IJAER/Feb-March 2013/Volume-2/Issue-2 ISSN: 2278-9677
International Journal of Arts & Education Research

HINDI AND URDU SHARE A GRAMMAR BUT NOT THE LEXICON

Abdul Hamid Ahmed

Principal

Halakura College

Dist. Dhubri, Assam

Abstract

Although Hindi and Urdu share a fundamental vocabulary and syntax, they are often incomprehensible to one another because they utilise different terms at higher registers and even in everyday settings. We provide computational translation proof of this atypical relationship (in contrast to the conventional pattern, in which related languages share advanced vocabulary but differ in the fundamentals). We took a GF resource grammar for Urdu and modified it for Hindi, modifying just the script (Urdu is written in Perso-Arabic, whereas Hindi is written in Devanagari) and the vocabulary where necessary. In testing, the Urdu grammar and its Hindi counterpart either properly translated an English phrase or failed in the same grammatical fashion, indicating that the two languages share a grammar. However, the Hindi and Urdu lexicons differed in 18% of fundamental words, 31% of tourist phrases, and 92 percent of school arithmetic concepts, according to the study.

Keywords: *Grammatical Framework, Resource Grammars.*

Introduction

India's official language is Hindi, and Pakistan's is Urdu, despite the fact that neither country's original language is spoken by a majority of its citizens.

'Hindi' is a broad phrase that encompasses a vast range of dialects. According to Wikipedia, Hindi has 422 million native speakers (Census-India, 2001). According to the census, there are 258 million native speakers of 'Standard Hindi.' Official Hindi is now Sanskritized, although Hindi has drawn from both Sanskrit and Perso-Arabic, resulting in a variety of dialects and making Standard Hindi difficult to define. To round out the image of India's "national language," keep in mind that Hindi is not spoken in many sections of the country (Agnihotri, 2007), and it competes with English as a lingua franca.

For numerous reasons, it is more convenient to discuss standard Urdu, which is the native language of 51 million Indians and 10 million Pakistanis (Census-India, 2001). (Census- Pakistan, 1998). Urdu's advanced vocabulary has always been derived only from Perso-Arabic, and it does not suffer from the same form issues as Hindi. It is Pakistan's official language and lingua franca, yet we should emphasise that Urdu's dominance is also challenged, if not despised, in certain sections of the country (Sarwat, 2006).

The structure and much of the fundamental vocabulary of ordinary speech are the same in Hindi and Urdu (Flagship, 2012). This common ground has been identified and is referred to as 'Hin-dustani' or 'Bazaar language' (Chand, 1944; Naim, 1999). However, "it has not been granted any importance in Indian or Pakistani society for attitudinal reasons" (Kachru 2006). After English, Mandarin, Spanish, and maybe Arabic, Hindi-Urdu is the fourth or fifth most commonly spoken language in the world, spoken by the

South Asian diaspora in North America, Europe, and South Africa.

History: Hindustani, Urdu, Hindi

From the 14th century onwards, a language known as Hindustani emerged by absorbing parts of the invaders' Perso-Arabic vocabulary into Khari Boli, a dialect of the Delhi area. Urdu is written in the Perso-Arabic script and originated from Hindustani via extensive borrowing from Persian and some Arabic. It was built in the late 1800s. Hindi developed from Hindustani in the late 1800s, but through borrowing from Sanskrit. It's written in a version of the Sanskrit Devanagari script.

However, the underlying character of Hindi/Urdu has been preserved: 'the common spoken variety of both Hindi and Urdu is close to Hindustani, i.e., free of major borrowings from either Sanskrit or Perso-Arabic' (Kachru, 2006).

One language or two

According to a tagline on the newspaper article, Hindi and Urdu are "one language, two scripts" (Joshi, 2012). According to the lexicons, neither Hindi nor Urdu can fulfil the phrase. By definition, Hindustani does, although it is restricted to the common parts of the Hindi and Urdu lexicons.

(Flagship, 2012) notes that Hindi and Urdu "have evolved as two independent languages in terms of script, higher vocabulary, and cultural ambience." '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 practise and usage each has developed into an individual language,' writes Gopi Chand Narang in the preface to (Schmidt, 2004).

However, lexical differences may not tell the complete picture. (Naim, 1999) cites some slight morphological differences as well as some notable phonetic differences between Hindi and Urdu. Most Hindi speakers have trouble pronouncing the Urdu sounds that appear in Perso-Arabic loan words, such as q (unvoiced uvular plosive), x (unvoiced velar fricative), G (voiced velar fricative), and some final consonant clusters, whereas Urdu speakers replace the (retroflex nasal) of Hindi with n and have difficulty pronouncing many Hindi consonant clusters.

Beginning with Hindi and Urdu simultaneously, according to Naim, does not aid learners. Those seeking a grasp of the written language, he claims, should start by learning the Urdu-specific rules.

To the best of our knowledge, there are numerous scholarly and differing perspectives on whether Hindi and Urdu are one or two languages, but nothing has been computationally proven. Our research shows that whereas Hindi and Urdu share a grammar, their lexicons differ dramatically beyond the basic and broad registers. Our exploratory tests have already provided tentative answers to issues such as 'How much do the lexicons of Hindi and Urdu differ?'

Overview the tool utilized in this experiment, Grammatical Framework, is described in Section 2, and the results are listed in Section 3. Section 4 compares and contrasts the Hindi and Urdu resource grammars, as well as the differences between them and how we deal with them. The generic and domain-specific lexicons utilized in this experiment are presented in Section 5. The findings of the evaluation are presented at the conclusion of Sections 4 and 5. Section 6 serves as a wrap-up and gives perspective.

This document employs an IPA-style alphabet, complete with standard values and standards. In Hindi and

Urdu, retroflexed sounds are transcribed with a dot beneath the letter; in Sanskritised Hindi and (a flap) are prevalent (though many dialects pronounce them n and n). In Hindi and Urdu, the palatalized spirant and aspirated stops, indicated as kh, are prevalent. A macron above a vowel signifies nasalization and a lengthy vowel. The macron is eliminated in Hindi and Urdu because e and o are usually lengthy. Finally, we utilise to denote the nasal homorganic with the consonant that follows.

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 trans- lation. These translations are not usually for arbitrary sentences, but for those restricted to a specific domain, such as tourist phrases or school mathematics.

Resource and Application Grammars in GF

The (English or Hindi) application grammars Phrasebook (Caprotti et al 2010, (Ranta et al., 2012) and MGL (Saludes and Xambó, 2010) respectively describe the sublanguages of English or Hindi that deal with these specific fields. However, the underlying English in both the English Phrasebook and the English MGL is the same (similarly for Hindi). The underlying English (or Hindi) syntax, morphology, prediction, modification, quantification, and so on are all contained in a resource grammar, which is a general-purpose module.

As a result, resource grammars are distributed as software libraries, and the GF resource grammar library presently contains resource grammars for over twenty-five languages (Ranta, 2009). Creating a resource grammar needs both GF expertise and linguistic understanding. Application grammars need domain knowledge but are unaffected by the complexity of expressing things in English or Hindi. The resource grammar, on the other hand, specifies how to speak the language, but the application grammar describes what may be said in a certain application domain.

Abstract and Concrete Syntax

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

Abstract sentence:

PQuestion (HowFarFrom (ThePlace Station)(ThePlace Airport))

Concrete English sentence: How far is the airport from the station?

Concrete Hindustani sentence: sṭešan se havāī aḍḍā kitnī dūr hæ? (3Zशन ų हवाई अ5ा fकतनी 5र ž? , ؟,.ɾ دورΊǒï2 اڈا Ίوائɾ ,.س شنiٹ.اس)

Hindustani word order: station from air port how-much far is?

The abstract sentence is a tree made up of components that have had functions applied to them. These aspects are made up of different types of inquiries, locations, and distances. For example, the concrete syntax for Hindi defines a mapping from abstract syntax to Hindi textual representation. That is, a concrete syntax provides rules for linearizing the abstract syntax's trees.

MGL examples would feature a variety of abstract functions and components. In general, the abstract syntax specifies which categories and functions are accessible, allowing for semantic compositions that are independent of the language.

It is feasible to have many concrete syntaxes for a single abstract by separating the tree-building rules (abstract syntax) from the linearization rules (concrete syntax). This allows you to parse text in one language and have it produced in any of the others.

To demonstrate the difference between resource and application grammars, compare the above tree to the resource grammar abstract tree for "How far is the airport from the station?"

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

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

Using an existing Urdu resource grammar, we first created a new Hindi grammar in the Grammatical Framework (GF) (Ranta, 2011). (Virk et al., 2010). Our new Hindi resource grammar is therefore the first item on which we report, despite the fact that it is not the primary topic of this work.

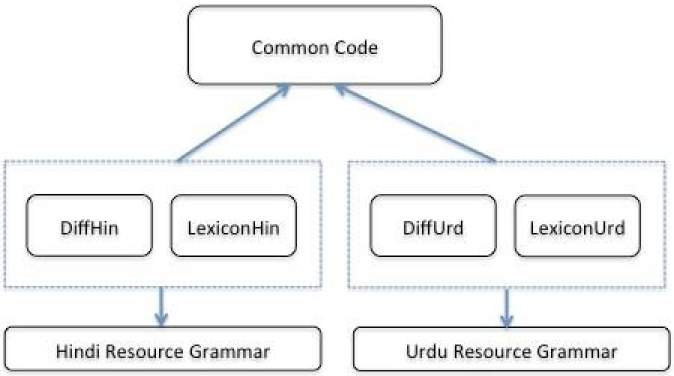

Figure 1: Hindi/Urdu Functor.

We developed Hindi and Urdu resource grammars using a functor-style approach, which allows two grammars to share similarities. This implementation technique is shown in Figure 1. The 'common code box' contains the majority of the syntactic code, whereas the 'DiffLang box' contains the small syntactical differences (described in Section 4). There is a lexicon for each resource grammar. This indicates that Hindi and Urdu have the same syntax and differ virtually entirely in their lexicons.

We tested 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 for correctness or incorrectness (see Section 6 for results).

Differences between Hindi and Urdu in the Resource Grammars

We began with the Urdu script-based GF resource grammar and converted it nearly totally for Hindi by simply re-coding from Urdu to Hindi script. The resource grammars come with a simple test vocabulary that may be altered as needed: it turns out that even in this basic vocabulary, Hindi and Urdu differ by up to 18%. The application lexicons are discussed in Section 5.

We don't go into depth on how to build these resource grammars in this article since the interesting pieces have already been covered in the previous section (Virk et al., 2010). However, we discuss resource level differences between Hindi and Urdu, as well as solutions for dealing with them, below.

Morphology

Every GF resource grammar comes with a 450-word basic test lexicon, for which the morphology is programmed via lexical paradigms, which are special functions. The current Urdu morphology is simply recoded for the Devanagari script in our Hindi morphology. The morphologies are not similar because to lexical differences; for example, Hindi employs a simple term whereas Urdu uses a complex word, and vice versa. However, since there are no patterns that occur in just one of the languages, the Hindi test lexicon works well.

The slight morphological differences highlighted in (Naim, 1999) may theoretically be implemented, but we chose not to. Our informants find the resultant Hindi to be quite normal, indicating that these differences are small.

Internal Representation: Sound or Script?

"How far is the airport from the station?" was written in IPA to represent the Hindi/Urdu language's sound. It has the same sound in both languages, therefore we may call it 'Hindustani.' One logical method to developing Hindi/Urdu grammars from scratch would be to represent the languages internally by sound, resulting in a single grammar, a common lexicon, and differentiated lexicons solely for words that sound differently in Hindi and Urdu. We would next map the IPA to the Hindi or Urdu characters for output.

However, we were beginning with (Virk et al., 2010), which employs a textual Urdu-based internal representation. Re-doing this in terms of speech would be a significant undertaking, but the end product would be readily reusable for Hindi and might also aid in capturing parallels to other South Asian languages. This re-modeling will be done in the future.

So, in this paper, we simply replaced written Urdu with written Hindi to convert the Urdu grammar to a Hindi grammar. This script alteration was likewise applied to the core lexicon, however certain words were uttered in a different way. As a result, our parallel grammars make no mention of the fact that Hindi and Urdu often sound the same.

One benefit of script-based representations is that they eliminate spelling issues. Several sound differences in Persian, Arabic, and Sanskrit are obliterated in Hindi-Urdu. Because Urdu accurately keeps the spelling of the original Perso-Arabic words while representing Sanskrit words phonetically, whereas Hindi does the opposite, a pho- netic transcription would not reflect these collapsing differences, but orthography does. Each language adheres to the same script as the original sources. We can see that mechanically converting a phonetic representation to a written one would be difficult.

Obviously, the greater the overlap between the Hindi and Urdu lexicons, the more wasted effort in the parallel technique However, as we'll see, the lexicons differ quite a little from one another. For use in a redesigned grammar, we created an enhanced phonetic representation that maintains account of spelling.

Idiomatic, Gender and Orthographic Deference's

Apart from spelling, Hindi and Urdu feature orthographic deferences that are seldom re-marked. Some ostensibly grammatical deferences are really idiomatic, gender, or orthographic deferences.

The lexicon may, for example, interpret the verb "to add" as "jon" in Hindi and "jame karn" in Urdu. The imperative statement "add 2 to 3" would thus be expressed in Hindi as "do ko tn se joo" and in Urdu as "do ko tn m jame karo." However, the decision between the post-positions "se" and "m" is governed by the post-positional idiom of the selected verb, "jon" or "jame karn," since each phrase works in either language.

With the word "war," which is pronounced "la" in Urdu, there is a gender distinction (fem.). This term also works in Hindi, but it has a stronger sense of "war," so we went with "saghar" (masc.). The change from feminine to masculine is based on word choice rather than language.

Orthographic differences next. "He will go" is "vo jāegā" in both languages; in writing, (वह जाएगा, ˡdb ,ائ� وہ), the final "gā" (गा, ˡdb) 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, (हमF चाय पी, ʹI� ,چائ ,ن ہم), the particle "ne" (F, ,ن) 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".

Evaluation and Results

With external informants

As previously stated, we constructed 80 abstract trees at random (40 from each of the Phrase-book and MGL) and linearized them in Hindi and Urdu. Three separate informants were then given these linearizations.

They assessed the Hindi and Urdu translations that our grammars produced. In both Hindi and Urdu, the information found 45 phrases to be valid. The other phrases were determined to be comprehensible but grammatically incorrect - in both Hindi and Urdu - and nothing the informants reported could be linked to a syntactic difference between the two languages. The purpose of this study is that all 80 phrases, both incorrectly and successfully translated, provide mechanical confirmation that Hindi and Urdu share a grammar.

For the record, the 35 grammatical errors provide the false appearance that the grammar is only "45/80" accurate. In truth, the grammar is considerably better: there are just a few different recognised structures that need to be fixed, such as negation and question word placement, although they appear often in the evaluation sentences.

We considerably enhanced the Urdu grammar of (Virk et al., 2010) while constructing the Hindi equivalent, which is not the topic of this research. As previously stated, there are still errors.

With internal informants

The second author is an Urdu native speaker, whereas the first is a Hindi native speaker. We could quickly do numerous more extensive informal reviews using ourselves as internal informants. We looked at 300 Phrasebook sentences, 100 MGL sentences, and 100 sentences derived using resource grammars directly. We can confirm that the Urdu and Hindi translations for all 500 English phrases were intelligible and in accordance with Urdu and Hindi grammar (barring the known errors noted by the external informants).

We found that randomly produced MGL phrases may be exceedingly involuted, and that the Hindi and Urdu translations in every instance had the same structure.

The Lexicons

As we saw in Section 1, Urdu has a standard form, but Hindi does not, despite the fact that official Hindi is progressively more Sanskritized. Hindustani counts as 'Hindi' and is a neutral form, however it only has a limited vocabulary, as previously stated (Chand, 1944). As a result, Hindi speakers must pick between one of the higher forms in order to progress. For example, elementary mathematics may be taught in Hindustani or Sanskritised Hindi, as documented by the NCERT publications (NCERT, 2012), or in English-ised Hindi, which can be heard in any high school or institution in the Hindi-speaking countries. When we had sources, such as the NCERT mathematics books or a government phrase book, we utilised these to arbitrate the choice of Hindi terms. Otherwise, we chose the most common options using (Snell and Weightman, 2003) and (Hindi-WordNet, 2012).

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.

The Phrasebook lexicon

There are 134 items in this dictionary, divided into 112 terms and 22 greetings. The Hindi and Urdu entries are identical for 92 terms, including 42 English borrowings for currency names, (European) nations and nationalities, and words like "tram" and "bus." So, although Hindi and Urdu share 50 native terms, they differ on 20 others, including days of the week (excluding Monday, which is "somvr" in both languages). There are 22 terms in the greetings lexicon, the most of which are difficult to translate. "Good morning," "bye," and other phrases may be translated, however they are often simply "hi" and "bye." Greetings are definitely more culture-specific: there are 17 locations where Hindi and Urdu differ.

An example not in the Phrasebook drives home the point about greetings: airport an- nouncements beginning “Passengers are requested ...” are rendered in Hindi as “yātriyõ se nivedan hæ ...” (याfTयK ų fिनbदान ž) and in Urdu as “musāfirõ se guza:riš kī jātī hæ ...” (,.ɾ Ίiˡ.� Ί2 زارشjↀb ,.س روں.ساف.م), which suggests that Hindi and Urdu have diverged even in situations almost tailored for ‘Bazaar Hindustani’!

Conclusion

Our findings show that although Hindi and Urdu share a grammar, their vocabulary is sufficiently different (even for travel and elementary education) that they are now considered separate languages in all save the most basic situations. Given the many linguistic, cultural, and political forces at play in India and Pakistan, it's safe to assume that the languages will continue to differ.

A regular Sanskrit substrate for Hindi technical terminology would reinforce the language's separation from Urdu while also giving it a more typical convergent connection with other Indian languages, with differences at the daily level but convergence at higher registers. Indeed, this circumstance may support for Sanskritized Hindi as a national language, since it may be simpler to grasp for non-native Indian speakers than Hindi with more Perso-Arabic vocabulary. "Indonesia, nearly alone among post-colonial states, has been successful in promoting an indigenous language as its national language," according to (Paauw, 2009). Pakistan may have found a similar solution to its national language challenge, with Urdu being the native language of a minority. Urdu, on the other hand, has extensive lexical and word-building resources, but Bahasa Indonesia has not. As a result, the Istilah committee has standardized hundreds of thousands of phrases throughout the years. India does not need many new words since it has a large common lexical resource in Sanskrit, which also has a large word-building potential. However, since the same Sanskrit term is often employed in different ways in different Indian languages, a standardising committee would be useful. A common pan-Indian vocabulary for technical terminology would make translation easier and might encourage the use of all Indian languages in science and technology.

References

1. Agnihotri, R. K. (2007). Hindi: An Essential Grammar. London/New York: Routledge.
2. Caprotti, O. and Saludes, J. (2012). The gf mathematical grammar library. In Conference on Intelligent Computer Mathematics /OpenMath Workshop.
3. 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.
4. Census-Pakistan (1998). Population by Mother Tongue. http://www.census.gov.pk/

MotherTongue.htm.

5. Chand, T. (1944). The problem of Hindustani. Allahabad: Indian Periodicals. www. columbia.edu/itc/mealac/pritchett/00fwp/sitemap.html.
6. 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/.
7. Hindi-WordNet (2012). Hindi Wordnet. 2012. Universal Word – Hindi Lexicon. http:
8. //www.cfilt.iitb.ac.in.
9. Joshi, M. M. (2012). Save Urdu from narrow minded politics. Bombay: The Times of India, 19 Jan 2012.
10. Kachru, Y. (2006). Hindi (London Oriental and African Language Library). Philadelphia: John Benjamins Publ. Co.
11. 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.
12. Naim, C. (1999). Introductory Urdu, 2 volumes. Revised 3rd edition. Chicago: University of Chicago.
13. NCERT (2012). Mathematics textbooks (English and Hindi). New Delhi: National Council for Educational Research and Training.
14. 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.
15. Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism.
16. Journal of Functional Programming, 14(2):145–189.
17. Ranta, A. (2009). The GF Resource Grammar Library. Linguistics in Language Technol- ogy, 2. http://elanguage.net/journals/index.php/lilt/article/viewFile/214/
18. 158.
19. Ranta, A. (2011). Grammatical Framework: Programming with Multilingual Grammars. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).
20. 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.
21. Saludes, J. and Xambó, S. (2010). MOLTO Mathematical Grammar Library. http:
22. //www.molto-project.eu/node/1246.
23. Sarwat, R. (2006). Language Hybridization in Pakistan (PhD thesis). Islamabad: National

University of Modern Languages.

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

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

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