Correlates between Performance, Prosodic and Phrase Structures in Bangla and Hindi: Insights from a Psycholinguistic Experiment

Kalika Bali, Monojit Choudhury, Diptesh Chatterjee,[1] Sankalan Prasad[1] and Arpit Maheswari[2]

Microsoft Research Lab India, Sadashivnagar, Bangalore 560080, India
[1]Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India
[2]Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
E-mail: {kalikab,monojitc}@microsoft.com

Abstract

Shallow parsers or Chunkers that create word groups based on local dependencies are considered essential for many NLP related tasks such as Machine Translation, Speech Synthesis and Information Retrieval. The linguistic stature of chunks is accepted even though their definition varies with the linguistic theory, specific task and even practical concerns of system building. The psychological reality of a chunk in linguistic performance is under-explored for Indian languages. In this study we conduct a set of psycholinguistic experiments in Bangla and Hindi in an attempt to understand how linguistically structured chunks correspond to a naïve native speaker's perception of chunk in an attempt to understand the correlations between performance and linguistic structures. The results reaffirm the cognitive basis of chunks and show some interesting patterns in their correspondence with linguistic structure. This may have consequences for how speech synthesis models prosodic phrasing and also the computationally appealing notion of a local dependency based chunk.

1 Introduction

The grouping of words into *chunks* has been long accepted as a useful first step towards syntactic parsing in Natural Language Processing. Shallow parsers or chunkers that create word groups based on local dependencies are considered *de rigueur* for Machine Translation, Speech Synthesis and even Information Retrieval (Watanabe *et al*, 2003, Atterer, 2003, Liu *et al*, 2004). Production data of speech utterances in languages such as English supports the cognitive reality of such chunks (Selkirk 1983). Similarly, there is evidence from a number of psycholinguistic experiments for English that correlate perceptual data to corresponding linguistic structures that sentence processing by humans involves chunking (Gee and Grosjean, 1983, Abney, 1995). However, we know of no such attempt to explore the cognitive evidence behind the linguistically motivated chunks for Indian languages.

Different linguistic theoretical frameworks and practical concerns can lead to different definitions, and hence, very different groupings of chunks. A chunk could thus be defined as a sub-tree within a syntactic phrase structure tree corresponding to Noun, Prepositional, Adjectival, Adverbial and Verb Phrases (Abney 1991, 1992, 1995) or as Noun Group and Verb Group based only on local surface information (Bharati *et al* 1995).

However, as pointed out in Abney (1991), all linguistic models are nothing but an attempt to explain real use of language. Therefore, a theoretical linguistic unit of language cannot be independent of its perceptual reality. The validation of any linguistic theory by necessity lies in its ability to explain linguistic performance.

In this study, we conduct a set of psycholinguistic experiments in Bangla and Hindi in an attempt to understand how linguistically structured chunks correspond to a naïve native speaker's perception of chunk as a word grouping. Corresponding prosodic data is also considered to find correlations between performance and linguistic structures. The

[1] This work was done during the author's internship at Microsoft Research India.

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

analysis of the data from these experiments supports the cognitive status of chunks in Bangla and Hindi, and shows some interesting patterns in their correspondence with linguistic structure. The results have direct consequences for how prosodic phrases are modeled in speech synthesis for a more natural speech output. It also examines the different approaches towards chunking in NLP tasks and there basis in actual cognitive parsing by the native speakers.

The rest of the paper is organized as follows: we discuss previous work done in related areas of psycholinguistics and computational linguistics in Section 2. Section 3 describes the experiment design and set-up in detail. The observations and results from the experimental data are discussed in Section 4. In Section 5 we present the implications of the observations that lead to certain conclusions.

2 Definition of a Chunk

One of the earliest attempts to define a linguistic chunk for a shallow parser is presented in Abney (1991). Abney's (1992) work follows from the sentence-parsing psychological experiments by Gee and Grosjean (1983). He replicates their experiment to study the "performance structure" of English as observed in naïve parsing done by native speakers and its manifestation in corresponding speech data.

Based on his experiments, Abney (1995) revised Gee and Grosjean's notion of a Φ-phrase to define a chunk. For Gee and Grosjean, a Φ-phrase with a boundary between a content word and the following function word, is best used to explain performance structure (naïve parses). However, as Abney points out, they fail to adequately explain certain syntactic clusters. For example, in Fig. 1, the nominal modifier and noun would result in two separate Φ-phrases, viz., [the old] and [man], in deviance from all experimental data. Abney proposes that chunks are connected sub-graphs of a parse tree not necessarily sharing constituency.

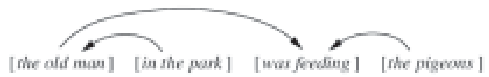


Fig 1: Chunks for English sentence (Abney 1992) where the chunks are marked with the brackets and the dependencies are shown through the arcs.

Each chunk therefore, consists of a dominant head which in this particular example (Fig.1) is "man". He goes on to show the syntactic validity of chunks as intermediate levels of analysis to a full parse structure.

Abney (1995), thus, defines a chunk as "the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head, but not including post-head dependents." He lists seven chunk categories and introduces the notion of *maximal chunk*, i.e., a chunk that is not contained inside another chunk. It may be noted that this approach allows for the existence of certain words, for example, coordinators and subordinators, to not be a part of any chunk. Also, a chunk may contain elements that exhibit tight bonding based on morphology, syntax and semantics, a chunk is strictly a syntactic entity.

Any work on defining chunks in order to create shallow parsers for Indian languages like Hindi and Bangla has to deal with the relative free word-order of these languages which allows for the movement of chunks within a sentence. NLP researchers have attempted to overcome such issues by defining chunks in Indian languages as having a fixed internal order with a sentence as "bag" of chunks that can move around in various permutations (Ray *et al* 2003).

In the context of Indian languages, we find two very different approaches to chunking or grouping of syntactically related words. Bharati *et al* (1995) defines a local word group as a non-hierarchical sequence of words which can be formed based on local information, but are sufficient for further processing of the sentence according to the *Paninian karaka theory*. Examples include noun group (noun + postpositions) and verb group (verb + auxiliary). They justify this definition by claiming that (a) the concept of verb phrase does not seem natural for Indian languages, and (b) from a computational point of view, recognition of noun and verb phrases is hard. Nevertheless, for agglutinative languages such as Tamil, where the case-markers are affixed to the noun roots, and aspectual, tense and gender markers are affixed to verb roots, there seem to be no scope of local word grouping; almost all the words will remain ungrouped or equivalently, as singleton groups. Bharati *et al* (1983) suggest a rule-based algorithm for identifying local word groups for Hindi. In Sangal *et al* (2007), several research groups reported their experiments on using machine learning approaches for developing chunkers as a part of part-of-speech tagging and chunking competition, for which the chunking scheme was similar in spirit to that of Bharati *et al*.

Das *et al* (2005) on the other hand introduced a concept of nested chunks. According to this approach, every word is a chunk to begin with; subsequently, two adjacent chunks can be joined to form a larger chunk. The process of chunking stops when a sentence is divided into multiple non-overlapping chunks all of which can be freely permuted within the sentence without changing the meaning. A chunk, therefore, is a sequence of words in a sentence which cannot be intruded by any word from outside it, but the words within a chunk and the chunks themselves are allowed to permute. They proposed a valence-theoretic rule-based algorithm for chunking.

Ray *et al* (2003) suggested a chunking scheme which lies in between the two extremes. It is non-hierarchical in nature, but identifies simple phrases which are typically longer and more complex than local word groups. The algorithm proposed for chunking is again rule-based.

3 Experiment Design

In his experiments, Abney (1995) asked naïve speakers of English to bisect sentences at their most natural break, and continue to bisect the resulting word clusters until they are left with a single word. Boundary strengths were assigned to each node based on the number of non-terminal nodes in that cluster, and the final boundary strength values were obtained by averaging across all speakers for each cluster. Corresponding speech data was also analyzed. Intra-sentential pause duration was used to mark major and minor prosodic boundaries to obtain prosodic phrases. The prosodic boundaries in this case were found to correlate very well with the boundaries of the naïve parsing.

In order to explore the cognitive reality of "chunks" for a shallow parse of a sentence in Bangla and Hindi we designed and conducted experiments similar to those of Abney (1995) for English. As has been stated earlier in the paper, we hope to achieve through these experiments a better understanding of how a native speaker's linguistic performance relates to the theoretical linguistic structures.

3.1 The Tasks

The experiments were divided into two tasks a) text parsing and b) prosodic parsing. Each task was carried out by 6 native speakers in Bangla and Hindi. None of the speakers had any linguistic expertise but were fluent in reading, writing and speaking the language concerned.

As in the earlier studies on English (Gee and Grosjean, 1983, Abney, 1995) the first task was aimed at eliciting performance structures from naïve speakers. In this task, native speakers of the language were required to bisect sentences iteratively into their most natural groupings until the users felt that the word groups could not be divided any further. For example (see Table 1(b) for transliteration and gloss),

[H5] ((खबर)(सुनते ही))((मैं)(तुंरत)((घर से)(भागा)))

All the sentences were also recorded in a separate session as read out by each speaker to elicit natural prosodic breaks. The task was carried out in both Hindi and Bangla. The speech recordings were made on a high quality PCM digital recorder in a normal office environment and the data analysed using the speech analysis software PRAAT (Boersma *et al* 2009).

3.2 Choosing the Sentences

The sentences for testing in both Bangla and Hindi are based on certain common structures of Indian languages. We sought to test through these sentences not only the role word-order plays in the perception of a chunk by naïve users, but also certain syntactic constructions where chunking is not always straightforward.

A set of 10 sentences for Bangla and 9 for Hindi were used to specifically cover the following syntactic structures:

a) Embedded clauses
b) Sentence and phrase-level adverbs
c) Conjuncts
d) Relative clauses
e) Participles
f) Noun Groups:
    a. Compound Nouns, Named Entities and Multiword Expressions
    b. (multiple) Adjectives+ Noun
    c. Qualifier + Adjectives + Determiner + Noun

   d. Noun + Postpositions

g) Complex Postpositions

h) Verb Groups[2]:

   a. Compound Verbs (Polar+Vector)
   b. Verb + (multiple) Auxiliaries + Particles (aspect/negation/wh-)
   c. Noun + Verb (group)

Table 1, at the end of the paper, includes all the sentences used for a) Bangla and b) Hindi. The sentences in two languages are near translations to make a comparative analysis easier.

4 Observations

The results from both the text parsing and the prosodic parsing experiments are presented in Table 1 (a) and (b). For every example, first line shows the example sentence, followed by its transliteration into roman, gloss and English translation. The transliteration has been annotated with POS tags (marked as 'word/TAG') and multiword expressions (marked by '[…]$_{TAG}$'). See Table 2 for description of the POS and multiword expressions tags. The next two lines show the boundaries obtained from prosodic parsing and text parsing, respectively.

The representation convention is as follows:

- A hyphen denotes a word.
- The order of words in the original sentence is maintained. Note that hyphenated words have been considered as a single word and punctuation marks have been ignored.
- The breaks are marked as major '|', and minor '*|' (see Section 4.1, 4.2 for explanation of major and minor breaks)

4.1 Prosodic-Parsing Experiment

Waveforms of each recorded sentence were analysed and intra-sentential pause durations labeled. The durations thus extracted were normalized for each speaker to rule out effects of reading rate and speaker variation.

The pause durations fell into two distinct classes: short pauses of <4.5 ms and long pauses of >7 ms. A minor boundary, marked with an *, was posited at each short pause and a major boundary at each long pause. The number of speakers who produced that pause was also enumerated at each boundary. Thus,

[B5] সিমলা হয়ে মানালী না গিয়ে সোজা দিল্লী থেকেই ফ্লাইট
নিয়ে যান ।
- 2*| - 6| - - - 6| - - - 6| - - -

implies that 2 of the subjects have a minor break after "*shimlA*" while all the subjects have a major break in the other places marked.

Again, as in the text parsing experiment, there is very good correlation within subjects.

Some points to note for both Bangla and Hindi:

a) There is always a major break at clause boundary, and major prosodic boundaries also correspond very well with the phrase structure of the sentence.

b) The sentential adverb is consistently separated by major boundaries by all speakers in both languages (B1, H1)

c) In general, postpositions are chunked with nouns. A consistent exception is H4 where "*sAhitya akAdemI*" and "*ne*" have a major boundary in between.

d) Similarly, compound nouns are always chunked together except "*trinamUl kAn.gres*" in H1 the two nouns by half the speakers. This might be because of unfamiliarity of the speakers with the compound itself.

e) The classification of coordinating and subordinating particles as not being part of any chunk (Section 2) is supported by this data as these particles are sometimes grouped with the preceding and sometimes with the following clause.

f) There also is some role of focus in determining the prosodic breaks. For example, in B7 "*laal, neel, sabuj pataaka duliye duliye*" was chunked differently based on whether the subject wanted to emphasise the colours of the flags or the act of waving them.

4.2 Text-Parsing Experiment

For the text-parsing experiment, we asked naïve native speakers of the language to construct a binary tree out of the sentence by appropriate bracketing, such as:

[B3] (তুমি)((রামকে (লাঠি দিয়ে))((এত জোরে)(মারলে কেন)))

[2] "Verb Group" here means a group of words with the main verb of the phrase as its head.

Then the brackets were numbered according to their height in the underlying binary tree. We show it for the above example where the bracket numbers are presented as subscripts.

$(_0$ তুমি $)_0$ $(_0$ $(_1$ রামকে $(_2$ লাঠি দিয়ে$)_2$ $)_1$ $(_1$ $(_2$ এত জোরে $)_2$ $(_2$ মারলে কেন $)_2$ $)_1$ $)_0$

Between every pair of words, the break is assigned a weight equal to the minimum of the closing or opening brackets at that point. The weight of a break between a word pair which are adjacent (i.e., the subject has not introduced any brackets between the words) is assigned a score of 1 more than the minimum of the nearest brackets. Thus, we get

তুমি 0 রামকে 1 লাঠি 3 দিয়ে 1 এত 3 জোরে 2 মারলে 3 কেন

The boundary scores are then averaged over 6 speakers. Let us assume that the above example shows the average score. The boundary score at each word is subtracted from the maximum average boundary score over all boundaries (which is equal to 3 for our running example). This is done in order to assign a higher score to a stronger boundary. This gives us

তুমি 3 রামকে 2 লাঠি 0 দিয়ে 2 এত 0 জোরে 1 মারলে 0 কেন

The scores thus obtained are then converted to a fraction of the maximum boundary score (i.e., 3, in this example). Hence,

তুমি 1 রামকে 0.67 লাঠি 0 দিয়ে 0.67 এত 0 জোরে 0.33 মারলে 0 কেন

We consider a score over 0.67 as a major boundary and one between 0.33 and 0.67 as a minor boundary. Therefore,

তুমি | রামকে *| লাঠি দিয়ে *| এত জোরে মারলে কেন
Final structure: - | - *| - - *| - - - -

Both the languages showed a good agreement amongst the subjects and it was observed that:

a) Noun is grouped with the following postpositions. H4 shows an exception where a relative clause is embedded between the Named Entity "*premchan.d*" and the postposition "*ke*". This holds true for spatio-temporal nouns as well (B10).
b) Named entities and MWE are always grouped together (B1, H1).
c) Adjectives, including multiple adjectives, are grouped with the noun (B7). However, for qualifier + adjective + determiner + noun, even though all subjects agree that these should be chunked together (H9), there is a lot of ambiguity on how to group words within the chunk. Fig. 2 illustrates this ambiguity through the groupings obtained for H9 from the 6 subjects.
d) Complex postpositions have been chunked into a single group and then chunked with the previous noun. (B10, H7)
e) There is high consistency in the verb chunks as polars–vectors, aspect markers, negatives and even question markers are consistently chunked with the main verb.
f) However, there is a difference in case of negation. (for example, H7 versus H2, H3, and B2 versus B5, B6)
g) For participles (B4) while there is a clause-level agreement, there is some confusion in chunking at the lower levels.
h) Sentential adverbs (B1, H1) while chunked as a part of the following Noun chunk, are marked separate by a minor break in both the languages.

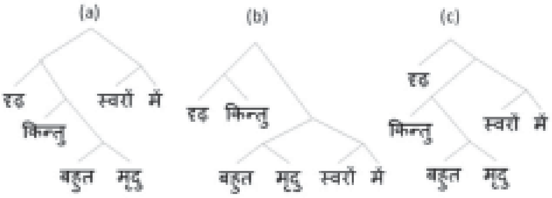


Fig 2: Various groupings of a noun phrase from H9 by the native speakers show that there is ample ambiguity in the perception of the chunks within a noun phrase.

4.3 Comparing Prosodic Structure to Performance Structure

The data from the two parsing experiments showed good correspondence between the prosodic structure and the performance structure, at least for the major boundaries. As the number of minor boundaries in the prosodic data was comparatively few and also as it was not possible to analyze the prosodic structure as hierarchical in nature, it is difficult to say much about the correspondence between the minor boundaries in the two datasets. However, if we look at comparable sentences, B1 and H1, for example, we find that the breaks are quite consistent.

A few anomalies can be observed between the prosody and text data. For example, B4, where the large number of prosodic breaks are not reflected in the text grouping. Though this cannot be conclusively determined from the current data, it is our intuition that the prosodic chunks are not entirely determined on the basis of syntactic structure alone but are influenced by phonological considerations like the number of syllables in a chunk and discourse-level phenomena such as topic and focus.

5 Discussion and Conclusion

In this study, we have made an attempt to understand the nature of chunks in Bangla and Hindi from a cognitive perspective. We have observed that the prosodic and performance structures agree with each other, which in turn matches with the phrase structure of the sentence. Moreover, there is a good amount of agreement between the speakers on the major prosodic and performance breaks. Even though the experiments have been conducted with a small set of sentences and few speakers, the systematic and consistent behavior of the three kinds of structures across the speakers and the languages makes us believe that it is possible to make some significant generalizations from the observations.

Firstly, we would like to highlight that chunking, at least from a cognitive perspective, is a sentence level phenomenon and therefore, local information is not sufficient for obtaining chunks. For instance, a very commonly used local rule for chunking is to introduce a chunk boundary after a postposition. However, we note significant deviations from this rule such as:

- [H1]: No chunk boundary after 'kAn.gres ke', because 'kAn.gres ke sadasyoM ne' forms noun a phrase.
- [H3]: No prosodic boundary after 'shimlA se', because 'shimlA se manAli' is being treated as a phrase.

This is more evident when we compare H5 and H8. While in the former case the verb 'bhAgA' forms a chunk (verb phrase) with the preceding postpositional phrase 'ghar se', in the latter case it forms a chunk with the adverb 'turant'. This suggests that there is an underlying notion of verb phrase in Indian languages, even though its syntactic structure is largely governed by context.

Bangla being a mildly agglutinative language, we do not find such instances. Nevertheless, while in B2 we observe a minor break between the genitive 'Taber' and the modified phrase 'bhA.nga TukRo', no such break is observed in B1 between the genitive 'ka.ngreser' and the modified 'sadasyarA'. This is due to the presence of the adjective 'bhA.ngA' in the former case, and therefore, indicates that chunk boundaries might also depend on factors such as length of the chunk and the distance between the modifier and the modified.

Secondly, we want to reemphasize the fact that while speakers seem to strongly agree on major breaks that correlate well with linguistic definition of phrases and clauses, there is not much agreement when it comes to chunks within phrases. For instance, in B2 the speakers have grouped 'EkTa bhAlo wAl haen.gin.g' variously, and therefore upon averaging we observe no breaks, though in the prosodic structure there is a major break after 'bhAlo'. Similar performance structure can be observed for Hindi (H1).

Thus, it is important that any chunker for speech applications makes use of global information to identify clause boundaries as well as major phrase boundaries. We also note that while the concept of local word grouping (Bharati *et al* 1995) is computationally appealing, there is no consistent correlation between such word groups and prosodic or performance structures. On the other hand, clause and phrase level groupings as suggested by (Das *et al*, 2005) seem to have a cognitive underpinning. However, it should be noted that while the scheme proposed by Das *et al*. is hierarchical, our analysis, by definition cannot elicit hierarchical structures out of the data.

References

A. Bharati, V. Chaitanya, and R. Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice

Akshay Singh, Sushma Bendre, and Rajeev Sangal. 2005. Hmm based chunker for Hindi. In *Proceedings of IJCNLP-05*. Jeju Island, Republic of Korea.

Dipanjan Das, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2005. An Affinity based Greedy approach towards Chunking for Indian Languages. *Proceedings of the International Conference on Natural Language Processing (ICON)*, pp. 55 - 62, Kanpur, India

Elisabeth O. Selkirk. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. The MIT Press, Cambridge, MA.

Feifan Liu, Qianli Jin, Jun zhao, and Bo Xu. 2004. Bilingual Chunk Alignment Based on Interactional Matching and Probabilistic Latent Semantic Indexing. In *Proc. of IJCNLP, 2004*.

Gee, James Paul, and Fran¸cois Grosjean 1983. Performance Structures: A Psycholinguistic and Linguistic Appraisal, Cognitive Psychology 15, 411–458.

Michaela Atterer. 2002. Assigning prosodic structurefor speech synthesis: a rule-based approach. In *Proc. of the Speech Prosody 2002 Conference*, Aix-en-Provence.

Paul Boersma, and David Weenink, 2009. Praat: doing phonetics by computer (Version 5.1.12) [Computer program]. Retrieved August 4, 2009, from http://www.praat.org/

P. R. Ray, Harish, V., S. Sarkar, and A. Basu. 2003. Part of Speech Tagging and LocalWord Grouping Techniques for Natural Language Parsing in Hindi. *Proceedings of International Conference on Natural Language Processing (ICON 2003)*. Mysore, India.

Rajeev Sangal, Sushma Bendre, Dipti Misra Sharma, Prashanth Reddy (ed.) 2007. *Proceedings of the IJCAI – 2007*. Workshop on Shallow Parsing for South Asian Languages (SPSAL-2007) Hyderabad, India

Steven Abney. 1991. Parsing by Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.

Steven Abney. 1992. Prosodic Structure, Performance Structure and Phrase Structure. In: Proceedings, Speech and Natural Language Workshop, pp.425-428. Morgan Kaufmann Publishers, San Mateo, CA.

Steven Abney.1995. Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. In: *Computational Linguistics and the Foundations of Linguistic Theory*. CSLI.

Taro Watanabe , Eiichiro Sumita , Hiroshi G. Okuno. 2003. Chunk-based statistical translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, p.303-310, July 07-12, 2003, Sapporo, Japan

Table 1 (a) Text and Speech Parsing data in Bangla

<table>
<tr><th>Ex#</th><th>Bangla sentences and the chunk boundaries</th></tr>
<tr><td>B1</td><td>গতকাল মহাত্মা গান্ধী রোডে তৃণমূল কংগ্রেসের সদস্যরা এক বিশাল পথ অবরোধের আয়োজন করেছিলো ।
gatakAl\RB [mahatma\NP gAndhi\NP rode\NP]$_{NE}$ [triNamul\NP kan.greser\NP] $_{NE}$ sadasyarA\NN Ek\DT viS-Al\JJ patha\NN avarodher\NN [Ayojan\NN korechhilo\VM]$_{CV}$
yesterday Mahatma Gandhi Road-loc Trinamool Congress-gen members a huge road block-gen organize do-past
Yesterday, the members of Trinamool Congress organized a huge roadblock on Mahatma Gandhi Road.

- 6| - - 2| - 6| - - 2| - 6| - 2*| - 2| - - 3| - -
- *| - - *| - | - - *| - | - - - - *| - -</td></tr>
<tr><td>B2</td><td>টবের ভাঙ্গা টুকরো গুলো ফেলে দিয়ো না, ওগুলো দিয়ে একটা ভালো ওয়াল হ্যাংগিং বানানো যেতে পারে ।
Tober\NN bhAn.gA\JJ Tukrogulo\NN phele\VM diyo\VA nA\RP, ogulo\PR diye\PP ekTa\DT bhAlo\JJ [wAl\NN hyan.gin.g\NN]$_{CN}$ bAnAno\VN jete\VA pAre\VA
Pot-gen broken piece plural throw give no those using one good wall hanging create go can
Don't throw the broken pieces of the pot; a good wall-hanging can be made out of those.

- 2*| - - - 5| - - - 6| - - 4| - - 5| - - 6| - - -
- *| - - - | - - *| - | - - *| - - - - | - *| - -</td></tr>
<tr><td>B3</td><td>তুমি রামকে লাঠি দিয়ে এত জোরে মারলে কেন ?
Tumi\PR rAmke\NP lAThi\NN diye\PP eto\JJ jore\RB mArle\VM kEna\PR
you Ram-acc stick with so hard beat-past why</td></tr>
</table>

	Why have you beaten Ram with a stick so hardly? - 1*| - 6| - - 6| - - 6| - - - | - *| - - *| - - - -
B4	জীবনানন্দ দাস, যিনি বনলতা সেন লিখে বিখ্যাত হয়েছিলেন, উনার প্রতি শ্রদ্ধা জানাতে বাংলা আকাদেমী একটা নতুন পুরস্কার ঘোষণা করেছে । [jibanAnanda\NP dAs\NP]$_{NE}$, jinni\PR [baNalatA\NP sen\NP] $_{NE}$ likhe\PL [vikhyAta\JJ hayechhilen\VM] $_{CV}$, onAr\PR prati\PP [shraddhA\NN jAnate\PL] $_{CV}$ [bAn.glA\NP AkAdemi\NP] $_{NE}$ ekTA\DT natun\JJ puraskAr\NN [ghoSanA\NN korechhe\VM] $_{CV}$ Jibananada Das who Banalata Sen write-participle famous be-past his towards respect show-inf Bangla Academy a new award announcement do The Bangla Academy has announced a new award in honour of Jibananda Das who is well known for composing Banalata Sen. - - 6| - *1| - 4| - 1| - 3| - 2| - 2| - 1*| - 1*| - 1| - 2| - 2*| - 6| - 2| - 2| - 3| - 4*| - - - | - - - - - - | - - - - - - - - - - -
B5	সিমলা হয়ে মানালী না গিয়ে সোজা দিল্লী থেকেই ফ্লাইট নিয়ে যান । shimlA\NP haye\PP mAnAli\NP nA\RP giye\PL sojA\RB dilli\NP thekei\PP flAiT\NN niye\VM jAn\VA Shimla via Manali no go-participle straight Delhi from flight take go Rather than going to Shimla via Manali, take a flight directly from Delhi. - 2*| - 6| - - - 6| - - - 6| - - - - - *| - *| - - | - *| - - *| - - -
B6	আমি ভেবে রেখেছি পুজোর আগেই এই কাজটা শেষ করে ফেলবো, কারণ পরে আর সময় পাব না । Ami\PR bhebe\PL rekhechhi\VA pujAr\NN Agei\NST ei\DT kAjTa\NN [sheS\JJ kare\VM]$_{CV}$ phelbo\VA, kAroN\CNJ pare\RB Ar\JJ samay\NN pAba\VM nA\RP I think-participle have prayer-gen before this work finish do throw-fut because later more time get-fut no I have decided to finish this work before the prayer because I will not find time later. - - - 6| - - 6| - - 5| - - - 6| - 6| - 1*| - - - - - *| - - | - - *| - - *| - - - | - | - *| - *| - *| - -
B7	লাল, নীল, সবুজ পতাকা দুলিয়ে দুলিয়ে বাচ্চারা ভারতীয় খেলোয়াড়দের স্বাগত জানালো। lAl\JJ, nil\JJ, sabuj\JJ patAkA\NN duliye\PL duliye\PL bAchchArA\NN bhAratiya\JJ khElowARder\NN [swAgata\NN jAnAlo\VM] $_{CV}$ red blue green flag waive-participle waive-participle children Indian players-gen welcome show-past Children welcomed the Indian players by waiving red, blue and green flags. - 6| - 6| - 3| - 3| - - 6| - 4| - - 2| - - - - - - *| - - | - *| - - *| - -
B8	দৃঢ়, অথচ মৃদু কন্ঠে তিনি এক অত্যন্ত প্রেরণাদায়ক বক্তৃতা দিয়েছিলেন । driR\JJ, athacha\CNJ mridu\JJ kanThe\NN tini\PR Ek\DT atyanta\JJ preranAdAyak\JJ [vaktritA\NN diyechhilen\VM] $_{CV}$ firm but soft voice-loc he a very motivating speech give-past In a firm yet soft voice he delivered a very motivating speech. - 3| - 3| - - 6| - 2| - 3| - 2| - 4| - 3| - - *| - - - | - *| - - - - -
B9	তুমি রামকে এতো জোরে লাঠি দিয়ে মারলে কেন ? Tumi\PR rAmke\NN Eto\JJ jore\RB lAThi\NN diye\PP mArle\VM kEno\PR you Ram-acc so had stick with beat-past why Why have you beaten Ram so hard with a stick?

<table>
<tr><td></td><td>- 1*| - 6| - - 6| - - 6| - -
- | - | - - *| - - *| - -</td></tr>
<tr><td>B10</td><td>মেঘের অনেক উপর দিয়ে উড়তে উড়তে যখন ক্লান্ত হয়ে পড়েছিলো বাজপাখিটা, তখনই হঠাৎ নীচের থেকে একটা আওয়াজ শুনতে পেলো ।
Megher\NN anek\JJ upar\NST diye\PP uRte\PL uRte\PL jakhan\RB klAnta\JJ haye\VM paRechhilo\VA bAjpAkhiTa\NN, takhanai\RB haThAt\RB nicher\NST theke\PP EkTa\DT AwAj\NN sunte\VM pelo\VA
cloud-gen lot up through fly-participle fly-participle when tired be-participle fall-past hawk then suddenly down-gen from a sound hear get-past
When the hawk got tired flying way above the clouds, suddenly it could hear a sound from below.

- 4| - 1*| - - 6| - - 6| - 2*| - - - 4| - 6| - 1*| - 4| - - 4| - - 3| - -
- *| - - - | - - | - *| - - - *| - | - *| - | - - *| - - *| - -</td></tr>
</table>

Table 1 (b) Text and Speech Parsing data in Hindi

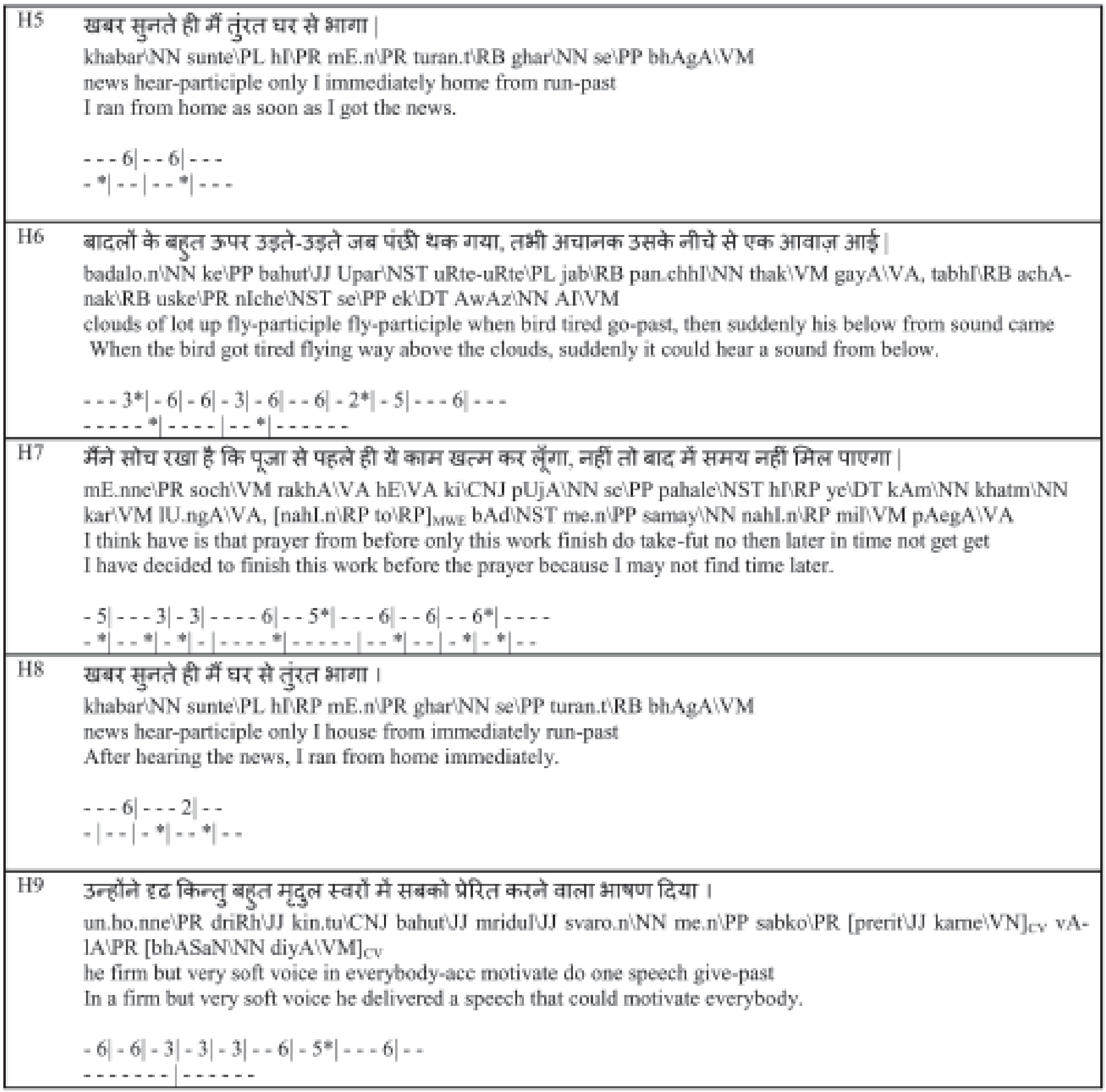
<table>
<tr><th>Ex#</th><th>Hindi sentences and the chunk boundaries</th></tr>
<tr><td>H1</td><td>कल महात्मा गॉधी रोड पर तृणमूल कांग्रेस के सदस्यों ने एक विशाल रास्ता रोको रैली का आयोजन किया |
Kal\RB [mahatma\NP gA.ndhI\NP roD\NP]$_{NE}$ par\PP [triNamUl\NP kAn.gres\NP]$_{NE}$ ke\PP sadasyo.n\NN ne\PP ek\DT vishAl\JJ [rAstA\NN rook\NN]$_{CN}$ rElI\NN kA\PP [Ayojan\NN kiyA\VM]$_{CV}$
yesterday Mahatma Gandhi Road on Trinamool Congress of members ergative a huge road block rally of organize do-past
Yesterday, the members of Trinamool Congress organized a huge roadblock rally on Mahatma Gandhi Road.

- 6| - - 4*| - - 6| - 3| - - - 1*| - 5|* - - 5*| - - 5*| - - 6| - -
- *| - - - - | - - - - - | - - - - - - - -</td></tr>
<tr><td>H2</td><td>गमले के टुकडों को फैक मत देना |
gamle\NN ke\PP TukRo.n\NN ko\PP phE.nk\VM mat\RP denA\VA
pot of pieces acc throw not give
Do not throw away the pieces of the pot.

- - 4| - - 6| - - -
- - *| - - | - | - -</td></tr>
<tr><td>H3</td><td>शिमला से मनाली ना जाकर सीधे दिल्ली से विमान ले लो |
shimlA\NP se\PP manAlI\NP nA\RP jAkar\PL sIdhe\RB dillI\NP se\PP vimAn\NN le\VM lo\VA
Shimla from Manali no go-participle straight Delhi from plane take take
Rather than going to Shimla via Manali, take a flight directly from Delhi.

- - - 3| - - 6| - - - 6| - - -
- - *| - *| - - | - *| - - | - *| - -</td></tr>
<tr><td>H4</td><td>प्रेमचंद, जिनकी प्रसिद्ध रचनाओं में गोदान प्रमुख है, के सम्मान में साहित्य अकादमी ने एक नए पुरस्कार की घोषणा की है |
premchand\NP, jinkI\PR prasiddh\JJ rachnAo.n\NN me.n\PP godAn\NP pramukh\JJ hE\VM, ke\PP sammAn\NN me.n\PP [sAhitya\NP aKAdamI\NP]$_{NE}$ ne\PP ek\DT nae\JJ puraskAr\NN kI\PP [ghoSaNA\NN kI\VM]$_{CV}$ hE\VA
Premchand whose famous creations in Godaan main is of honour in Sahitya Academy ergative one new award of announcement do is
Sahitya Academy has announced a new award in honour of Premchand, whose famous creations include Godaan.

- 6| - - - - 6| - - - 6| - - - 6| - - 3| - 6| - 2*| - - - 6| - - -
- | - *| - - - *| - *| - - | - - - | - - - | - - - *| - - *| - -</td></tr>
</table>



H5 खबर सुनते ही मैं तुंरत घर से भागा |
khabar\NN sunte\PL hI\PR mE.n\PR turan.t\RB ghar\NN se\PP bhAgA\VM
news hear-participle only I immediately home from run-past
I ran from home as soon as I got the news.

- - - 6| - - 6| - - -
- *| - - | - - *| - - -

H6 बादलों के बहुत ऊपर उड़ते-उड़ते जब पंछी थक गया, तभी अचानक उसके नीचे से एक आवाज़ आई |
badalo.n\NN ke\PP bahut\JJ Upar\NST uRte-uRte\PL jab\RB pan.chhI\NN thak\VM gayA\VA, tabhI\RB achAnak\RB uske\PR nIche\NST se\PP ek\DT AwAz\NN AI\VM
clouds of lot up fly-participle fly-participle when bird tired go-past, then suddenly his below from sound came
When the bird got tired flying way above the clouds, suddenly it could hear a sound from below.

- - - 3*| - 6| - 6| - 3| - 6| - - 6| - 2*| - 5| - - - 6| - - -
- - - - - *| - - - - | - - *| - - - - - -

H7 मैंने सोच रखा है कि पूजा से पहले ही ये काम खत्म कर लूँगा, नहीं तो बाद में समय नहीं मिल पाएगा |
mE.nne\PR soch\VM rakhA\VA hE\VA ki\CNJ pUjA\NN se\PP pahale\NST hI\RP ye\DT kAm\NN khatm\NN kar\VM lU.ngA\VA, [nahI.n\RP to\RP]$_{MWE}$ bAd\NST me.n\PP samay\NN nahI.n\RP mil\VM pAegA\VA
I think have is that prayer from before only this work finish do take-fut no then later in time not get get
I have decided to finish this work before the prayer because I may not find time later.

- 5| - - - 3| - 3| - - - - 6| - - 5*| - - - 6| - - 6| - - 6*| - - - -
- *| - - *| - *| - | - - - - *| - - - - - | - - *| - - | - *| - *| - -

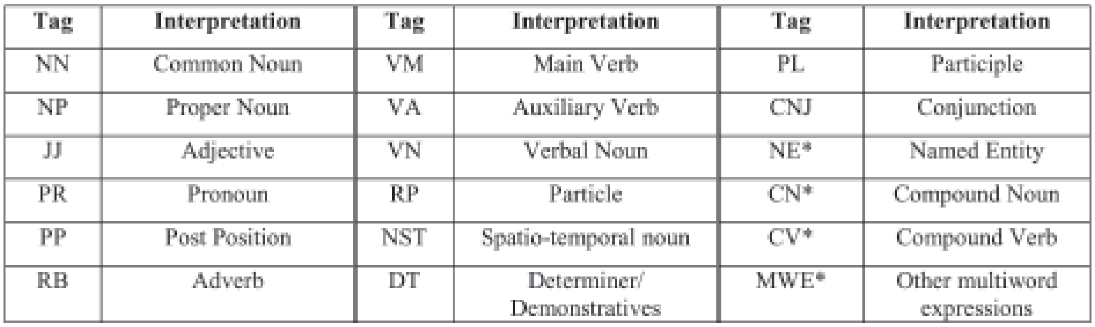
H8 खबर सुनते ही मैं घर से तुंरत भागा ।
khabar\NN sunte\PL hI\RP mE.n\PR ghar\NN se\PP turan.t\RB bhAgA\VM
news hear-participle only I house from immediately run-past
After hearing the news, I ran from home immediately.

- - - 6| - - - 2| - -
- | - - | - *| - - *| - -

H9 उन्होंने दृढ किन्तु बहुत मृदुल स्वरों में सबको प्रेरित करने वाला भाषण दिया ।
un.ho.nne\PR driRh\JJ kin.tu\CNJ bahut\JJ mridul\JJ svaro.n\NN me.n\PP sabko\PR [prerit\JJ karne\VN]$_{CV}$ vAlA\PR [bhASaN\NN diyA\VM]$_{CV}$
he firm but very soft voice in everybody-acc motivate do one speech give-past
In a firm but very soft voice he delivered a speech that could motivate everybody.

- 6| - 6| - 3| - 3| - 3| - - 6| - 5*| - - - 6| - -
- - - - - - - | - - - - - -

Table 2 List of POS Tags and Multi-word expression (*) tags used in Table 1 (a) and (b)



Tag	Interpretation	Tag	Interpretation	Tag	Interpretation
NN	Common Noun	VM	Main Verb	PL	Participle
NP	Proper Noun	VA	Auxiliary Verb	CNJ	Conjunction
JJ	Adjective	VN	Verbal Noun	NE*	Named Entity
PR	Pronoun	RP	Particle	CN*	Compound Noun
PP	Post Position	NST	Spatio-temporal noun	CV*	Compound Verb
RB	Adverb	DT	Determiner/ Demonstratives	MWE*	Other multiword expressions