Towards Hindi/Urdu FrameNets via the Multilingual FrameNet

Shafqat Mumtaz Virk[1] and K.V.S. Prasad[2]
[1]Språkbanken, Department of Swedish, University of Gothenburg, Sweden
[2]Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
shafqat.virk@svenska.gu.se, prasad@chalmers.se

Abstract

The Multilingual FrameNet Project (MLFN, 2017) is using translations of Ken Robinson's popular TED talk (Robinson, 2006) to study universal and cross lingual aspects of frame annotation. There are no FrameNets yet for Hindi and Urdu, but we are annotating the Hindi and Urdu translations of Robinson's talk using the frames of the English FrameNet. (Surprisingly, there was no Hindi translation, so we did that ourselves). Preprocessing is needed: the word-segmentation and POS tagging tools available for Hindi and Urdu were satisfactory, the full-form lexicons less so. The web-based multi-layer frame annotation tool allows additions to the lexicon, so we simply added each form as a new "word", our goal here being only to look at the frames and frame elements—we plan to look at grammatical function and phrase type later. While some sentences show that the frame analysis of English or Portuguese will not carry over to Hindi or Urdu for cultural or linguistic reasons, others are harder to be definite about. Partly, this is because there are so many possible translations. An expected observation is that a choice of word can steer the focus from one frame to another. Our annotations will help when we start building framenets for Hindi and Urdu.

Keywords: Frame semantics, FrameNet, Multilingual FrameNet, Lexico-Semantic Resources

1. Background: Frame Semantics

Frame semantics, developed by Charles Fillmore and others (Fillmore, 1976; Fillmore, 1977; Fillmore, 1982), thinks of language as creating scenes, in which we understand what a word or phrase means by the role it plays in the scene. E.g., using frame semantics we model a kidnapping *situation* as a structure called a *frame*, a script-like description in which *frame elements* (FEs) such as `Perpetrator, Victim, Purpose, Time` and `Place` play their various roles. Words like `kidnap, abduct, nab` and `snatch` *trigger* this frame. Frames similarly model events, objects, and relations.

Based on frame semantics, a lexico-semantic *FrameNet* (Baker et al., 1998) has been developed since 1998, for English. Descriptions of real world situations are stored as frame scripts in FrameNet, along with the frame elements and triggers that evoke the frame. Each frame is given example sentences, actually occurring text, and there is also a *frame annotated* corpus. The frames are linked by relations to make a FrameNet (henceforth FN). E.g., the frame `Invading` *inherits* from `Attack`, is a *subframe* of `Invasion_scenario`, and *precedes* `Conquering` and `Repel`.

These resources (the FrameNet, the example sentences, and the annotated corpus) have been used for automatic shallow semantic parsing (Gildea and Jurafsky, 2002), itself used in tasks such as information extraction (Surdeanu et al., 2003), question-answering(Shen and Lapata, 2007), coreference resolution (Ponzetto and Strube, 2006), paraphrase extraction (Hasegawa et al., 2011), and machine translation (Wu and Fung, 2009; Liu and Gildea, 2010).

2. MultiLingual FrameNet

FrameNets have since been built for several languages (Chinese, French, German, Hebrew, Korean, Italian, Japanese, Portuguese, Spanish, and Swedish), and have helped explore various semantic characteristics of the individual languages, but the cross linguistic and universal aspects of the FN model are largely yet to be studied. So a MultiLingual FN (MLFN) is now being built by aligning FNs of the individual languages. As a first step, translations of Ken Robinson's popular TED talk (Robinson, 2006) are being annotated using the frames of the Berkeley English FrameNet. An example annotation is shown in Fig. 1

Annotators for each language mark the frame-elements (FE), the grammatical function (GF), and the phrase type (PT) of the marked FEs. (See Sec. 5. for a brief description of these layers, and (Ruppenhofer et al., 2006) for more details). Fig. 1 shows the annotations of two frames, `Conditional_occurrence` and `Questioning`, in the sentence "But if you ask about their education, they pin you to the wall". These are triggered respectively by the lexical units `if` and `ask`. The text blocks "you" and "about their education" have been marked as FEs `Speaker` and `Topic` respectively. The GF and PT of the marked FEs have been labeled at their corresponding layers. (Ruppenhofer et al., 2006) explains the PT and GF labels for English.

Annotators choose from a given list of frames and their FEs. If an annotator does not find a suitable frame from the given list, they select the best alternative (if any), note why the frame is unsuitable, and suggest a better frame. The PT and GF are language dependent, and a list of PTs and GFs for each language has to be provided by the annotators.

Fig. 1 also shows the Portuguese translation of the sentence with the same two frames. Note that a different FE, `Message`, used for asking "What is this", is chosen instead of `Topic`, used for "asked about train times". Whether the choice is appropriate is up to the

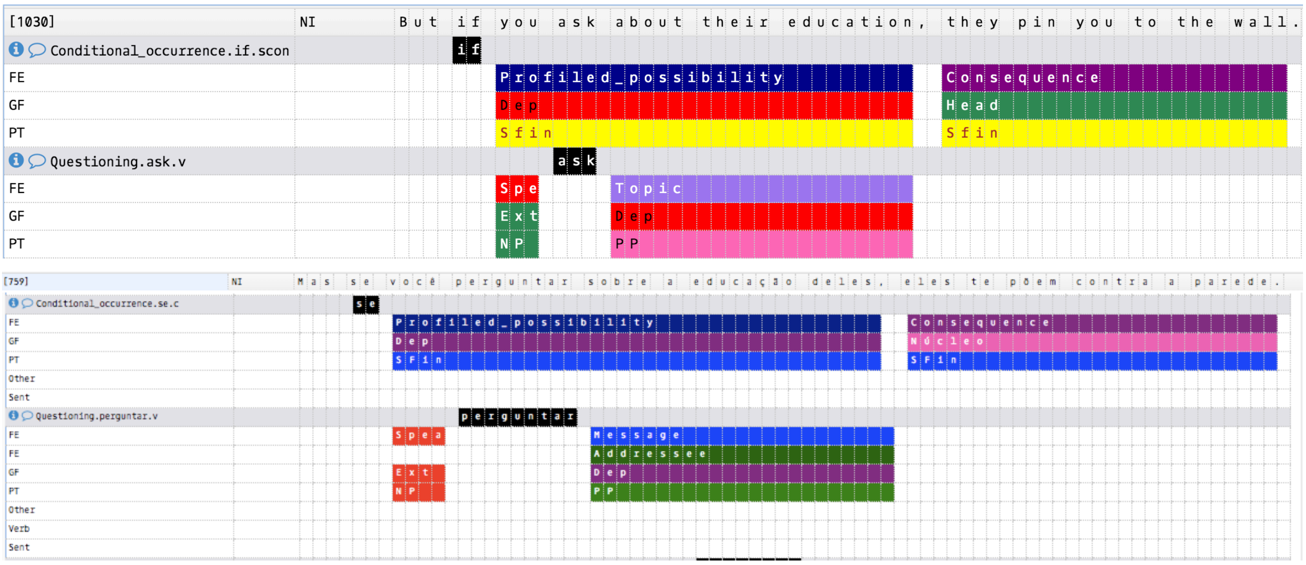

Figure 1: Partial frame-annotation of an example sentence in English and Portuguese.

annotator.

Once this multi-lingual annotation is completed, the challenges faced by annotators, the common and uncommon frames chosen, and the attached notes, will be collated and reported. Similarly, variations in PTs and GFs of the various FEs. These reports will be used to learn about the difficulties and challenges in trying to align existing framenets, and building a multilingual framenet.

This paper reports our experience of annotating the Hindi and Urdu translations of Robinson's talk.

3. Background: Hindi and Urdu

'Hindi'[1] has ca. 400 m (million) speakers, of whom 250 m are native. Urdu[2] has ca. 250 m speakers, of whom 60 m are native. Only English, Mandarin, Spanish and Arabic have more speakers than Hindi-Urdu.

Hindi and Urdu 'share the same grammar and most of the basic vocabulary of everyday speech', but are 'two separate languages in terms of script, higher vocabulary, and cultural ambiance' (Flagship, 2012; Prasad and Virk, 2012). They are thus different standard registers of one language (Bhat et al., 2016). Indeed, we used a tool (Apertium, 2017) that translates efficiently between the two, doing mostly only lexical substitution.

Hindustani. The 'Hindi' of films and songs is 'the common spoken variety, devoid of heavy borrowings from either Sanskrit or Perso-Arabic' (Kachru, 2006). We call this form *Hindustani* (Chand, 1944; Bailey et al., 1950). India's 'Hindi' belt speaks more Hindustani than Hindi. But Hindustani has no 'status in Indian or Pakistani society' (Kachru, 2006). We study only Urdu and (standard) Hindi[3] here.

Scripts. Hindustani[4] began ca. 1400 as a Delhi dialect with some Perso-Arabic vocabulary. Urdu, ca. 1750, is Hindustani with copious Perso-Arabic borrowings. Both are written in Perso-Arabic script.

By 1900, some began to write Hindustani in Devanagari[5], the script giving it an identity, *Hindi*, distinct from Urdu, and an impetus to progressively use Sanskrit vocabulary instead of Perso-Arabic.

The Hindi lexicon. Hindi and Urdu 'share the same Indic[6] base' (Schmidt, 2004), and a phonology (UH) that breaks up the consonant clusters of Sanskrit, and drops short vowels at the end of syllables.

Phonology plays no role in frame analysis, but that the phonologies of Sanskrit and UH are at odds is a feature of the Hindi lexicon, which does matter.

E.g., suppose we replace the Indic Hindi-Urdu word सूरज `sūraj` "sun" with the Sanskrit सूर्य `sury`. In Sanskrit, सूर्य is pronounced `surya`, easy to say, but UH drops the final `a` in speech, producing a hard-to-say word-final consonant cluster. (Dropped vowels remain in the script, and re-appear as schwas in song).

Other awkward Sanskrit words are e.g. यदि `yadi` "if", परन्तु `parantu` "but", शक्ति `šakti` "power", with their short vowel endings, a feature foreign to UH.

Unadapted Sanskrit words make Hindi more "national", but sound odd. Older Indic literary languages with UH phonology and adapted Sanskrit borrowings are not 'in direct linguistic antecedence to [...] Hindi.

[1] By Hindi, we mean standard Hindi. We take 'Hindi' more broadly, including its many dialects, some being arguably distinct languages. Multiple lexicons give 'Hindi' multiple *forms* (Kachru, 2006).

[2] Urdu has always drawn its advanced vocabulary only from Perso-Arabic, and has basically just one form.

[3] In Hindustani and in the 'Hindi' belt, "sky" is आसमान `āsmān`, a Persian word. In Hindi and other Indian languages, it is आकाश `ākāš`, a Sanskrit word. The preference for Sanskrit makes Hindi better understood nationally.

[4] Also called 'Hindi' then, but we reserve this term for the modern language, to reduce confusion.

[5] The script used for Sanskrit.

[6] i.e., with no Perso-Arabic words.

The one language that is antecedent [is] Urdu ...' (Masica, 1991). The lexical future will be interesting.

4. Translating the Urdu text to Hindi

An Urdu translation of Robinson's talk was available when we started, but surprisingly, not a Hindi one. We produced one ourselves (one of us speaks Hindi, but is not native), starting by pushing the Urdu text through Apertium, which fortunately has an Urdu-Hindi pair implemented. The output included much text that was just a transcription from Perso-Arabic script to Devanagari, as well as some Urdu text where even the transcription failed. These might be seen as shortcomings, but we think they are outweighed by the sensible behaviour of the tool in keeping going—the user will have to edit the output anyway, and these errors are easy to spot.

The manual corrections needed took several days full time, though experience with other languages suggests this is still less time than a translation from scratch would take. Finally, our text was validated and improved by a native Hindi speaker.

5. Pre-processing

We do frame annotations using the MLFN version of the Berkeley English FrameNet webtool. It allows us to attach syntactic and semantic annotation layers to the subject text. To set up the tool for a given language, the following data files are needed. Given the size of Hindi-Urdu, it is odd that we sometimes didn't find the needed resources. Those working with other South Asian (SA) languages may face similar situations.

1. A sentence segmented UTF text. We could find no publicly available sentence segmentors for either Hindi or Urdu, so we used a program to split the text at particular punctuation symbols, and then validated the results by hand.

2. A file listing all word forms of all the lexemes in the text together with the part of speech (POS) tag of each lexeme. For this, we used the smart morphological paradigms of GF (Virk et al., 2010). These take a word, and based on word endings and other clues, attempt to find suitable word-formation functions to build inflection tables. However, they are still occasionally error-prone and also have limited coverage. Fortunately, the MLFN tool allows additions to the lexicon, so we simply added each surface form as a new "word" as we went along.

3. We used the universal POS tagger for Hindi to tag the text, and the tags were then mapped to the FrameNet POS tagset [7]. For Urdu POS tagging, we used curlp Urdu POS tagger[8].

4. A list of annotation labels to be used for each language. For this experiment, Frame Element (FE), Phrase Type (PT), and Grammatical Function (GF), layers are to be added. Details can be found in the FrameNet book (Ruppenhofer et al., 2006). We briefly describe the only three annotation layers needed at this stage.

   Frame Element (FE) Here, annotators choose a suitable FE label. E.g., `Topic` in Fig. 1. Labels are taken from FrameNet data release 1.7, and annotators are not allowed to change them.

   Phrase Type (PT) Here, annotators classify the text that makes up each FE. The set of PTs is language dependent, will be chosen by the annotation team. For Hindi and Urdu, we opted to start with the English PTs, and add/edit types as needed (the MLFN tool allows these actions).

   Grammatical Function (GF) Annotators assign a GF to each FE, saying how the FE satisfies its grammatical requirements (Ruppenhofer et al., 2006). The set of GFs too is language dependent, but we opted to start with the English GF labels.

[7]FN tagset: 'A' = Adjective, 'ADV' = Adverb, 'ART' = Article, 'AVP' = Adverbial Preposition, 'C' = Conjunction, 'INTJ' = Interjection, 'N' = Noun, 'NUM' = Number, 'PREP' = Preposition, 'PRON' = Pronoun, 'V' = Verb.

[8]For a demo,see http://182.180.102.251:8080/tag

6. Annotation Status

Table 1 shows statistics of the annotations done so far both for Hindi and Urdu. For Hindi, a total of 84 frames and 154 frame-elements were annotated from the first 25 sentences of the talk. As can be noted, most of the lexical units (i.e. triggers) are from the noun and verb class followed by adjectives and adverbs. The remaining lexical units are conjunctions, prepositions and numbers. For Urdu, a total of 42 frames and 76 frame-elements were annotated from the first 27 sentences of the talk.

	Hindi	Urdu
Sentence	25	27
Frames	84	42
Frame-Elements	154	76
Noun Triggers	25	17
Verb Triggers	22	16
Adjective Triggers	13	6
Num Triggers	3	2
Adverb Triggers	8	1
Prep Triggers	3	-
Conjunction Triggers	10	-

Table 1: Annotation Statistics

7. Observations and Lessons

Some example sentences from Robinson's talk, where cross-lingual annotation is expectedly problematic: idiom ("good morning"), slang ("I've been blown away"), and metaphor ("themes running through").

1. Good morning.

   In Hindi, this is नमस्ते `namaste` "Greetings". There are no separate greetings for times of day, or even to say "hello" or "bye". The occasion may be marked by other sentences.

In Urdu,

صبح بخیر

`subah buxair` "Good morning"

2. I've been blown away by the whole thing.

   In Hindi, this is मेरी तो बुद्धि ही उड गयी है
   `merī to buddhi hī uṛ gayī hai`
   "my mind itself has been blown away".

   In Urdu,

   مجھے تو اس سب نے ہلا کر رکھ دیا ہے

   `mujhe to is sab ne hilā kar rakh diyā hai`
   "As for me, all this has left me shaken".

   A slang expression, this is hard to translate. In both English and Hindi, the verb `blown` evokes the frame `Motion`, but the FE `Theme` changes from `me` to `my mind`. Urdu changes the frame to `Cause_to_move_in_place`, but the FE `Theme` is again `me`.

3. There have been three themes running through the conference.

   In Hindi, सम्मेलन में तीन विषय उभर कर आ रहे हैं
   `sammelan mẽ tīn viṣay ubhar kar ā rahe hãĩ`
   "in the conference, three things are coming up".

   The English `running` evokes the frame `Fludic_motion`, with FEs `Fluid` "three themes" and `Area` "through the conference". The Hindi `ubhar kar ā` evokes `Coming_to_be` with FEs `Place` "in the conference", `Entity` "three things" and `Time` "are ...ing". Both are idiomatic expressions, and a different Hindi translation might have used the image of three streams flowing.

Most of the few dozen sentences we have annotated so far pose more interesting questions since the differences are not as easily explained away as in the above examples. Unfortunately, these few dozen are not enough to observe patterns in bulk. For when we have a larger number, we anticipate a few features and challenges.

Causation Where the intransitive verb "shake" evokes `Motion`, the transitive verb evokes `Cause_to_move_in_place`, as in example 2. In Hindi-Urdu this shift is done morphologically, by making causative verbs out of intransitive ones. Thus `hilnā` "to shake (intr.)" becomes the `hilānā` "to shake (tr.)" of example 2.

Examples abound: `khānā` "to eat" and `khilānā` "to feed", `sonā` "to sleep" and `sulānā` "to put to bed", etc., where English uses a different verb or an auxiliary causative verb.

Hindi-Urdu also have verbs for indirect causation. `hilvānā, khilvānā, sulvānā` mean to get somebody else to shake (tr.), feed, and put to bed. Even when the basic verb is transitive, such as "sell" `becnā` with its causal version `bicvānā` "get sb. to sell", there may be a kind of back-formation to the intransitive verb: `biknā`, used to say something sells well/badly, or is available for sale.

These regular causative links can perhaps be reflected in FN; of interest because this feature appears in other SA languages.

Abstract or concrete? A sentence in the talk is "Because it's one of those things that goes deep with people", where `deep` evoked frame `Measurable_attributes`. The Urdu text maintained the abstraction: "it goes into the depth in people". Our Hindi informant preferred "it lives in the depths of the heart(s) of people", more concrete and evoking the frame `Body_parts`.

A similar example is "a future that we can't grasp", where the verb evoked `Grasp`. Again, our Hindi text is more concrete: "that is outside our imagination", evoking `Image_schema`.

It is unlikely that such cases will show a systematic variation in frame choice going from English to Hindi-Urdu, beyond suggesting many new frames (heart, imagination, etc.).

Verb or noun? The Urdu text for "future we can't grasp" is "doesn't come into our grasp", a verb-noun variation that may be systematic. The frames evoked are different, but the meaning is the same, suggesting we look for higher level frames. Hindi-Urdu has a range of nouns that come from verbs, and vice-versa, as does English. Frame connections even within Hindi-Urdu may be interesting, as with causation.

Complex lexemes "Come into grasp" can be seen as a *complex lexeme*, a verb-based multi-word expression (Hook, 1974; Masica, 2005). In the English FN lexicon, there are many lexical units which will correspond to such complex lexemes in Hindi/Urdu. The status of these constructions as lexical or grammatical is debated and they are generally under-researched (Schultze-Berndt, 2006; Butt, 2010; Slade, 2016)

8. Conclusions and Outlook

We have started annotating Hindi/Urdu using the MLFN tool, and have reported on our experience so far. We are some way from being able to note systematic changes in annotation going from, say, English to Hindi or Urdu, and we have even further to go to construct FrameNets for Indic languages. But we can already say confidently that despite the shortage of resources, our exercise has been worthwhile and we would encourage similar work on other SA languages. Two lessons to note:

First, translation and frame annotation teach us much about the target languages.

Second, provided the target language has at least rudimentary dictionaries and enough text online to help the novice writer, a translator can start with not much more than an ability to speak the language. They can learn as they go. Indeed, the TED translations are crowd-sourced. This is one way to rapidly add publicly available texts, a big help for poorly resourced languages. The quality will be variable, but can be improved afterhand. Meanwhile, the crowd-sourcing builds up an even more valuable resource: a community with greater competence in the target languages.

Annotation needs access to the tool, and some training, but not too much. Here too, one might be able to use volunteers to help, thus building up a FrameNet, and a full form lexicon.

For future work, we list some features of Hindi-Urdu, many shared with all SA languages, both Indo-European and Dravidian. We want to know how these features affect frame analysis. The data we gather will help us build FNs for Hindi and Urdu.

Reduplication is a prominent feature of all SA languages. It can mean greater intensity, or longer duration, but also distribution: "give the children two-two pencils" means "give each child two pencils".

States of mind. In SA languages, "I am hungry" and "I like spinach" are both expressed "to me, hunger affects" and "to me, spinach liking affects". Note that the English verbs can be transitive or intransitive. Is there a regular change in frames and triggers?

Clitics Examples are `hī` and `to` in Example 2 of Section 6, translated crudely as "itself" and "as for". These are function words, and it is hard to see of hand how they might affect choice of frame, but they do change the meaning of a sentence.

PTs and GFs. We have yet to work these out.

Incompatible lexicons. Hindi and Urdu differ only in lexicons, but the words are not one-to-one equivalents, at best they overlap largely. Some social or religious structures don't map at all, and the words have to be borrowed. An apparently equivalent word might require a different grammatical structure. So we expect the FrameNets to be affected by these factors, and cast light on them.

Cultural factors. Translating from English to SA languages involves a huge change of culture, and we can expect interesting new frames and compromises in translations. E.g., the Hindi "good morning" in Sec. 6. For more spectacle, consider weddings in the various Western and Asian communities.

9. Acknowledgements

We thank the anonymous referees for their detailed comments, which we have incorporated into the text above.

The work presented here has been financially supported by the University of Gothenburg, its Faculty of Arts and its Department of Swedish, through their truly long-term support of the Språkbanken research infrastructure, the Swedish Research Council through its funding of the projects *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (2015–2019; contract 421-2014-969), *Swe-Clarin* (2014–2018; contract 821-2013-2003), and the Department of Computer Science and Engineering, Chalmers University of Technology, Sweden.

References

Apertium. (2017). Apertium Wiki main page. `http://wiki.apertium.org/wiki/Main_Page`.

Bailey, T. G., Firth, J. R., and Harley, A. H. (1950). *Teach yourself Hindustani.* English Universities Press, London. Available from archive.org.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bhat, R. A., Bhat, I. A., Jain, N., and Sharma, D. M. (2016). A House United: Bridging the Script and Lexical Barrier between Hindi and Urdu. In *COLING*, pages 397–408. ACL.

Butt, M. (2010). The light verb jungle: Still hacking away. In Mengistu Amberber, et al., editors, *Complex predicates: Cross-linguistic perspectives on event structure*, page 48–78. Cambridge University Press, Cambridge.

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

Fillmore, C. J. (1976). Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137, Seoul, South Korea. Hanshin Publishing Co.

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

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, September.

Hasegawa, Y., Lee-Goldman, R., Kong, A., and Akita, K. (2011). Framenet as a resource for paraphrase research. *Constructions and Frames*, 3(1):104–127.

Hook, P. (1974). *The Compound Verb in Hindi.* Michigan series in South and Southeast Asian languages and linguistics. University of Michigan, Ann Arbor.

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

Liu, D. and Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of COLING 2010*, COLING '10, pages 716–724, Beijing. ACL.

Masica, C. (1991). *The Indo-Aryan languages.* Cambridge University Press.

Masica, C. (2005). *Defining a Linguistic Area: South Asia.* Chronicle Books.

MLFN. (2017). Multilingual FrameNet Project. `framenet.icsi.berkeley.edu`.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of HLT-NAACL 2006*, pages 192–199, New York, June. ACL.

Prasad, K. V. S. and Virk, S. (2012). Computational evidence that Hindi and Urdu share a grammar but not the lexicon. In *3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), collocated with COLING 12.*

Robinson, K. (2006). Do schools kill creativity? *TED: Ideas worth spreading.* `https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity/up-next`.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice.* International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Schmidt, R. L. (2004). *Urdu: An Essential Grammar.* London/ New York: Routledge. See the preface by Gopi Chand Narang.

Schultze-Berndt, E. (2006). Taking a closer look at function verbs: Lexicon, grammar, or both? In Felix K. Ameka, et al., editors, *Catching language: The standing challenge of grammar writing*, page 359–391. Mouton de Gruyter, Berlin.

Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL 2007*, pages 12–21, Prague, June. ACL.

Slade, B. (2016). Compound verbs in Indo-Aryan. In Hans Henrich Hock et al., editors, *The languages and linguistics of South Asia: A comprehensive guide*, pages 559–567. De Gruyter Mouton.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of COLING 2003*, pages 8–15, Sapporo, July. ACL.

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, August. Coling 2010 Organizing Committee.

Wu, D. and Fung, P. (2009). Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of HLT-NAACL 2009*, NAACL-Short '09, pages 13–16, Boulder. ACL.