoured with the Bharat Ratna	भारतरत्न से सम्मानित किया गया
honoured with the Bharat Ratna	को भारतरत्न से सम्मानित किया गया

Computing Phrase Translation Probabilities

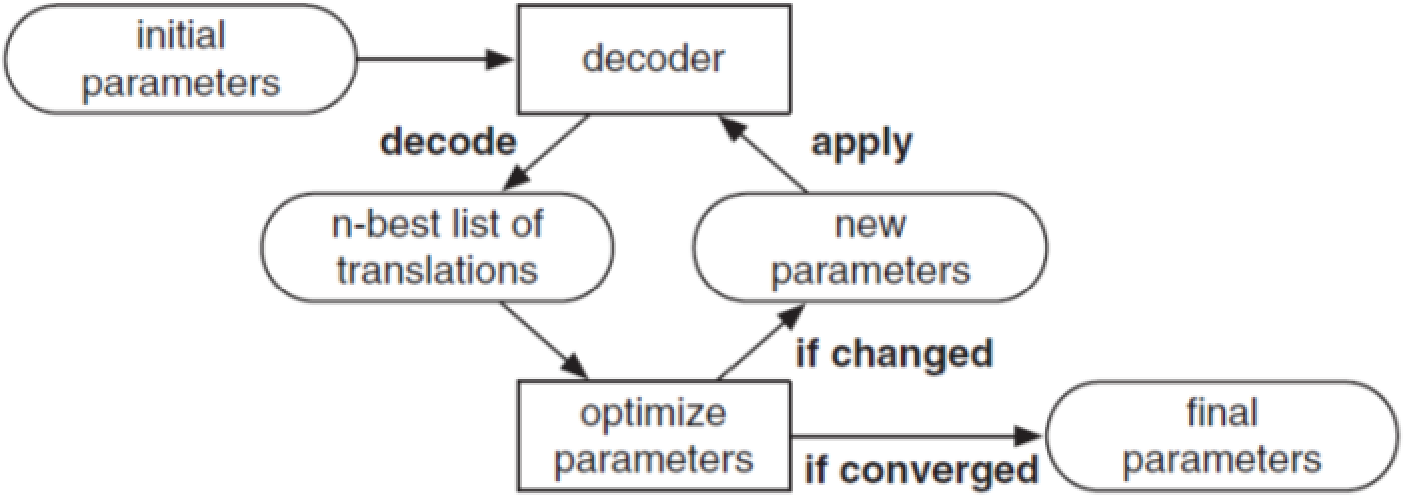
- Estimated from the relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e},\bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e},\bar{f}_i)}$$

Prime Minister of India	भारत के प्रधान मंत्री India of Prime Minister	***0.75***
Prime Minister of India	भारत के भूतपूर्व प्रधान मंत्री India of former Prime Minister	***0.02***
Prime Minister of India	प्रधान मंत्री Prime Minister	***0.23***

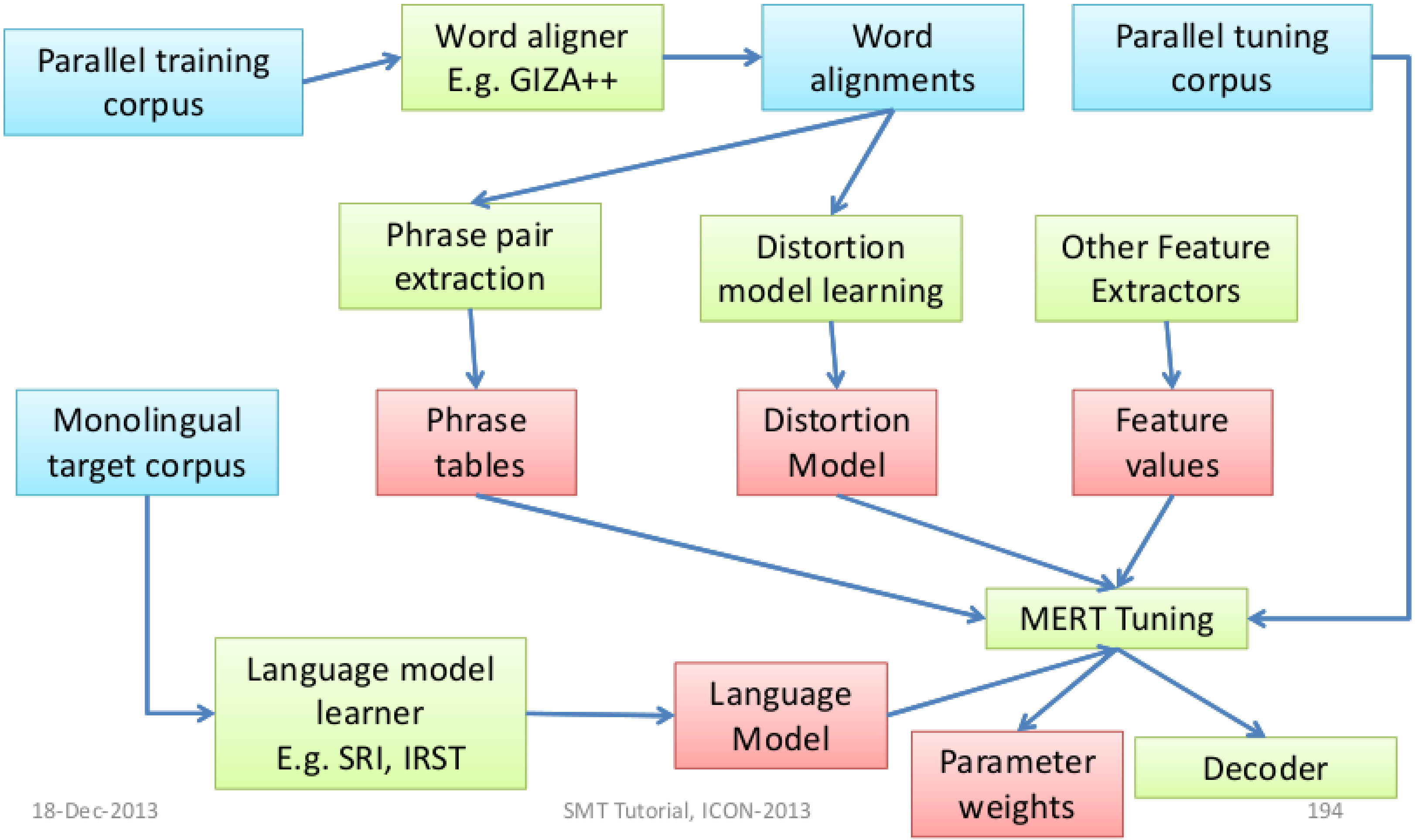
Tuning

- Learning feature weights from data – λ_i
- Minimum Error Rate Training (MERT)
- Search for weights which minimize the translation error on a held-out set (tuning set)
  - Translation error metric : (1 – *BLEU)*



Source: SMT, Phillip Koehn

Overall Training Process for PB-SMT



Parallel training corpus

Word aligner E.g. GIZA++

Word alignments

Parallel tuning corpus

Phrase pair extraction

Distortion model learning

Other Feature Extractors

Monolingual target corpus

Phrase tables

Distortion Model

Feature values

MERT Tuning

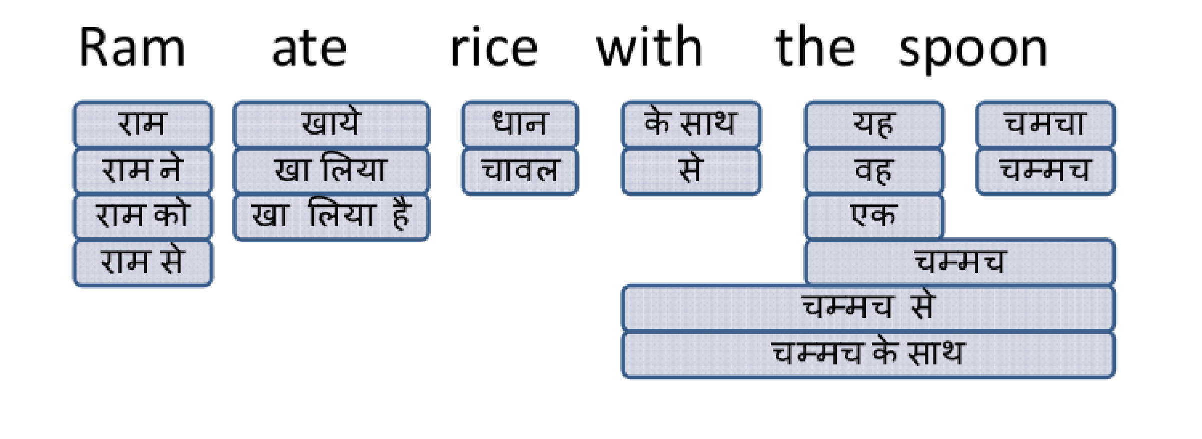
Language model learner E.g. SRI, IRST

Language Model

Parameter weights

Decoder

Decoding



Ram ate rice with the spoon

राम
राम ने
राम को
राम से

खाये
खा लिया
खा लिया है

धान
चावल

के साथ
से

यह
वह
एक

चमचा
चम्मच

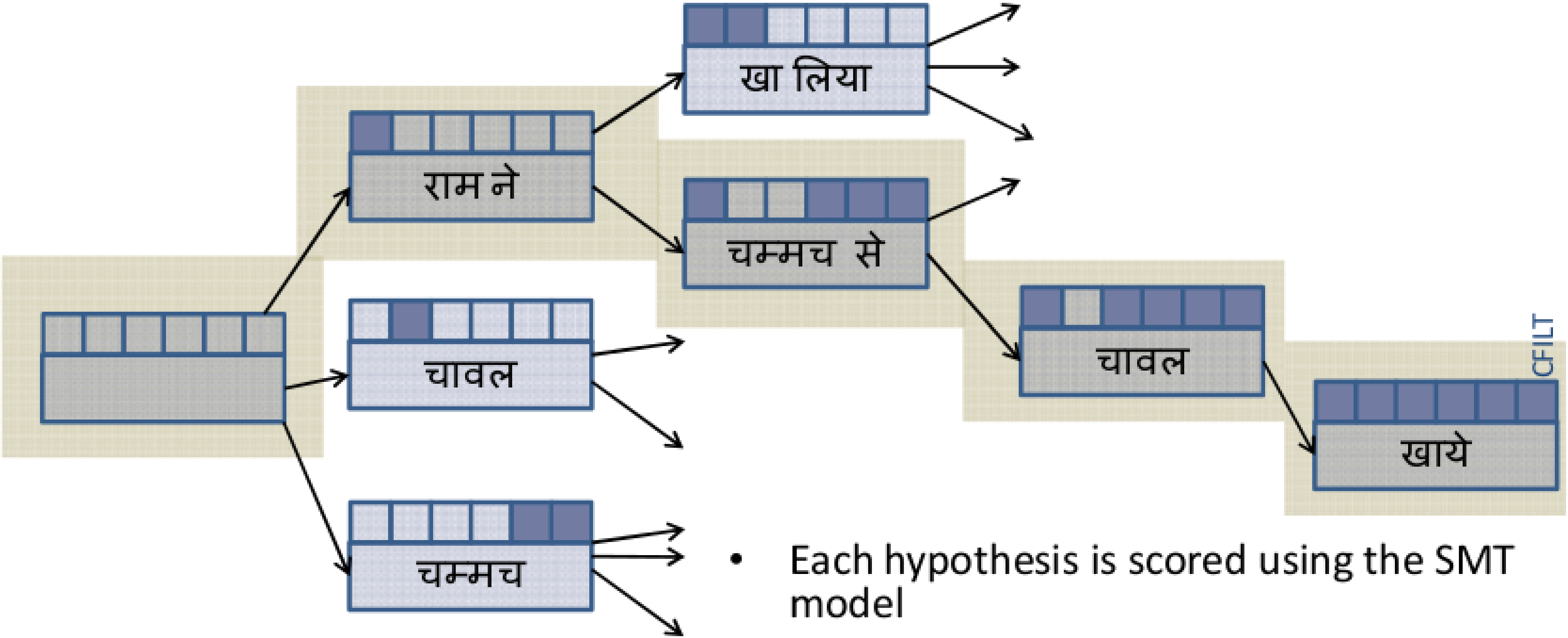
चम्मच

चम्मच से

चम्मच के साथ

- Find best translation among a very large number of possible translations
- NP-hard problem: 10-word sentence, 5 translations per word: 10^{5*}10! ~ 362 billion possibe translations
- Look for approximate solutions
  - Restrict search space: *some word orders are not possible*
  - Incremental construction and scoring
  - Remove candidates that are unlikely to eventually generate good translations

Search Space and Search Organization

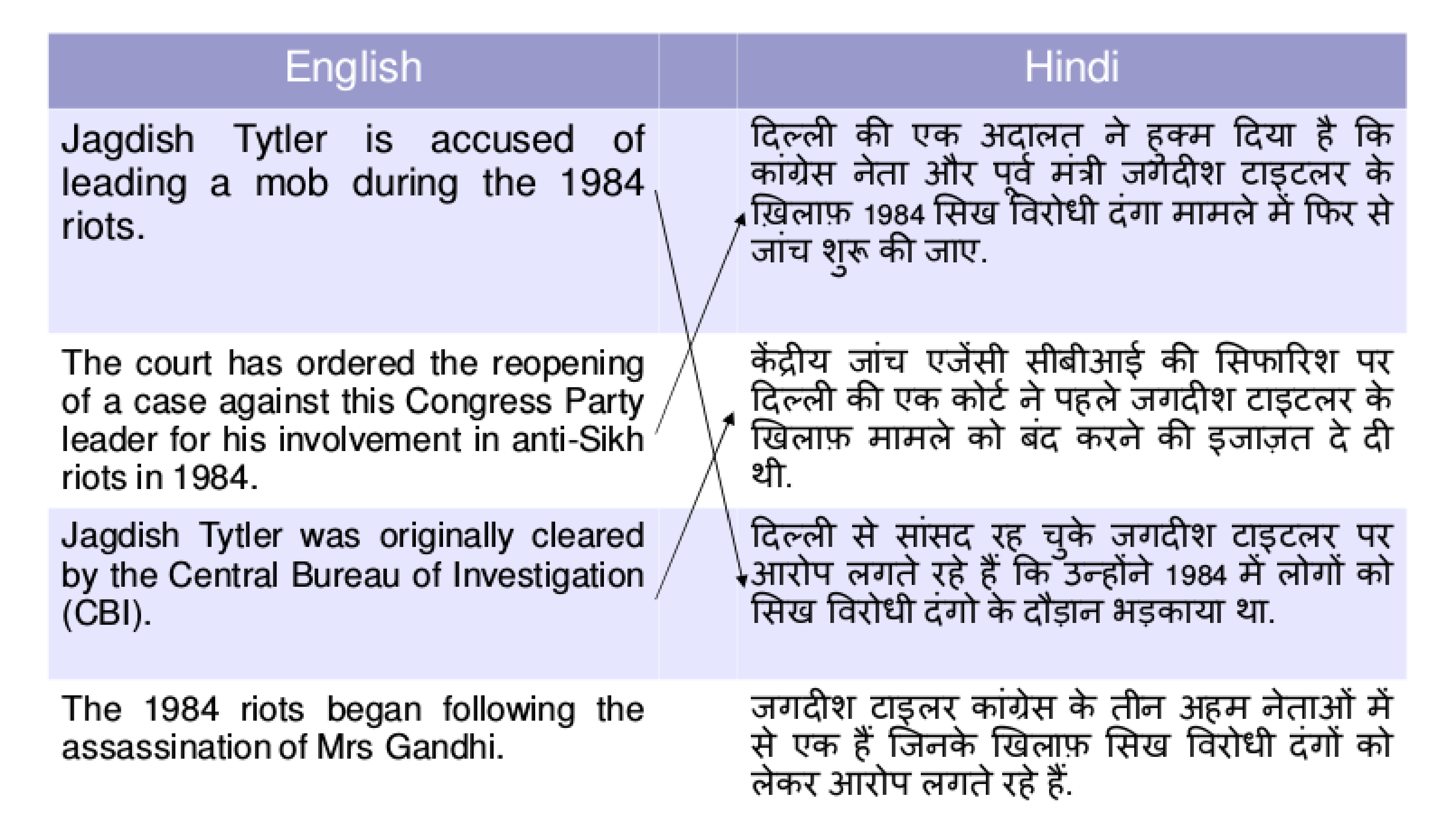


- Each hypothesis is scored using the SMT model
- Hypotheses are maintained in a priority queue (called stack decoding historically)
- Limit to the reordering window for efficiency

Richer Translation Models

- Syntax based SMT
- Factor based SMT
- Whole Document Translation

Building Parallel Corpus from Comparable Corpus



English		Hindi
Jagdish Tytler is accused of leading a mob during the 1984 riots.		दिल्ली की एक अदालत ने हुक्म दिया है कि कांग्रेस नेता और पूर्व मंत्री जगेदीश टाइटलर के ख़िलाफ़ 1984 सिख विरोधी दंगा मामले में फिर से जांच शुरू की जाए.
The court has ordered the reopening of a case against this Congress Party leader for his involvement in anti-Sikh riots in 1984.		केंद्रीय जांच एजेंसी सीबीआई की सिफारिश पर दिल्ली की एक कोर्ट ने पहले जगदीश टाइटलर के खिलाफ़ मामले को बंद करने की इजाज़त दे दी थी.
Jagdish Tytler was originally cleared by the Central Bureau of Investigation (CBI).		दिल्ली से सांसद रह चुके जगदीश टाइटलर पर आरोप लगते रहे हैं कि उन्होंने 1984 में लोगों को सिख विरोधी दंगो के दौड़ान भड़काया था.
The 1984 riots began following the assassination of Mrs Gandhi.		जगदीश टाइलर कांग्रेस के तीन अहम नेताओं में से एक हैं जिनके खिलाफ़ सिख विरोधी दंगों को लेकर आरोप लगते रहे हैं.

We could go further …. Unsupervised translation

Speech-to-Speech Translation

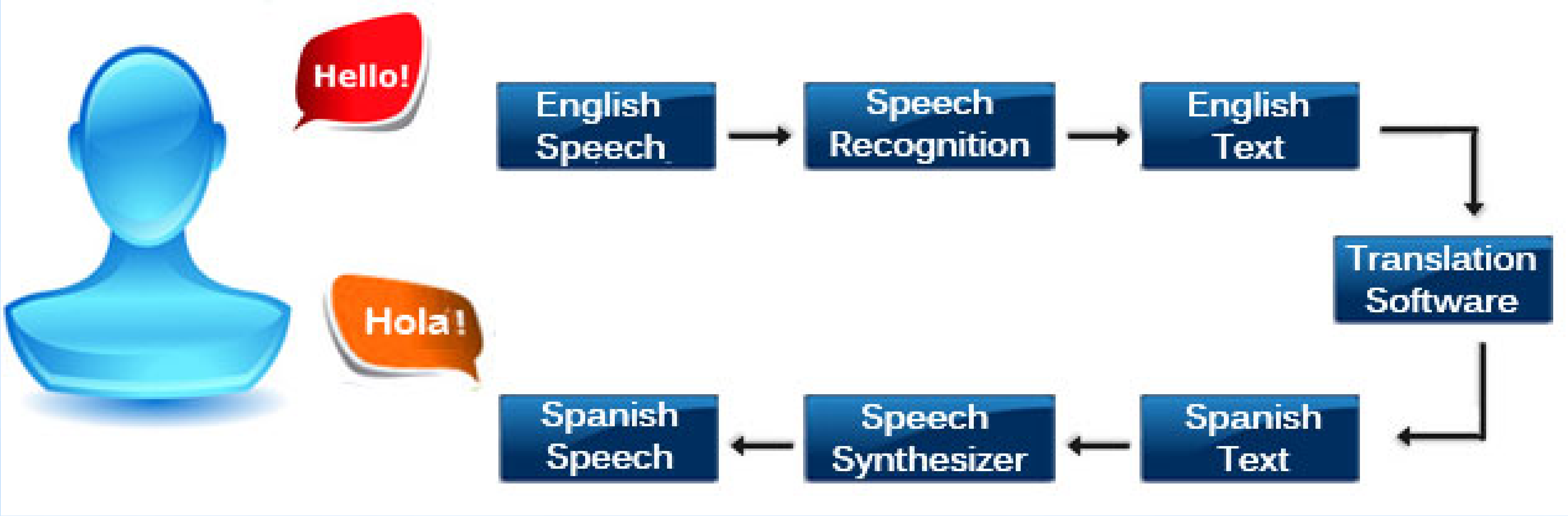
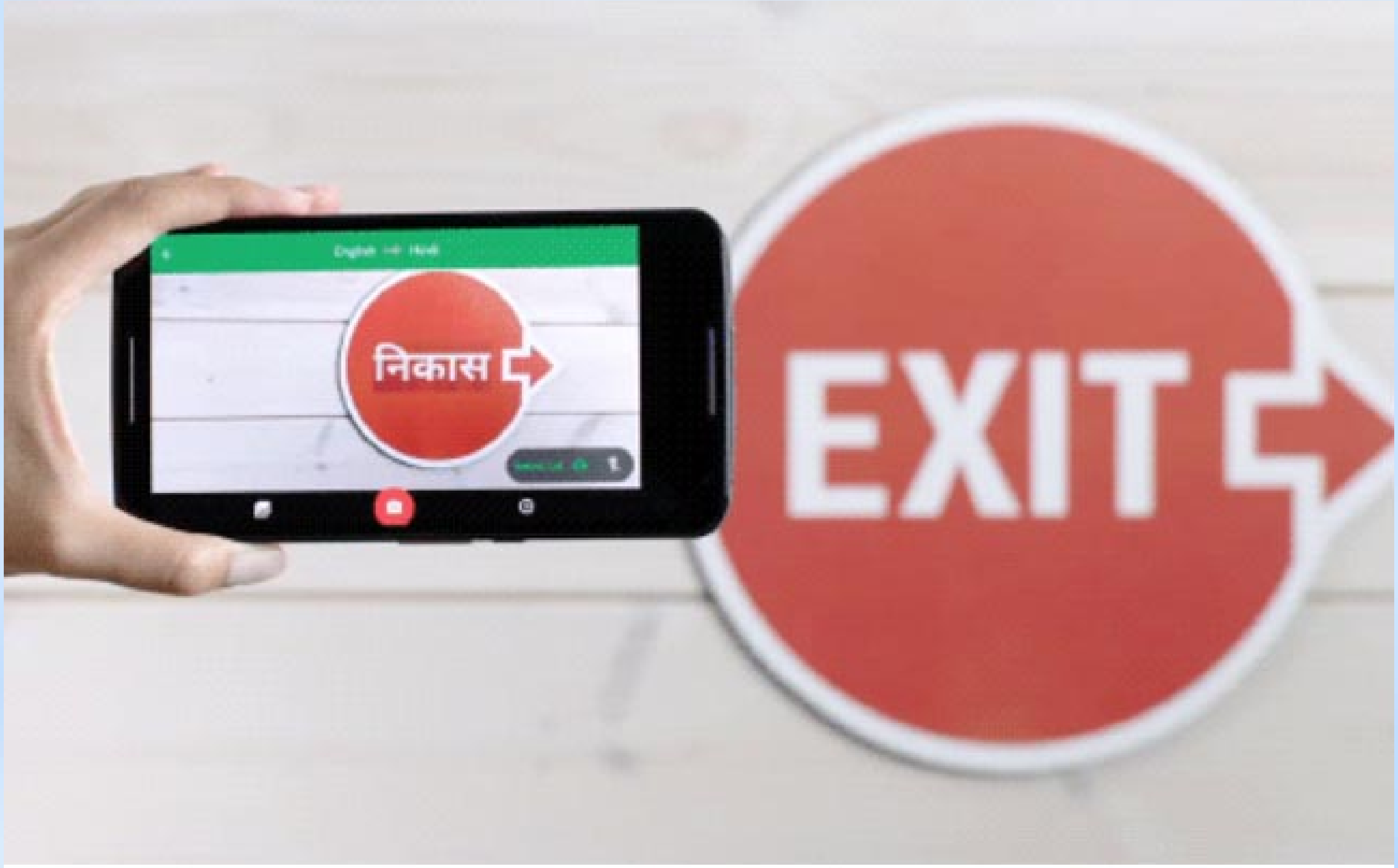


Image Text to Image Text Translation



Translation on smaller devices

Some more Interesting Problems

- Translation among Related Languages
- Scaling to larger corpora
- Deep learning and Machine Translation

References

- Introductory textbooks
  - Hutchins, William John, and Harold L. Somers. An introduction to machine translation. Academic Press, 1992.
  - Pushpak Bhattacharyya. Machine Translation. CRC Press, 2015.
- Other introductory material
  - *Kevin Knight's MT workbook*

    www.isi.edu/natural-language/mt/wkbk.pdf
  - *ICON 2013 tutorial on Statistical Machine Translation* *http://www.cfilt.iitb.ac.in/publications/icon_2013_smt_tutorial_slides.pdf*

References (2)

- Getting hands on
  - Moses

    http://www.statmt.org/moses/
  - Google Translate API

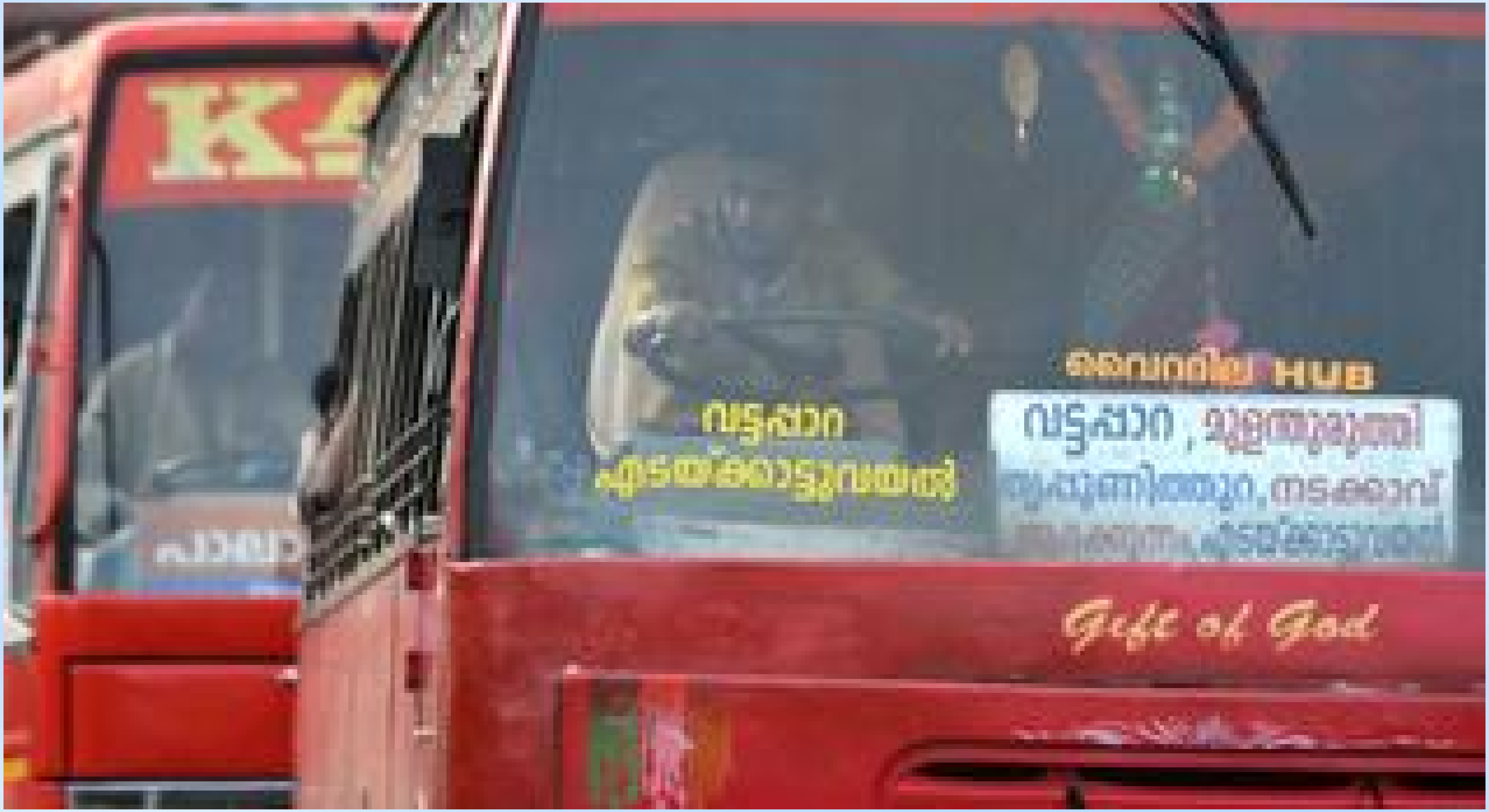
    https://cloud.google.com/translate/
  - Indic NLP Library

    https://github.com/anoopkunchukuttan/indic_nlp_library
  - IITB SMT tools for Indian languages

    http://www.cfilt.iitb.ac.in/static/download.html

Transliteration

You are in Kerala … waiting to travel by bus



Not a hypothetical situation …. Read this: http://www.thehindu.com/todays-paper/tp-national/tp-kerala/call-to-bring-on-board-bus-signage-in-three-languages/article5224039.ece

How do you translate Xi Xinping?

Xi Jinping is the President of China

शी चिनफिंग चीन के राष्ट्रपति है

Ok, we got lucky here … but there are so many names you will not find in any corpus

Transliteration can simplify Translation

Hindi

यदि श्वास प्रणालिका में सूजन आ जाये तब भी रक्त मुँह के रास्ते बाहर आने लगता है ।

Punjabi Translation

ਜੇਕਰ ਸਾਹ ਪ੍ਰਣਾਲੀ ਵਿਚ ਸੋਜ ਆ ਜਾਵੇ ਤਦ ਵੀ ਖੂਨ ਮੂੰਹ ਦੇ ਰਾਸਤੇ ਬਾਹਰ ਆਉਣ ਲਗਦਾ ਹੈ ।

जेकर साह प्रणाली विच सोज आ जावे तद वी खून मू□ह दे रासते बाहर आउण लगदा है ।

Hindi-Punjabi Transliteration

ਆਦਿ ਸਾਹ ਪ੍ਰਣਾਲੀ ਮੈਂ ਸੂਜਨ ਆ ਜਾਵੇ ਤਦ ਵੀ ਰਕਤ ਮੂੰਹ ਦੇ ਰਾਸਤੇ ਬਾਹਰ ਆਉਣ ਲਗਦਾ ਹੈ ।

आदि साह प्रणाली मैं सूजन आ जावे तद वी रकत मू□ह दे रासते बाहर आउण लगदा है ।

Some Concepts

Natural Language: A system of communication among humans with sound

Script: A system of symbols for representing language in writing

- *A language can have multiple scripts:*

  Sanskrit is written in many scripts (Devanagari, Malayalam, Tamil, Telugu, Roman, etc.)
- *A script can be used for multiple languages*

  Devanagari is used to write Sanskrit, Hindi, Marathi, Konkani, Nepali

Phoneme: basic unit of sound in a language that is meaningful

Grapheme: basic distinct unit of a script

- *A phoneme can be represented by multiple graphemes*

  cut, dirt
- *A grapheme can be used to represent multiple sounds*

  cut, put

What is transliteration?

Transliteration is the conversion of a given name in the source language (from source script) to a name in the target language (target script), such that the target language name is:

- phonemically equivalent to the source name

  मुम्बई → Mumbai

- conforms to the phonology of the target language

  ऩरेन्द्र → ਨਰੇਂਦਰ (नरेंदर)

- matches the user intuition of the equivalent of the source language name in the target language, considering the culture and orthographic character usage in the target language

  ആലപ്പുഴ (aalappuzha)→ Alappuzha

Isn't it easy to just map characters from one script to another?

- Local spelling conventions

  लता in Roman: Latha (South India) vs Lata (North India)

  Laxmi → लक्ष्मी

- Missing sounds

  കോഴിക്കോട് (kozhikkoT) → कोषिक्कोड (koShikkod)

- Transliterate or translate

  കോഴിക്കോട് (kozhikkoT) → Calicut

- Transliteration variants

  मुंबई, मुम्बई

Why English spellings caused trouble in school ...

Ambiguity in character to sound mapping

ionize vs nation

fish can be pronounced as *ghoti*

gh as in *tough*

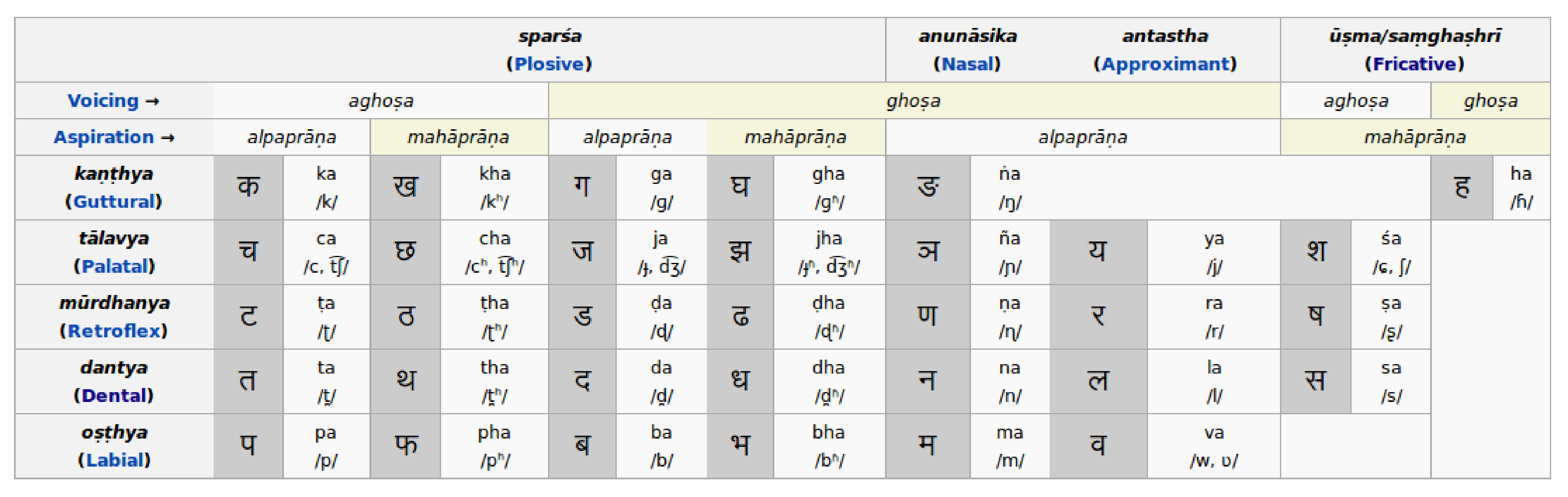
o as in women

ti as in nation

… and Hindi spellings didn't

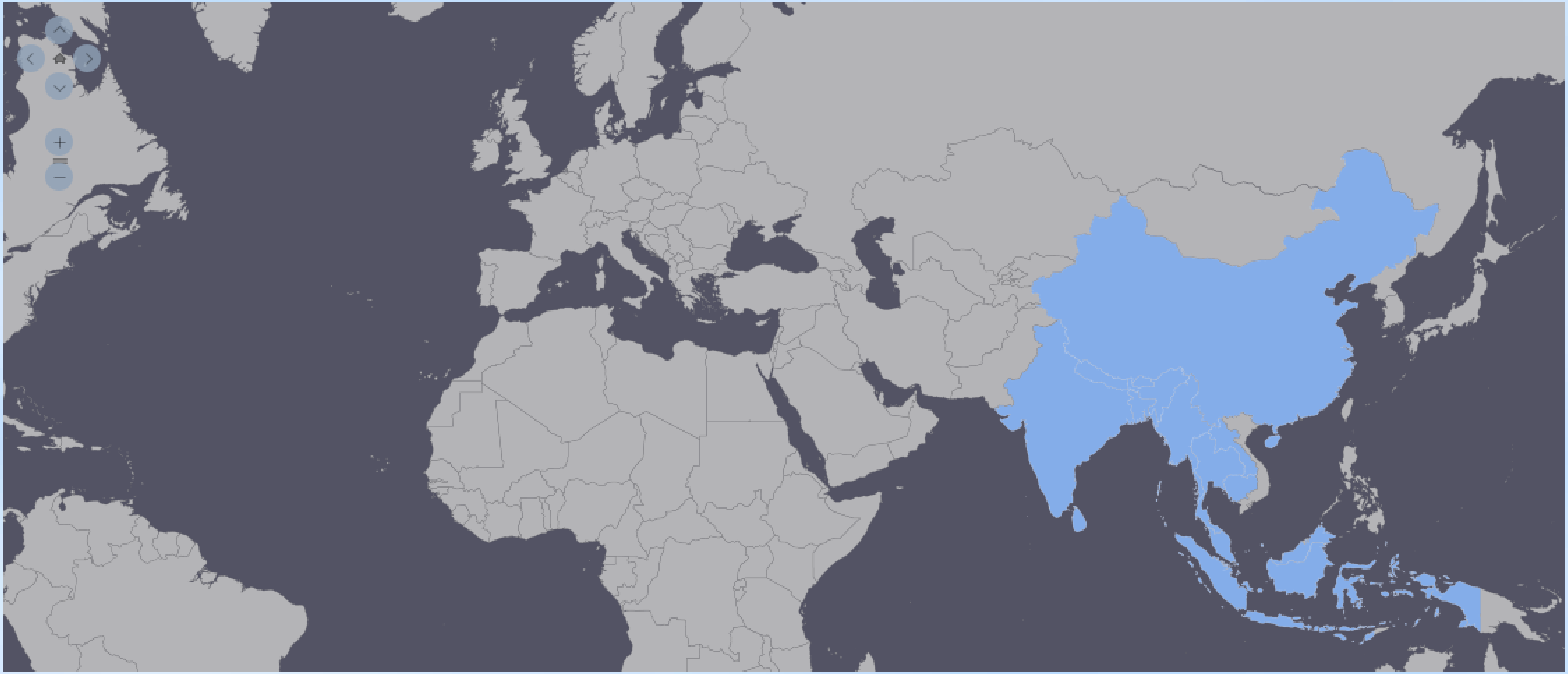
Unambiguous mapping from character to sound

*Rememember the **varnamala**? – organized according to scientific principles*



	sparśa **(Plosive)**								***anunāsika*** **(Nasal)**		***antastha*** **(Approximant)**		***ūṣma/saṃghaṣhrī*** **(Fricative)**			
Voicing →	*aghoṣa*				*ghoṣa*								*aghoṣa*		*ghoṣa*	
Aspiration →	*alpaprāṇa*		*mahāprāṇa*		*alpaprāṇa*		*mahāprāṇa*		*alpaprāṇa*				*mahāprāṇa*			
kaṇṭhya **(Guttural)**	क	ka /k/	ख	kha /kʰ/	ग	ga /g/	घ	gha /gʱ/	ङ	ṅa /ŋ/					ह	ha /ɦ/
tālavya **(Palatal)**	च	ca /c, t͡ʃ/	छ	cha /cʰ, t͡ʃʰ/	ज	ja /ɟ, d͡ʒ/	झ	jha /ɟʱ, d͡ʒʱ/	ञ	ña /ɲ/	य	ya /j/	श	śa /ɕ, ʃ/		
mūrdhanya **(Retroflex)**	ट	ṭa /ʈ/	ठ	ṭha /ʈʰ/	ड	ḍa /ɖ/	ढ	ḍha /ɖʱ/	ण	ṇa /ɳ/	र	ra /ɾ/	ष	ṣa /ʂ/		
dantya **(Dental)**	त	ta /t̪/	थ	tha /t̪ʰ/	द	da /d̪/	ध	dha /d̪ʱ/	न	na /n/	ल	la /l/	स	sa /s/		
oṣṭhya **(Labial)**	प	pa /p/	फ	pha /pʰ/	ब	ba /b/	भ	bha /bʱ/	म	ma /m/	व	va /w, ʋ/				

The extent of Devanagari-like scripts



How do we solve the transliteration problem?

- Transliteration is very similar to translation
- Instead of words, we have characters
- However, it is much simpler
  - No reordering
  - Small vocabulary (except Chinese and Japanese Kanji)
  - Regular grammar
- Similar to Vouquois triangle, you can transliterate at different levels:
  - Phoneme (like transfer based MT)
  - Grapheme (like direct MT)

References

- About Scripts
    - Omniglot: http://www.omniglot.com/
    - Wikipedia pages on Devanagari & Brahmi script
- About Transliteration
    - Karimi, Sarvnaz, Falk Scholer, and Andrew Turpin. "Machine transliteration survey." ACM Computing Surveys. 2011.
- Hands on
    - Google Transliterate

      http://www.google.com/inputtools/
    - Brahmi-Net: IITB's transliteration system

      http://www.cfilt.iitb.ac.in/brahminet/

Thank You!

Acknowledgments

- Some of slides & images have been borrowed from
  - ICON 2013 Tutorial on Statistical Machine Translation. Pushpak Bhattacharyya, Anoop Kunchukuttan, Piyush Dungarwal, Shubham Gautam
  - Wikipedia