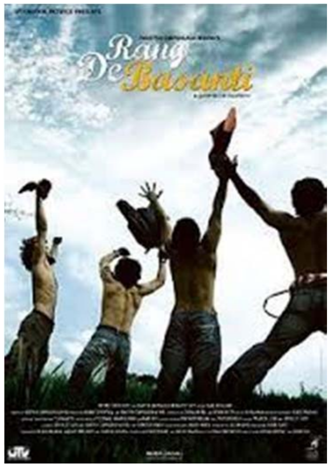
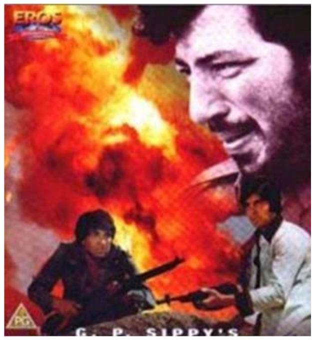
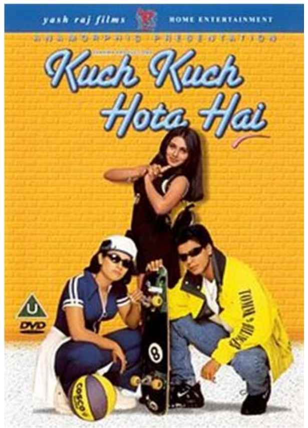
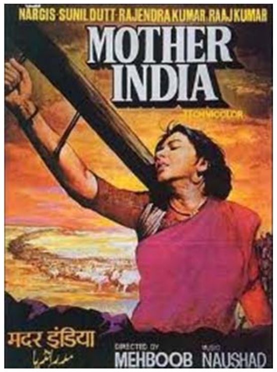
Evaluating the Effect of Phrase Set in Hindi Text Entry







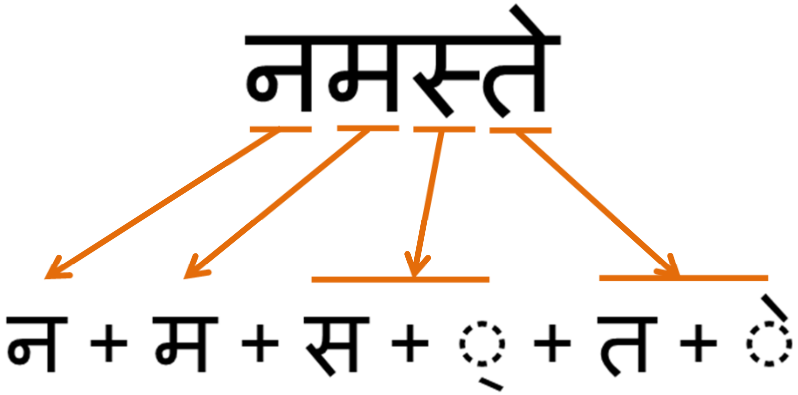


Mohit Jain
IBM Research India

Namaste

नमस्ते

Namaste



नमस्ते

न + म + स + ् + त + े

About Hindi

Disconnected

requires two or more letters to be combined together to form a *character*

क(k) + ई(i) = की(ki)

About Hindi

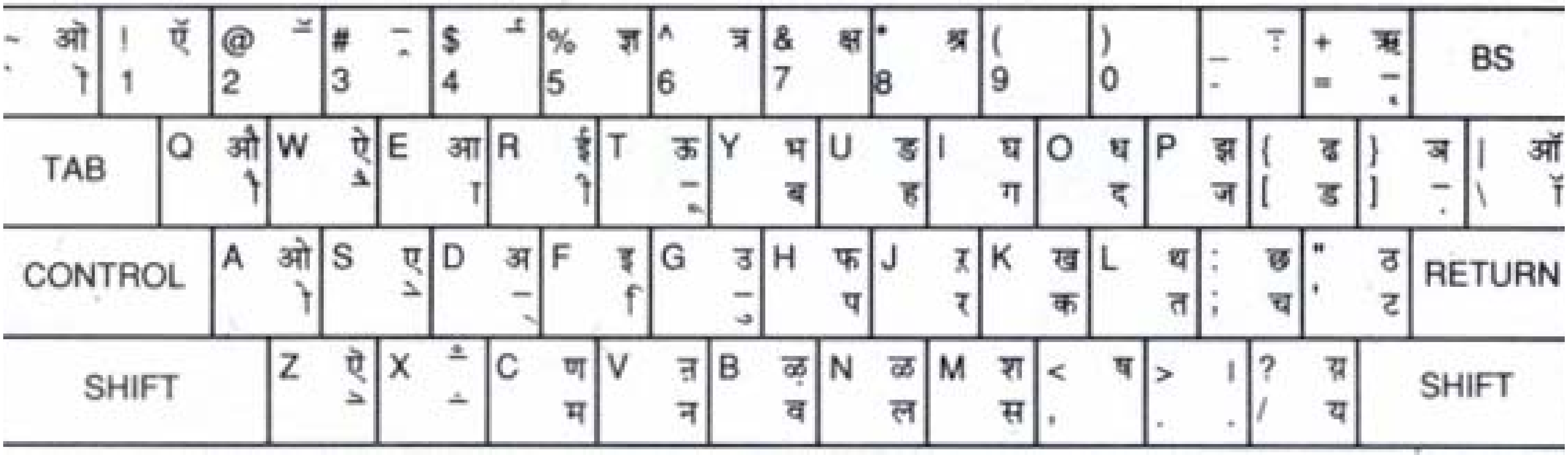
Disconnected

requires two or more letters to be combined together to form a *character*

क(k) + ई(i) = की(ki)

Complex

53 base letters – 34 consonants, 11 vowels and 8 diacritic marks



About Hindi

Disconnected

requires two or more letters to be combined together to form a *character*

क(k) + ई(i) = की(ki)

Complex

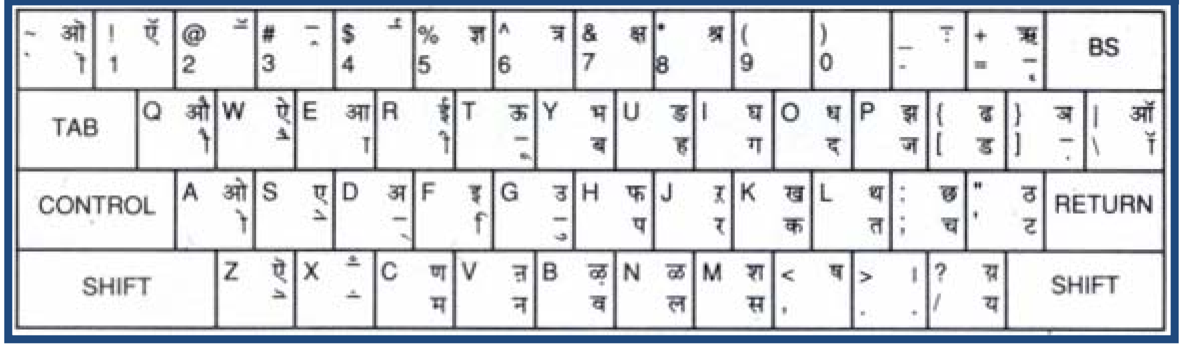
53 base letters – 34 consonants, 11 vowels and 8 diacritic marks

Phonetic vs Visual Sequence

differences between the phonetic sequence of letters and the visual sequence of writing the letters

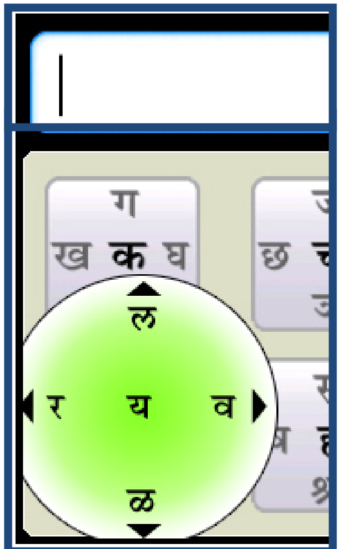
प(p) + इ(i) = पि(pi)

Keyboard



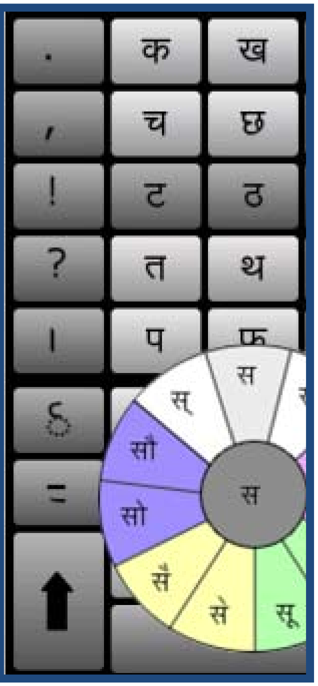
Inscript (Indian Script)

Keylekh

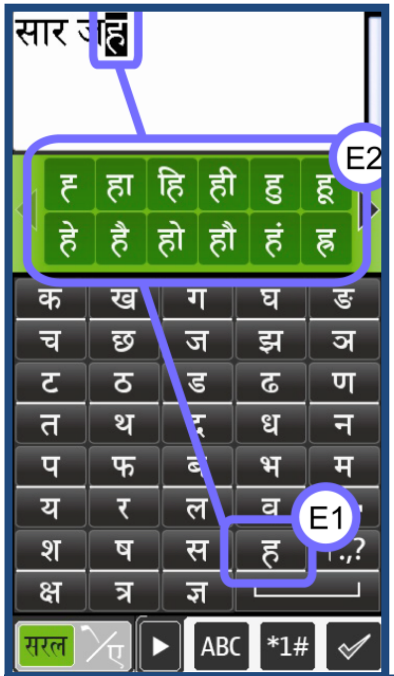


DISHA

Gesture Keyboard



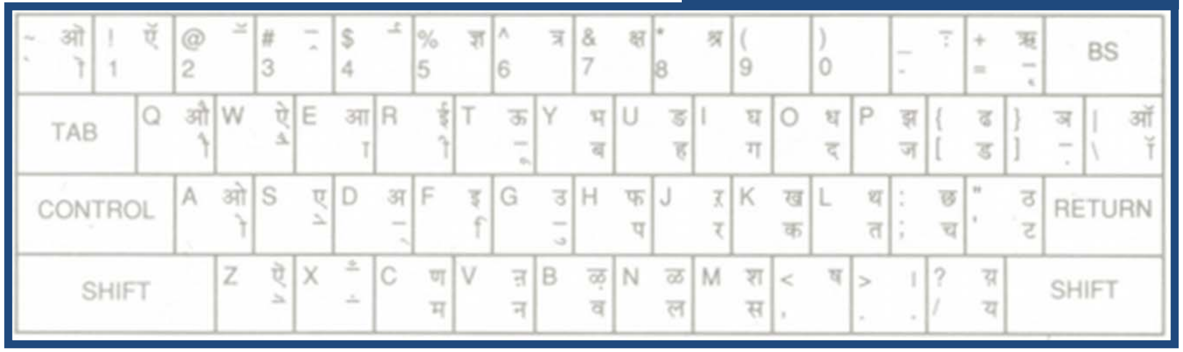
Swarachakra



Adaptive Keyboard

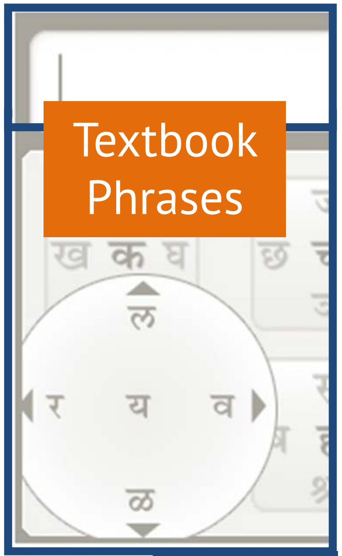
Problem

Inscript (Indian Script)



Keylekh

Random Phrases

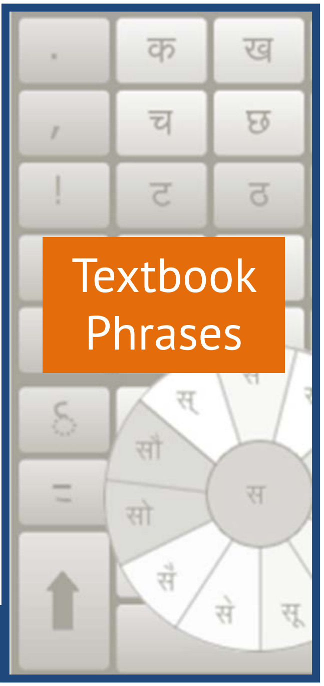


Textbook Phrases

DISHA

News Phrases

Gesture Keyboard



Textbook Phrases

Swarachakra

Hindi Films Phrases

Adaptive Keyboard

Research Objective

Develop and evaluate three different types of Hindi phrase sets

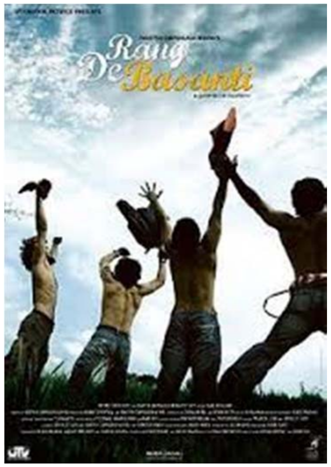
Films

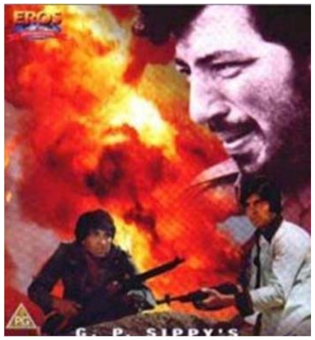
Textbooks

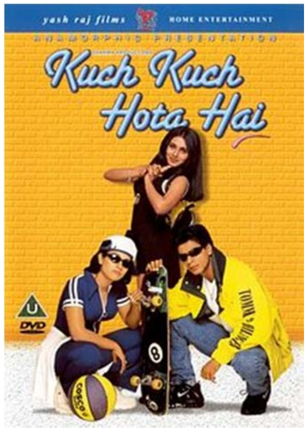
Translated MacKenzie and Soukoreff's Phrases

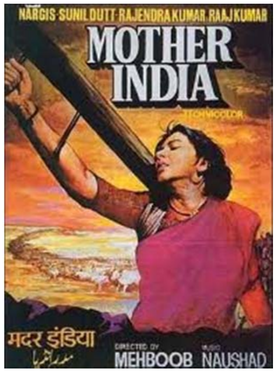
to study effects of their characteristics on performance

Evaluating the Effect of Phrase Set in Hindi Text Entry



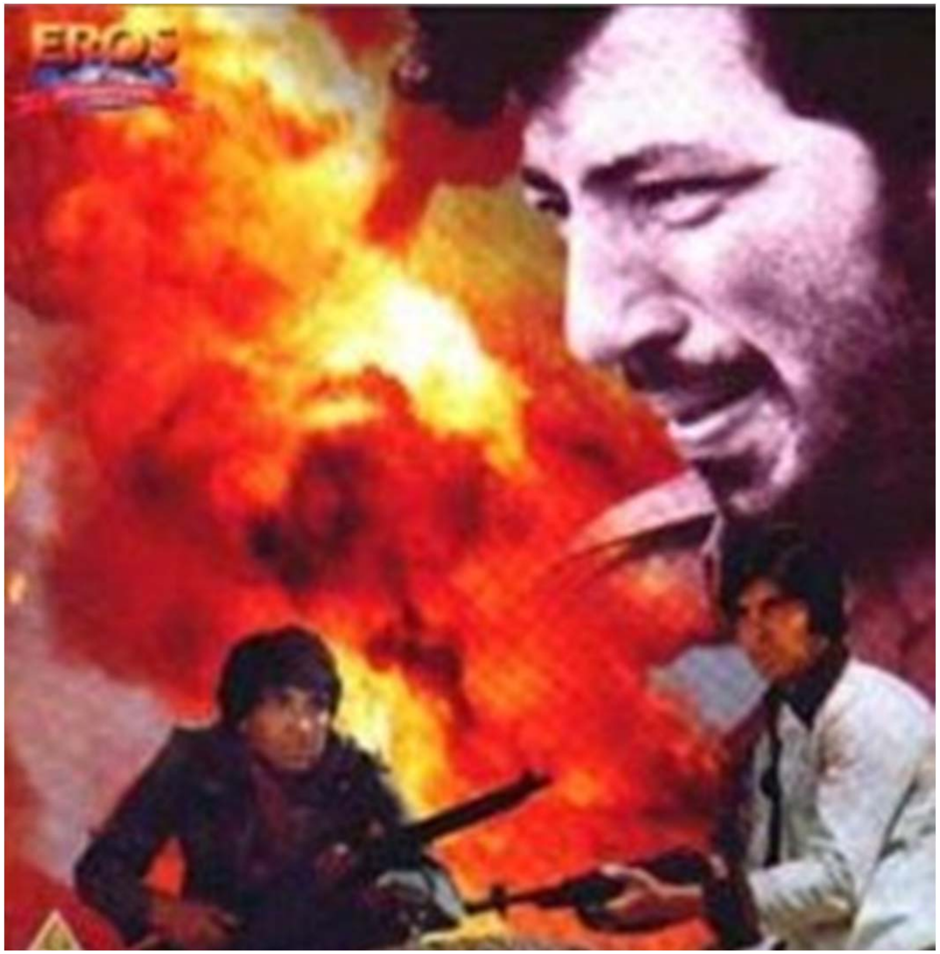






Mohit Jain
IBM Research India

Phrases from Films (FP)



Phrases from Films (FP)

Process

Randomly selected from online forums and blogs

Benefit

Very familiar

Example

डान को पकड़ पाना मुश्किल ही नहीं नामुमकिन है

(to catch Don is not only hard but impossible)

Phrases from Textbooks (TP)

Process

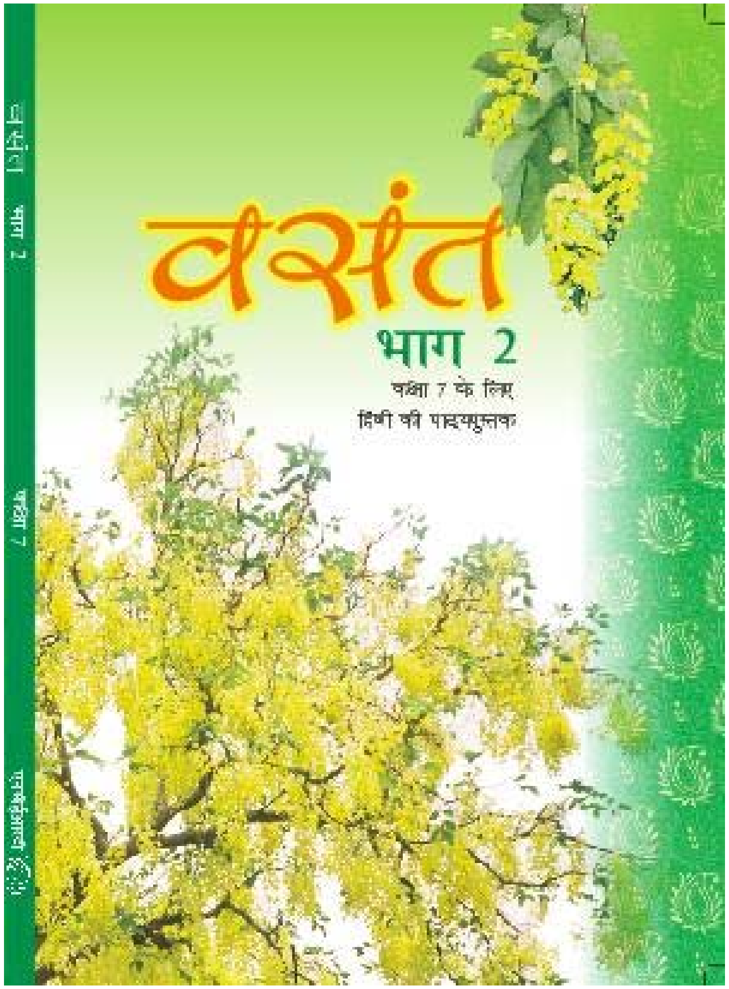
Randomly selected from Grade VII Hindi textbook

Benefit

Topical relationship between consecutive phrases

Example

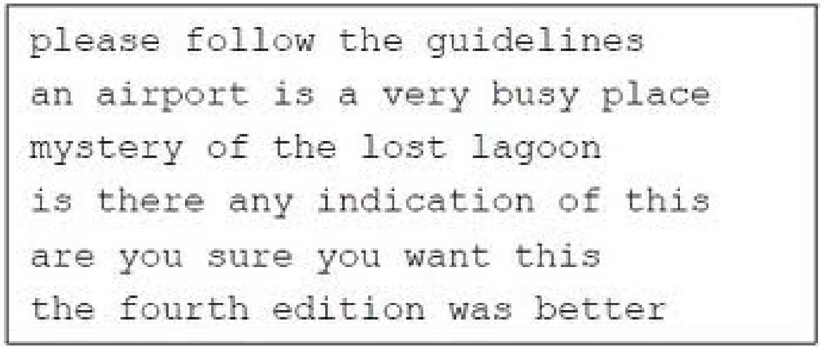
दिव्या अनिल कि छोटी बहन है
(Divya is Anil's younger sister)



Translated MacKenzie & Soukoreff's Phrase Set (MSP)

Process

Translated the phrase set into Hindi using context-appropriate words



please follow the guidelines
an airport is a very busy place
mystery of the lost lagoon
is there any indication of this
are you sure you want this
the fourth edition was better

Benefits

Standard

Used extensively for evaluation

Example

प्यार के कई मतलब है

(*love means many things*)

Linguistic Analysis

Metrics	EMILLE/CIIL Corpus	FP	TP	MSP	MS English Set
Number of phrases/sentences	737528	60	50	150	500
Number of words	12295677	490	673	881	2712
Number of unique words	202042	267	382	464	1163
Minimum word length	2	2	2	2	1
Maximum word length	33	10	13	14	13
Min. phrase length (# words)	1	4	3	3	3
Max. phrase length (# words)	888	14	39	11	9
Min. phrase length (# letters)	1	16	10	12	16
Max. phrase length (# letters)	4752	58	167	49	43
Single-letter correlation	-	0.97	0.98	0.98	0.95
Word-based correlation	-	0.70	0.68	0.75	0.85
Readability	m=10.34 sd=6.76	m=5.36 sd=2.4	m=8.0 sd=3.82	m=5.68 sd=2.46	m=4.17 sd=3.88
Words per phrase	m=16.67 sd=13.27	m=8.16 sd=2.4	m=13.46 sd=7.45	m=5.87 sd=1.6	m=5.4 sd=1.1
Letters per phrase	m=83.34 sd=67.4	m=35.45 sd=10.15	m=61.44 sd=34.63	m=26.82 sd=7.08	m=28.61 sd=5.02
Letters per word	m=4.06 sd=2.16	m=3.46 sd=1.44	m=3.63 sd=1.65	m=3.73 sd=1.72	m=4.46 sd=2.4

Linguistic Analysis

Metrics	MS English Set	EMILLE Hindi Corpus	FP	TP	MSP
Single-letter correlation	0.95	-	0.97	0.98	0.98
Word-based correlation	0.85	-	**0.70**	**0.68**	**0.75**
Readability	4.17	10.34	**5.36**	**8.0**	**5.68**
Words per phrase	5.4	16.67	**8.16**	**13.46**	**5.87**

Hypothesis

H1 Use of MSP and FP will result in faster text entry and a lower error rate than TP

Reason MSP and FP have lower readability and lower words per phrase

Hypothesis

H1 Use of MSP and FP will result in faster text entry and a lower error rate than TP

Reason MSP and FP have lower readability, higher word correlation, and lower words per phrase

H2 MSP will be preferred over FP and TP

Reason MSP's high word-based correlation to the corpus

Demographics

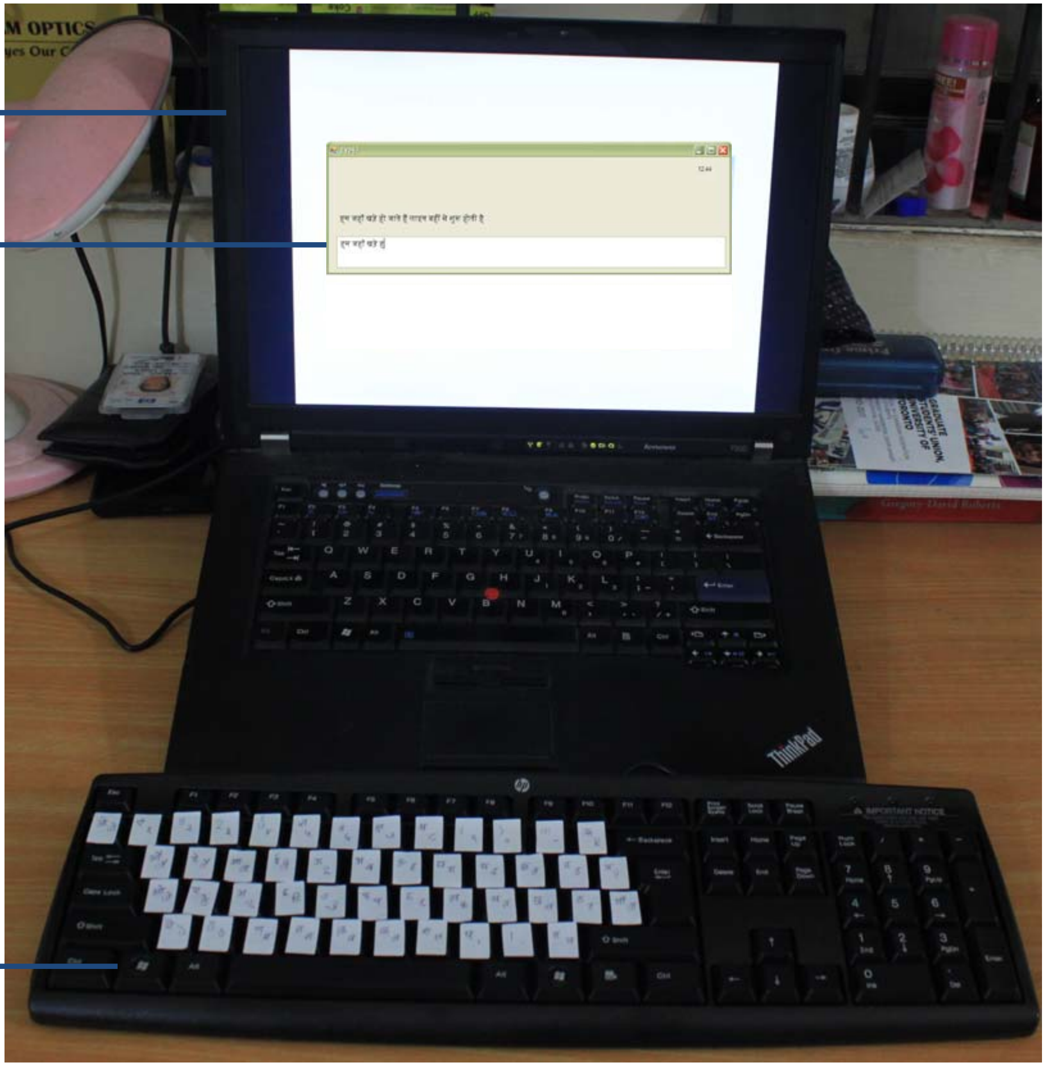
18 participants (12 males, 4 females, mean age=21.8)

Criteria: Must know how to read, write, and speak in Hindi, but have never used an Inscript (Indian Script) keyboard before

All undergraduate Computer Science students (average 10.16 years with QWERTY)

Paid Rs 100 (~$2) per session; Prize money of Rs 1000 and Rs 500 for the two fastest

Apparatus



15.4 inches laptop screen

Custom software in C# (test phrase at the top of the screen and participant typing the same phrase into a text box below it)

Inscript keyboard

Procedure

Within-subject three 45-min session study

A session consisted of two 20-minute typing blocks with a break of 3-5 minutes between the blocks

Asked to enter text as quickly and as accurately as possible

Ordering of the phrase sets was counterbalanced

After each session, participants were required to rate the phrase set in terms of memorability, understandability, phrase length, and frequency of usage on a 5-point Likert scale

Results: Speed

Words per minute (wpm): (letters per second)*60/5, with the definition that a word consists of 5 letters

Results: Accuracy

Keystrokes per Letter (KSPL): Number of keystrokes required to input a letter in Hindi

Minimum String Distance (MSD): between the presented and transcribed phrase

KSPL measures the corrected errors as every correction adds multiple keystrokes, while MSD accounts for the uncorrected errors in the final transcribed text

Note: For Hindi, ideal KSPL for Inscript keyboard is 1.12

Results: Speed & Accuracy

	FP	TP	MSP	H1

Speed (wpm)	m=6.22
	sd=2.16
		m=7.28
		sd=2.62
			m=7.22
			sd=2.48
				$F_{2,34}$=2.5
				p=0.1

Accuracy (KSPL)	m=1.41
	sd=0.13
		m=1.40
		sd=0.1
			m=1.43
			sd=0.22
				$F_{2,34}$=1.3
				p=0.3

Accuracy (MSD)	m=0.028
	sd=0.01
		m=0.046
		sd=0.03
			m=0.03
			sd=0.01
				$F_{2,34}$=2.4
				p=0.1

Results: Questionnaire

Participant preferred MSP

(because it was short, easy to understand

and memorable phrases)

Friedman $\chi^2(2)=14.7$, $p<0.01$ ***(H2)***

Results: Questionnaire

Understandability

FP (m=4.6, sd=0.8) > TP (m=3.1, sd=0.9) **$p<0.0001$**

MSP (m=4.17, sd=1) > TP **$p=0.01$**

Length

Phrases from TP were too long, whereas phrases from FP were just right, thus ~8 words per phrase seems acceptable

Memorability

"Phrases should be interesting, so that we enjoy typing." – FP

FP (m=4.2, sd=0.2) > TP (m=2.5, sd=0.1) **$p<0.0001$**

MSP (m=3.9, sd=0.2) > TP **$p<0.0001$**

Limitations & Future Work

Limited demography (only undergraduate students)

- **Study w/ wider demography; demography-based phrases?**

Only three sessions long study

- **Longitudinal study is needed to show that there is perhaps no significant difference between any sets of phrases**

Only studied on one type of keyboard

- **Results might differ for other input method**

Conclusion

Three phrase sets – FP, TP and MSP, with different linguistics characteristics

No performance difference, but MSP most preferable

Readability, memorability and phrase length should be considered

In future, use our phrase sets for more consistency across studies, to produce generalizable results

Thank you!

http://www.dgp.toronto.edu/~mjain/HindiTextEntry.zip

Mohit Jain **IBM Research, India**

mohitjain@in.ibm.com

Khushboo Tekchandani **DA-IICT, India**

Khai N. Truong **Univ of Toronto, Canada**