South Asian Languages

K. V. S. Prasad (Chalmers University)

Suma Bhat (University of Illinois)

History in GF

- The work of Shafqat Virk, starting from an earlier morphology by Muhammad Humayoun.
  - Urdu
  - Punjabi
  - Persian
  - Sindhi
  - Nepali
  - Hindi
- We will return to these works, but first a general introduction

“South Asia” !

- Present day India, Pakistan, Bangladesh, Sri Lanka, Afghanistan, Nepal, Bhutan, Maldives, …
- A quarter of humanity
- Historically, mostly “India” (the land beyond the Sindhu = “Indus”, as called by the Greeks).
  - “India”, “Indian”, “Hindi”, “Hindu”, “Hind” all similar
    - Until quite recently, some of these terms meant little in India, a universe unto itself
    - In law, “Hindu” = Indian not self-identifying as Muslim, Christian, Buddhist, etc.
  - But the Greeks were not too far wrong
    - there was a shared culture, carried mostly by Sanskrit

The language families: 1, Indo-Aryan

- Indo-Aryan (Indo-Iranian)
  - Nepali, Bengali, Assamese, Oriya, Konkani, Marathi, Gujarati, Sindhi, Marwari, Punjabi, Kashmiri, Dogri
  - Hindi/Urdu
    - Braj, Awadhi, Maithili, Chattisgarhi, Haryanvi, Mewati, Bundeli, Kannauji, Bhojpuri, … all loosely called “Hindi”
      - Many of these are seen by their speakers as local languages
        - » They use Hindi for education and official use
      - even Punjabi follows this pattern, to some extent
    - Bombay and Kolkata have Hindi pidgins
    - Dialects of Urdu: Hyderabad (Dakhani) and Bangalore.

The language families: 2, Dravidian

- In South India, many dialects of each
  - Telugu
  - Tamil
  - Kannada
  - Malayalam
  - Tulu
- in Baluchistan
  - Brahui (4 m speakers)

Other families

- Tibeto-Burman
  - Bodo, Manipuri
- Sino-Tibetan
  - Kokborok
- Munda (AustroAsiatic)
  - Santhali
- Extended Iranian
  - Pashto, Balochi, Dari

In this talk, only the major families

- Indo-Aryan
  - Sanskrit
  - Hindi/Urdu
- Dravidian
  - Telugu
  - Kannada
  - Tamil

Sounds and scripts

- All Indic scripts derive from Brahmi, an “abugida” or “alphasyllabary”.
  - A “letter” is most often a CV
    - Progressively less often a CCV, V, CCCV, or C.
  - The Urdu script is an alphabet, Perso-Arabic, not Indic.
- The order of presentation
  - V, Ca, CV, CCa, CCCa
  - Formalised at least by Panini’s time
  - Still used to teach all Indian children.
- It is simplest to begin with the Unicode (in roman) for the Devanagari script used for Sanskrit, and add the few letters needed for the other language unicodes.

Extended Sanskrit vowels

a a: i i: u u: r. r.: l. l.: e e: e+ o o: o+ m. h.

Capitals A, A:, … mean V, whereas a, a: etc. mean the V in a CV, CCV or CCCV.

The two short vowels e and o are needed for the Dravidian languages. Indeed Telugu has yet another, more open, e vowel, but that is not represented in the script, so we ignore it.

Extended Sanskrit consonants

	V-A-	V-A+	V+A-	V+A+	N	
Velar	k	k’	g	g’	n-	q
Palatal	c	c’	j	j’	n*	
Retroflex	T	T’	D	D’	N	
Dental	t	t’	d	d’	n	
Labial	p	p’	b	b’	m	

Continuants y r l v L r+ Spirants s* S s h

Fricatives f z x G Z

Non-Sanskrit consonants in Urdu

- The uvular stop q and the fricatives f, z, x and G, all sounds from Persian or Arabic.
  - Many Indians and some Pakistanis too replace these by k, p’, j, k’ and g, respectively, but we need them for spelling.
  - Indeed, we need z1..z4 and h1..h4 etc., since multiple Arabic sounds are collapsed into z or h in Urdu, but we ignore that in this talk.
  - In the South, q is consistently pronounced k’

Non-Sanskrit consonants in Dravidian

- L (retroflex continuant) in all Dravidian languages and Marathi, Z (retroflex approximant) and r+ in Tamil.
- The aspirates are typically only needed for Sanskrit borrowings.
- Pure Tamil does not even need to indicate voicing – this is allophonic variation, voiceless word-initially and voiced intervocalically – but there are enough borrowings in modern usage that all stops are shown e.g. in classical lyrics.

Suggestion – use Roman internally

- A good way to see common patterns and to exploit large shared vocabulary.
- Even printing out finally in Roman has its uses
  - Otherwise smart people don’t see that a script is easy to learn (except Urdu), so they miss out on texts they might enjoy.

Readability

- My proposal is like the phonetic and popular standard roman for Indian languages; also popular typescript.
  - These show the sound, not the script.
  - The popular typscript for a: is A
    - but we use A for V, and a and a: for the V in a CV, CCV or CCCV.
- They show vowels as they sound. We can’t.
  - All Indic scripts show consonants with a built-in “a”. So “pa” and “p” both show “pa”. To get “p”, we have to remove the built-in vowel.
    - In GF, we have to write “pa_” or some such.
    - Sadly, “prasa:d” has to be “pa_rasa:d” internally.
    - But this can be filtered for printout.

Grammar common to Indic languages

- All are SOV
- Free phrase order, where a NP includes a case–marker or postposition
- Eight cases in Sanskrit
  - Most modern Indian languages have two or three genuine cases, but postpositions to cover the eight of Skt.

Nouns in Hindi/Urdu

- Morphology – see Humayoun’s presentation
  - Nouns – {Number (Sg|Pl) => Case => Str; Gender}
    - Three cases: Nom, Obl, Voc
    - But see postpositions and “cases” in Shafqat p 25
    - Two genders, Masc|Fem, mostly grammatical
  - They found 15 paradigms
- Note that their romanisation is aimed only at Urdu spelling, not the sound
  - Sound -> Urdu, or Devanagari -> Urdu
    - Has been studied, see Coling 2012

Nouns in Telugu

Case = Nom|Acc |Inst|Dat|Abl|Gen|Loc|Voc ;

-- Nom, Gen, Voc will do, but 8 to pivot from Skt

Gender = Masc | Fem | Neut ;

-- logical gender

Number = Sg | Pl ;

So far, only the logical gender differs from Hin/Urd

Telugu Noun classes

- Classified by plural formation
  - These rules involve internal sandhi
    - E.g., ra:muDu + lu -> ra:muLLu
  - And sometimes vowel harmony
    - pilli + lu -> pillulu
  - By supplying the plural explicitly for now, these issues can be postponed
  - Most of the time, we need just the nominative and genitive, in singular and plural
    - Later on, we might be able to guess all these from the nominative singular for many nouns.

Telugu “Declension”

```
DeclTable : Type = Number => Case => Str;
N : Type = {s : DeclTable ; g : Gender} ;
```

This is mostly a matter of ending a postposition (or suffix) (a “case-ending”) to the genitive.

(Note that in Hin/Urd too, postpositions other than the case-markers take the genitive: e.g., andar, ni:ce, u:par, pi:ce, pa:s)

The rarely used vocative case prevents us from saying this is all there is.

Vocative messes up endings

```
Postpos : Type = {so: StemObl; ce :Str};

postpos :  Number -> Case -> Postpos =
     \num, c -> case <num, c> of
     {<num, Nom> => {so = Stem; ce = ""};
      <num, Gen> => {so = Obl; ce = ""};
      <Sg, Voc> => {so = Stem; ce = ":"};
      <Pl, Voc> => {so = Obl; ce = ":ara:"};
      <num, c>   => {so = Obl; ce = caseendings c}
      };
```

The case endings for nouns

```
caseendings: Case -> Str =
    \c -> case c of
      {Nom => "";
     Acc => "ni";
     Inst => "ceta";
     Dat => "ki";
     Abl => "num.Di";
     Gen => "";
       Loc =>  "lo:" ;
     Voc => ""
    };
```

Pronouns show there is more

- I, mine, we, our = ne:nu, na:, me:mu, ma: BUT
  - Acc na:ni -> nannu,   ma: + ni -> mammalni
    - From an old genitive, mammula
  - Dat na:ki -> na:ku   *(Bangalore Telugu na:ki)*
- You, your, pl = nuvvu, ni:, mi:ru, mi:
  - Acc ni:ni -> ninnu,   mi:ni -> mimmalni
- The ending vowels i or u often disappear in connected speech (external sandhi).
- Also, in many cases, either will do, at the cost of sounding old-fashioned.

Hin/Urd: Verbs

- Classified by: do intr., tr., and causatives exist?
  - banna:, bana:na:, banva:na: (become, make, get someone to make), or even
  - kaTna:, ka:Tna:, kaTa:na:, kaTva:na: (be cut, cut, get cut, have someone cut)
- Most Dravidian transitive verbs X can take a morpheme (to get someone to do X), so this classification is irrelevant.

Hin/Urd verbs

- Conjugated by
  - person, number, gender ("agr")
  - Tense
- The agr endings are just that.
  - Indeed in Urdu orthography, the "copula" ga:, gi:, ge: has to be written as a separate word (it is not in Hindi).
- The tense can be analysed (?) as a marker added to the root
  - The API does not fit the tense system of Hin/Urd

A first analysis of Telugu verbs

```
ConjTable: Type = PolTense => Agr => Str;

  conjtablefn : VStemsStr -> PolTense -> VClass -> Agr -> Str =
   \vss,poltense,vc,agr ->
    let stem = vss.pt ! poltense;
        tense = poltense.t
    in
    stem + (tensesuffix stem tense) ! agr + persuffix vc agr;

  mkConjtable: VStemsStr -> VClass -> ConjTable =
   \vss, vc ->
     table {poltense => table {agr => conjtablefn vss poltense vc agr}
 };
```

Personal suffixes: tense independent

```
persuffix : VClass -> Agr -> Str =  \vc,agr -> case agr of
    {Ag n p g => case n of
     {Sg=>case p of
      {Per1=> "nu”; Per2=> "vu";
       Per3=>case g of
          {Masc=>"Du";
           _ => case vc of {Un => "di";
                      Le => "du”} } };
     Pl=>case p of
      {Per1=> "mu”; Per2=> "ru";
       Per3=>case g of
         {Neut=> case vc of {Un => "yi";
                        Le => "vu"};
          _ => "ru”}  } }};
```

only in Per3 does the verb class “is/is not” complicate matters

Tense suffixes

```
tensesuffix: Str -> Tense -> (Agr => Str) =
\s, t -> case t of
       {Perf => case s of
            {"an" | "un" | "tin" | "vin" | "kon" | "kan" |
"ku:rcun" | "nilcun"
            => table {Ag Sg Per3 (Fem|Neut)  => "na”;
                                                     _ => "na:"};
            _ => table {Ag Sg Per3 (Fem|Neut)  => "im.”;
                                                    _ => "a:”} };
        Pres => table{_ => ”uta:”} };
```

But the root changes too for tense

- Speak, utter, articulate = paluku
- palukuta:nu, etc. for present tense
- palika:nu, etc. for past tense
- So the information flow is a bit back-and-forth
  - Go back and see the first analysis of verbs

Agglutination

- Dravidian languages are acknowledged to be agglutinative, as the analysis also shows
- I think that some features of Hindi are best understood as fossilised agglutination, baked into “words”
- Typically, both external and internal sandhi play a role in Dravidian languages
  - Where to break utterances into words is a bit arbitrary

Example: external and internal sandhi, and agglutination

- paDagoTTincuko:dalucukunnavaTagada: ?

paDu	to fall
koTTu	to hit => to demolish, by deriv. morph.
incu	causative morpheme
ko:	self-benefit morpheme
dalucu	to think, to decide or determine
kuni	”having done/acquired” morpheme
unna:	to be, present tense
vu	Per2 (fam., informal) morpheme
Ta	“hearsay” morpheme
ka:du	to not be so (to be false), present tense
a:	yes/no question morpheme

- “I hear you have decided to get (it) demolished (for your benefit), is that true?”

To be or not to be

- Dravidians never say “no”
  - A delightful people, true, but not because they accept everything you propose
    - Negation is a conjugation
    - The form depends on the verb implicit in the question
  - true = avunu (lit. it will become)
  - false = ka:du (lit. it will not become)
  - is there = um.di
  - is not there = le:du

Positive tenses Kan/Tel

- Present or future habitual ma:Dutte:ne ce:sta:nu I do
- Past ma:Didenu ce:sa:nu I did
- Future ma:Duvenu ce:sta:nu I will do
- Contingent ma:Diye:nu ce:ddunu I would do
- Imperative ma:Du ceyyi do!
- Hortative ma:Do:Na ce:dda:m. Let us do
- Pres. cont. mADuttidde:ne ce:stunna:nu I am doing
- Durative stative only for gottu/baruttade/sa:ku/ telusu/vaccu/caalu/erugu knows German, can swim
- Polite Imp ma:Duviri am.te ce:dduru ga:ni (then) you can do

More positive verb forms

- Pres part ma:Dutta: ce:stu (while) doing
- Conditional ma:Didare ce:ste If I do
- Past verbal adjective ma:Dida ce:sina (which was) done   krdanta
- Durative Verbal adjective ma:Duttidda ce:stunna (which is) doing
- Past participle ma:Di ce:si having done
- Infinitive ma:Dalu ceyya to do
- Verbal noun ma:Duvudu ceyyyaTam. (the) doing
- Future habitual verbal adjective ma:Duva ce:se: (that which) will do

Negative tenses and forms

- Negative ma:Denu ceyyanu I don’t/won’t do
- Neg Imp ma:Dabe:Da ceyyaku don’t do!
- Neg. Participle ma:Dade ceyyaka not having done
- Negative verbal adjective ma:Dada ceyyani (that which) did not do
- Contrafactual ma:Duttidde ce:se:va:Dini I would have done
- Concessive ma:Didaru: ce:sina: even though I did

Dravidian functor

- After we have individually done at least parts of Tel/Kan/Tam

- Ergativity in Hin (see Shafqat)
- Conjunction and or in Tel/Kan
- Prefixing not productive in Dravidian except mun, hin, etc. – viharincu (borrowing baked form)

Adj -> noun -> Adj in Tel/Kan

i: (this)	a: (that)	e: (which)
idi (this thing)	adi	e:di
ivi (these things)	avi	e:vi
vi:Du (this boy)	va:Du	evaDu
vi:ru (these people)	va:ru	*ve:ru (evaru)

i:	a:	ya:va
idu	adu	ya:vadu
ivaru	avaru	*evaru (ya:ru)
ivaLu (this girl)	avaLu	*evaLu (ya:vaLu)

These nouns then give adjectives via genitive

di:ni	da:ni	de:ni	
vi:Ti	va:Ti	*ve:Ti	
vi:Di	va:Di	*ve:Di	evaDi
Vi:ri	va:ri	*ve:ri	evari

idara	adara	ya:vadara?	
ivara	avara	*evaru	(ya:ra)
ivaLa (this girl)	avaLa	*evaLu	(ya:vaLa)

Works for Hindi too

i	u	ka	ja (rel.)
yah	vah	*kah (ko+n)	*jah jo:
yaha:~	vaha:~	kaha:~	jaha:~
id’ar	ud’ar	*kad’ar kid’ar	jid’ar
itna:			
e+sa:			
ab	*ub tab	kab	jab

Higher structures: Reduplication

- “Why are you walking slowly-slowly?”
- “Wherever-wherever there are cows therever-therever there will be calves”
- “Give everyone two-two apples”
- sam.b’ava:mi yuge: yuge: (from the b’agavadgi:ta, “I will be born in age-age”)
- bande: barta:ne
- The function is often described in computational ling. circles as intensification, we think it is firstly distribution.

Echoing

- k’a:na: is “food”, often “meal”, but k’a:na: va:na: is “food appropriate now”, probably not a full meal
- tim.Di is “food”, but tim.Di:tippalu: is “food and other such necessities”, usually followed by le:ka “without”
- Or as in Yiddish -> Amer.Eng., dismissive
  - “fancy-schmancy” (I feel silly in such a place)
  - “baby-schmaby” (he’s already five years old)

Dropped elements

- Pronouns - As in Italian, since the verb conjugates fully for person and number at least.
- Copula in Dravidian languages
  - adi illu ”that (is) a house”
  - anaganaga: oka ra:ju ”once upon a time (there was) a king”
- But also other verbs. Guess the verb!
  - nanage b’aya, nanage sa:ru, u:Takke sa:ru

  to me fear (happen), to me soup (want), to dinner soup (is)

Vocabulary and culture

- Kinship terms, food and drink, greetings and phatic communication
  - Cumulatively problematic enough – do we need a new flavour of API?

Diglossia

- Dravidian languages are all seriously multi-glossic, though efforts have been made over the last 100 years to write as one speaks.
  - How much of this can be captured in GF?

One grammar, diverging lexicons: the case of Hindi/Urdu

K. V. S. Prasad and Shafqat Virk
Dept of Computer Science
Chalmers Univ and Univ of Göteborg
Sweden

Sources of examples; Acknowledgements

- NCERT and Tamilnadu Govt. textbooks
- Phrasebook
  - ”Controlled Language for Everyday Use”
  - Aarne Ranta, Ramona Enache, Gregoire Detrez
- MGL
  - ”The GF Mathematics Library”
  - Jordi Saludes and Sebastian Xambo
- We acknowledge the Molto project

  molto-project.eu

  for support. The systems reported above have been developed further under Molto.

A taste of GF and Hindustani

- Abstract sentence:
  - PQuestion (HowFarFrom
    (ThePlace Station)(ThePlace Airport))
- Concrete English sentence
  - How far is the airport from the station?
- Concrete Hindustani sentence
  - *sTeshan se hava:i: aDDa: kitni: du:r hae?*
  - स्टेशन से हवाईअड्डाकितनीदूर है? اسٹیشن
    سے ہوائی اڈا کتنی دور ہے؟
  - Hindustani word order *station from air port how-much far is?*

One language or two? Scholarly views.

- One language, two scripts
  - Tarachand and many others, on Hindustani (pre 1947)
  - Murli Manohar Joshi (2012)

- Two languages
  - Gopi Chand Narang (2004)
    - Hindi and Urdu share the same Indic base, but
    - In practice and usage are different languages
    - Urdu lexicon extensively from Arabic and Persian
    - Hindi from Sanskrit
  - C. M. Naim (1999, but reprinted from late 60’s)
    - Does not help to learn Urdu and Hindi together
    - Has translated books from Hindi to Urdu

Airport announcements: situation tailored for Bazaar Hindustani?

- “Passengers are requested ...”

- H: *yātriyõ se nivedan hæ ...* यात्रयिों से नविदन है

- U: *musāfirõ se guzāriš kī jātī hæ ...* مسافروں سے گزارش کی جاتی ہے

- One grammar, differing content words.
  - Green – identical grammar (function words)
  - Blue – function words common to H/U
    - different syntax because of lexical choices

- Hindi/Urdu greetings differ. ”Culture”. Here?

A School Geometry Theorem

If a perpendicular is drawn from the vertex of a right angled triangle to its hypotenuse, then the triangles on each side of the perpendicular are similar to the whole triangle.

The same theorem in Hindi and Urdu

- Hindi: *yadi kisī samkoṇ tribhuj ke samkoṇ vāle šīrṣ se karṇ par lañb ḍālā jāe to is lañb ke donõ or bane tribhuj sañpūrṇ tribhuj ke samrūp hote hæ̃.*
- Urdu: *agar kisī qāyam zāvī mašallaš ke qāyam zāviyā vāle rās se ek amūd us ke vitar par ḍālā jāe to amūd ke donõ taraf banne vāle mašallaš asal mašallaš ke or āpas mẽ mušābā hõge.*
- Green – identical grammar (function words)
- Blue – function words common to H/U; different syntax because of lexical choices
- Red – stylistic choice of function words, unique to one of H or U (or becoming so)
- Black – content words; unintelligible in both directions.

School Mathematics Lexicon

- 260 entries. Hindi and Urdu differ on 245.
  - The overlapping 15 include function words used in a technical mathematical sense, “such that”, “where”…
  - Content word examples:
    perpendicular (lamb लंब, amu:d عمود),
    right-angled (samkoN समकोण, qa:yam za:vi: قائم زاوی),
    triangle (tribhuj त्रिभुज, mashallash مثلث), hypotenuse (karN कर्ण , vitar وتر), vertex (shi:rS शीर्ष, ra:s راس)
- Why so different?
  - Different resource languages to borrow from.
    - Hindi: only Sanskrit, Urdu only Persian/Arabic.

Computationally confirmed
Same grammar

- Made parallel computational grammars
  - Took our Urdu grammar
  - Changed Urdu script to Hindi, lexicon too as needed
- Tested both grammars
  - Translated English sentences to Hindi /Urdu
  - Showed the results to native Hindi/Urdu speaking informants
- Results. Each of 80 English sentences was either
  - Correctly translated to both Hindi and Urdu (45 cases)
  - Or attracted the same corrections from both Hindi and Urdu informants (35 cases)
- We ourselves checked (without external informants) 500 more sentences, and found the grammars agreed on all.

What we found: Diverging Lexicons

- Base lexicon of 350 words, 18% different
- Phrasebook for tourists
  - 112 words including 42 English borrowings
    - 50 native words same, 20 different
    - Even days of the week are different, except Monday
  - 22 greetings phrases, 17 different
    - Include ”Good Morning” etc, not natural in Hin/Urd
    - Hard to know what to pick for Hindi
    - Used government sources, dictionaries if no such

Hindi/Urdu lexical divergence is not a matter of register

- A layperson hearing a technical talk
  - ”Of the snarfs, a boojum is made”
    - Gets the grammar, but not the content
  - But if they understand one talk in English on mathematics
    - they will understand another on the same topic at the same level!

- A Hindi speaker who understands school math and airports
  - Will fail to understand the same topic in Urdu and vice-versa
  - Urdu sounds like technical talk to a Hindi speaker and vice-versa

- But they are not hearing technical talk!
  - Takes great skill and calibrated use of the ”other” words to communicate across this gulf – listen to Shamim Hanfi’s talk at the Hindi-Urdu flagship

Lexical divergence is unusual: Related languages usually converge

- Example: Telugu and Kannada
  - Closely related Dravidian (South Indian) languages

Telugu/Kannada, mutually foreign: isomorphic grammar, different lexicon

- ”Will you stay home tomorrow, or are you thinking you must go to work?”
- T: *re:pu mi:ru iNTlo:ne: uNTa:ra: le:ka paniki veLLa:lanukoNTunna:ra:*
- K: *na:Le ni:vu maneyalle: irti:ra: athava: kelasakke ho:gabe:kanukoNDutiddi:ra:?*
- Common to Telugu and Kannada
  - Morpheme order
    - tomorrow you at-home-(only) will-be? or to-work must-go-having-said-(to-self)-are?
  - Emphasis *e:* yes/no? *a:* having-said-(to-self) *anukoNTu*
- One-to-one correspondences of function words
- Different base lexicon: ’home’, ’work’

Converging lexicons (to Sanskrit) of Telugu, Kannada *and Hindi*

- Telugu: *oka lam.bako:Na trib'ujamulo: lam.bako:Namu gala šīrṣamununñḍi karṇamunaku lañbamu gīcina, ā lañbamunaku iruvaipula gala tribhujamulu pūrṇa tribhujamunaku sarūpamuganuñḍunu*
- Kannada: *oñdu lañbakōna tribhujada oñdu šruñgadiñda adara vikarNakke lañbarēkheyannu eLedare, lañbarēkheya prati kaDeyalli iruva tribhujagaLu pūrNa trib'ujakke samarūpavāgiruttave*
- Hindi: *yadi kisī samkoṇ tribhuj ke samkoṇ vāle šīrṣ se karṇ par lañb ḍālā jāe to is lañb ke donõ or bane tribhuj sañpūrṇ tribhuj ke samrūp hote hæ̃*

Converging lexicons to Sanskrit, contd.

- Red: Identical content words
- Blue: Closely related content words
  - Would benefit by Istilah style standardisation
- the phonology and morphology around the borrowed words differ (of course)

How unnatural is Sanskritised Hindi?

- Quite, to native speakers of Hindustani
  - Though they are getting used to airports and news
    - What else can the airport say?
  - The Hindustani speaking child asks
    - What does ”a:ka:sh” mean (Sanskritised Hindi for ’sky’)
    - The answer may be ”a:sma:n” (Urdu from Persian)
  - But for the non-Hindi speaking child struggling with their compulsory Hindi, ”a:sma:n” is foreign, and ”a:ka:sh” familiar.

Compared to Bahasa Indonesia

- Bahasa has technical terms invented and standardised to compensate for their absence
- Urdu has rich lexical resources (Perso-Arabic)
- Hindi has them, if it taps Sanskrit
  - Makes terms understandable in other languages too
  - Need Istilah committee to standardise these terms across India? (similar=samaru:pa in Skt/Hin, but saru:pa in Tel.)
  - Would ease translation of technical texts across India

Conclusions

- Apart from Hindustani base, Urdu and Hindi are becoming mutually unintelligible
- Given socio-politico-cultural factors, a guess is they will continue to diverge
  - Especially if they try to develop science terms
- Future work: pursue preliminary experiments translating arbitrary English texts, using a ”robust parser” (Angelov)

History of Hindi/Hindustani/Urdu

- Spoken language
  - Organic growth, from 12c
    - Local grammar and base lexicon. For administrative terms, borrowings from invader languages.
  - Went by various names including ’Hindi’
    - By 19c and 20c, called ”Hindustani”
  - Yamuna Kachru:
    - Hindustani has no status now in India or Pakistan
    - For attitudinal reasons

As Literary Languages

- Literary Urdu
  - Dakhani from 16c
    - Delhi then still Persian for literary use, ’Hindi’ spoken
  - Delhi Urdu from 18c
    - Resource languages largely Persian/Arabic/Turkish
    - No form problem – apart from code switching English/Urdu
- Modern Hindi from 19c
  - Reaction to rise of English/fall of Urdu/decline of Braj
    - (Braj was then active literary dialect related to Hindustani)
  - Mostly Sanskrit resource
    - Particularly recently, in official Hindi, and textbooks
  - Massive form problem
    - Hindustani? (Urdu? Persian?) Sanskrit? English?

As National Languages: Urdu

- Administrative Urdu in Delhi – from 1830 (1857?)
- Urdu in Pakistan
  - Generally accepted
    - though only 10% population are native speakers
  - Language of education and public discourse
  - Popular and prestigious as literary language
  - Associated with Islam and Pakistan Movement
- Religion link: false but persistent
  - Millions of Muslims in South Asia do not speak Urdu
  - Millions of Urdu speakers are not Muslims
    - though many of these may now claim to be Hindi speakers.

As National Language: Hindi

- Hindi in India
  - Native language of ~40%
    - depending on how you count
  - Where successful, mostly by organic growth
    - commercial (films/songs, advertisements, business)
  - Unclear effects of enormous push/imposition of Hindi
    - Regional radio/TV channels now challenge the previous domination of Hindi via central broadcasts
  - Largely failed as a language of science or higher education
    - No other Indian language has really succeeded either
  - Other Indian languages have millennia of literary history
    - Hindi can only claim less, that too via Urdu, Braj, Maithili, etc.
- Recent development – all want English medium