Computational evidence that Hindi and Urdu share a grammar but not the lexicon

K. V. S. Prasad[1] and Shafqat Mumtaz Virk[2]
(1) Department of Computer Science, Chalmers University, Sweden
(2) Department of Computer Science and Engineering,University of Gothenburg, Sweden
and Department of Computer Science and Engineering UET, Lahore
prasad@chalmers.se, virk.shafqat@gmail.com

Abstract

Hindi and Urdu share a grammar and a basic vocabulary, but are often mutually unintelligible because they use different words in higher registers and sometimes even in quite ordinary situations. We report computational translation evidence of this unusual relationship (it differs from the usual pattern, that related languages share the advanced vocabulary and differ in the basics). We took a GF resource grammar for Urdu and adapted it mechanically for Hindi, changing essentially only the script (Urdu is written in Perso-Arabic, and Hindi in Devanagari) and the lexicon where needed. In evaluation, the Urdu grammar and its Hindi twin either both correctly translated an English sentence, or failed in exactly the same grammatical way, thus confirming computationally that Hindi andUrdu share a grammar. But the evaluation also found that the Hindi and Urdu lexicons differed in 18% of the basic words, in 31% of tourist phrases, and in 92% of school mathematics terms.

Keywords: Grammatical Framework, Resource Grammars, Application Grammars.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 1–14,
COLING 2012, Mumbai, December 2012.

1 Background facts about Hindi and Urdu

Hindi is the national language of India and Urdu that of Pakistan, though neither is the native language of a majority in its country.

'Hindi' is a very loose term covering widely varying dialects. In this wide sense, Hindi has 422 million speakers according to (Census-India, 2001). This census also gives the number of native speakers of 'Standard Hindi' as 258 million. Official Hindi now tends to be Sanskritised, but Hindi has borrowed from both Sanskrit and Perso-Arabic, giving it multiple forms, and making Standard Hindi hard to define. To complete the 'national language' picture, note that Hindi is not understood in several parts of India (Agnihotri, 2007), and that it competes with English as lingua franca.

It is easier, for several reasons, to talk of standard Urdu, given as the native language of 51 million in India by (Census-India, 2001), and as that of 10 million in Pakistan by (Census-Pakistan, 1998). Urdu has always drawn its advanced vocabulary only from Perso-Arabic, and does not have the same form problem as Hindi. It is the official language and lingua franca of Pakistan, a nation now of 180 million, though we note that Urdu's domination too is contested, indeed resented in parts of the country (Sarwat, 2006).

Hindi and Urdu 'share the same grammar and most of the basic vocabulary of everyday speech' (Flagship, 2012). This common base is recognized, and known variously as 'Hindustani' or 'Bazaar language' (Chand, 1944; Naim, 1999). But, 'for attitudinal reasons, it has not been given any status in Indian or Pakistani society' (Kachru 2006). Hindi-Urdu is the fourth or fifth largest language in the world (after English, Mandarin, Spanish and perhaps Arabic), and is widely spoken by the South Asian diaspora in North America, Europe and South Africa.

1.1 History: Hindustani, Urdu, Hindi

From the 14th century on, a language known as Hindustani developed by assimilating into Khari Boli, a dialect of the Delhi region, some of the Perso-Arabic vocabulary of invaders. Urdu evolved from Hindustani by further copious borrowing from Persian and some Arabic, and is written using the Perso-Arabic alphabet. It dates from the late 18th century. Hindi, from the late 19th century, also evolved from Hindustani, but by borrowing from Sanskrit. It is written in a variant of the Devanagari script used for Sanskrit.

But the Hindi/Urdu has base retained its character: 'the common spoken variety of both Hindi and Urdu is close to Hindustani, i.e., devoid of heavy borrowings from either Sanskrit or Perso-Arabic' (Kachru, 2006).

1.2 One language or two?

Hindi and Urdu are 'one language, two scripts', according to a slogan over the newspaper article (Joshi, 2012). The lexicons show that neither Hindi nor Urdu satisfies that slogan. Hindustani does, by definition, but is limited to the shared part of the divergent lexicons of Hindi and Urdu.

(Flagship, 2012) recognizes greater divergence: it says Hindi and Urdu 'have developed as two separate languages in terms of script, higher vocabulary, and cultural ambiance'. Gopi Chand Narang, in his preface to (Schmidt, 2004) stresses the lexical aspect: 'both

Hindi and Urdu share the same Indic base ... but at the lexical level they have borrowed so extensively from different sources (Urdu from Arabic and Persian, and Hindi from Sanskrit) that in actual practice and usage each has developed into an individual language'.

But lexical differences are not quite the whole story. (Naim, 1999) lists several subtle morphological differences between Hindi and Urdu, and some quite marked phonological ones. Most Hindi speakers cannot pronounce the Urdu sounds that occur in Perso-Arabic loan words: q (unvoiced uvular plosive), x (unvoiced velar fricative), G (voiced velar fricative), and some final consonant clusters, while Urdu speakers replace the ṇ (retroflex nasal) of Hindi by n, and have trouble with many Hindi consonant clusters.

Naim does not think it helps learners to begin with Hindi and Urdu together. Those who seek a command of the written language, he says, might as well learn the conventions exclusive to Urdu from the beginning.

Thus there are many learned and differing views on whether Hindi and Urdu are one or two languages, but nothing has been computationally proved, to the best of our knowledge. Our work demonstrates computationally that Hindi and Urdu share a grammar, but that the lexicons diverge hugely beyond the basic and general registers. Our as yet first experiments already give preliminary estimates to questions like 'How much do Hindi and Urdu differ in the lexicons?'.

Overview Section 2 describes Grammatical Framework, the tool used in this experiment, and Section 3 lists what we report. Section 4 describes the Hindi and Urdu resource grammars, some differences between them, and how we cope with these differences. Section 5 presents the general and domain-specific lexicons used in this experiment. Evaluation results are given at the ends of Sections 4 and 5. Section 6 provides context and wraps up.

This paper uses an IPA style alphabet, with the usual values and conventions. Retroflexed sounds are written with a dot under the letter; ṭ, ḍ, and ṛ (a flap) are common to Hindi and Urdu, while ṇ and ṣ occur in Sanskritised Hindi (though many dialects pronounce them n and š). The palatalised spirant š and aspirated stops, shown thus: k^h, are common to Hindi and Urdu. A macron over a vowel denotes a long vowel, and ˜, nasalisation. In Hindi and Urdu, e and o are always long, so the macron is dropped. Finally, we use ñ to mean the nasal homorganic with the following consonant.

2 Background: Grammatical Framework (GF)

GF (Ranta, 2004) is a grammar formalisim tool based on Martin Löf's (Martin-Löf, 1982) type theory. It has been used to develop multilingual grammars that can be used for translation. These translations are not usually for arbitrary sentences, but for those restricted to a specific domain, such as tourist phrases or school mathematics.

2.1 Resource and Application Grammars in GF

The sublanguages of English or Hindi, say, that deal with these specific domains are described respectively by the (English or Hindi) *application grammars* Phrasebook (Caprotti et al 2010, (Ranta et al., 2012) and MGL (Saludes and Xambó, 2010). But the English Phrasebook and English MGL share the underlying English (similarly for Hindi). The underlying English (or Hindi) syntax, morphology, predication, modification, quantification, etc., are captured in a common general-purpose module called a *resource grammar*.

Resource grammars are therefore provided as software libraries, and there are currently resource grammars for more than twenty five languages in the GF resource grammar library (Ranta, 2009). Developing a resource grammar requires both GF expertise and knowledge of the language. Application grammars require domain expertise, but are free of the general complexities of formulating things in English or Hindi. One might say that the resource grammar describes how to speak the language, while the application grammar describes what there is to say in the particular application domain.

2.2 Abstract and Concrete Syntax

Every GF grammar has two levels: abstract syntax and concrete syntax. Here is an example from Phrasebook.

1. Abstract sentence:
   `PQuestion (HowFarFrom (ThePlace Station)(ThePlace Airport))`
2. `Concrete English sentence: How far is the airport from the station?`
3. `Concrete Hindustani sentence: sṭešan se havāī aḍḍā kitnī dūr hæ?`
   (स्टेशन से हवाई अड्डा कितनी दूर है? , اسٹیشن سے ہوائی اڈا کتنی دور ہے؟)
4. `Hindustani word order: station from air port how-much far is?`

The abstract sentence is a tree built using functions applied to elements. These elements are built from categories such as questions, places, and distances. The concrete syntax for Hindi, say, defines a mapping from the abstract syntax to the textual representation in Hindi. That is, a concrete syntax gives rules to linearize the trees of the abstract syntax.

Examples from MGL would have different abstract functions and elements. In general, the abstract syntax specifies what categories and functions are available, thus giving language independent semantic constructions.

Separating the tree building rules (abstract syntax) from the linearization rules (concrete syntax) makes it possible to have multiple concrete syntaxes for one abstract. This makes it possible to parse text in one language and output it in any of the other languages.

Compare the above tree with the resource grammar abstract tree for "How far is the airport from the station?" to see the difference between resource and application grammars:

```
PhrUtt NoPConj (UttQS (UseQCl (TTAnt TPres ASimul) PPos (QuestIComp (CompIAdv
(AdvIAdv how_IAdv far_Adv))(DetCN (DetQuant DefArt NumSg) (AdvCN (UseN
airport_N)(PrepNP from_Prep (DetCN(DetQuant DefArt NumSg)(UseNstation_N))
))))))NoVoc
```

3 What we did: build a Hindi GF grammar, compare Hindi/Urdu

We first developed a new grammar for Hindi in the Grammatical Framework (GF) (Ranta, 2011) using an already existing Urdu resource grammar (Virk et al., 2010). This new Hindi resource grammar is thus the first thing we report, though it is not in itself the focus of this paper.

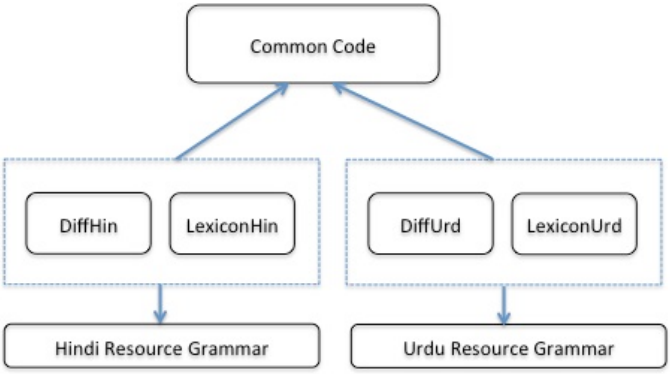

Figure 1: Hindi/Urdu Functor.

We used a functor style implementation to develop Hindi and Urdu resource grammars, which makes it possible to share commonalities between two grammars. Figure 1 gives a picture of this implementation style. Most of the syntactic code resides in the 'common code box', and the minor syntactical differences (discussed in Section 4) are placed in each of the 'DiffLang box'. Each resource grammar has its own lexicon. This mechanically proves that Hindi and Urdu share a grammar and differ almost only in the lexicons.

We evaluated our claim by (1) porting two application grammars to Hindi and Urdu: a Phrasebook of tourist sentences (Ranta et al., 2012), and MGL, a mathematical grammar library for school mathematics (Caprotti and Saludes, 2012), (2) randomly producing 80 abstract trees (40 from each of the Phrasebook, and MGL), (3) linearizing them to both Hindi and Urdu, and finally checking them either for correctness, or badness (see Section 6 for results).

4 Differences between Hindi and Urdu in the Resource Grammars

We started from the script based GF resource grammar for Urdu, and adapted it for Hindi almost entirely just by re-coding from Urdu to Hindi script. A basic test vocabulary accompanies the resource grammars, and this was changed as needed: it turned out that Hindi and Urdu differ up to 18% even in this basic vocabulary. Section 5 deals with the application lexicons.

We do not give any implementation details of these resource grammars in this paper, as the interesting bits have already been explained in (Virk et al., 2010). But we describe below resource level differences between Hindi/Urdu, and strategies to deal with them.

4.1 Morphology

Every GF resource grammar provides a basic test lexicon of 450 words, for which the morphology is programmed by special functions called lexical paradigms. Our Hindi morphology simply takes the existing Urdu morphology and re-codes it for the Devanagari script. Lexical differences mean that the morphologies are not identical; e.g., Hindi some-

times uses a simple word where Urdu has a compound word, or vice-versa. But there are no patterns that occur in only one of the languages, so the test lexicon for Hindi works with few problems.

We could in principle implement the subtle morphological differences noted in (Naim, 1999), but we ignored them. That these differences are minor is shown by the fact that our informants find the resulting Hindi entirely normal.

4.2 Internal Representation: Sound or Script?

The translation of "How far is the airport from the station?" was written in IPA, representing the sound of the Hindi/Urdu. It sounds identical in the two languages, and thus we could label it 'Hindustani'. An obvious approach to writing grammars for Hindi/Urdu from scratch would be to represent the languages internally by sound, so that we would get just one grammar, one shared lexicon, and differentiated lexicons only for those words that sound different in Hindi and Urdu. For output, we would then map the IPA to the Hindi or Urdu script.

But we were starting from (Virk et al., 2010), which uses an internal representation based on written Urdu. It would be a fair sized task to re-do this in terms of speech, though the result would then be immediately re-usable for Hindi and might also help capture similarities to other South Asian languages. We reserve this re-modelling for future work.

So, in the present work, we changed the Urdu grammar to a Hindi grammar merely by replacing written Urdu by written Hindi. This script change was also done for the basic lexicon, though here some words were indeed different even spoken. Our parallel grammars therefore give no indication that Hindi and Urdu often sound identical.

One compensating advantage is that script-based representations avoid spelling problems. Hindi-Urdu collapses several sound distinctions in Persian, Arabic and Sanskrit. A phonetic transcription would not show these collapsed distinctions, but the orthography does, because Urdu faithfully retains the spelling of the original Perso-Arabic words while representing Sanskrit words phonetically, while Hindi does the reverse. Each language is faithful to the sources that use the same script. We see that it will not be entirely trivial to mechanically go from a phonetic representation to a written one.

Obviously, the more the Hindi and Urdu lexicons overlap, the more the wasted effort in the parallel method. But as we shall see, the lexicons deviate from each other quite a bit. We have designed an augmented phonetic representation that keeps track of spelling, for use in a remodelled grammar.

4.3 Idiomatic, Gender and Orthographic Differences

In addition to spelling, Hindi and Urdu also have orthographic differences, not often remarked. Indeed some apparently grammatical differences result from in fact idiomatic, gender or orthographic differences.

For example, the lexicon might translate the verb "to add" as "`joṛnā`" in Hindi, and as "`jame karnā`" in Urdu. The imperative sentence "add 2 to 3" would then be rendered "`do ko tīn se joṛo`" in Hindi, and "`do ko tīn mẽ jame karo`" in Urdu. But the choice between the post-positions "`se`" and "`mẽ`" is determined not by different grammars

for Hindi and Urdu, but by the post-positional idiom of the chosen verb, "joṛnā" or "jame karnā", as can be seen because either sentence works in either language.

A gender difference appears with "war", rendered in Urdu as "laṛāī" (fem.). This word works in Hindi as well, but has more a connotation of "battle", so we chose instead "saṅgʰarṣ" (masc.). The shift from feminine to masculine is driven by the choice of word, not language.

Orthographic differences next. "He will go" is "vo jāegā" in both languages; in writing, (वह जाएगा, وہ جائے گا), the final "gā" (गा, گا) is written as a separate word in Urdu but not in Hindi. Similarly, "we drank tea" is "hamne cāy pī" in both languages, but in writing, (हमने चाय पी, ہم نے چائے پی), the particle "ne" (ने, نے) is written as a separate word in Urdu but not in Hindi.

These differences were handled by a small variant in the code, shown below. To generate the future tense for Urdu, the predicate is broken into two parts: finite (fin) and infinite (inf). The inf part stores the actual verb phrase (here "jāe"), and the fin part stores the copula "gā" as shown below.

```
VPFut=>fin=(vp.s!  VPTense VPFutr agr).fin; inf=(vp.s!  VPTense
VPFutr agr).inf
```

For Hindi, these two parts are glued to each other to make them one word. This word is then stored in the inf part of the predicate and the fin part is left blank as shown below.

```
VPFut=>fin=[]; inf=Prelude.glue ((vp.s! VPTense VPFutr agr).inf)
((vp.s! VPTense VPFutr agr).fin)
```

Similarly in the ergative "hamne cāy pī" ("we drank tea"), Urdu treats "ham" and "ne" as separate words, while Hindi makes them one. We used for Urdu, `NPErg => ppf ! Obl ++ "ne"` and for Hindi, `NPErg => glue (ppf ! Obl) "ne"`.

4.4 Evaluation and Results

With external informants

As described earlier, we randomly generated 80 abstract trees (40 from each of the Phrasebook, and MGL), linearized them to both Hindi and Urdu. These linearizations were then given to three independent informants.

They evaluated the Hindi and Urdu translations generated by our grammars. The informants found 45 sentences to be correct in both Hindi and Urdu. The other sentences were found understandable but failed grammatically - in exactly the same way in both Hindi and Urdu: nothing the informants reported could be traced to a grammatical difference between Hindi and Urdu. For this paper, the point is that all 80 sentences, the badly translated as well as the correctly translated, offer mechanical confirmation that Hindi and Urdu share a grammar.

We note for the record that the 35 grammatical failures give a wrong impression that the grammar is only "45/80" correct. In fact the grammar is much better: there are only a

handful of distinct known constructs that need to be fixed, such as placement of negation and question words, but these turn up repeatedly in the evaluation sentences.

A result that has not been the focus of this paper is that we greatly improved the Urdu grammar of (Virk et al., 2010) while developing the Hindi variant. Errors remain, as noted above.

With internal informants

The second author is a native Urdu speaker, while the first speaks Hindi, though not as a native. With ourselves as internal informants, we could rapidly conduct several more extensive informal evaluations. We looked at 300 Phrasebook sentences, 100 MGL sentences, and 100 sentences generated directly from the resource grammars. We can confirm that for all of these 500 English sentences, the corresponding Urdu and Hindi translations were understandable and in conformance with Urdu and Hindi grammar (barring the known errors noted by the external informants).

We note particularly that randomly generated MGL sentences can be extremely involuted, and that the Hindi and Urdu translations had the same structure in every case.

5 The Lexicons

As we noted in Section 1, Urdu has a standard form, but Hindi does not, though official Hindi increasingly tends to a Sanskritised form. Hindustani itself counts as 'Hindi', and is a neutral form, but has only basic vocabulary, a complaint already made in (Chand, 1944). So to go beyond this, Hindi speakers have to choose between one of the higher forms. Elementary mathematics, for example, can be done in Hindustani or in Sanskritised Hindi, attested by the NCERT books (NCERT, 2012), or in English-ised Hindi, which can be heard at any high school or university in the Hindi speaking regions.

We arbitrated the choice of Hindi words thus: when we had sources, such as the NCERT mathematics books or a government phrase book, we used those. Otherwise, we used (Snell and Weightman, 2003) and (Hindi-WordNet, 2012) to pick the most popular choices.

5.1 The general lexicon

Out of 350 entries, our Hindi and Urdu lexicons use the same word in 287 entries, a fraction of 6/7 which can easily be changed by accepting more Urdu words as Hindi' or by avoiding them. We note in passing that the general lexicon is any case often tricky to translate to Hindi-Urdu, as the cultural ambience is different from the European one where GF started, and which the test lexicon reflects. Many words ("cousin", "wine", etc.) have no satisfactory single equivalents, but these lexical items still help to check that the grammars work.

5.2 The Phrasebook lexicon

This lexicon has 134 entries, split into 112 words and 22 greetings. For 92 of the words, the Hindi and Urdu entries are the same; these include 42 borrowings from English for names for currencies, (European) countries and nationalities,and words like "tram" and "bus". So Hindi and Urdu share 50 of 70 native words, but differ on 20, including days of the week (except Monday, "`somvār`" in both Hindi and Urdu). The greetings lexicon has 22 entries,

most of which are hard to translate. “Good morning” etc. can be translated though they are often just “hello” and “bye”. Greetings are clearly more culture dependent: Hindi and Urdu differ in 17 places.

An example not in the Phrasebook drives home the point about greetings: airport announcements beginning “Passengers are requested ...” are rendered in Hindi as “yātriyõ se nivedan hæ ...” (यात्रियों से निवेदान है) and in Urdu as “musāfirõ se guza:riš kī jātī hæ ...” (مسافروں سے گزارش کی جاتی ہے), which suggests that Hindi and Urdu have diverged even in situations almost tailored for ‘Bazaar Hindustani’!

5.3 The Mathematics lexicon

Our MGL lexicon, for use with high school mathematics, has 260 entries. Hindi and Urdu differ on 245 of these. The overlapping 15 include function words used in a technical mathematical sense, “such that”, “where”, and so on.

As examples of the others, here are some English words with their Hindi and Urdu equivalents in parentheses: perpendicular (lañb लंब, amūd عمود), right-angled (samkoṇ समकोण, qāyam zāvī قائم زاوی), triangle (trib^{h}uj त्रिभुज, mašallaš مثلث), hypotenuse (karṇ कर्ण , vitar وتر), vertex (šīrṣ शीर्ष, rās راس).

This total divergence comes about because Urdu borrows mathematical terms only from Perso-Arabic, and Hindi, only from Sanskrit. There would be more overlap in primary school, where Hindi uses more Hindustani words, but the divergence is already complete by Class 6. The parallel English, Hindi and Urdu texts (NCERT, 2012), from which we got the list above, show that the grammar of the Hindi and Urdu sentences continue to be identical modulo lexical changes, even when the lexicons themselves diverge totally.

Since it often happens in mathematics that every Hindi content word is different from its Urdu counterpart, the languages are mutually unintelligible. Even function words can differ. Either “yadi” or “agar” can mean “if” in Hindi, but the Sanskrit “yadi” is often chosen for reasons of stylistic unity with the Sanskrit vocabulary. Urdu never uses “yadi”.

5.3.1 More on Hindi mathematical terms

Our Hindi words were taken mostly from the NCERT books, which particularly in the later classes use Sanskritised Hindi. They make good use of the regular word-building capacity of Sanskrit. For example, “to add” is “joṛnā” in the lower classes, but “addition” becomes “yog” in the higher classes. This allows constructs like (yogātmak, additive), which is like (guṇātmak, multiplicative), (b^{h}āgātmak, divisive) and so on.

One might think the NCERT books overly Sanskritised, but it is hard to find other solutions, short of massive code switching between English and Hindi. NCERT books are widely used all over India. We have no sales figures for the NCERT mathematics books in Hindi, but there are not many widely available alternatives. If Hindi is to become a language for mathematics, these books might be a major lexical source.

5.4 Contrast: the converging lexicons of Telugu/Kannada

Hindi and Urdu make a very unusual pair, agreeing so completely at the base and diverging so much immediately after. Related languages usually go the other way. An example is the pair Telugu/Kannada, two South Indian languages.

Telugu/Kannada do not share a base lexicon, and so are mutually unintelligible for everyday use, unlike Hindi/Urdu.

But at higher registers, where Hindi/Urdu diverge, Telugu/Kannada converge. So where a Hindi speaker listening to technical Urdu would understand the grammar but not the content words, the Telugu speaker listening to technical Kannada would recognise all the content words but not the grammar.

For mathematics, Telugu/Kannada use a Sanskrit-based lexicon essentially identical to that of Hindi. We do not list the exact Telugu and Kannada versions, but do note that the convergence Hindi-Telugu-Kannada would be improved by deliberate coordination. For completeness, we mention that a smaller part of the higher vocabulary, mostly administrative terms, is shared with Urdu.

Further,Telugu/Kannada are in fact grammatically close, so a Telugu speaker who knows no Kannada would need only a brief reminder of grammar and a basic lexicon to read mathematics in Kannada—the mathematical terms would be familiar. A hypothetical "Scientific Kannada for Telugu Speakers" need only be a slim volume. It is the general reading in Kannada that would need a bigger lexicon. This parallels the situation of an English speaking scientist trying to read French—the scientific reading is easier!

But for a Hindi-speaking scientist trying to read Urdu, it is the everyday texts that are easier, not the scientific ones.

5.5 Summary of lexical study

Our figures suggest that everyday Hindi and Urdu share 82% of their vocabulary, but this number drops if we move to a specific domain: for tourist phrases, to 69%, and for very technical domains, such as mathematics, to a striking 8%.

An English speaker who knows no mathematics might hear mathematics in English as built of nonsense words that function recognizably as nouns, adjectives, verbs and so on. This is how mathematics in Urdu would sound to a Hindi speaking mathematician (and the other way around), even though Hindi and Urdu share a base lexicon and the grammar.

The mathematics lexicons of Hindi, Telugu and Kannada suggest that a Sanskrit based vocabulary makes a powerful link across India. That vocabulary also makes Urdu the odd language out amongst Indian languages, despite its close relation to Hindi.

6 Discussion

Our results confirm that Hindi and Urdu share a grammar, but differ so much in vocabulary (even for travel and primary school) that they are now different languages in any but the most basic situation. With the various linguistic, cultural and political factors obtaining in India and Pakistan, a good guess is that the languages will diverge further.

A regular Sanskrit base for Hindi technical terms would cement this divergence from Urdu, but would give Hindi a more usual convergent relationship with other Indian languages, differing at the everyday level but coming together at higher registers. Indeed this situation might argue for Sanskritised Hindi as a national language, because for non-native Indian speakers this may be easier to understand than Hindi with more Perso-Arabic words.

(Paauw, 2009) says "Indonesia, virtually alone among post-colonial nations, has been suc-

cessful at promoting an indigenous language as its national language." Pakistan may have similarly solved its national language problem, with a parallel situation of Urdu being the native language of a minority. A difference is that Urdu already has rich lexical and word-building resources, whereas Bahasa Indonesia did not. So the *Istilah* committee has over the decades standardised hundreds of thousands of terms. India does not need that many new terms, since it too has a rich shared lexical resource in Sanskrit, one that moreover has tremendous word-building capacity. But a standardising committee may help, since often the same Sanskrit word is used in different ways in different Indian languages. A standard pan-Indian lexicon for technical terms would allow for ease of translation, and might spur the usabilty of all Indian languages for science and technology.

Future Work

We hope to develop our Phrasebook and MGL tools, aiming for practical use. We also need to fix the remaining errors in our grammars, to do with continuous tenses, word order for some questions and negations, and the translation of English articles. Fixing these might be non-trivial. We have stated two other goals, to rebuild our resource grammars on a phonetic basis, and to do a progressive mathematics lexicon. We have started work on this last, which we believe will show an increasing divergence between Hindi and Urdu as we go to higher classes. The NCERT books are available in both Hindi and Urdu, so we have a ready made source for the lexicons.

Currently, popular articles and TV programs that need advanced vocabulary (e.g., music competitions or political debates) in Hindi take the terms needed from English, Urdu and Sanskrit sources, though these elements sit uncomfortably together, at least as of now. More examples are worth studying.

Acknowledgements

We thank our informants Anurag Negi, Vinay Jethava, and Azam Sheikh Muhammad for their painstaking comments.

Our paper originally used French and English as an example of the usual relationship: shared technical vocabulary but differing everyday words. We thank one of our anonymous referees for pointing out that we should rather take a pair closer to home - they suggested Malay-Indonesian (Paauw, 2009), but we chose Telugu-Kannada both because the first author speaks these and because we can link them to Hindi via Sanskrit.

References

Agnihotri, R. K. (2007). *Hindi: An Essential Grammar*. London/New York: Routledge.

Caprotti, O. and Saludes, J. (2012). The gf mathematical grammar library. In *Conference on Intelligent Computer Mathematics /OpenMath Workshop*.

Census-India (2001). *Abstract of Speakers' Strength of Languages and Mother Tongues*. Government of India. `http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm`.

Census-Pakistan (1998). *Population by Mother Tongue*. `http://www.census.gov.pk/MotherTongue.htm`.

Chand, T. (1944). *The problem of Hindustani. Allahabad: Indian Periodicals*. `www.columbia.edu/itc/mealac/pritchett/00fwp/sitemap.html`.

Flagship (2012). *Undergraduate program and resource center for Hindi-Urdu at the University of Texas at Austin*. `http://hindiurduflagship.org/about/two-languages-or-one/`.

Hindi-WordNet (2012). *Hindi Wordnet. 2012. Universal Word – Hindi Lexicon*. `http://www.cfilt.iitb.ac.in`.

Joshi, M. M. (2012). Save Urdu from narrow minded politics. *Bombay: The Times of India, 19 Jan 2012*.

Kachru, Y. (2006). *Hindi (London Oriental and African Language Library)*. Philadelphia: John Benjamins Publ. Co.

Martin-Löf, P. (1982). Constructive mathematics and computer programming. In Cohen, Los, Pfeiffer, and Podewski, editors, *Logic, Methodology and Philosophy of Science VI*, pages 153–175. North-Holland, Amsterdam.

Naim, C. (1999). *Introductory Urdu, 2 volumes. Revised 3rd edition*. Chicago: University of Chicago.

NCERT (2012). *Mathematics textbooks (English and Hindi)*. New Delhi: National Council for Educational Research and Training.

Paauw, S. (2009). One land, one nation, one language: An analysis of Indonesia's national language policy. *University of Rochester Working Papers in the Language Sciences*, 5(1):2–16.

Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. *Journal of Functional Programming*, 14(2):145–189.

Ranta, A. (2009). The GF Resource Grammar Library. *Linguistics in Language Technology*, 2. `http://elanguage.net/journals/index.php/lilt/article/viewFile/214/158`.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Ranta, A., Détrez, G., and Enache, R. (2012). Controlled language for everyday use: the molto phrasebook. In *CNL 2012: Controlled Natural Language*, volume 7175 of *LNCS/LNAI*.

Saludes, J. and Xambó, S. (2010). MOLTO Mathematical Grammar Library. `http://www.molto-project.eu/node/1246`.

Sarwat, R. (2006). *Language Hybridization in Pakistan (PhD thesis)*. Islamabad: National University of Modern Languages.

Schmidt, R. L. (2004). *Urdu: An Essential Grammar*. London/ New York: Routledge.

Snell, R. and Weightman, S. (2003). *Teach Yourself Hindi*. London: Hodder Education Group.

Virk, S. M., Humayoun, M., and Ranta, A. (2010). An open source Urdu resource grammar. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 153–160, Beijing, China. Coling 2010 Organizing Committee.

PDF | Hindi and Urdu share a grammar and a basic vocabulary, but are often mutually unintelligible because they use diï¬€erent words in higher registers | Find, read and cite all the research you need on ResearchGate.Â Download full-text PDF. Computational evidence that Hindi and Urdu share a grammar but not the lexicon. Conference Paper (PDF Available) Â· December 2012 with 232 Reads. How we measure 'reads'. Hindi and Urdu share a grammar and a basic vocabulary, but are often mutually unintelligible because they use diï¬€erent words in higher registers and sometimes even in quite ordinary situations. We report computational translation evidence of this unusual relationship (it diï¬€ers from the usual pattern, that related languages share the advanced vocabulary and diï¬€er in the basics).Â @inproceedings{Prasad2012ComputationalET, title={Computational evidence that Hindi and Urdu share a grammar but not the lexicon}, author={K. V. S. Prasad and Shafqat Mumtaz Virk}, booktitle={WSSANLP@COLING}, year={2012} }. K. V. S. Prasad, Shafqat Mumtaz Virk. Published in WSSANLP@COLING 2012. Computer Science. A grammar formalism determines the architecture of the grammar.Â Dictionaries, linguistic postgraduate theses and informants (who speak the language and/or are linguists) formed the data source for the lexicon and descriptive grammar. Linguists were used in cleaning, authenticating the data and through elicitation, they generated morphology and syntax of the categories that were missing in the Descriptive grammar from corpora. The elicitation was performed either through language analysis of the corpus through linguist judgment or by translation from English to the specific Bantu language as proposed by Chelliah [28].