Transliteration involving English and Hindi languages using Syllabification Approach

DDP Stage3 Presentation
January 8, 2010

By

Ankit Aggarwal (03d05009)

Guided by

Dr. Pushpak Bhattacharyya

Roadmap

- Introduction to Transliteration
- Problem Definition
- Existing Statistical Approaches
- Noisy-Channel Model
- Moses Tool
- Baseline
- Syllabification Approach
- Statistical Machine Syllabification
- Statistical Machine Transliteration
- Error Analysis: Transliteration
- **Hindi-English Transliteration**
- Conclusion & Future Work
- References

What is Transliteration?

- Practice of transcribing a word or text written in one writing system into another writing system.
- E.g., 'school' will be transliterated to 'स्कूल'.
- Different from translation:
  - 'school' will be translated to 'पाठशाला' (*'paathshaala')*.
- Forward Transliteration: The process of transforming source language phonemes into target language approximations.
- Back Transliteration: The reverse of forward transliteration; transforming target language approximations back into their original source language.

मनोरमा → manorama (forward)

jitesh → जितेश (back)

computer → कंप्यूटर (forward)

Need of Transliteration

- OOV words are a common search item in CLIR.

- Applications in SMT:
  - Inability to have complete coverage of proper names.
  - Unknown words either get dropped or are copied unchanged.

Shilpa Shetty takes a keen interest in Indian football.

↓

Shilpa शेट्टी भारतीय फुटबॉल में विशेष रुचि लेती/लेता है.

Challenges & Problem Definition

- More than one valid forward transliterations:
  - गौतम → gautam, goutam, gowtam, gautham
  - मीनाक्षी → meenakshi, meenaxi, minaakshi
- Back Transliteration is even more challenging.
  - The information losing aspect of forward transliteration makes it hard to invert.
  - Less flexibility in the reverse direction.

gautam → गौताम (X)

minaakshi → मिनाक्शी (X)

- **Problem Definition:**
  - Given a word (an Indian origin name) written in English (or Hindi) language script, the system needs to provide five-six most probable Hindi (or English) transliterations of the word, in the order of higher to lower probability.

Existing Statistical Approaches

- The statistical transliteration approaches can be classified into:

  - **phoneme-based:** A *phoneme* is the smallest unit of speech. E.g., /t/ sound found in ***t***ip, s***t***and, ca***t***.
    - Source -Channel Model (Knight and Graehl, 1998).

  - **grapheme-based:** A *grapheme* is the fundamental unit in written languages. Includes characters of the alphabet, numerals and punctuation marks.
    - A grapheme can map to multiple phonemes: ***t*** can map to phonetic equiv. of ट and त.
    - It treats transliteration as a statistical machine translation problem at the surface level.

  - **hybrid approach:** Attempts to utilize both phoneme and grapheme information for transliteration.

SOURCE-CHANNEL MODEL

- **Source-Channel Model** *(Knight and Graehl, 1998)*

  - Generative Model
  - Japanese (katakana) to English back-transliteration
  - Given a Japanese 'katakana' string:
    - An English string is written (generated).
    - A translator pronounces it in English.
    - The pronunciation is modified to fit the Japanese sound inventory.
    - The sounds are converted to 'katakana'.
    - 'katakana' is written.
  - The idea is to transform both source and target names into comparable phonemes so that the phonetic similarity between the two names can be measured easily.

SOURCE-CHANNEL MODEL

- A Japanese 'katakana' string o is given to the system and it tries to find the English word 'w' that maximizes P(w|o).
- Using Bayes' theorem, this problem can be re-written as:

$$\arg\max_{w} P(w|o) = \arg\max_{w}\{P(w) \cdot P(e|w) \cdot P(j|e) \cdot P(k|j) \cdot P(o|k)\}$$

- So, the new problem is to find an English word 'w' that maximizes the sum , over all e, j and k, of the above (RHS).

  - P(w) – generates written English word sequences.
  - P(e|w) – pronounces English word sequences.
  - P(j|e) – converts English sounds into Japanese sounds.
  - P(k|j) – converts Japanese sounds to katakana writing.
  - P(o|k) – introduces misspellings caused by optical character recognition (OCR).

Noisy-Channel Model

- **Noisy-channel Model**
  - E.g., Foreign language word (*f*) - English (*e*) back-transliteration
  - Using Bayes' Theorem:

$$P(e|f) = \frac{P(e) \cdot P(f|e)}{P(f)}$$

  - Since P(*f*) is independent of *e*, we can rewrite above equation as:

$$\hat{e} = \arg\max_{e} P(e) \cdot P(f|e)$$

  - The language model P(*e*) is simplified as n-gram model of English characters and is trained with an English name corpus.
  - The transliteration model P($f|e$) is estimated from a parallel corpus of English names and their Foreign language transliterations.
  - The decoder combines the language model and transliteration model to generate English back-transliterations for Foreign language names.

Our Project

- We are interested in:
  - English to Hindi (Devanagari) back-transliteration and Hindi (Devanagari) to English forward-transliteration
- We intend to use the noisy channel model.
- Requirements:
  - Bilingual parallel training data
  - A tool to automatically train from the training data and decode the test strings based on the noisy channel model.
- **Data:**
  - English-Hindi bilingual data is not readily available.
  - The data available on the website of Election Commission of India is not good in quality and requires a lot of mending.
  - We will discuss the sources of data sets used in our project in detail later.
- **Tool:**
  - ***Moses*** is leveraged to implement the noisy channel model for grapheme-based machine transliteration.

Moses Tool

- A phrase-based statistical machine translation tool
- Used for grapheme based machine transliteration without *reordering*.
- A collection of tools is used by Moses.
- **SRILM**
  - Used to build statistical language models.
  - Primarily used in speech recognition, statistical tagging and segmentation.
  - Available from: http://www.speech.sri.com/projects/srilm/
- **GIZA++**
  - Used to perform word (characters in our case) alignments over parallel corpora.
  - Available from http://www.fjoch.com/GIZA++.html
- **Special Features of Moses**
  - **beam-search:** an efficient search algorithm that quickly finds the highest probability translation among the exponential number of choices.
  - **phrase-based:** the state-of-the-art in SMT allows the translation of short text chunks.

Baseline Approach

- **Training Format:** Space-delimited characters in the words (names).

Source	Target
s u d a k a r	स ु द ा क र
c h h a g a n	छ ग ण
j i t e s h	ज ि त े श
n a r a y a n	न ा र ा य ण
s h i v	श ि व
m a d h a v	म ा ध व
m o h a m m a d	म ो ह म ् म द
j a y a n t e e d e v i	ज य ं त ी द े व ी

- **Moses Settings** *(Serve as default settings throughout the project unless specified otherwise)*
  - **Language Model (LM) N-gram Order:** 5
  - **LM Smoothing & Interpolation:** Kneser-Ney (Kneser and Ney, 1995), Interpolate
  - **Model Weights:**
    - Translation Model: 0.2, 0.2, 0.2, 0.2, 0.2
    - Language Model: 0.5
    - Distortion Model: 0.0

Experiment and Evaluation

- **Data Set:**
  - 23k parallel names, split into training and testing set.
  - Testing set consists of 4,500 names.

- **Evaluation Metric:**

$$Top-n\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1\ if\ \exists j:\ r_i = c_{i,j} \\ 0 \qquad otherwise\end{cases}$$

  - N : Total Number of names (source words) in the test set
  - r_i : Reference transliteration for i-th name in the test set
  - $c_{i,j}$: j-th candidate transliteration (system output) for i-th name in the test set ($1 \le j \le 5$)

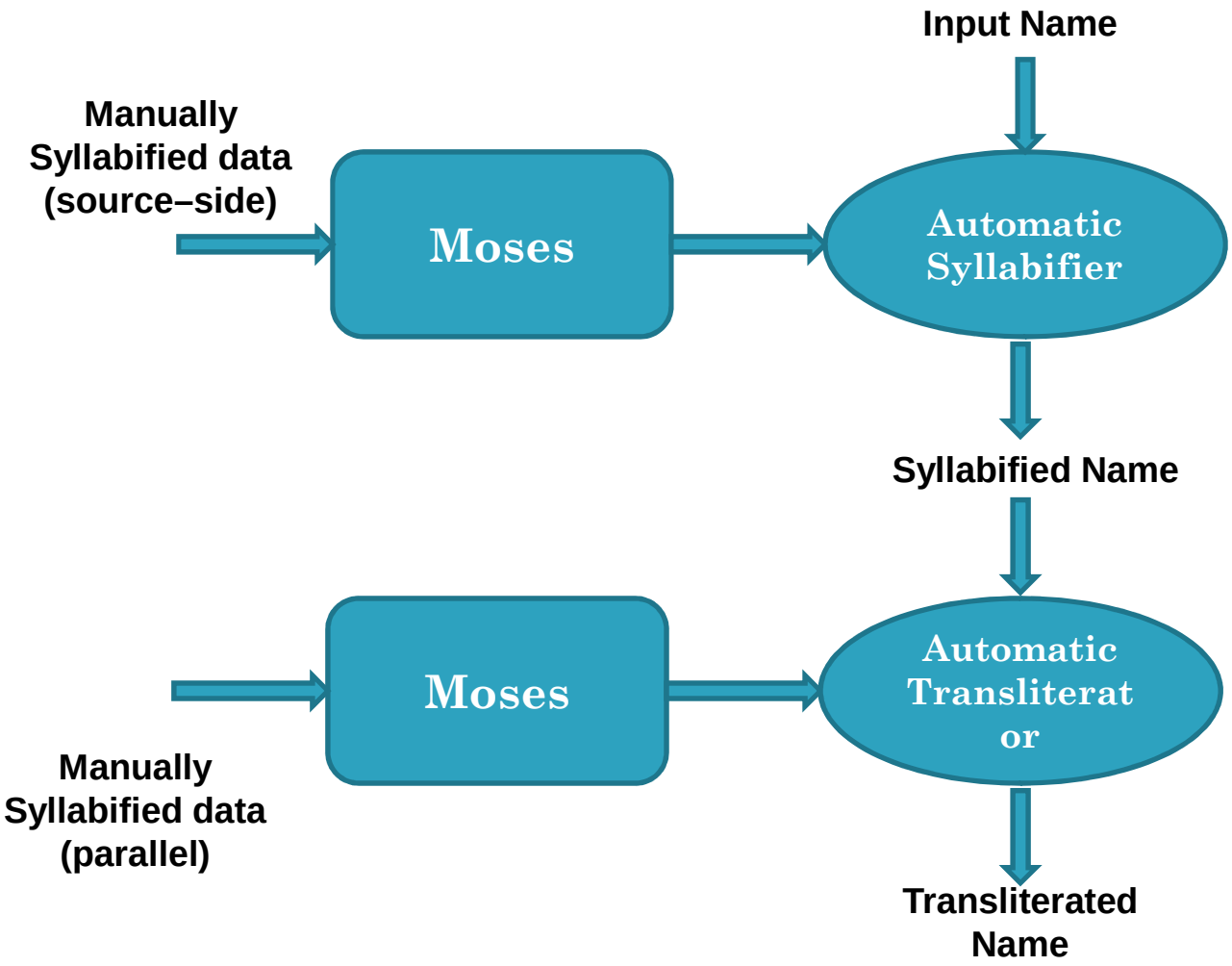
Baseline Performance

- **Results**

Top-n	Correct	Correct %age	Cumulative %age
1	1,868	41.5%	41.5%
2	520	11.6%	53.1%
3	246	5.5%	58.5%
4	119	2.6%	61.2%
5	81	1.8%	63.0%
Below 5	1,666	37.0%	100.0%
	4,500		

- The Top-5 Accuracy is just 63.0% which is much lower than required.

Syllabification Approach



Statistical Machine Syllabification

Data

- **Sources of data**

1. **Election Commission of India (ECI) Name List :** This web source provides native Indian names written in both English and Hindi.
2. **Delhi University (DU) Student List :** This web sources provides native Indian names written in English only. These names were manually transliterated for the purposes of training data.
3. **Indian Institute of Technology, Bombay (IITB) Student List:** The Academic Office of IITB provided this data of students who graduated in the year 2007.
4. **Named Entities Workshop (NEWS) 2009 English-Hindi parallel names :** A list of paired names between English and Hindi of size 11k is provided.

- We start with 8,000 randomly chosen names from the ECI Name list and syllabify them manually.

Statistical Machine Syllabification

Effect of Training Format

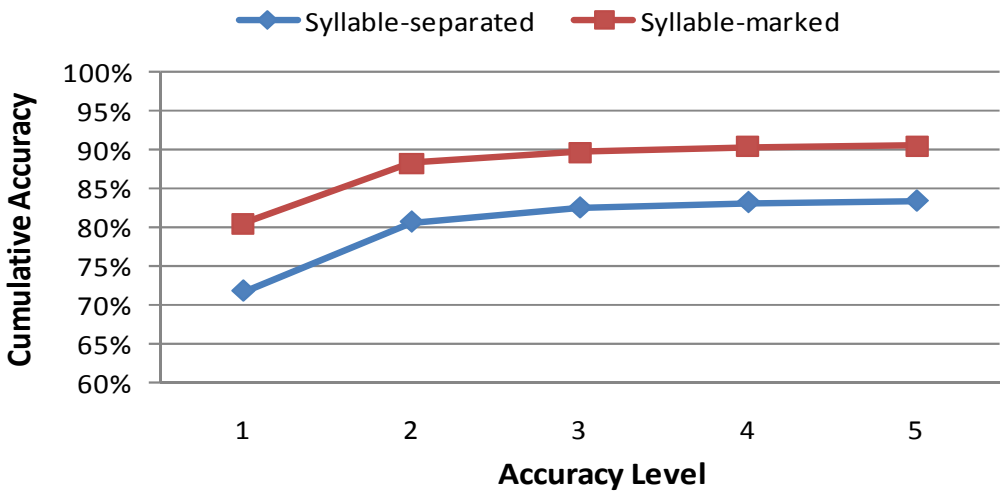
- *Syllable-separated Format*

Source	Target
s u d a k a r	su da kar
c h h a g a n	chha gan
j i t e s h	ji tesh
n a r a y a n	na ra yan
s h i v	shiv
m a d h a v	ma dhav
m o h a m m a d	mo ham mad
j a y a n t e e d e v i	ja yan tee de vi

- *Syllable-marked Format*

Source	Target
s u d a k a r	s u _ d a _ k a r
c h h a g a n	c h h a _ g a n
j i t e s h	j i _ t e s h
n a r a y a n	n a _ r a _ y a n
s h i v	s h i v
m a d h a v	m a _ d h a v
m o h a m m a d	m o _ h a m _ m a d
j a y a n t e e d e v i	j a _ y a n _ t e e _ d e _ v i

Statistical Machine Syllabification



- Top-1 Accuracy: 71.8% (Syllable-separated); 80.5% (Syllable-marked)
- Top-5 Accuracy: 83.4% (Syllable-separated); 90.4% (Syllable-marked)

Statistical Machine Syllabification

- *Syllable-separated*
  - The system needs to learn the alignment. For e.g., 'sudakar'

    s u d a k a r → su da kar ('s u d' -> 'su', 'a k' -> 'da' & 'a r' -> 'kar')

    s u d a k a r → su da kar

    s u d a k a r → su da kar

    .....

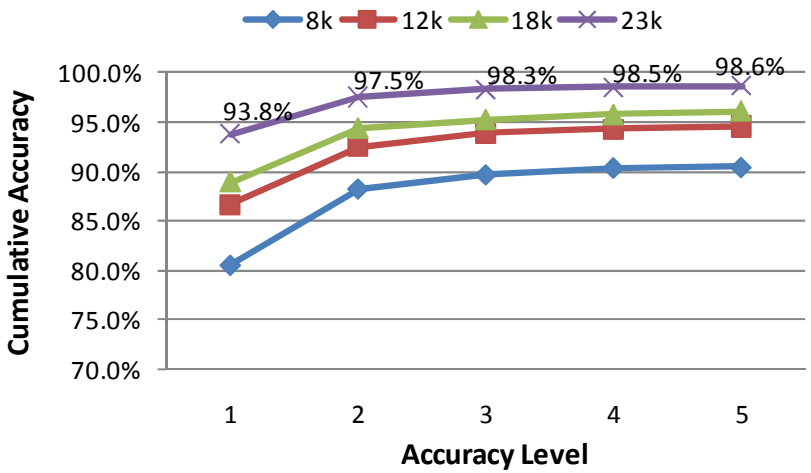
- *Syllable-marked*
  - The system generates a target sequence.
  - To estimate the score of this sequence, it looks back upto 'n' number of characters from any *underscore* and calculates the probability of this *underscore* being at the right place.
  - Thus, it avoids alignment task and performs better.

Statistical Machine Syllabification

Effect of Data Size

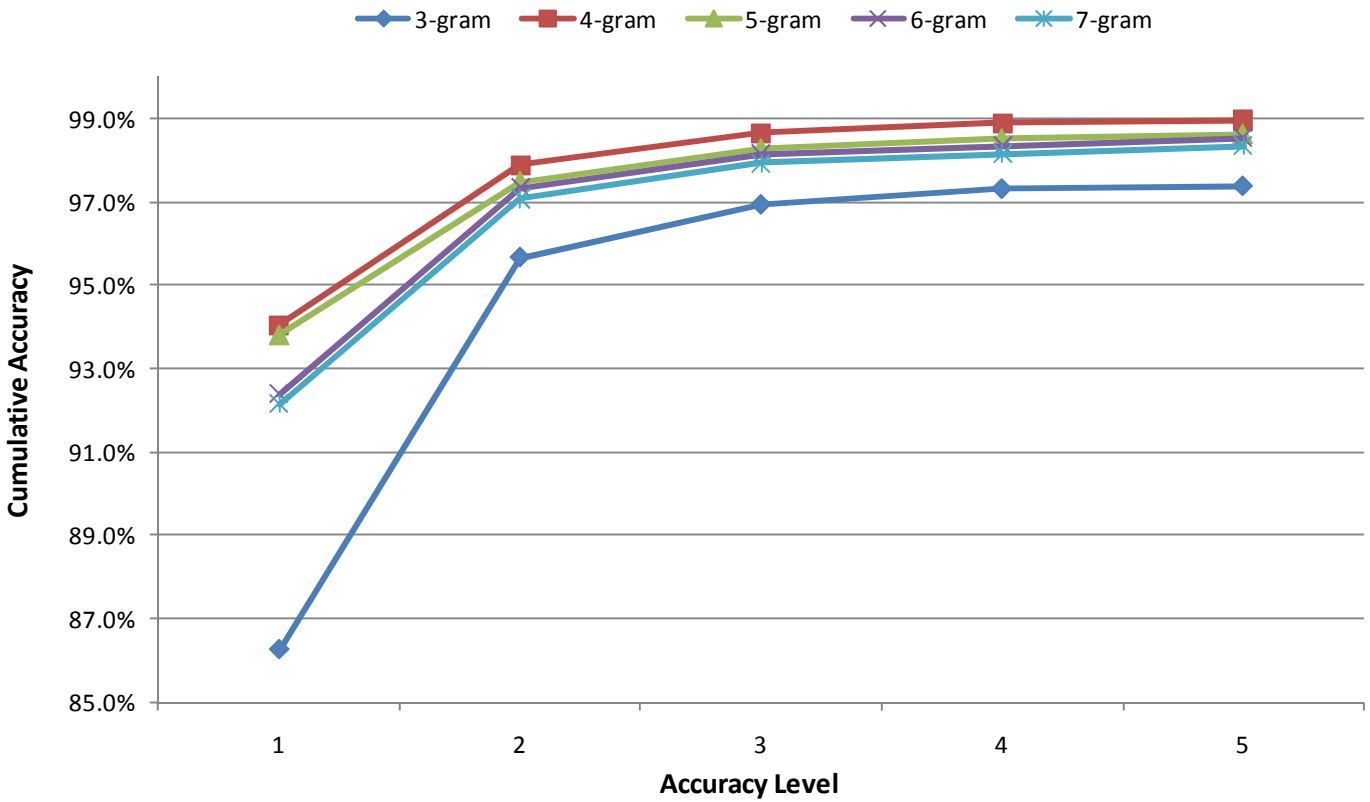
- 4 different experiments were performed:

1. **8k**: Names from the ECI Name list.
2. **12k**: An additional 4k names were syllabified to increase the data size.
3. **18k**: IITB Student List and the DU Student List was included and syllabified.
4. **23k**: Some more names from ECI Name List and DU Student List were syllabified and this data acts as the **final data** for us.



Statistical Machine Syllabification

Effect of Language Model n-gram Order



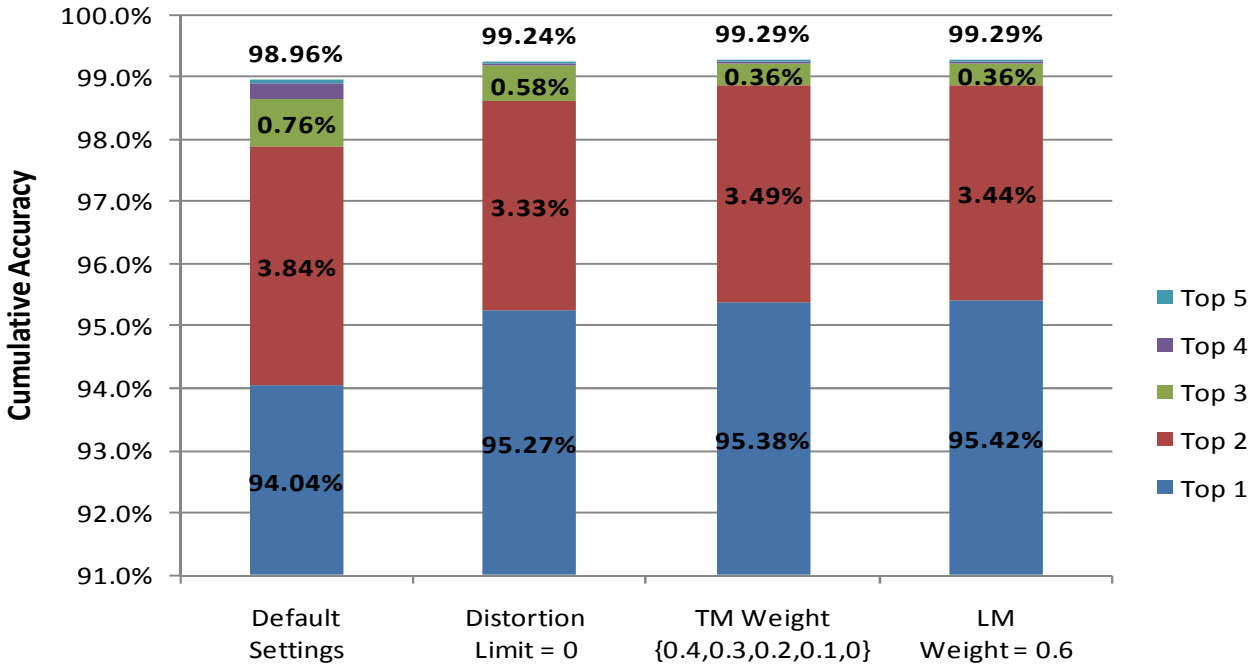
Statistical Machine Syllabification

- **'n' = 2**
  - Top-1 Accuracy: 23.3%
  - The score of a generated target sequence determined only on the basis of a single English character (one of them being *underscore*).
- **'n' >= 3**
  - Big jump in the performance. Top-1 Accuracy: 86.2%
  - The performance increases and then decreases.
  - Maximum performance is achieved at 'n' = 4.
    - Top-1 Accuracy: 94.0%; Top-5 Accuracy: 99.0%
- **Explanation**
  - *Average number of characters per word:* **7.6**
  - *Average number of syllables per word:* **2.9**
  - *Average number of characters per syllable:* **2.7 (= 7.6/2.9)**
  - Estimate: Integer (2.7) + 1 (for *underscore*) = 4

Statistical Machine Syllabification

Tuning the Model Weights

- **Distortion Limit:** 0
- **Translation Model Weight:** {0.4, 0.3, 0.2, 0.1, 0.0}
- **Language Model Weight:** 0.6



Statistical Machine Transliteration

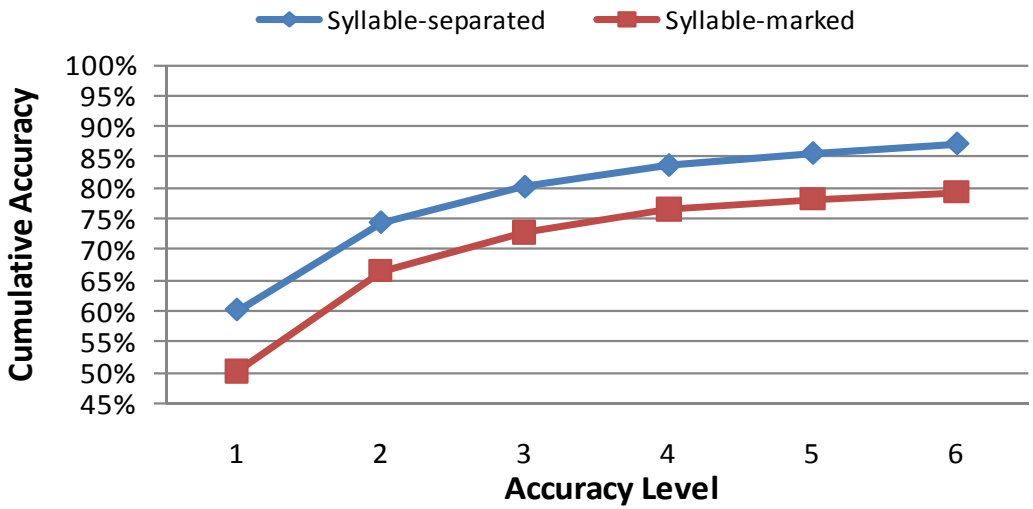
- **Data Set:** 23k data as described previously.
- **Training Format**
  - *Syllable-separated*

Source	Target
su da kar	सु दा कर
chha gan	छ गण
ji tesh	जि तेश
na ra yan	ना रा यण
shiv	शिव
ma dhav	मा धव
mo ham mad	मो हम् मद
ja yan tee de vi	ज यं ती दे वी

  - *Syllable-marked*

Source	Target
s u _ d a _ k a r	स ु _ द ा _ क र
c h h a _ g a n	छ _ ग ण
j i _ t e s h	ज ि _ त े श
n a _ r a _ y a n	न ा _ र ा _ य ण
s h i v	श ि _ व
m a _ d h a v	म ा _ ध व
m o _ h a m _ m a d	म ो _ ह म ् _ म द
j a _ y a n _ t e e _ d e _ v i	ज य ं _ त ी _ द े _ व ी

Statistical Machine Transliteration



- Top-1 Accuracy: 60.1% (Syllable-separated); 50.2% (Syllable-marked)
- Top-6 Accuracy: 87.2% (Syllable-separated); 79.3% (Syllable-marked)
- *Reason*: Most of the syllables that are seen in the training corpora are present in the testing data as well.
- But, it is accompanied with the problem of failure to transliterated unknown syllables.

Statistical Machine Transliteration

Effect of Language model n-gram Order

		n-gram Order					
		2	**3**	**4**	**5**	**6**	**7**
Level-n Accuracy	**1**	58.7%	60.0%	60.1%	60.1%	60.1%	60.1%
	2	74.6%	74.4%	74.3%	74.4%	74.4%	74.4%
	3	80.1%	80.2%	80.2%	80.2%	80.2%	80.2%
	4	83.5%	83.8%	83.7%	83.7%	83.7%	83.7%
	5	85.5%	85.7%	85.7%	85.7%	85.7%	85.7%
	6	86.9%	87.1%	87.2%	87.2%	87.2%	87.2%

Statistical Machine Transliteration

Tuning the Model Weights

- **Distortion Limit:** 0
- **Translation Model Weight:** {0.4, 0.3, 0.15, 0.15, 0.0}
- **Language Model Weight:** 0.5

Top-n	Correct	Correct %age	Cumulative %age
1	2,780	61.8%	61.8%
2	679	15.1%	76.9%
3	224	5.0%	81.8%
4	177	3.9%	85.8%
5	93	2.1%	87.8%
6	53	1.2%	89.0%
Below 6	494	11.0%	100.0%
	4,500		

Error Analysis: Transliteration

- **Unknown Syllables**
  - Unknown syllables are left as is.
  - Error reduced with increase in data size.
  - E.g., “jodh”, “vish”, “dheer”, “srish” etc.

- **Incorrect Syllabification**
  - The names with incorrect syllabification (Top-1 Result only) are prone to incorrect transliteration.
  - E.g., “shyamadevi” is syllabified as “shyam a devi”. “shweta” is syllabified as “sh we ta”.
  - Exceptionally, there are cases when incorrectly syllabified name gets correctly transliterated.
  - E.g., both “ga yat ri” and “ga ya tri” are transliterated to गायत्री.

Error Analysis: Transliteration

- **Foreign Origin**
  - Names of foreign origin, but widely used in India.
  - E.g., “mickey”, “prince”, “baby”, “dolly”, “cherry”, “daisy”.

- **Half Consonants**
  - Half consonants present in the name get wrongly transliterated as full consonants.
  - *Reason*: Probability of latter is more than the former.
  - E.g., “himmat” → हिममत (Correct: हिम्मत)

Error Analysis: Transliteration

- **Error in *maatra* (मात्रा)**
  - In the names with 3 or more *maatrayein*, the system might place the desired output very low in probability. E.g., 'bakliwal':
  - **1st a:** अ / आ **i:** इ / ई **2nd a:** अ / आ
  - 8 possibilities: बाकलीवाल, बकलीवाल, बाकलिवाल, बकलिवाल, बाकलीवल, बकलीवल, बाकलिवल, बकलिवल

- **Multi-mapping**

English Letters	Hindi Letters
t	त, ट
th	थ, ठ
d	द, ड, ड़
n	न, ण, ं
sh	श, ष
ri	रि / ऋ
ph	फ / फ़

Error Analysis: Transliteration

Error Type	Number	Percentage
Unknown Syllables	45	9.1%
Incorrect Syllabification	156	31.6%
Foreign Origin	54	10.9%
Half Consonants	58	11.7%
Error in 'maatra'	51	10.3%
Multi-mapping	56	11.3%
Others	74	15.0%

Error Percentages in Transliteration

REFINEMENTS

Solution to *Unknown Syllables* and *Incorrect Syllabification* errors

1. Take the 1st output of the automatic syllabifier and pass it to the automatic transliterator. Store Top-6 transliteration outputs and their scores.
2. Take the 2nd output of the automatic syllabifier and pass it to the automatic transliterator. Store Top-6 transliteration outputs and scores.
3. Transliterate the name as per the baseline approach. Store Top-6 transliteration outputs of this system and the scores.
4. If the outputs of Step 1 contain English characters, then we know that the word contains unknown syllables.
    1. We then apply the same step to the outputs of Step 2.
    2. If the problem still persists, the system throws the outputs of Step 3.
    3. If the problem is resolved, but the scores of transliteration are low, it shows that the syllabification is wrong. In this case as well, we use the outputs of Step 3 only.
5. In all the other cases, we consider the best output (different from Step 1 outputs) of both Step 2 and Step 3. If we find that these best outputs have a very high score as compared to the scores of 5th and 6th outputs of Step 1, we replace the latter with these.

Final Results: Transliteration

Top-n	Correct	Correct %age	Cumulative %age
1	2,801	62.2%	62.2%
2	689	15.3%	77.6%
3	228	5.1%	82.6%
4	180	4.0%	86.6%
5	105	2.3%	89.0%
6	62	1.4%	90.3%
Below 6	435	9.7%	100.0%
	4,500		

COMPARISON

- English-Hindi Language Pair
  - *A Language-Independent Transliteration Schema Using Character Aligned Models At NEWS 2009*: English – Hindi transliteration; Using Moses; Top-1 Accuracy of 40.7%.
  - *English to Hindi Machine Transliteration System at NEWS 2009*: English – Hindi transliteration; Joint source-channel model; Top-1 Accuracy of 47.1%.
  - *English - Hindi Transliteration Using Context-Informed PB-SMT - the DCU System for NEWS 2009*: English – Hindi Transliteration; Top-1 Accuracy of 44.5%.
  - *Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009*: English – Hindi Transliteration; Top-1 Accuracy of 49.0%.
  - *Machine Transliteration using Target-Language Grapheme and Phoneme: Multi-engine Transliteration Approach*: English – Hindi Transliteration; Top-1 Accuracy of 48.1%.

COMPARISON

- Other Language Pairs
  - *Machine Transliteration of Proper Names, David Matthews*: English – Chinese forward transliteration; Moses-based; Top-5 Accuracy of 37.8%.
  - *Machine Transliteration of Proper Names, David Matthews*: Arabic – English forward transliteration; Moses-based; Top-5 Accuracy of 43.0%.
  - *Cluster Specific Named-Entity Transliteration*: English – French transliteration; Top-1 Accuracy of 56.8%; Top-5 Accuracy of 63.6%.
  - *English to Arabic Transliteration for Information Retrieval*: English – Arabic Transliteration; Top-1 Accuracy of 43.4%; Top-10 Accuracy of 81.1%.

HINDI-ENGLISH TRANSLITERATION

- Baseline Approach:

Top-n	Correct	Correct %age	Cumulative %age
1	2,719	60.4%	60.4%
2	732	16.3%	76.7%
3	278	6.2%	82.9%
4	136	3.0%	85.9%
5	95	2.1%	88.0%
6	58	1.3%	89.3%
Below 6	482	10.7%	100.0%
	4,500		

HINDI-ENGLISH TRANSLITERATION

- Syllable Approach:

Top-n	Correct	Correct %age	Cumulative %age
1	4,038	89.7%	89.7%
2	259	5.8%	95.5%
3	70	1.6%	97.0%
4	34	0.8%	97.8%
5	13	0.3%	98.1%
Below 5	86	1.9%	100.0%
	4,500		

Results of Syllabification system

Top-n	Correct	Correct %age	Cumulative %age
1	3,106	69.0%	69.0%
2	564	12.5%	81.6%
3	128	2.8%	84.4%
4	82	1.8%	86.2%
5	25	0.6%	86.8%
6	22	0.5%	87.3%
Below 6	573	12.7%	100.0%
	4,500		

Results of Syllable-based Transliteration system

HINDI-ENGLISH TRANSLITERATION

- Error Analysis:

Error Type	Number	Percentage
Unknown Syllables	43	7.5%
Incorrect Syllabification	362	63.2%
Low Probability	56	9.8%
Foreign Origin	54	9.4%
Error in full 'maatra'	32	5.6%
Others	26	4.5%

- Unknown Syllables: हेश
- Incorrect Syllabification: कोमलता => कोम ल ता
- Low Probability:
- Foreign Origin: मिक्की, डेज़ी, प्रिंस, बेबी, डॉली
- Error in full 'maatra': रूपाली => rupali, rupalee, rupaali, rupaalee, roopali, roopalee, roopaali, roopaalee

HINDI-ENGLISH TRANSLITERATION

- Refinement:
  - Take outputs of the baseline approach and compute average Moses score of all 4500 Top-1 results, do the same for Top-2, Top-3 and so on.
  - Take outputs of the syllable approach and compute average Moses score of all 4500 Top-1 results, do the same for Top-2, Top-3 and so on.
  - In each Top-1 Moses score of syllable-approach, add the difference of Top-1 averages of the 2 approaches, do the same for Top-2, Top-3 and so on.
  - Merge all the results and pick best 6.

Top-n	Correct	Correct %age	Cumulative %age
1	3,130	69.6%	69.6%
2	635	14.1%	83.7%
3	179	4.0%	87.6%
4	104	2.3%	90.0%
5	70	1.6%	91.5%
6	53	1.2%	92.7%
Below 6	329	7.3%	100.0%
	4,500		

Final results of Hindi-English transliteration system

Conclusion & Future Work

- **Conclusion**
  - We took a look at the English to Hindi transliteration problem.
  - Explored various techniques used for transliteration between other language pairs.
  - Took a look at the statistical approach of syllabification.
  - Later, we passed the output of statistical syllabifier to the transliterator and found that this *syllable*-based approach outperforms baseline approach.
- **Future Work**
  - For the completion of project, we still need to do the following:
    1. We need to carry out similar experiments for Hindi to English transliteration. This will involve statistical syllabification model and transliteration model for Hindi.
    2. We need to create a working single-click working system interface which would require CGI programming.

References

1. Nasreen AbdulJaleel and Leah S. Larkey. Statistical transliteration for English-Arabic Cross Language Information Retrieval. In *Conference on Information and Knowledge Management*, pages 139–146, 2003.
2. Ann K. Farmer Andrian Akmajian, Richard M. Demers and Robert M. Harnish. *Linguistics: An Introduction to Language and Communication*. MIT Press, 5th edition, 2001.
3. K. Knight and J. Graehl. Machine transliteration. In Computational Linguistics, pages 24(4):599–612, Dec. 1998.
4. Dan Mateescu. English Phonetics and Phonological Theory. 2003.
5. Della Pietra P. Brown and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. In Computational Linguistics, page 19(2):263Ű311, 1990.
6. P Shishtla S V Ganesh S Subramaniam and V Varma. A Language-Independent Transliteration Schema Using Character Aligned Models At NEWS 2009. In *Named Entities Workshop (NEWS) 2009*
7. A Das A Ekbal M Tapabrata and S Bandyopadhyay. English to Hindi Machine Transliteration System at NEWS 2009. In *Named Entities Workshop (NEWS) 2009*

References

8. R Haque S Dandapat A K Srivastava S K Naskar and A Way. English - Hindi Transliteration Using Context-Informed PB-SMT - the DCU System for NEWS 2009. In *Named Entities Workshop (NEWS) 2009*
9. M K Chinnakotla and Om P Damani. Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009. In *Named Entities Workshop (NEWS) 2009*
10. David Matthews. Machine Transliteration of Proper Names. Master Thesis. *University of Edinburgh*. 2007
11. Philipp Koehn. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL, 2007*
12. Jong-Hoon Oh K Uchimoto and K Torisawa. Machine Transliteration using Target-Language Grapheme and Phoneme: Multi-engine Transliteration Approach. In *Named Entities Workshop (NEWS) 2009*

Appendix

English Phonology

- **Phonology**
  - Study of the structure and systematic patterning of sounds in human language.
  - Refers to a description of the sounds of a particular language and the rules governing the distribution of these sounds.

- **English Phonology**
  - No. of speech sounds in English varies from dialect to dialect.
  - *Longman Dictionary*: 24 consonant phonemes (c.p.), 23 vowel phonemes (v.p.), additionally 2 c.p. & 4 v.p. for foreign words.
  - *American Heritage Dictionary*: 25 c.p., 18 v.p., additionally 1 c.p. & 5 v.p. for foreign words.

Consonant Phonemes

- 25 consonant phonemes found in most dialects of English.
- Categorized under **six** different categories (on the basis of their sonority level, stress, way of pronunciation etc.):

  - **Nasal**: Acoustically, nasal stops are sonorants, meaning they do not restrict the escape of air and cross-linguistically are nearly always voiced.

  - **Plosive**: Produced by stopping the airflow in the vocal tract (the cavity where sound is filtered).

  - **Affricate**: Affricate consonants begin as stops (such as /t/ or /d/) but release as a fricative (such as /s/ or /z/) rather than directly into the following vowel.

Consonant Phonemes

- **Fricative**: Produced by forcing air through a narrow channel made by placing two articulators close together. These are the lower lip against the upper teeth in the case of /f/.

- **Approximant**: In the articulation of approximants, articulatory organs produce a narrowing of the vocal tract, but leave enough space for air to flow without much audible turbulence. Examples: /l/, as in 'lip', and approximants like /j/ and /w/ in 'yes' and 'well' which correspond closely to vowels.

- **Lateral**: Laterals are "L"-like consonants pronounced with an occlusion made somewhere along the axis of the tongue, while air from the lungs escapes at one side or both sides of the tongue.

Consonant Phonemes

Nasal	/m/, /n/, /ŋ/
Plosive	/p/, /b/, /t/, /d/, /k/, /g/
Affricate	/ʧ/, /ʤ/
Fricative	/f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /h/
Approximant	/r/, /j/, /ʍ/, /w/
Lateral	/l/

Vowel Phonemes

- 20 vowel phonemes found in most dialects of English.
- Categorized under different categories (on the basis of their sonority level).

Short Monophthongs	/ɪ/, /e/, /æ/, /ɒ/, /ʌ/, /ʊ/, /ə/
Long Monophthongs	/iː/, /ɑː/, /ɔː/, /ɜː/, /uː/
Diphthongs	/eɪ/, /ɑɪ/, /ɔɪ/, /ɪə/, /eə/, /ʊə/, /əʊ/, /aʊ/

Vowel Phonemes

- **Monophthong**: “monophthongos” ≡ single note. “pure” vowel sound.
  - Articulation at both beginning and end is relatively fixed.
  - Does not glide up or down towards a new position of articulation.
  - Categorized in *Short* and *Long* vowels.
  - **Short**: Perceived for a shorter duration. For e.g., /ə/, /e/ etc.
  - **Long**: Comparatively longer duration. For e.g., /i:/, /u:/ etc.

- **Diphthong**: “two tones”. Vowel combination involving quick but smooth movement from one vowel to another.
  - Often interpreted by listeners as a single vowel sound.
  - Two target tongue positions.
  - Represented by two symbols. For e.g., /eə/

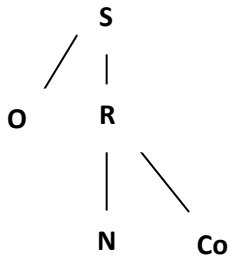
What are Syllables?

- 'Something which *syllable* has three of'.
- Flow of human voice is not a monotonous and constant one.
  - Variations in intensity, loudness, quantity etc.
  - We make a distinction between sounds that are more sonorous than others.
  - And thus, we have a consciousness of vowel structure.

- Vowels can then be said to be the "purest" sounds human beings produce when they talk.

Syllable Structure

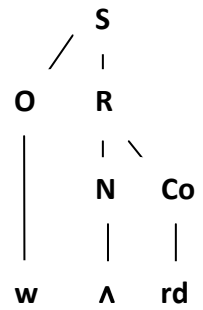
- Count of no. of syllables in a word is roughly/intuitively the no. of vocalic segments in a word.
- Thus, presence of a vowel is an obligatory element in the structure of a syllable. This vowel is called *"nucleus".*
- Basic Configuration: **(C)V(C)**.
- Part of syllable preceding the nucleus is called the *onset*.
- Elements coming after the nucleus are called the *coda*.
- Nucleus and coda together are referred to as the *rhyme*.



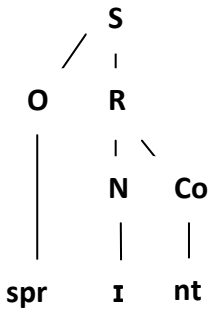
S ≡ Syllable, O ≡ Onset
R ≡ Rhyme, N ≡ Nucleus
Co ≡ Coda

Syllable Structure: Examples

- **‘word’**



- **‘sprint’**



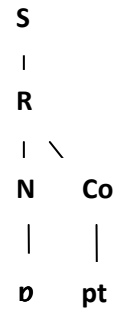
Syllable Structure: Examples

- **‘may’**



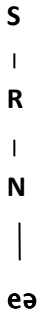
← No Coda.

- **‘opt’**



← No Onset.

- **‘air’**



← No Coda, No Onset.

Syllable Structure

- *Light Syllable*: A syllable which is open and ends in a short vowel.
  - General Description – CV.
  - Example, ‘may’.
- *Closed Heavy Syllable*: All closed syllables are heavy syllables.
  - Example, ‘opt’.
- *Open Heavy Syllable*: An open syllable with a nucleus which is long or diphthong is also a heavy syllable.
  - Example, ‘air’.

Syllabification: Determining Syllable Boundaries

- Given a string of syllables (word), what is the coda of one and the onset of another?
- In a sequence such as VCV, where V is any vowel and C is any consonant, is the medial C the coda of the first syllable (VC.V) or the onset of the second syllable (V.CV)?
- To determine the correct groupings, there are some rules, two of them being the most important and significant:
  - Maximal Onset Principle,
  - Sonority Hierarchy

Maximal Onset Principle

- The consonants that form a word-internal onset are the maximal sequence that can be found at the *beginning of words*.
- English permits only 3 consonants to form an onset.
  - Once 2nd and 3rd consonants are determined, only 1 consonant can appear in the 1st position.
  - Second = /p/, Third = /r/. Then First can only be /s/. E.g., 'spring'.

- **Working: '*constructs*'**
  - Consonant sequence: n-s-t-r
  - Either '*con structs*' OR '*cons tructs*' OR '*const ructs*' OR '*constr ucts*'.
  - As, 'str' can serve as the onset of a syllable, that's why the correct syllabification will be '*con structs*'.

Sonority Hierarchy

- **Sonority**: A perceptual property referring to the loudness of a sound relative to that of other sounds with the same length.

- **Sonority Hierarchy**: Ranking of speech sounds (or phonemes) by amplitude.
  - For e.g., if you say the vowel /e/, you will produce louder sound than if you say the plosive /t/.
  - It suggests that nuclei are the peaks of sonority and segments on either side of the peak show a decrease in sonority w.r.t. peak.

- Plosives → Affricates → Fricatives → Nasals → Laterals → Approximants → Vowels (Increasing order of sonority).

CONSTRAINTS: PHONOTACTICS

- **Phonotactics**
  - Determines possible comb. of onsets and codas which can occur.
  - Deals with restriction on the permissible comb. Of phonemes.
  - Defines permissible syllable structure, consonant clusters and vowel sequence by means of phonotactical constraints.
- In general, rules operate around the sonority hierarchy.
- Fricative /s/ is lower on the sonority hierarchy than the lateral /l/, so the combination /sl/ is permitted in onsets and /ls/ is permitted in codas. Opposite is not allowed.
- Thus, '*slips*' and '*pulse*' are possible English words.
- '*lsips*' and '*pusl*' are not possible.

Constraints on Onsets

- One-consonant: Only /ŋ/ can't be distributed in syllable-initial position.
- Two-consonant: We refer to the scale of sonority.
  - Sequence 'rn' is ruled out since there is a decrease of sonority.
  - *Minimal Sonority Distance*: Distance in sonority between the first and the second element in the onset must be of at least 2 degrees.
  - Thus, on the basis of *Sonority Hierarchy* and *Minimal Sonority Distance*, only a limited no. of possible two-consonant clusters.
- Three-consonant:
  - Restricted to licensed two-consonant onsets preceded by /s/.
  - Also, /s/ can only be followed by a voiceless sound.
  - Therefore, only /spl/, /spr/, /str/, /skr/, /spj/, /stj/, /skj/, /skw/, /skl/, /smj/ will be allowed. (splinter, spray, strong etc.)
  - While /sbl/, /sbr/, /sdr/, /sgr/, /sθr/ will be ruled out.

Constraints on Onsets

Plosive plus approximant other than /j/	/pl/, bl/, /kl/, /gl/, /pr/, /br/, /tr/, /dr/, /kr/, /gr/, /tw/, /dw/, /gw/, /kw/	play, blood, clean, glove, prize, bring, tree, drink, crowd, green, twin, dwarf, language, quick
Fricative plus approximant other than /j/	/fl/, /sl/, /fr/, /θr/, /ʃr/, /sw/, /θw/	floor, sleep, friend, three, shrimp, swing, thwart
Consonant plus /j/	/pj/, /bj/, /tj/, /dj/, /kj/, /gj/, /mj/, /nj/, /fj/, /vj/, /θj/, /sj/, /zj/, /hj/, /lj/	pure, beautiful, tube, during, cute, argue, music, new, few, view, thurifer, suit, zeus, huge, lurid
/s/ plus plosive	/sp/, /st/, /sk/	speak, stop, skill
/s/ plus nasal	/sm/, /sn/	smile, snow
/s/ plus fricative	/sf/	sphere

Possible 2-consonat clusters in an Onset

Constraints on Coda

The single consonant phonemes except /h/, /w/, /j/ and /r/ (in some cases)	
Lateral approximant + plosive: /lp/, /lb/, /lt/, /ld/, /lk/	help, bulb, belt, hold, milk
In rhotic varieties, /r/ + plosive: /rp/, /rb/, /rt/, /rd/, /rk/, /rg/	harp, orb, fort, beard, mark, morgue
Lateral approximant + fricative or affricate: /lf/, /lv/, /lθ/, /ls/, /lʃ/, /ltʃ/, /ldʒ/	golf, solve, wealth, else, Welsh, belch, indulge
In rhotic varieties, /r/ + fricative or affricate: /rf/, /rv/, /rθ/ /rs/, /rʃ/, /rtʃ/, /rdʒ/	dwarf, carve, north, force, marsh, arch, large
Lateral approximant + nasal: /lm/, /ln/	film, kiln
In rhotic varieties, /r/ + nasal or lateral: /rm/, /rn/, /rl/	arm, born, snarl
Nasal + homorganic plosive: /mp/, /nt/, /nd/, /ŋk/	jump, tent, end, pink

Constraints on Coda

Nasal + fricative or affricate: /mf/, /mθ/ in non-rhotic varieties, /nθ/, /ns/, /nz/, /ntʃ/, /ndʒ/, /ŋθ/ in some varieties	triumph, warmth, month, prince, bronze, lunch, lounge, length
Voiceless fricative + voiceless plosive: /ft/, /sp/, /st/, /sk/	left, crisp, lost, ask
Two voiceless fricatives: /fθ/	fifth
Two voiceless plosives: /pt/, /kt/	opt, act
Plosive + voiceless fricative: /pθ/, /ps/, /tθ/, /ts/, /dθ/, /dz/, /ks/	depth, lapse, eighth, klutz, width, adze, box
Lateral approximant + two consonants: /lpt/, /lfθ/, /lts/, /lst/, /lkt/, /lks/	sculpt, twelfth, waltz, whilst, mulct, calx
In rhotic varieties, /r/ + two consonants: /rmθ/, /rpt/, /rps/, /rts/, /rst/, /rkt/	warmth, excerpt, corpse, quartz, horst, infarct
Nasal + homorganic plosive + plosive or fricative: /mpt/, /mps/, /ndθ/, /ŋkt/, /ŋks/, /ŋkθ/ in some varieties	prompt, glimpse, thousandth, distinct, jinx, length
Three obstruents: /ksθ/, /kst/	sixth, next

Other Constraints

- Nucleus: The following can occur as nucleus:
  - All vowel sounds (monophthongs as well as diphthongs).
  - /m/, /n/ and /l/ in certain situations (for example, '*bottom*', '*apple*')

- Syllabic:
  - Both the onset and the coda are optional (as seen previously).
  - /j/ at the end of an onset (/pj/, /bj/, /tj/, /dj/, /kj/, /fj/, /vj/, /θj/, /sj/, /zj/, /hj/, /mj/, /nj/, /lj/, /spj/, /stj/, /skj/) must be followed by /ʊɪ/ or /ʊə/.
  - Long vowels and diphthongs are not followed by /ŋ/.
  - /ʊ/ is rare in syllable-initial position.
  - Stop + /w/ before /ʊɪ, ʊ, ʌ, aʊ/ are excluded.