Syllables as Linguistic Units?

Amitabha Mukerjee
Computer Science & Engineering
Indian Institute of Technology, Kanpur
amit@cse.iitk.ac.in

Prashant Jalan
Computer Science & Engineering
Indian Institute of Technology, Kanpur
prasant@cse.iitk.ac.in

Abstract

While there continues to be a debate in linguistics and speech processing as to the nature of an atomic unit, in computational approaches the atomic unit is universally taken to be the orthographic, space-demarcated "word". We argue that for many richly inflected languages such as Indo-European languages, syllable-based approaches together with semantic grounding may provide certain advantages. As a demonstration, we consider a language-acquisition system for Hindi, and propose a text syllabification technique, and show that syllable-based models perform somewhat better than the traditional word-based approach in building up a noun lexicon based on unannotated commentaries on video. We suggest further exploration of this potentially important idea in other domains.

1 Introduction

For resource-poor languages that are also morphologically complex - which applies to most of the languages in India - we suggest that instead of the orthographic word, the syllable may provide some advantages in terms of discovering structures in the language. This is motivated by the fact that the high number of morphological variants make it difficult to align verbs if we consider orthographic word boundaries. However, a syllabic approach is more likely to find overlaps with alternately inflected forms.

While a number of approaches seeking to discover morphological structure have worked with syllable-like structures (Creutz and Lagus, 2007, Clark, 2001), these approaches, like the rest of NLP, assume that the linguistic input occurs isolated from the extra-linguistic situation of the utterance. Thus, while the term "morpheme" is usually defined based on semantics (the smallest meaning-bearing unit), past work in NLP at the syllabic level are almost invariably based solely on linguistic input.

The primary contribution of this work is that to our knowledge this may be among the early works on syllabic-unit approach to text analysis that also operates in a grounded manner for discovering semantic codes in NLP. The main difference with traditional parsing driven models is that here semantics, in the form of visual or non-textual inputs, is used to segment the input into maximal syllable-sequences. Thus, semantics is being invoked from the very earliest stages. Later interpretations can then fall back on such sensorimotor models of meaning for interpreting novel structures such as metaphor, etc. Here we consider the language acquisition problem, where we attempt to map a lexical item from a textual description stream to its referent in a visual input stream. The proposal involves three steps:

1. Syllabification from text input without the knowledge of word-boundary. This is known to be relatively simpler in most Indian languages (Kishore et al., 2002; Patil et al., 2013) than some others (e.g. English (Marchand et al., 2009)).

2. Association of syllables with concept structures that are learned independently, e.g. using contrasting concepts. (Nayak and Mukerjee, 2012; Semwal et al., 2014).

3. Attention to relations between patterns of syl-

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 258–266,
Goa, India. December 2014. ©2014 NLP Association of India (NLPAI)

lables across different utterances (e.g. object verb agreement).

1.1 Problem with POS tags

We observe that while the semantics in this present approach is based on visual input, the same can also apply to formal semantic models for the data. However, annotating an input with such semantic labels is as difficult a problem as creating a treebank, and possibly subject to even greater ambiguities and disagreements. Also, a formal semantic tagger is based on some kind of a parse structure (typically a constituency tree or a dependency graph) - and the accuracy of such semantic models is dependent on the parse. The best POS taggers today perform at approx. 97%, but at a phrasal level, POS tags are accurate for only 50% of the sentences (Manning, 2011). Further, discrete, atomic word sense categories cannot diffuse into one another like a continuum model based on sensorimotor abstraction. Thus, the disjoint word senses often have conceptual overlaps which restrict accuracy severely (Jurgens and Stevens, 2011). Another difficulty is that standardizing on a single tag-set seems impossible, with each group suggesting its own set of tags; this is because all intermediate level tagsets constitute a compromise. Also, discrete partitions on the input space, hide the overlaps and similarities that actually exist across concepts, and are disambiguated by considering other information such as that from perception or other modalities (Fig. 1) (Pezzulo et al., 2011).

Just as continuum models of semantics provide a finer decomposition of the meaning space, so also a syllabic-unit model of the input itself provides a finer discrimination of the target map.

1.2 Grounded language models

The traditional ideas of formal semantics has also been applied to the notion of "grounding" linguistic structures. Grounding as used in computational modelling may define "meaning" in terms as an intermediate level of formal descriptions (Matuszek et al., 2012; Liang et al., 2009; Chen and Mooney, 2011), or it may attempt to relate the elements of language directly to clusters discovered on perceptual or motor data (Steels and Loetzsch, 2012). The former, in which good results have been obtained for sentential data, require rich training databases of sentence/meaning pairs, as well as supervised training datasets for learning classifiers that will work on visual input (Matuszek et al., 2012). The deeper problem with such systems is that the set of predicates used often have overlaps or gaps and just as with intermediate levels in other NLP situations, may not scale well to discourse modelling.

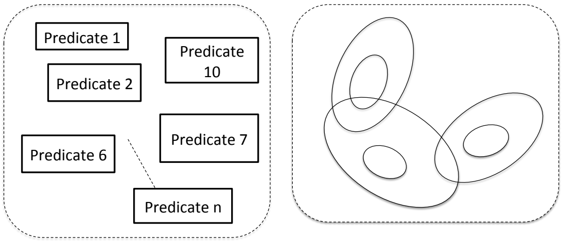


Figure 1: *Boolean vs. Continuous models of meaning*. Formal descriptions use boolean predicates to partition the world. Continuous models discover distributions in the input. The latter permits notions of similarity and conceptual overlap which are extremely difficult in boolean models.

When mapping language directly to sensorimotor data, the scenes are often isolated for a particular concept, and the description may be short or even a single word (Stramandinoli et al., 2013; Steels and Loetzsch, 2012). At other times, the scenes may have many concepts marked already (Reckman et al., 2011). Unconstrained sentential commentary, together with dynamically changing scenes, make for considerable ambiguity in association, and have been investigated very little. Yet, this is the type of input that children learn language from.

In theory, approaches based on grounding should work with any language. However, in practice, many approaches use knowledge of parsers or other intermediate levels (Chen and Mooney, 2011). At the very least, most approaches use morphological knowledge in the form of stemmers; even this can be problematic for richly-inflected languages.

Drawing on ideas in developmental cognition which indicate that infants are aware of conceptual distinctions well before they come to language (Mandler, 2007), our goal in this paper is to investigate the present-day limits of what we call *Uninformed symbol grounding* for morphologically rich languages. The attempt is to discover grounded lexemes in a system that associates an unparsed video and a set of unconstrained raw commentaries (sentential text).

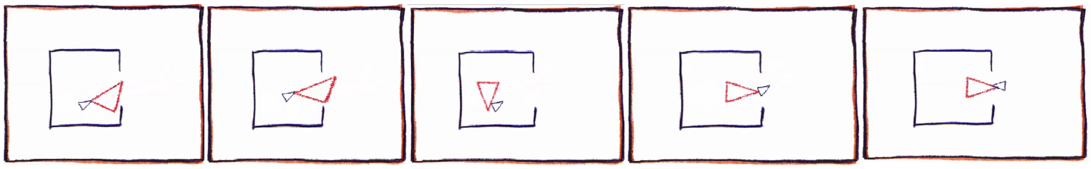


Figure 2: *Stills from the video.* The big triangle is pulling the little one, trying to get it outdoors. Eventually the big one pushes it out, and blocks it from coming back in. Finally, the two circle around each other playfully. The blocks are clustered by shape into C1(big red triangle) and C2(small blue triangle).

1.3 Grounding of syllable sequence

When using syllables, we search for maximal syllable sequences that associate strongly with a chunk of extra-linguistic context. Later, when the same syllable-gram occurs in a slightly different context, its semantics is broadened to include the new situation. Thus, the lexicon is a dynamic entity that changes with experience.

In this work, we attempt to ground syllable-grams from the input to structures in the video. Given the ambitious nature of the project, we chose a simple schematic video, one that has been developed for psychological experiments to detect autism (Sarah J. White and Frith, 2011)(Fig 2).

The only perceptual priors we assume are for identifying objects as coherently moving connected blobs. We use no other priors in either the visual or the linguistic processing. Thus, we use no knowledge of objects or shapes or actions, nor any language model. In order to be able to handle richly-inflected or agglutinative languages, we ignore word boundaries and start from syllable level and discover putative words that may be matched with the various concepts.

These perceptual priors are then mapped to syllable-clusters in the language stream. Hindi is a richly inflected language, with rich inflectional paradigms. We try to find descriptors for objects using a number of association measures.

Owing to lack of other experimentation in similar processes, we are not able to compare the results with other work. The commentaries are being made available at (Jalan, 2012).

The rest of the paper is organised as follows: in the next section we present the idea of using syllables as linguistic units followed by the explanation of the psychological video, the commentaries collected and the process of finding syllables in a sequence. Later, we present the noun discovery model followed by the results obtained.

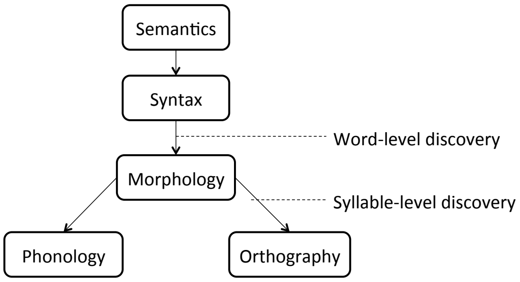


Figure 3: *Word vs. syllable.* Linguists are not sure if the morphology-syntax boundary can be defined clearly enough. Possibly a better place for starting the analysis may lie near the phonology-morphology boundary, where syllables occur.

2 Syllables as Units for NLP

A syllable is the segment of speech uttered in a single impulse of air. The nucleus of a syllable is usually the high-energy vowel sound, surrounded by co-articulations of lower energy occurring near the syllable boundary. As with all aspects of language, there is no single best mechanism for identifying syllables in text. However, for languages such as Hindi, syllabification is relatively easier than for other systems.

While linguists have been very careful in analyzing what is a ”unit” for different levels of analysis, computational approaches have overwhelmingly gone with the orthographic word as its dominant unit. What some linguists (Cahill and Gazdar, 1997) said more than two decades ago - “morphemes also exist, but only as second class citizens” - holds even more strongly today. This word-focus obscures the structures hidden within words and makes it particularly difficult for highly inflected languages. However, as “word” is difficult to define for linguists, so also is “morpheme”. A typical word in computational linguistics today (e.g. “boys” or किया (*kiyā*)) often has some structure hidden inside it.

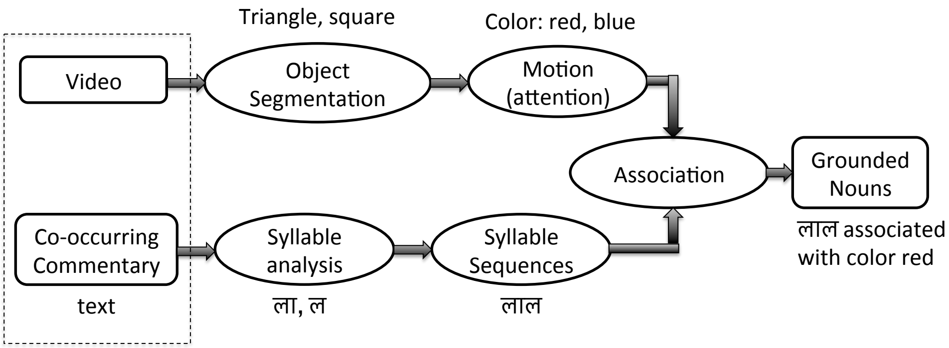


Figure 4: *Association approach for grounded noun learning*. The syllable sequences are potential words. Syllables ला (*lā*), ल (*la*) will form a syllable sequence लाल (*lāla*). All three of these would be strongly associated with the extra-linguistic context "color: red", but since लाल is a superset of the other two, we take this to be a word.

The morpheme-level syntax is hard to distinguish from word-level syntax (is "was clean" two words and "cleaned" only one?). Thus, linguists prefer to look at morphosyntax rather than word-level syntax (Fig 3). In lightly-inflected languages like English - only eight inflective variations, mostly at start or end of a "word" - one can get by with "stemmers" to identify a root. Some computational morphologists have pointed to the emphasis on English for the very poor development of morphological approaches in NLP (Clark, 2001). However, in Hindi, करवाया (*karavāyā*) is much harder to relate to किया (*kiyā*). The main idea here is that in such situations, syllables may be a better place to start than at word boundaries.

Morphemes arise at a level of segmentation that is smaller than the word, but larger than a letter (Goldsmith, 2010). This is also the space occupied by syllables. However, morpheme boundaries need not coincide with syllable boundaries, but in this work, our primary goal is not to discover morphemes per se, but words, so we assume an edge constraint, for which syllables are appropriate. Another model for this analysis could have been at the level of larger structures called feet, but since analysis of feet usually involves stress, and our work is based on text, we have kept it at the syllable level.

While syllables are not free from debate, they are certainly easier to identify than words, and also more useful as atomic constituents than phonemes (or in orthography, characters). Syllables are phonological clusters that may combine to make morphemes. While syllable boundaries are not aligned with morpheme boundaries, the more detailed analysis is still useful for identifying familial relations between a group of words. Thus, inflectional variants may retain some core syllables while changing others, enabling a syllable-based n-gram approach to identify these core elements.

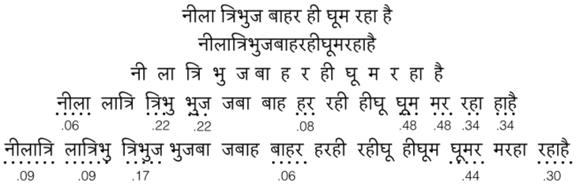
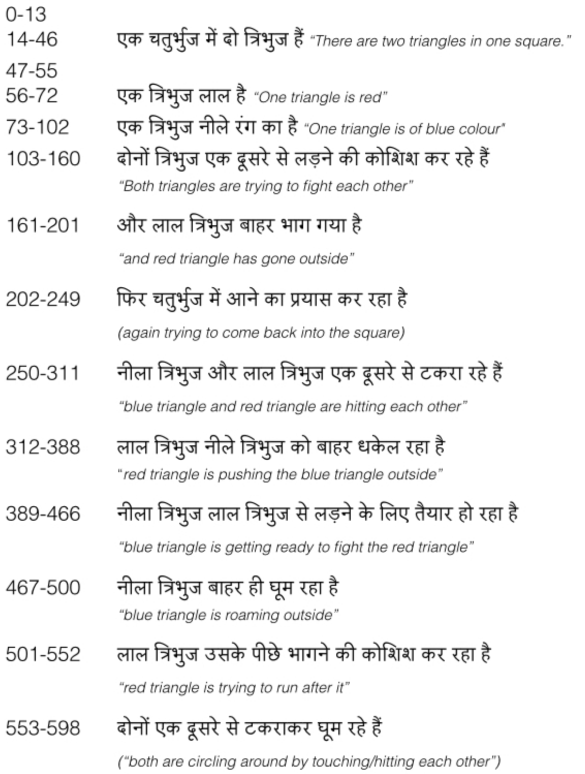

नीला त्रिभुज बाहर ही घूम रहा है
नीलात्रिभुजबाहरहीघूमरहाहै
नी ला त्रि भु ज बा ह र ही घू म र हा है
नीला लात्रि त्रिभु भुज जबा बाह हर रही हीघू घूम मर रहा हाहै
.06 .22 .22 .08 .48 .48 .34 .34
नीलात्रि लात्रिभु त्रिभुज भुजबा जबाह बाहर हरही रहीघू हीघूम घूमर मरहा रहाहै
.09 .09 .17 .06 .44 .30

Figure 5: *Syllabic analysis of text*. The text is re-segmented based on meaning associations rather than on orthographic cues. The last two sentences show the two and three syllable grams. The numbers shown are the mutual information score evaluated only for the narration shown in fig: 6. The process discovers structures such as नी (*nī*) which can now be related to other forms such as नीले (*nīle*) and नीला (*nīlā*). 3-syllable-grams such as घूमर (*ghūmara*) would have a strength almost same as the 5-syllable-gram घूमरहा (*ghūmarahā*) and hence only the latter would be taken as an unit, not the shorter fragment.

In the following, we begin by removing all word boundaries in the text, (preserving gaps at punctuations) and describe the input as a sequence of syllables. This is similar to the analysis in Chinese or Thai, which do not use spaces in orthography. The linguistic units then emerge as meaning-mapped substrings on the syllable space. Because similarity in meaning is used to learn these syllable se-



0-13	
14-46	एक चतुर्भुज में दो त्रिभुज हैं *"There are two triangles in one square."*
47-55	
56-72	एक त्रिभुज लाल है *"One triangle is red"*
73-102	एक त्रिभुज नीले रंग का है *"One triangle is of blue colour"*
103-160	दोनों त्रिभुज एक दूसरे से लड़ने की कोशिश कर रहे हैं *"Both triangles are trying to fight each other"*
161-201	और लाल त्रिभुज बाहर भाग गया है *"and red triangle has gone outside"*
202-249	फिर चतुर्भुज में आने का प्रयास कर रहा है *(again trying to come back into the square)*
250-311	नीला त्रिभुज और लाल त्रिभुज एक दूसरे से टकरा रहे हैं *"blue triangle and red triangle are hitting each other"*
312-388	लाल त्रिभुज नीले त्रिभुज को बाहर धकेल रहा है *"red triangle is pushing the blue triangle outside"*
389-466	नीला त्रिभुज लाल त्रिभुज से लड़ने के लिए तैयार हो रहा है *"blue triangle is getting ready to fight the red triangle"*
467-500	नीला त्रिभुज बाहर ही घूम रहा है *"blue triangle is roaming outside"*
501-552	लाल त्रिभुज उसके पीछे भागने की कोशिश कर रहा है *"red triangle is trying to run after it"*
553-598	दोनों एक दूसरे से टकराकर घूम रहे हैं *("both are circling around by touching/hitting each other")*

Figure 6: *Samples from narratives in Hindi*.

quences, they are closer in spirit to the definition of morpheme in linguistics. Fig 4 represents the model for grounded noun learning.

We demonstrate this process on a small example of lexeme acquisition from an unparsed unannotated corpus (fig. 5). Note that some forms such as रहाहे (*rahā-hai*) has a strong semantic relevance and is suggested as an unit, which is plausible, (and so are कररहाहे (*kar-rahā-hai*) and some larger strings). Also, syllable-grams such as त्रिभु (*tribhu*) which are appearing as frequently as the longer syllable-gram त्रिभुज (*tribhuj*) would be discarded as candidate units. On the other hand, लालत्रिभुज (*lāltribhuja*) is not an unit because लाल (*lāl*) and त्रिभुज (*tribhuj*) are independently associated with items "red" and "triangle" in the semantic inventory. Thus, लालत्रिभुज (*lāltribhuja*) is recognized as a compound of two words.

3 Video and Co-occurring Narrative Dataset

For the lexeme semantics acquisition task, extra-linguistic context is obtained from a short video (Fig 2). Language commentaries were recorded from university students (ages 21-24), by showing them the Frith-Happe video, in which two abstract shapes (triangles) engage in what may be called *coaxing* (Sarah J. White and Frith, 2011)(Fig 2). The two agents are a big red triangle (बड़ा लाल त्रिभुज; *baṛā lāla tribhuja*) and a small blue triangle (छोटा नीला त्रिभुज; *chōṭā nīlā tribhuja*). A total of 21 commentaries were collected.

Each subject was given the following instructions:

You will be shown this 39 seconds video thrice.

For the first two times you can just see the video and gain an understanding of what is happening in the scene.

The third time you have to describe whatever is going on in the video in Hindi without involving yourself in the description.

You should not metaphorize the objects in the video.

A small, variable time lag between the action shown on screen and the subjects description can be introduced. Being familiar with the video (showing it thrice) reduced this time lag.

Three corrupt narratives with extensive English and Hindi code-mixing were removed from the dataset for instruction non-following. The last instruction was added after a subject metaphorized the triangles, describing the small one as चोर(thief) and the big as पुलिस (police).

The spoken narratives were manually transcribed as text, maintaining a standardised spelling. The sentences were manually time-stamped, and sentences were broken at pauses, roughly longer than 10 frames (0.67 seconds), or non-language sounds or breathing breaks.

Another difficulty was that the two main objects were a large red triangle (C1) and a small blue triangle (C2). Our speakers divided into three groups - one described the triangles predominantly in terms of size (six narratives), other in terms of colour (ten narratives) and few of them described in terms of both size and color or very vaguely without ample information (five narratives). Given the lack of prior knowledge of any kind, it is important that we have a coherent narrative for lexicon discovery. So the results are reported here only for the larger of these two sets - the colour distinguishing narratives.

3.1 Syllabic Analysis of Commentary corpus

The morphological root (lemma or stem) may be difficult to identify owing to variations. Associations are then diluted across a large number of

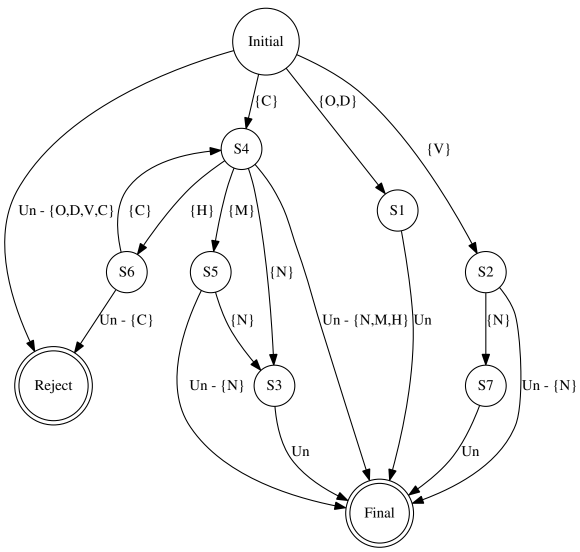


Figure 7: *FSM to identify syllables in Hindi*. Here $\{U_n\}$ = set of all unicode characters; $\{C\}$ = consonants (क, ख, ग,...); $\{V\}$ = set of all vowels (अ, आ, इ, ई, ...); $\{M\}$ = vowels on consonants (*mAtrAs*) (का, की, ...); $\{O\}$ = character ॐ ; $\{N\}$ = nasals (ः ँ) ; $\{H\}$ = halant; $\{D\}$ = digits.

morphological variants. To avoid this, it may be useful to discover smaller units driven by the association itself.

A character in Indian language scripts is close to syllable and can be typically of the following form: C, V, CV, CCV and CVC, where C is a consonant and V is a vowel (Kishore et al., 2002).

In our **syllabic analysis**, we consider the narratives without any knowledge of the word boundaries, and discover syllable sequences with good correlation with the concepts. The syllables are based on the orthography and are defined as a unit of speech having one vowel sound, with or without surrounding consonants - implemented as a Finite State Machine (Fig. 7), based on the automata defined by Nikhil Joshi (Mukerjee and Joshi, 2011). We made a few modifications to include the nasal sounds and other characters such as digits and the character ॐ. S_1 to S_7 are the intermediate states with 'Reject' and 'Final' as two accept states. At the 'Reject' state the scanned input till now is rejected and the machine goes back to the 'Initial' state. At the 'Final' state, we declare the whole sequence of Unicode characters except the last one as a syllable and start searching for next syllable with the last character observed. Note that punctuation symbol, non recognised foreign characters and illegal Hindi characters are rejected ('Reject state).

Finally, the learning agent is not operating in a linguistic void. This enables it to identify the very common words that occur in a wide range of contexts. This is done by computing the frequencies on a large Hindi corpus (CFILT, 2010). The top 1000 words are considered frequent, and not analysed.

4 Object name association : Noun discovery

The object concepts being associated here are the two objects that are the primary agents in the scene. It is assumed that at any time, only the moving objects will be in attentive focus, and are more likely to be spoken about. Thus, the sentences that are spoken when the *big red triangle* (C1) is moving are associated with it, whereas those that are associated with the *small blue triangle* (C2) are considered as the contrastive set. We consider all lexical candidates occurring in sentences that overlap with the interval when either C1 or C2 is in motion as candidates for association with these concepts.

4.1 Association measures

For a label (lexeme) l, concept c, speaker s and time t, we define following probabilities.

Attention probability of the concept c for the speaker s at time t

$$P(c|s,t) = \begin{cases} 1 & \text{if } c \text{ is attended by speaker } s \\ & \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

$$P(l|s,t) = \begin{cases} 1 & \text{if } l \text{ is uttered by speaker } s \\ & \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

We define the Joint probability of a label l and an object category c as

$$J(l,c) = \frac{1}{T * |S|} * \sum_{t=1}^{T} \sum_{s \in S} P(c|s,t) * P(l|s,t)$$

Similarly, we define the concept probability of a concept c as

$$P(c) = \frac{1}{T * |S|} * \sum_{t=1}^{T} \sum_{s \in S} P(c|s,t)$$

	Contrastive Mutual Information (Syllabic)	Relative Frequency (Syllabic)	Contrastive Mutual Information (Word)	Relative Frequency (Word)
C1	लाल (*lāla*), "red" - 108.5	लाल (*lāla*) "red" - 9.5	अभी (*abhī*) "now" - 2.3	लाल (*lāla*) "red" - 6.0
	नीलात्रि (*nīlātri*) "'blue [frag]" - 28.9	त्रिभुज (*tribhuja*) "triangle" - 6.0	उसने (*usanē*) "he/it" - 1.9	घूम (*ghūma*) "roam"- 5.0
	नीलात्रिभुज (*nīlātribhuja*) "blue triangle" - 27.6	नीलात्रिभुज (*nīlātribhuja*) "blue triangle"- 4.5	चतुर्भुज (*caturbhuja*) "square" - 1.4	त्रिभुज (*tribhuja*) "triangle" - 3.5
C2	नीला (*nīlā*) "blue" - 42.3	नीला (*nīlā*) "blue" - 7.3	अभी (*abhī*) "now" - 4.9	त्रिभुज (*tribhuja*) "triangle" - 13.5
	लालि (*lālatri*) "red+[frag]" - 25.9	लालत्रिभुज (*lālatribhuja*) "red triangle"- 6.5	बक्से (*baksē*) "box(s)" - 2.9	नीला (*nīlā*) "blue" - 10.0
	लालत्रिभुज (*lālatribhuja*) "red triangle" - 25.2	लाल (*lāla*) "red"- 6.0	रंग (*raṅga*) "color" - 2.8	अंदर (*andara*) "inside" - 2.5

Table 1: *Hindi lexeme association*. The top three associated lexemes (ranked based on the score obtained) for both the concepts (C1: big red triangle and C2: small blue triangle) are presented for syllabic analysis and word analysis. Word analysis makes use of the word boundary knowledge. Meaningful results are obtained in the syllabic analysis.

The label probability of a label l is given as

$$P(l) = \frac{f(l)}{\sum_l f(l)}$$

where $f(l)$ is the frequency of label l in the narrative corpus.

Based on the above, we used three association measures to identify the label maximally associated with a perceptual category.

1. *Conditional Probability* for a label l given a concept c is $P(l|c) = J(l,c)/P(c)$. However, this fails to penalise labels which co-occur with multiple categories; in practice it gave poor results and is not being reported in the results.

2. *Contrastive Mutual Information*. Mutual information is given as

   $$MI(l,c) = J(l,c) * log(\frac{J(l,c)}{P(c) * P(l)})$$

   It favours rare concepts and rare labels having sufficient degree of co-occurrence. Contrastive mutual information is the ratio of mutual information between label and concept for a binary contrastive situation.

3. *Relative Frequency*: This is a ratio of the label frequency when concept c is in focus (object is moving), versus the frequency (l) when c is not in focus.

4.2 Results

We report the top three associated lexemes for the concepts C1 (big red triangle) and C2 (small blue triangle) for both syllable analysis and for space demarcated orthographic words (Table 1). In syllable analysis, only k-grams occurring more than once are considered as candidate words, but for whole words, all words are candidates.

We observe that the key discriminants, "red" and "blue" are discovered as being more relevant for the large red triangle or the small blue triangle in the syllabic approach whereas such discovery is not made for the word analysis. Both Relative frequency and contrastive Mutual Information works reasonably well for syllabic analysis. Plain conditional probability results were poor and is not reported.

5 Conclusion

The main intent of this work was to investigate the possibility of computation with something smaller than orthographic words. This was motivated by the idea that in highly inflected languages such as Hindi, such structures may hold certain advantages, particularly for finding stems etc. It is certainly able to do this, but for our purposes, it also finds structures such as रहाहे (*rahA-hai*) which may be considered as an compound auxiliary unit. While the empirical demonstration here is very primitive and only scratches the surface of the problem, the results do suggest that this is an idea that deserves being investigated further as an alternative approach that shifts the boundaries at the very base of the model, and hence for the entire computational superstructure.

Here we have attempted to learn lexical associations with perceptual data, in an *Uninformed symbol grounding* approach. This implies that we discover any intermediate structures that arise, and minimize priors for the visual data. This is an ambitious task, and we have attempted this based on a meagre 39 second video, albeit a simple schematic one.

This work derives from cognitive ideas, but we do not consider many aspects such as shared attention, prosody and the simpler constructs in child-directed speech. It is possible that if one could collect corpora of this kind, we may obtain somewhat improved results. Nonetheless, it is surprising that even with such meagre input, many “correct” phrases emerge.

Once a few words are learned, the initial semantic models corresponding to these (often called *image schema*) become pivots around which other words can be learned (Kuhl, 2004). The lexeme learned serves as an index or a handle, so that future exposure to it invokes the same image schema, which is thereby defined more crisply and associated with a host of other concepts. Further, a few pivot words in an utterance helps the recovery of meaning for nearby elements.

This preliminary investigation suggests that the conviction that the orthographic word can be the only possible unit for computations in NLP may be worth revisiting. Many of the processes in language, particularly those involving acquisition without prior biases such as grammars and parse structures may be easier if we move down the scale from a word to a syllable. It is hoped that this preliminary exercise may induce others to take up this exploration so that such a process may expand and become an important part of NLP in times to come.

Acknowledgments

We would like to thank Dr. Uta Frith and Sarah White for providing and permitting us to use their psychological videos (Sarah J. White and Frith, 2011). We would like to thank Mr. Diwakar Chauhan for his help in collecting the narrations.

References

Lynne Cahill and Gerald Gazdar. 1997. The inflectional phonology of german adjectives, determiners, and pronouns. *Linguistics*, 35(2):211–246.

IIT Bombay CFILT. 2010. Ciil hindi corpus. `http://www.cfilt.iitb.ac.in/Downloads.html`.

David L. Chen and Ray J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. *Association for the Advancement of Artificial Intelligence (AAAI), Cambridge, MA*.

Alexander Simon Clark. 2001. *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, September.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):3.

John A Goldsmith. 2010. Segmentation and morphology. *The Handbook of Computational Linguistics and Natural Language Processing*, 57:364.

Prashant Jalan. 2012. Dataset. `http://www.cse.iitk.ac.in/users/grounded-lang/prashant/syllable/`.

David Jurgens and Keith Stevens. 2011. Measuring the impact of sense similarity on word sense induction. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 113–123.

SP Kishore, Rohit Kumar, and Rajeev Sangal. 2002. A data-driven synthesis approach for indian languages using syllable as basic unit. In *Proceedings of Intl. Conf. on NLP (ICON)*, pages 311–316.

P.K. Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings ACL ’09 / IJCNLP Volume 1*, pages 91–99.

Jean M. Mandler. 2007. Actions organize the infant's world. In K. Hirsh-Pasek & R. M. Golinkoff, editor, *Action meets word: How children learn verbs*, pages 111–133. Oxford University Press, New York.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Yannick Marchand, Connie R Adsett, and Robert I Damper. 2009. Automatic syllabification in english: A comparison of different algorithms. *Language and Speech*, 52(1):1–27.

Cynthia Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. 2012. A joint model of language and perception for grounded attribute learning. *Arxiv preprint arXiv:1206.6423*.

Amitabha Mukerjee and Nikhil Joshi. 2011. Word and phrase learning based on prior semantics. In *RANLP*, pages 616–621.

Sushobhan Nayak and Amitabha Mukerjee. 2012. Grounded language acquisition: A minimal commitment approach. In *COLING*, pages 2059–2076. Citeseer.

Hemant A Patil, Tanvina B Patel, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, GR Kasthuri, T Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, et al. 2013. A syllable-based framework for unit selection synthesis in 13 indian languages. In *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, pages 1–8. IEEE.

G. Pezzulo, L.W. Barsalou, A. Cangelosi, M.H. Fischer, K. McRae, and M.J. Spivey. 2011. The mechanics of embodiment: a dialog on embodiment and computational modeling. *Frontiers in psychology*, 2.

Hilke Reckman, Jeff Orkin, and Deb Roy. 2011. Extracting aspects of determiner meaning from dialogue in a virtual world environment.

Rosannagh Rogers Sarah J. White, Devorah Coniston and Uta Frith. 2011. Developing the frith-happe animations: A quick and objective test of theory of mind for adults with autism. Technical Report 149-154, Institute of Cognitive Neuroscience.

Deepali Semwal, Sunakshi Gupta, and Amitabha Mukerjee. 2014. Continuum models of semantics for language discovery. In *ICON*.

Luc Steels and Martin Loetzsch. 2012. The grounded naming game. *Experiments in Cultural Language Evolution. Amsterdam: John Benjamins*.

Francesca Stramandinoli, Davide Marocco, and Angelo Cangelosi. 2013. Grounding abstract action words through the hierarchical organization of motor primitives. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on*, pages 1–2.